

THE UNIVERSITY OF MICHIGAN
COLLEGE OF LITERATURE, SCIENCE, AND THE ARTS
Department of Communication Sciences

Technical Report

SIMULATION OF GENETIC POPULATIONS
WITH BIOCHEMICAL PROPERTIES

Richard Rosenberg

ORA Project 08333

supported by:

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE
NATIONAL INSTITUTES OF HEALTH
GRANT NO. GM-12236-03
BETHESDA, MARYLAND

administered through:

OFFICE OF RESEARCH ADMINISTRATION ANN ARBOR

June 1967

This report was also a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in The University of Michigan, 1967.

ABSTRACT

It is only with the advent of high-speed, large-storage digital computers that simulations of some complexity have become possible. The development of the present simulation, which differs in many important respects from other genetic simulations, and its application to a few specific problems is described. It constitutes a broad, flexible system for the study of evolutionary problems related to a closed, small population. Of particular importance are the many similarities to natural populations such as a time-varying phenotype, a dynamic interaction between the population and its environment, and mating rules, mortality rules, and selection rules of wide variety. The definition of the phenotype on the basis of chemical concentrations, whose levels are under genetic control via enzyme action, is the central feature of the model.

The relation between the number of crossover positions in the chromosome pair to the rate of evolution of the population is explored. The importance of the dynamic relationship between a population under selection pressure and its environment is underlined. An important result is achieved in the final set of experiments in which the selection of crossover probabilities is demonstrated. In this effort it is necessary to apply strong selection pressure to combat the disruptive stochastic effects in a small population. This pressure is manifested both in more powerful mortality procedures and in an increased number of allowable offspring per mating.

This system has been developed at least two viewpoints in mind.

The first is the desire to construct a research tool of use to biologists, especially geneticists. The second, which is dependent on the first, is to illustrate its usefulness by exploring a few problems of interest to geneticists. With the number of issues discussed and the importance of the final successful result, it is hoped that the potential of the system as a useful research tool is indicated.

ACKNOWLEDGEMENTS

I wish to express my gratitude to Professors J.H. Holland, A. Rapoport, W.J. Schull, and H.H. Swain for serving on my doctoral committee and for their interest in my work. I am indebted to Professor Holland for suggesting this area of research and for giving his time and effort without reservation. He has been a continual source of stimulation and most important a good friend. To Professor Schull I extend additional thanks for his willingness to bridge an interdisciplinary gap and to provide a much-needed genetic background.

I wish to thank the members of the Logic of Computers Group for their assistance in providing a congenial atmosphere for my research. In particular, I am indebted to Mrs. Kuniko Misawa for much of the typing of this dissertation. I also wish to thank Mr. Richard Weiland and Mr. Lee Erman for writing most of the program. To Richard Laing I extend warm appreciation for his continual supply of articles on genetics (and other matters) from the New York Times.

This work was supported by the National Institutes of Health while I was with the Logic of Computers Group at the University of Michigan.

TABLE OF CONTENTS

	<u>Page</u>
ACKNOWLEDGEMENTS	ii
LIST OF FIGURES	iv
LIST OF TABLES	vi
CHAPTER 1: INTRODUCTION.....	1
CHAPTER 2: A MODEL FOR CHEMICAL KINETICS	7
2.1. Definitions and the Fundamental Equation.....	7
2.2. Computation Procedure.....	10
CHAPTER 3: BIOLOGICAL MODEL	15
3.1. Genetic Structure	15
3.2. Enzymes and Chemistry.....	17
3.3. Permeation	24
3.4. Mating and the Production of Offspring	27
3.5. Mutation.....	33
3.6. Selection	34
3.7. Summary and Comments	38
CHAPTER 4: SOME SPECIFIC ASPECTS OF THE SYSTEM	41
4.1. Reaction Graphs and the Genotype	48
4.2. The Development of a Model	54
4.3. Some Statistical Aids	70
4.4. Chemical Dynamics	74
4.5. Built-in Selection	82
4.6. Summary	92
CHAPTER 5: SELECTION AND THE NUMBER OF CROSSOVER POSITIONS	94
5.1. Preliminaries	94
5.2. Crossover Positions and Selection	103
5.3. Structure of the Environment	112
5.4. Increased Mating Fraction	119
5.5. Summary	123
CHAPTER 6: SELECTION OF CROSSOVER PROBABILITIES	126
6.1. The First Model	129
6.2. The Second Model	135
6.3. A Theoretical Analysis	152
6.4. First Successful Results	167
6.5. Final Experiments and Conclusions	183
CHAPTER 7: CONCLUSIONS	194
7.1. General Considerations	196
7.2. Self-Regulatory Mechanisms	200
7.3. Morphogenesis	201
7.4. Random Fixation and the Founder Principle	202
7.5. Adaptive Systems Theory	203
7.6. Final Remarks	204
APPENDIX A: PROGRAM DESCRIPTION	205
REFERENCES (BIBLIOGRAPHY)	217

LIST OF FIGURES

<u>Figure</u>	<u>Title</u>	<u>Page</u>
3.1	Chromosome Pair.....	16
3.2	(a) Reaction Graph.....	22
	(b) Reaction Matrix.....	22
3.6-1	Offspring Generation Function.....	36
4.1-1	Reaction Graph.....	52
4.1-2	Reaction Matrix.....	52
4.2-2	Reaction Matrix.....	55
4.2-1	Reaction Graph.....	56
4.2-3a	Frequency Distribution of Crossover Probabilities.....	63
4.2-3b	Frequency Distribution of Crossover Probabilities.....	63
4.2-3c	Frequency Distribution of Crossover Probabilities.....	64
4.2-3d	Frequency Distribution of Crossover Probabilities.....	64
4.2-4	Offspring Frequency Distribution.....	67
4.4-1	Experiment 5.....	76
4.4-2	Comparison of Experiments 5 and 6.....	81
4.5-1	Experiment 7 & 8.....	86
4.5-2	Comparison of Experiments 9, 10, & 11.....	88
4.5-3	Experiments 11 & 7.....	90
5.1-1	Initial Chemistry for Offspring.....	97
5.1-2	Comparison of $\bar{F}_2$ for Experiments 1, 2, 3, 4.....	100
5.1-3	The Variation in $\bar{F}_2$ over 20 Time Steps, for the Initial Population.....	102
5.2-1	Comparison of $\bar{F}_2$ for Experiments 5, 6, 7.....	108
5.2-2	Comparison of $\bar{H}$ for Experiments 5, 6, 7.....	109
5.3-1	Chemical Flow in the System.....	113
5.3-2	Revised Chemical Flow in the System.....	117
5.4-1	Comparison of $\bar{F}_2$ between Experiments 9 and 10, and 6 and 7..	121
5.4-2	Comparison of $\bar{H}$ between Experiments 9 and 10, and 6 and 7..	122
6.1-1a	Reaction Graph A.....	130
6.1-1b	Reaction Matrix A.....	130
6.1-1c	Enzyme Table A-1.....	130
6.1-2	Comparison of $\bar{F}_1$ and $\bar{H}$ for Experiments 1 and 2.....	133
6.1-3a	Distribution of Crossover Probabilities for Experiments 1 and 2: Experiment 1.....	134
6.1-3b	Distribution of Crossover Probabilities for Experiments 1 and 2: Experiment 2.....	134
6.2-1a	Reaction Graph B.....	137
6.2-1b	Reaction Matrix B.....	137
6.2-1c	Enzyme Table B-1.....	137
6.2-2	Comparison of $\bar{F}_2$ in Experiments 3, 4, and 5.....	140
6.2-3	$\bar{F}_2$ and $\bar{H}$ for Experiment 6.....	147
6.2-4	Distribution of Crossover Probabilities for Experiment 6..	148
6.2-5	Heterozygote Statistics for Experiment 6.....	150
6.3-1	$\bar{F}_2$ and $\bar{H}$ for Experiment 7.....	156
6.3-2	Heterozygote Statistics for Experiment 7.....	157
6.3-3	Matings Between Classes 1 and 2.....	161
6.3-4	Matings Within Class 1.....	162
6.3-5	Matings Within Class 2.....	162

LIST OF FIGURES (Continued)

<u>Figure</u>	<u>Title</u>	<u>Page</u>
6.3-6	Matings Between Classes 1 and 3.....	163
6.3-7	Matings Between Classes 2 and 3.....	163
6.3-8	Matings Within Class 3.....	164
6.4-1	Mortality Independent of Population Pressure.....	169
6.4-2	Mortality Due to Population Pressure.....	169
6.4-3	Variation in $\bar{F}_2$ for Experiment 8.....	172
6.4-4	Population Size and Number of Double Heterozygotes for Experiment 8.....	173
6.4-5	Variation in $\bar{F}_2$ for Experiment 9.....	177
6.4-6	Population Size and Number of Double Heterozygotes for Experiment 9.....	178
6.4-7	Distribution of Crossover Probabilities for Experiment 9...	179
6.4-8	Ratio of Number of Actual Crossovers to Potential Number in Experiment 9.....	181
6.4-9	Distribution of Crossover Probabilities at Position 1-2 in Experiment 9.....	182
6.5-1	Variation in $\bar{F}_3$ for Experiments 10 and 11.....	187
6.5-2	Distribution of Crossover Probabilities for Experiment 10..	188
6.5-3	Population Size and Number of Double Heterozygotes for Experiment 10.....	189
6.5-4	Population Size and Number of Double Heterozygotes for Experiment 11.....	190
APPENDIX A		
A-1	MAIN.....	206
A-2	BIOPER.....	207
A-3	GEN.....	208

LIST OF TABLES

<u>Table</u>	<u>Title</u>	<u>Page</u>
4.1-1	Models 1, 2, & 3	51
4.1-2	Comparison in the Value of f for the Three Models	53
4.2-1	Enzyme Table	58
4.2-2	Crossover Probabilities	61
4.2-3	Offspring Generating Functions	66
4.4-1a	Allele Frequencies in the Initial Population	75
4.4-1b	Allele Frequencies at Generation 51 in Experiment 5	75
4.4-1c	Allele Frequencies at Generation 51 in Experiment 6	75
4.4-2	Specific Allele Frequencies	82
4.5-1	A Random Offspring Function	84
4.5-2	Allele Frequencies for Experiments 7 & 11	91
5.1-1	Age Dependent Mortality Distribution	95
5.1-2	Crossover Probabilities	98
5.1-3	Offspring Generating Functions	99
5.2-1	Crossover Positions for Experiments 5, 6, and 7	106
5.2-2	Allele Frequencies and Entropy for Experiments 5, 6, and 7..	110
5.2-3	Allele Frequencies at Gene 2 for Experiments 5, 6, and 7....	111
5.3-1	Distribution of Chemicals in the Environment of Experiments 5, 6, and 7	115
5.3-2	Distribution of Chemicals in the Environment of Experiment 8	118
6.2-1	Offspring Function L	138
6.2-2	NEWCON-4 and NEWCON-5	139
6.2-3	Differences in Experiments 3, 4, and 5	139
6.2-4	Ordering of Genotypes by Phenotypic Values	141
6.2-5	Enzyme Table B-2	142
6.2-6	Crossover Probabilities	144
6.2-7	Offspring Function N	145
6.3-1	NEWCON-6	153
6.3-2	Enzyme Table B-3	154
6.3-3	Offspring Function O	155
6.3-4	Genotype Classes	158
6.3-5	Mating Classes	159
6.3-6	Gamete Production	159
6.3-7	Distribution of Genotypes in Experiment 7 Compared with Random Distribution	160
6.3-8	Distribution of Double Heterozygote Offspring from Double Heterozygote Parents	166
6.4-1	Initial Age Distribution	170
6.4-2	Offspring Function Q	171
6.4-3	Offspring Function R	176
6.5-1	Comparison of f_2 and f_3	184
6.5-2	Offspring Function S	184
6.5-3	Crossover Probabilities	185

CHAPTER 1

INTRODUCTION

In recent years computers have been directed to areas previously unexplored. These investigations are possible only with such a tool as the high speed and powerful digital computer. By far the most common use of computers in the biological sciences is in statistics and data reduction. The role of computers in such calculations has been to carry them out at a vastly greater rate than possible with desk calculators and human operators. It is not that objectives have changed; rather, these are more readily achieved. For example, means, variances, covariances, and confidence levels are still the most commonly computed parameters.

With high speed computers, much greater quantities of data can be analyzed and in some instances this can lead to a better understanding of certain phenomena, a necessary prerequisite to the formulation of theories. There is a second area in which computers have made an important contribution and this is in the numerical solution of complex equations. In many cases the mathematical analysis of problems in biology has produced equations of such difficulty that they could not be solved by standard techniques in closed form. It is only the modern computer which has encouraged the development of numerical methods to attack previously intractable problems. Solutions can be found over wide ranges of parameter values in order to explore many important issues.

In some sense, the role of the computer in this situation is to extend and not to innovate. A theory is developed and this produces equations which in many cases have hitherto been unsolvable. The computer allows solutions over many conditions and in this sense permits the theory

to be more fully investigated. It is in the present research that a method of study is discussed which depends entirely on the existence of high-speed, large-storage digital computers. The research falls into the general category of simulations.

It is not the purpose here to develop a theory of simulation applied to biology or to trace its history but rather to present, by means of an actual simulation, some reasons for viewing it as a useful research tool in genetics. There have been several simulations done in population genetics a few of which are discussed in some detail in Chapters 4 and 6. There are, very broadly speaking, two classes of simulations, deterministic and stochastic. Deterministic simulations are very much akin to the numerical solution of sets of equations, usually large numbers of highly interrelated equations. By their very nature, problems in population genetics and in the theory of evolution are stochastic. Random influences enter at many points: in the survival of individuals of a population, in their reaching maturity and mating, in the numbers, genetics and viability of the offspring and in the ever-variable environment.

The population consists of a group of single-cell individuals, each having a single pair of chromosomes. The cell medium consists of a set of idealized chemicals. Reactions among these chemicals can occur only if the associated enzymes are present. In effect, the presence or absence of an enzyme corresponds to the presence or absence of an allele. The cells are situated in an environment which consists of the same chemicals. There exists a continuous permeation of chemicals through the cell membrane which will usually involve only a subset of the total number. At regular intervals, sexual mating will take place among some of the members of the population. In general, the nature of the mating

may be either random or related to some properties of the cells. The number of offspring resulting from any mating may again be a function of certain cell properties. Due to storage limitations of the computer, there will be an upper bound on the population, which will be maintained by various procedures to be discussed later.

This system was developed with at least two viewpoints in mind. The first was the desire to construct a research tool of use to biologists, especially geneticists. The second which is dependent on the first is to illustrate its usefulness by exploring a few problems of interest to geneticists. The choice of the term explore was carefully made. As the experiments illustrate, there is great difficulty in determining the contributions of the large number of interacting parameters. Also, because of the complexity of the system involving large numbers of calculations, the experiments require rather large periods of computer time. Thus, at present, some experiments cannot be repeated a sufficient number of times to guarantee complete validity. Nevertheless, they do provide a significant measure of support for the hypotheses being tested.

This simulation was conceived in a spirit which differs substantially from other simulations. Much of the detail related to this difference and the implications for the research which follows will be given in the beginning of Chapter 3, but a brief introduction seems worthwhile at this point. It is the goal of the present simulation to model a population which has many properties found in natural populations. Two of these are a biochemistry under the control of a set of enzymes defined by the genotype, and a cell membrane whose properties are also under genetic control, which regulates the flow of chemicals between the cell and its environment. There is a dynamic interaction between the population and

its environment, a situation which does not exist to any degree in other simulations.

The important feature in this simulation which will be stressed on several occasions is the time-dependent nature of the phenotype. In the earlier simulations, the phenotype is determined by a simple numeric transformation on the genotype which is essentially a set of zeros and ones. However, in the present situation the phenotype, which on a first level is a set of chemical concentrations, is a product of the interactions among various chemicals controlled by the cell enzymes. A process of growth as well as periods of mating are defined. The cells are aged and are subject to death on the basis of their age or fitness. In many respects, the population behaves in a somewhat realistic fashion.

One of the areas of study is the development of a system of chemical reactions suitably controlled by a carefully defined genotype. Various mating models, mortality models and selection models are used with a view to their influence on the rate and direction of the evolution of the population. Such questions as the relation between the percent of population turnover per generation on the evolution of the population, as well as the influence of the number of points at which crossing-over can occur, are studied.

The organization of this study is given as follows. Chapters 2 and 3 describe in great detail specific features of the chemical and biological parts of the system. The chemical model is a kinetic one devised in terms of collisions and decompositions of molecules. It is developed as a discrete time model in light of its implementation on a digital computer. The biological organization is structured in terms of its genetics not unlike a classical genetic model. That is, there are

chromosomes, genes and crossover probabilities, but in addition there is a model for gene action. At present, this model is very much unlike current models which describe the synthesis of proteins on the basis of DNA and RNA such as the one developed by Jacob and Monod [5]. However, while not presenting a detailed micro-behavior of the action, it does represent a reasonable overview. Genes define enzymes which act as gates to allow reactions to proceed with specified rates. A particular genotype is reflected in terms of a particular set of chemical concentrations.

In Chapter 4, a rather detailed examination of several aspects of the system is carried out. The main purpose is to indicate the generality of the system as well as the ease of substituting entire procedures. Also, the consequences of several methods for assigning the initial chemistry of offspring and eliminating cells to make room for offspring are discussed. Of particular interest is the development of the entropy concept as a measure of genetic variation. Finally, the concept of control experiments as applied to simulations is described together with an example of an unrecognized difficulty discovered by means of a control experiment.

It is in Chapter 5 that the first meaningful experiments are conducted along the lines suggested previously. Important among these is the confirmation of a dynamic relationship between the population and the environment. To maintain an overall stable condition between a population under a selection pressure and an environment, it is necessary for the environment to exhibit a dynamic rather than a passive behavior. Included in this chapter is a discussion of the concept of genetic space and its relation to the phenomenon of crossover.

An extensive set of experiments is described in Chapter 6, whose aim is to demonstrate the selection of crossover probabilities. The crossover

probabilities are assigned randomly to the initial population, and it will be shown that under selection pressure at a particular site, there will be a strong shift towards a low probability of crossing over. Many problems arise in such a study and in one instance the solution to a problem revealed a condition which had a parallel in the organism called paramecium.

No other experiments are discussed but it should be evident by the end of Chapter 6 that a system has been developed which has real potential in future research. Continuing areas for study are briefly outlined in Chapter 7 with a view to stressing the generality of the system.

We now begin with the description of the chemical model.

CHAPTER 2

A MODEL FOR CHEMICAL KINETICS

This model is very much like that defined by the differential equations derived from the mass action law. As will be shown later, it differs in that one factor related to a rate constant is time-varying. This formulation is from the outset a discrete one and thus avoids the difficulties inherent in the transformation of differential equations into difference equations. There are also certain computational advantages which will be made clear in a later chapter.

2.1 DEFINITIONS AND THE FUNDAMENTAL EQUATION

As mentioned above, the derivation will be in discrete terms. The chemicals in a cell can be numbered from 1 to n . Then M_i^t is defined to be the number of molecules of chemical i at time t and

$$M^t = \sum_{i=1}^n M_i^t$$

is the total number of molecules at time t , for the n chemicals. To each chemical, one can assign a molecular weight, W_i . Then for a given cell, the total weight of all the chemicals, W , is given by:

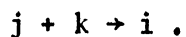
$$W = \sum_{i=1}^n W_i M_i^t .$$

Since the cell is a closed system (for the moment, ignore permeation), the total chemical weight W remain fixed, while the total number of molecules M^t is a time function.

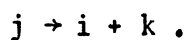
The number of molecules of chemical i , M_i^t can be changed in the following ways:

A. Increases in Number

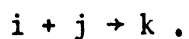
- (1) Molecules of chemical j react with molecules of chemical k to produce molecules of chemical i. This reaction will be diagrammed as



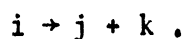
- (2) Molecules of chemical j decompose and among the products are molecules of chemical i. This is diagrammed as

B. Decreases in Number

- (1) Molecules of chemical i react with molecules of chemical j to produce molecules of chemical k, or



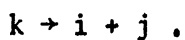
- (2) Molecules of chemical i decompose



Inherent in the above formulation is the assumption that the probability of three or more molecules colliding simultaneously is negligible as is the probability of a molecule decomposing into more than two molecules. The former assumption is quite reasonable, while the latter is made to preserve symmetry, i.e., to use one reaction prototype, namely



and



Based on the above ways of changing the number of molecules of a given chemical, the following recursive equation may be written:

$$\begin{aligned}
M_i^{t+1} = M_i^t + \lambda^t & \left[\sum_{\substack{j,k: \\ j+k \rightarrow i}} \lambda_{jk|i} \frac{M_j^t}{M^t} - \frac{M_k^t}{M^t} - \sum_{\substack{j: \\ i+j \rightarrow k}} \lambda_{ij|k} \frac{M_i^t}{M^t} \cdot \frac{M_j^t}{M^t} \right] \\
& + \sum_{\substack{j \\ j \rightarrow i+k}} \lambda_j|ik M_j^t - \sum_{\substack{j,k \\ i \rightarrow j+k}} \lambda_i|jk M_i^t
\end{aligned} \tag{2.1-1}$$

where λ^t is the total number of collisions in the time interval.

$\lambda_{jk|i}$, $\lambda_{ij|k}$, $\lambda_j|ik$, $\lambda_i|jk$ are constants less than 1 (corresponding to rate constants) which will be defined later. The terms in this equation can be labelled in the following way:

$$A(1) : \lambda^t \sum_{\substack{j,k \\ j+k \rightarrow i}} \lambda_{jk|i} \frac{M_j^t}{M^t} \cdot \frac{M_k^t}{M^t}$$

$$A(2) : \sum_{\substack{j \\ j \rightarrow i+k}} \lambda_j|ik M_j^t$$

$$B(1) : \lambda^t \sum_{\substack{i,j \\ i+j \rightarrow k}} \lambda_{ij|k} \frac{M_i^t}{M^t} \cdot \frac{M_j^t}{M^t}$$

$$B(2) : \sum_{\substack{j,k \\ i \rightarrow j+k}} \lambda_i|jk M_i^t$$

As explanation for their derivation, a simple probability model may be used. If there are M_j molecules of chemical j , then the probability of picking one j -molecule is $\frac{M_j}{M}$. Thus the probability of picking a j -molecule and a k -molecule is the product $\frac{M_j}{M} \cdot \frac{M_k}{M}$ (this is really $\frac{M_j}{M} \cdot \frac{M_k}{M-1}$ but since $M \gg 1$, the 1 can be ignored). If the total number of

collisions is λ , then the number of those collisions which involve j and k -molecules is $\frac{M_j}{M} \cdot \frac{M_k}{M}$. Finally, if the fraction of j - k collisions which produce an i -molecule is $\lambda_{jk|i}$, then the total number of i -molecules produced is $\lambda_{jk|i} \frac{M_j}{M} \cdot \frac{M_k}{M}$. In a more precise derivation, $\lambda_{jk|i}$ would be defined as the expected value of the fraction of successful collisions for a probability distribution of collisions. [Summing this over all j and k -molecules which produce an i -molecule results in the term $A(1)$.] By a similar derivation, the number of i -molecules lost through reaction with other molecules is the term $B(1)$.

The fraction of j -molecules which decompose into molecules i and k is $\lambda_{j|ik}$. Thus the number of i -molecules produced in this fashion is $\lambda_{j|ik} M_j$. Summing this over all different j -molecules which produce at least one i -molecule through decomposition results in the term $A(2)$. Similarly, the number of i -molecules lost through decomposition is the term $B(2)$.

It must now be shown that M_j^{t+1} can never be less than zero. While the number of molecules of a particular chemical may approach zero, it can never be less than zero. This will be demonstrated next in the course of discussing the computation procedure.

2.2 COMPUTATION PROCEDURE

For the actual computation, a time step is divided into two parts. The first part involving the contributions to the change in the number of molecules due to decompositions (terms $A(2)$ and $B(2)$) and the second those due to compositions ($A(1)$ and $B(1)$). This division of the time step is based on the following discussion.

For a chemical i , there are M_i^t particles at time t . This number can

be divided into two parts, the first being the number of molecules which decompose in the time interval and hence cannot take part in any collisions which may occur during this time. It follows that the products of these decompositions also cannot take part in any reactions in the same interval in which they were produced. Then subtracting the number of molecules of chemical i which decompose from M_i^t leaves those available for collisions and perhaps composition reactions. The preceding sentence thus implies that the number of decompositions of any chemical must be less than the actual number of molecules; a reasonable requirement.

One important fact is that not all collisions result in a product being formed. A successful collision is one in which the two colliding molecules react to form a single new molecule; an unsuccessful one occurs when the two molecules retain their identities after the collision. A collision will be unsuccessful for two reasons. First, the two molecules which collide cannot react because of an arbitrarily specified physical incompatibility, i.e., there does not exist an appropriate defining equation (see Section 3.2). Second, even with compatible molecules, not all collisions are fruitful because of lack of sufficient energy. This statement merely suggests a physical reason; the model is not directly concerned with energy calculations.

It is the role of enzymes in this system to prescribe that fraction of collisions of a particular pair of chemicals which are productive. A brief consideration of the requirement that the number of molecules never be allowed to be negative is now presented. What follows will be a 'worst' case analysis in the sense that some of the requirements will be sufficient without being necessary.

The computation is carried out in two parts; the decomposition part

will be discussed first. Consider that part of Equation 2.1-1, dealing with the decomposition terms. By the previous discussion, the results of the $A(2)$ term are not accounted for until the composition reactions have been completed, i.e., new molecules produced in decomposition reactions are not themselves available for composition in the same time step.

Therefore for the decomposition step, it is necessary that the term $\sum_{\substack{j,k \\ i \rightarrow j+k}} \lambda_{i|jk} M_i^t$ be less than M_i^t . This will be the case if $\sum_{\substack{j,k \\ i \rightarrow j+k}} \lambda_{i|jk} < 1$.

Thus it is necessary that for each chemical i , the sum of all the decomposition constants be less than 1. This is in fact a necessary and sufficient condition for the above requirement. Let $M_i^{t'} = M_i^t (1 - \sum_{\substack{j,k \\ i \rightarrow j+k}} \lambda_{i|jk})$

where t' denotes that the second half of the time step is being carried out.

From previous considerations, $M_i^{t'} > 0$. Then the terms of interest T_i are written as

$$T_i = M_i^{t'} - \lambda^{t'} \left(\sum_{\substack{j \\ i+j \rightarrow k}} \lambda_{ij|k} \frac{M_i^{t'}}{M^{t'}} \cdot \frac{M_j^{t'}}{M^{t'}} \right)$$

$$\text{or} \quad T_i = M_i^{t'} \left[1 - \lambda^{t'} \left(\sum_{\substack{j \\ i+j \rightarrow k}} \lambda_{ij|k} \frac{M_j^{t'}}{(M^{t'})^2} \right) \right] \quad 2.2-1$$

It must now be shown that the factor in the square brackets is greater than zero or equivalently that

$$\lambda^{t'} \left(\sum_{\substack{j \\ i+j \rightarrow k}} \lambda_{ij|k} \frac{M_j^{t'}}{(M^{t'})^2} \right) \leq 1. \quad 2.2-2$$

It should be noted that t' occurs on all the variables, indicating that these are determined after the decomposition reactions have been calculated. As defined previously, $\lambda^{t'}$ is the number of collisions occurring

among the $M^{t'}$ molecules during a time period. The number of collisions will depend on many factors, but as a simplifying assumption, the following relation will be used

$$\lambda^{t'} = kM^{t'} \quad 2.2-3$$

where k is a constant. Assume that the summation in 2.2-2 extends over m reactions where $1 \leq m \leq n$. Then substituting Equation 2.2-3 into the left side of inequality 2.2-2 gives

$$\frac{k}{M^{t'}} [\lambda_{ij_1|k_1} M_{j_1}^{t'} + \dots + \lambda_{ij_m|k_m} M_{j_m}^{t'}]. \quad 2.2-4$$

Define $X_i^{t'} = M_i^{t'} / M^{t'}$ to be the concentration of chemical i at time t' and let

$$\lambda_{ij|} = \text{Max}(\lambda_{ij_1|k_1}, \dots, \lambda_{ij_m|k_m}).$$

Then

$$\frac{k}{M^{t'}} [\lambda_{ij_1|k_1} M_{j_1}^{t'} + \dots + \lambda_{ij_m|k_m} M_{j_m}^{t'}] \leq k\lambda_{ij|} [X_{j_1}^{t'} + \dots + X_{j_m}^{t'}]. \quad 2.2-5$$

For $m = n$, the right side of 2.2-5 is $k\lambda_{ij|}$ since the sum of all the concentrations equals 1. Thus for 2.2-2 to hold it is necessary that

$$k\lambda_{ij|} \leq 1.$$

From previous discussion $\lambda_{ij|}$ must be less than or equal to 1 since it is defined to be the fraction of successful collisions. Finally, this implies that $k \leq 1$.

In summary, sufficient conditions for the left side of equation 2.1-1 to be non-negative are

$$(1) \sum_{\substack{j,k \\ i \rightarrow j+k}} \lambda_{i|jk} \leq 1$$

$$(2) \lambda_{ij|k} \leq 1 \text{ for all } j \text{ such that } i + j \rightarrow k$$

(3) $k \leq 1$

These conditions must hold of course for each chemical i .

As stated at the outset, this has been a worst case analysis. It is unlikely that the situation $m = n$ will ever occur and hence the sum of concentrations factor in the right side of inequality 2.2-5 will generally be less than 1. A further discussion of this point will be made in Chapter 4 including some reasons for violating condition 2.2-2. On this latter point it is found that for $k \leq 1$, the number of biochemical time steps needed for sufficient change in the concentrations is quite large. This number can be reduced by using a large k with preparations made in those cases for which negative concentrations occur.

This completes the discussion of the chemical model. In the following chapter the biological model will be presented in detail.

CHAPTER 3

BIOLOGICAL MODEL

In this chapter, the biological aspects of the present model will be discussed. This includes the genetic structure and the mechanism for representing crossover probabilities. The actual procedure for carrying out the genetic control of the cell chemistry is discussed in terms of the enzyme table. Various models for the permeation of chemicals into the cell are outlined. The remainder of the chapter consists of a presentation of many of the conditions which affect a genetic population such as mating rules, the production of offspring, mutation and the modes of selection forces.

3.1 GENETIC STRUCTURE

Each individual in the population, henceforth to be called a cell, has a single pair of chromosomes. There can be up to 20 genes per chromosome and up to 16 alleles per gene. This allows for a very large number of different genotypes. In general, for n genes and m alleles per gene, the number of genotypes, $G_{n,m}$ is

$$G_{n,m} = \left[\frac{m(m+1)}{2} \right]^n.$$

The number of the gene and the corresponding allele number are the co-ordinates of an entry in the enzyme table. Each allele, therefore, represents a unique enzyme whose operation will be discussed later.

The procedure for specifying crossover probabilities has been derived to provide the needed flexibility for many of the proposed experiments. It is based on associating the crossover probability at any position with the corresponding alleles in the following way. Consider

Fig. 3.1 below.

	1	2	3	
Chromosome 1	a ₁	b ₁	c ₁	d ₁
	x ₁₁	x ₁₂	x ₁₃	
Chromosome 2	a ₂	b ₂	c ₂	d ₂
	x ₂₁	x ₂₂	x ₂₃	

Fig. 3.1 Chromosome Pair

In this diagram, a pair of chromosomes are shown with the genotype

$\frac{a_1 \ b_1 \ c_1 \ d_1}{a_2 \ b_2 \ c_2 \ d_2}$. There are three possible places that crossover can occur

and these are labeled 1, 2, and 3. At each position the crossover probability is determined by a pair of numbers, called crossover levels, e.g., at position 1 this probability is determined by x_{11} and x_{21} , at position 2 by x_{12} and x_{22} and so on. Each of these numbers is associated with the corresponding allele in the sense that an allele can now be viewed as an ordered pair. Thus at Gene 1, the allele (a_1, x_{11}) has the double function of referencing an enzyme through a_1 and defining a "linkage" factor through x_{11} . The actual probability of crossing over at position 1 is a function of x_{11} and x_{21} .

In most genetic simulations, the crossover probabilities are the same at each crossover point and are fixed for the entire run [Fraser 1, 2]. In this system, it is possible to have variable crossover probabilities which can be selected for as part of the "fitness" of the alleles. More detail will be given in Chapters 4, 5 and 6.

A typical gamete produced from the pair of chromosomes in Fig. 3.1

would appear as

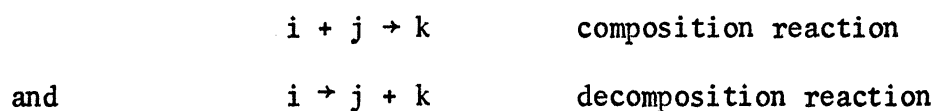
a_2	b_2	c_1	d
x_{21}	x_{22}	x_{13}	

In terms of the ordered pair referencing, the crossover level always remains with the corresponding allele even if crossing over occurs.

As far as other genetic mechanisms are concerned, epistasis is implicit in the chemical interactions although it is difficult to measure quantitatively. The following section will relate the epistatic effects to the actual cell biochemistry. Dominance can be included in the model but this has not yet been done explicitly. The difficulty is in a priori assigning dominance and recessiveness to alleles independent of knowledge of their interactions. There is a very natural way for dominance to arise and this is related to the nature of the enzymes. As with epistasis, the mechanism of dominance as manifested in the cell biochemistry will be discussed in the next section.

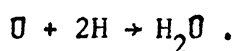
3.2 ENZYMES AND CHEMISTRY

The chemicals in a cell are represented by the integers 1, 2, ..., n. As stated in the previous section, the prototype reactions are



which may be called ternary forms. To represent the set of possible reactions among the n chemicals, a matrix form is used. For this matrix R, called the reaction matrix, the element r_{ij} is the product of a reaction between chemicals i and j. The product r_{ij} is itself an element

of the set of chemicals. Whereas the construction of the matrix R appears to be random, there is a definite restriction which arises. This is best discussed by reference to actual chemical reactions. For example, the equation representing the production of water from hydrogen and oxygen is



In this equation, the sum of the atomic weights on either side are the same. In fact, an equation written in this form is called the stoichiometric form, i.e., one concerned with the relations between the masses of components. What is important in the model is that it be consistent with respect to stoichiometry. Thus when the equation $i + j \rightarrow k$ is used, this means that the weight of k is equal to the sum of the weights of i and j . It is, therefore, impossible for k ever to appear on the left side of another composition equation where either i or j is on the right. If the reaction matrix is derived from an actual chemical system (in which all the reactions are of ternary form) this restriction is automatically preserved. However, in constructing an abstract reaction matrix, of some complexity, it may be quite difficult to check this condition. One method that may be used will now be discussed.

Consider an alphabet whose letters are $a_1, a_2, \dots, a_m$. A word on the alphabet is a string of letters. Thus, if u is a word then u is of the form $u = a_{i_1} a_{i_2} \dots a_{i_N}$ where the length of u , i.e., the number of letters in u , is N . A reaction is defined to be the concatenation of any two words over the alphabet. Thus if

$$u = a_{i_1} a_{i_2} \dots a_{i_N} \quad (\text{length} = N)$$

and

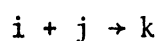
$$v = a_{j_1} a_{j_2} \dots a_{j_P} \quad (\text{length} = P)$$

then the word $w = uv$ is given by

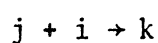
$$w = a_{i_1} a_{i_2} \dots a_{i_N} a_{j_1} a_{j_2} \dots a_{j_P} \quad (\text{length} = N+P)$$

It is obvious that if word length corresponds to weight, the equation $w = uv$ is weight-preserving.

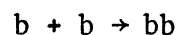
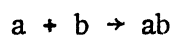
In a chemical reaction equation, the order of the reactants is arbitrary, i.e.,



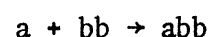
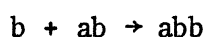
and



are equivalent statements of the fact that when chemicals i and j react, chemical k is produced. This result implies that the reaction matrix is symmetric. Another observation is that if the words are ordered according to length, after a certain length, none of the words of this length or greater can appear on the left side of a composition equation. Similarly, none of the alphabet letters can appear on the right side of a composition reaction. There is one practical consideration with respect to the construction of such reaction matrices which should be noted. Not all reactions need be included in the reaction matrix. For example given the alphabet a, b , the word abb can be produced in the following ways:



or



As the word length increases so do the number of ways of producing any word. In the following chapter an example of this situation will be discussed in detail. A discussion of the enzymes will now follow.

In its simplest form, an enzyme defines a set of pairs of chemicals or single chemicals and a constant. Perhaps it is worthwhile to

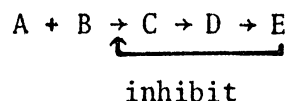
distinguish between composition and decomposition enzymes. A composition enzyme defines a set of pairs of chemicals, each pair of which is allowed to react. Included is the constant $\lambda_{ij|k}$, i.e., the fraction of collisions of compatible molecules which are successful in reacting. A decomposition enzyme defines a set of chemicals into which a given chemical is allowed to decompose. Again, the constant $\lambda_i|jk$ is included.

Essentially a "one gene — one enzyme" theory is being modeled, with one enzyme involved in a set of reactions rather than a single one. Actually a more correct statement is "one gene — one polypeptide chain" for when a protein contains more than one polypeptide chain, each chain is made separately [3]. This distinction is not relevant in this model since, as outlined above, the enzymes act as decision elements in the sense of allowing certain reactions to proceed as well as controlling their rate. The fact that the enzymes do not appear explicitly in the equations for the chemical reactions puts this type of chemical simulation into the general category, called by Garfinkel [4], the simulation of metabolic pathways without the explicit mention of enzymes. Since the main interest is in the overall change in chemical concentrations, this type of study is quite appropriate.

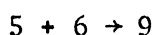
It should be mentioned that the present studies indicate that the gene controls not only the production of the enzyme (protein) but also, and this is most important, the rate of production [3]. Thus not only does the presence of an enzyme control reactions but also the amount of the enzyme present has a crucial effect. In our model a similar effect is achieved through the use of the rate constant. This is, higher rate constants can be interpreted as a faster production of enzymes. Although the model is not faithful on a fine level, it does preserve the main

features of cellular biochemical reactions.

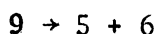
Additional complications which are not difficult to include involve inhibition or excitation in various parts of the metabolic pathways by certain chemicals. The following simple example is illustrative:



E can begin inhibiting this reaction after having achieved a certain concentration level. This type of inhibition is almost universal in metabolic systems. It has also been discovered to occur at the level of enzyme production. That is, certain chemicals upon reaching critical concentrations may either inhibit or excite the production of enzymes in their own or other metabolic pathways [3, 5]. An example of a simple reaction graph and the associated reaction matrix will now be discussed. In the reaction graph shown in Fig. 3.2a, there are 10 chemicals connected through 6 composition reactions. The numbers in parentheses are the actual words in terms of the alphabet 1, 2, 3 and 4 and the other numbers are the arbitrary chemical names. The decomposition reactions are the natural inverses, e.g.,

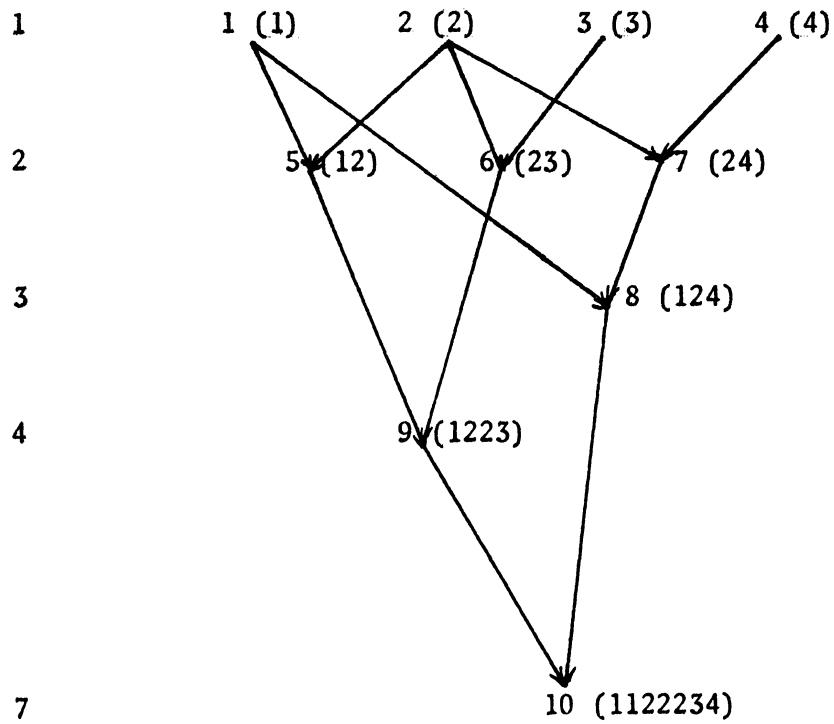


and therefore



Molecular weight corresponds to word length, i.e., the number of alphabet letters, and is designated by the numbers of the left. Thus chemical 9 has a weight of 4.

The reaction matrix expresses the information of the graph in another form. The zeroes in the matrix indicate that reactions between the chemicals designated by the row and column are not allowed. Blanks appear elsewhere for the same purpose, in each row, following the last



(a) REACTION GRAPH

	1	2	3	4	5	6	7	8	9	10
1	0	5	0	0	0	0	8			
2		0	6	7						
3			0							
4				0						
5					0	9				
6						0				
7							0			
8								0	10	
9									0	
10										0

(b) REACTION MATRIX

Fig. 3.2

non-zero number or zero diagonal element. This form is used in the computer program.

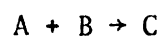
Before ending this section, we return to the discussion of epistasis and dominance. It is clear even in this simple reaction graph that the reactions are interconnected. Hence, if each reaction is specified by a single gene, the overall chemistry is determined by the epistatic interaction of all the genes. Although it is difficult to measure specifically the contribution of any gene as well as its influence on other genes, that influence is certainly evident.

With respect to dominance and recessiveness, Watson [3] points out that recessive genes frequently fail to produce any functional enzyme. For heterozygotes it might be expected that with one gene failing to produce the proper enzyme, such phenotypes would be intermediate between the two homozygous phenotypes. For one of the two following reasons this does not usually happen. To quote Watson, p. 241 [3]

Either there are still enough good enzyme molecules to catalyze the metabolic reaction even though the total number is reduced, or the recessive gene is not noticeable, because control mechanisms cause the wild-type gene in a heterozygote to produce more gene products than does each wild-type gene in a homozygote.

In general, dominance and recessiveness are largely a matter of context in the sense that the gene-products they elaborate may be "good" or "bad" in relation to the other gene products and their interaction in some environment.

To discuss the above quotation in terms of our model, consider the following reaction.



Suppose this reaction is controlled at a gene position for which

the two alleles specify rate constants k_1 and k_2 , respectively. Carrying out the state equation for this reaction for each allele results in the equations,

$$\begin{aligned}M_A^{t+1} &= M_A^t - k_1 X_A^t X_B^t - k_2 X_A^t X_B^t \\M_B^{t+1} &= M_B^t - k_1 X_A^t X_B^t - k_2 X_A^t X_B^t \\M_C^{t+1} &= M_C^t + k_1 X_A^t X_B^t + k_2 X_A^t X_B^t .\end{aligned}$$

These equations are equivalent to a set for the single rate constant, $k_1 + k_2$. One way for the allele associated with rate constant k_2 to be recessive to that associated with k_1 is for $k_2 \ll k_1$. Then due to linked reactions the effect of a total rate constant of $2k_1$ must be nearly equivalent to k_1 alone.

This model allows for dominance to arise naturally in a manner outlined above.

3.3 PERMEATION

For conceptual purposes, one can think of the physical space of this model as being a plane uniformly divided into squares. Some of these squares will be occupied by individuals of the population, but most will be part of the environment. The fineness of the square grid will be controlled by a program variable. A fine grid implies a sparseness of the population with respect to the environment.

The geometry of each individual corresponds to a square with permeation possible through the four sides, each side presenting one unit of 'surface' area. Any one of the chemicals is allowed to permeate on the basis of a simple quantity gradient. Thus the number of molecules

of chemical i which flow into a particular cell per time period is

$$M_i = 4\delta_i (E_i/G - M_i)$$

where E_i is the number of molecules of chemical i in the environment,

G is the grid size of the environment, i.e., $G = 1000$ means the environment is divided into 1000 squares,

δ_i is the permeation constant for chemical i .

The factor 4 appears because of the 4 cell sides open to permeation.

A first approximation is for δ_i to be a constant which is read in as data.

An immediate extension is to make δ_i a function of the concentrations of other chemicals (a phenotypic effect akin to membrane modification). This will result in the flow of chemicals through the cells being intimately linked to the genetic control of enzymes. By not allowing a particular chemical to either leave or enter the cell, one has in effect allowed the cell to accumulate this chemical in quantities much greater than the neighboring environment squares. In this way the cell membrane can be made active for certain chemicals (c.f. results of sodium and potassium pumps in actual cells as well as more complex molecules such as sugars, amino acids and purines).

One way to vary permeation constants will be outlined. Let $\bar{\delta}_i$ be the mean value of the permeation constant and δ_i the actual value. Then a functional dependence can be established among δ_i , $\bar{\delta}_i$ and a subset of chemicals P_i such as

$$\delta_i = \bar{\delta}_i + g(P_i)$$

where g can be constructed to have whatever characteristics are desired of δ_i . For example, if we want more of chemical i to flow in when chemical

j decreases below concentration $\bar{X}_j$ and vice versa, then the form of the equation might be

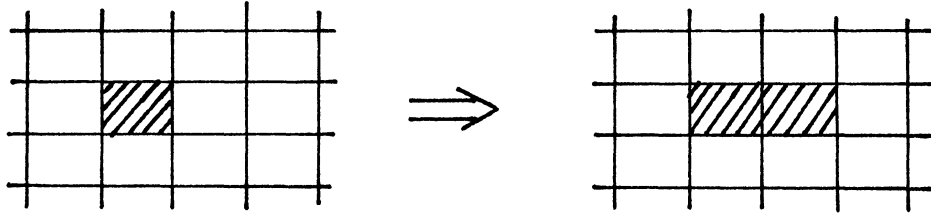
$$\delta_i = \bar{\delta}_i + a_i(\bar{X}_j - X_j)$$

where a_i controls the rate of change. The actual method used to vary these constants will depend on the goals of the experiment.

A few additional comments are in order. Homogeneity of chemicals both in the individual members of the population and in the environment is being assumed. Although cells are very 'compartmentalized,' on the level of this model, homogeneity is a necessary assumption with the alternative being very great computational problems. In the actual computation procedure, the nature of serial computation (rather than the more desirable parallel computation for this type of model) requires that chemical concentrations remain constant until all computations are complete. With respect to permeation, the environment concentrations are kept constant until each cell has been considered, after which the changes for the environment are computed. These changes affect the environment as a whole, resulting in every square having the same chemical constitution. Thus the necessity for homogeneity requires that there be almost instantaneous diffusability of chemicals in the environment. An alternative assumption is that the environment is large with respect to the cells so that cell products modify the environment in a negligible way in a given time step. Otherwise there would be insurmountable storage problems in attempting to record the distinct changes in each environment square. These problems are also present in any attempt to more carefully model the internal composition of the cell.

One final point concerns the possibilities of allowing a cell to change its shape. If the quantity of chemicals within a given cell

increases past a certain point, the cell might be allowed to change its shape in the following simple manner (the shaded square represents the cell).



This has the effect of increasing the number of cell sides to 6 instead of four and increasing the rate of permeation.

3.4 MATING AND THE PRODUCTION OF OFFSPRING

There are at present no sexes in the population although the mating is sexual. Thus any two individuals are able to mate. Since there is only one pair of chromosomes per individual, a sex difference would mean that all characteristics would be sex-linked. At present there is no reason for including sexes in this model, although if such a reason should arise they could be accommodated.

The most commonly used mating model is the random one in which the probability of mating is equal for all individuals. Kempthorne [6] outlines the following factors which he says should be included in the definition of random mating:

1. Actual mating at random occurs, i.e., every male in the population has the same probability of mating with every female, regardless of their respective genotypes.
2. Every mating has the same distribution of number of offspring.
3. The offspring are a random sample of the formal genetic possibilities of the cross.

4. Offspring are equally viable; for example, there are no lethal or semi-lethal genes.
5. There is no selection among parents or among offspring.

These aspects lead to two relationships:

The probability that $A_i A_j$ mates with $A_k A_l$ is equal to the probability that the individual is $A_i A_j$ multiplied by the probability that the individual is $A_k A_l$,

$$\text{and the offspring array of the mating } A_i A_j \times A_k A_l = \\ 1/4 A_i A_k + 1/4 A_i A_l + 1/4 A_j A_k + 1/4 A_j A_l.$$

When the term random mating is used in the model, it will include all the above factors except the second, for it is in the determination of the number of offspring for a given mating that selection may act.

An alternate form of mating which may be used and which is non-random is called assortative mating. If the mated pairs tend to be of the same genotype more often than would occur by chance this is called positive assortative mating and if less often it is called negative assortative (or sometimes disassortative) mating [7]. In this model, the mating is either random or based on phenotypic properties rather than genotypic ones and so positive assortative mating will mean that it is more likely for two individuals with similar phenotypes to mate. At a first level, the set of chemical concentrations is the phenotype. At the next level, the chemicals can be grouped into subsets each defining a property. The degree to which the concentrations of the subset are equal to preset values is the degree to which the property is present. In this case, the set of properties is the phenotype. Consider

the following example.

If the set of chemicals previously given as $\{1, 2, 3, \dots, n\}$ is denoted by C , let any subset of C be denoted by C_i . Then the function f_i defined as

$$f_i = \sum_{j \in C_i} (x_j - \bar{x}_j)^2$$

where $\bar{x}_j$ is the present concentration of chemical j ,

x_j is the actual concentration of chemical j ,

can be called a property. The closer f_i is to zero, the closer the individual is to having this property. On the basis of this property, individuals can be rated according to their values for f_i . Then under a positive assortative mating scheme, pairs will be selected on the basis of their f_i -values, i.e., the probability of selecting an individual to mate is inversely proportional to the magnitude of f_i . When several properties are employed, the probability of selecting an individual is a function of all of them. At present, only one property is being used, the selection of parents being carried out in the following manner.

For n cells, the set of their f -values will be $F = \{f_1, f_2, \dots, f_n\}$. The subscript here refers to the cell number, as only one function is being used. It should also be noted that the order is quite arbitrary. Then the set of inverses is $F^{-1} = \{1/f_1, 1/f_2, \dots, 1/f_n\}$. If by chance some $f_i = 0$, it is given a very small non-zero value before taking the inverse.

We now form the set $G = \{g_1, g_2, \dots, g_n\}$ where

$$g_1 = 1/f_1$$

and

$$g_i = 1/f_i + g_{i-1} \quad (i = 2, \dots, n)$$

and finally the set $H = \{h_1, h_2, \dots, h_n\}$ where

$$h_i = g_i/g_n. \quad (i = 1, \dots, n)$$

Then H consists of a set of fractions in increasing monotonic order between 0 and 1 with the last element $h_n = 1$. By this means, the interval 0 - 1 is divided into n segments, the length of segment i being inversely proportional to the f -value of cell i . An individual is selected for mating by generating a random number r uniformly over the interval 0 - 1, and determining a value i such that $h_{i-1} < r \leq h_i$. Thus the larger the segment represented by an individual, the greater is its probability of being chosen as a parent.

Regardless of which mating scheme is used, cells are allowed to mate only once per generation, but can mate in successive generations if they survive. The procedures used to eliminate cells from the population to make room for new offspring will be discussed in a later section.

The actual number of matings, limited by the number of individuals in the population, is itself an important selection force. Before looking into this question, the nature of a generation should be considered. There are in this model two ways in which a generation can be defined. If the entire population must mate at the same time with parents dying, a generation is clearly a new set of offspring. The mating thus occurs in a synchronous fashion, each generation clearly distinguished from the previous one. A variation on this is to have only a fraction of the entire population mating. Then the parents die and the offspring together with those which did not mate form the new population. In this version, generations are co-existent and may be so over a long period.

The time at which mating occurs depends on the number of biochemical

time steps allowed per generation. After the system has carried out this number, mating based on either of the schemes outlined above, follows. The next question is how many offspring can be produced and what is their constitution both genetic and chemical.

In the case of positive assortative mating, the number of offspring is assigned in a random fashion by using a particular probability distribution, which in some measure reflects that observed in nature. The negative binomial is used by some authors [8]. When mating is random, selection is directly operative in the determination of the number of offspring. This will be discussed in detail in Section 3.6.

The genetic constitution of offspring, independent of how many are produced, is determined through the application of the principles of classical genetics. Each parent produces gametes and a randomly chosen pair of gametes, one from each parent, constitutes the genetics of the offspring. The production of gametes by an individual involves the process of meiosis wherein a diploid cell of $2n$ chromosomes divides to yield $4n$ haploid cells. In the present case $n = 1$ and under ideal conditions 2 cells should each have 1 of the chromosomes and the other two cells the other chromosome. However, certain disruptive forces interfere, the most prominent one being the phenomenon of crossing-over, in which sections of chromosomes change positions on respective strands of the chromosome pair. The importance of crossing-over as a means of providing genetic recombination can hardly be overestimated. It will be considered in more detail in the next section.

As a final consideration, the initial chemical concentrations of the offspring can be specified in several ways. It seems reasonable that all offspring of one mating should have the same chemical constitution.

This constitution can depend entirely on the chemistry of one or both parents, be partially dependent on them, or be totally independent. Various models will be discussed in Chapter 4. Once the form of the chemistry has been decided, the source of the chemicals must next be determined. In the experiments to be discussed, the source of the chemicals is the environment. That is, offspring are given their initial chemicals from the environment but this does not result in a gradual depletion of the resources of the environment since with the death of cells, their chemicals are returned to it. Thus a kind of equilibrium is set up between the environment and the population with respect to the flow of chemicals. It may seem strange that the initial chemicals for the offspring are taken from the environment rather than from the parents. The reason for doing this is to eliminate the period of time between infancy and maturity. That is, to grow, the cell will take in resources from the environment and in fact the major part of its weight must come from the environment. Since we are not, at present, interested in this period of cell life, we arrange conditions so that an offspring, after a single period of biochemical development, is able to mate.

It should be stressed again that it is very important to specify carefully the initial offspring chemistry. Although the source of the chemicals, in all cases, is the environment, the amount of each chemical will in general be independent of the state of the environment or the state of the population. One interesting method for assigning the initial chemistry is to choose several schemes beforehand and have each keyed to one of the alleles of some gene. This gene would only serve to specify initial chemistry and not take part in any of the reactions. Thus the form of this chemistry can be left up to selective forces if

it is desired to have this feature as an experimental parameter.

3.5 MUTATION

Mutation is of course the ultimate source of genetic variation and its inclusion in any genetic model would appear to be important. There is the problem in determining how often and where it should act. Genetic experimentation has shown that the frequency of mutation varies from gene to gene and is generally a constant over time for a particular gene. The result of a mutation is the production of a new allele, or in this system, a new enzyme. This could be accomplished by randomly changing the characteristics of a present enzyme and adding this new enzyme to the enzyme table. The change could be in the rate constant or in one or more of the chemicals defined by the enzyme. However, a problem arises in the fact that the enzyme table, which is common to all the cells, has a limited capacity and hence cannot be indefinitely expanded. This leads to the following alternative. Let the present enzyme table represent all the enzymes that will ever be used by the system. Then the role of mutation is not to create new alleles but to change one allele into another in a random fashion. Mutation then provides the opportunity for alleles to be continually present in the population.

Since the number of cells in the population is a small number (about 200), random drift (see Section 7.4) as well as inbreeding are important phenomena and can lead to the fixation or loss of alleles without reflecting the effect of the selective forces. Thus in this context mutation serves its biological function as the source of genetic variation.

Questions with respect to the frequency and sites of the mutation will be left open as experimental parameters. An initial procedure is to

choose as the only parameter a mutation frequency, say 1 per cent. Then if there are 200 cells, each with 10 genes, the number of genetic loci is $200 \times 2 \times 10 = 4000$. With a mutation frequency of 1 per cent, 40 loci will be randomly changed each generation. By randomly it is meant that the locus is selected randomly and the allele to be put in at this locus is again randomly determined. It is possible for a locus to remain unchanged.

3.6 SELECTION

This will summarize the discussion in 3.4 of where and how selection acts. As discussed above, selective forces are in evidence in the number of matings allowed and in the number of offspring produced per mating. It is this latter explicit form of selection that we wish to discuss now.

When selective forces are operative in the determination of parents, the number of offspring assigned to a mating is random. Conversely, when the parents are chosen randomly selection operates in the determination of the number of offspring. One scheme, which is used in the experiments to be discussed, will now be outlined.

Define the variable s , the argument of the offspring function, as

$$s = \frac{f_i + f_j}{2\bar{f}}$$

where f_i, f_j are the function values for cells i and j (parents)

and $\bar{f}$ is the mean value of the function for the population.

We will now show that $\bar{s}$, the mean value of s , is 1. The number of values s can take is $\binom{n}{2}$ where n is the number of cells in the population. Then

$$\bar{s} = \frac{\sum_{i=1}^n s_i / \binom{n}{2}}{2 \sum_{j=1}^n f_j / n}.$$

It is clear that in the $\binom{n}{2}$ terms of the numerator each f_i appears $(n-1)$ times. Therefore

$$\begin{aligned} \bar{s} &= \frac{n \sum_{i=1}^n (n-1) f_i}{\binom{n}{2} 2 \sum_{j=1}^n f_j} \\ &= \frac{n(n-1)}{2 \binom{n}{2}} \cdot \frac{\sum_{i=1}^n f_i}{\sum_{j=1}^n f_j} \\ &= 1 \end{aligned}$$

This suggests that the desired mean number of offspring per mating should occur for a value of s near 1. Fig. 3.6-1 shows a typical graph for specifying the number of offspring to be assigned to a given mating. The form of this graph is experimentally determined by trying to achieve a reasonable frequency distribution for offspring, i.e., vary the form of the graph until the resulting offspring distribution is acceptable.

Selective forces are also present in the limitation on population size as well as in the choice of cell properties. It is obvious that the evolution of the population is dependent on the choice of these properties. The point to be made is that all these instances must be considered in analyzing the evolution of the population. In addition, other points at which selection can act should be considered. For

example, the fact that the population size is fixed means that cells must be eliminated to make room for new offspring. It will be shown in the next chapter that the method used for replacing cells does indeed exert a strong selective force. Three methods have been used and these will now be discussed.

Assume all cells begin 'life' at age zero and are aged by 1 unit after each mating. All three methods have in common the fact that cells are 'killed' only when space is needed. If the total number of offspring

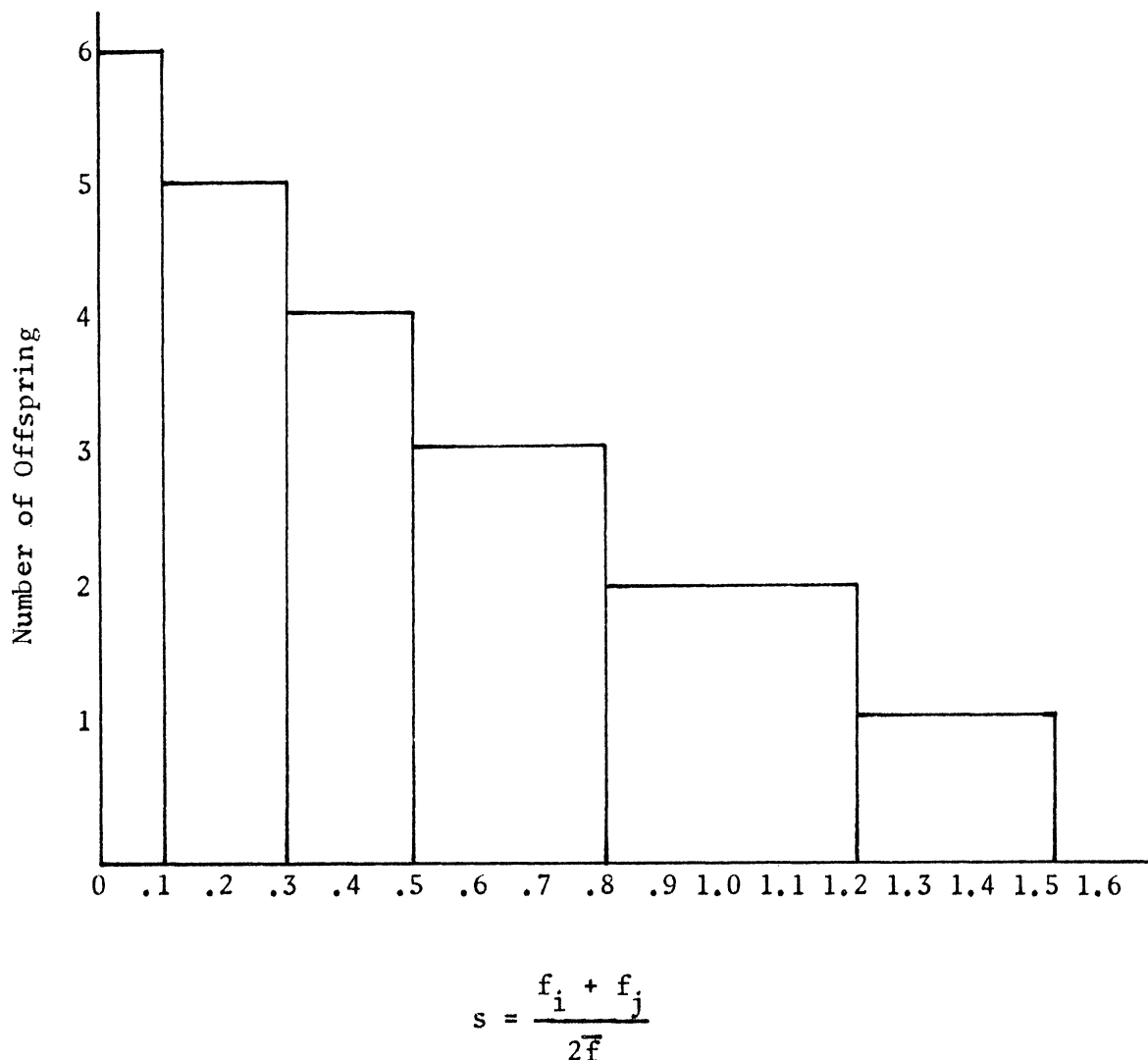


Fig. 3.6-1 Offspring Generation Function

to be produced were known, this number of cells could be eliminated and space thereby made available. However, this number is not known and therefore it is necessary to eliminate cells as the need arises.

The simplest procedure is to choose randomly those cells which are to be replaced without regard to age. The other two methods depend on the ages of the cells. In one case, the oldest cells are removed first and if there are more cells in this oldest age group than places required by the offspring, the needed places are chosen randomly.

For the other case, a probability distribution is used to specify a differential mortality for the age groups. Thus the age group from which the cell is to be eliminated is chosen on the basis of the mortality curve and the cell within this age group is chosen through a uniform probability distribution. The highest age in the mortality curve is the highest age any cell may achieve, for it is automatically eliminated at the time that an attempt is made to increase its age.

An additional refinement is not to select the cell randomly within the age group after having selected the age group on the basis of a mortality distribution. Rather, cells should be selected for elimination on the basis of their f -values so that those which are less 'fit' are more likely to be removed. This creates another site for selection forces.

As will be seen from the experimental results, where and to what degree selection acts in this model is of prime importance and must be carefully recognized and studied. It is especially important to uncover those instances of built-in selection.

3.7 SUMMARY AND COMMENTS

If it is not apparent now, it should be pointed out that an important feature of this model is the ability to distinguish between phenotypic and genotypic characters. The genotype is seen to be the pair of chromosomes and genes while on the first level the phenotype is the set of concentrations of the chemicals. The introduction of properties creates a second level of phenotypic characters. Selection acts on the phenotype, the changes in the population being manifested in the evolving genotype. Through this model, the degree of selection can be varied and the genetic constitution of the population carefully studied.

With respect to the genetic constitution of an individual, a very important concept should be introduced. This is best discussed in a chapter titled "The Unity of the Genotype" in the book Animal Species and Evolution by Ernst Mayr [9]. On page 295, Mayr writes:

The phenotype is the product of the harmonious interaction of all genes. The genotype is a "physiological team" in which a gene can make a maximum contribution to fitness by elaborating its chemical "gene product" in the needed quantity and at the time when it is needed in development. There is extensive interaction not only among the alleles of a locus but also between loci. The main locale of these epistatic interactions is the development pathway. Natural selection will tend to bring together those genes that constitute a balanced system. The process by which genes are accumulated in the gene pool that collaborate harmoniously is called "integration" or "coadaptation." The result of this selection has been referred to as "internal balance." Each gene will favor the selection of that genetic background on which it can make its maximum contribution to fitness. The fitness of a gene thus depends on and is controlled by the totality of its genetic background. [my underlining]

It is the implication of this final sentence which has been ignored in many genetic simulation models. Assigning fitness values to each allele and then selecting on the basis of a weighted sum of these values completely ignores the integrated aspect of the genotype. This model's

main strength is its recognition and implementation of this principle.

Finally, although the model of gene action being employed is somewhat classical as opposed to the newer "operon" theory of Jacob and Monod [5], it is nevertheless sufficient to carry out the proposed study. This is because the present study is not intended to be an adequate reflection of current theories of gene action, but rather an investigation of the evolutionary pathways in a closed, small population. The structure of the simulation program, however, allows models of increasing complexity to be readily substituted. The most direct extensions of the present model, which are beyond the scope of this report, are towards a description of gene action similar to the aforementioned "operon" theory. According to this theory, gene function is distributed amongst structural, operator and regulator genes with the production of proteins depending on their interaction as manifested through the use of various cytoplasmic substances.

Before concluding this chapter we list some of the main features and characteristics of this model in terms of its capacity for the investigation of

- (1) Realistic phenotypic selection
- (2) Various levels of phenotypic description
- (3) The integrated action of the gene complex
- (4) Direct interaction of the environment and the population
- (5) A wide range of environments
- (6) Dominance and the evolution of dominance
- (7) Models of gene action
- (8) Models of cellular biochemistry
- (9) Models of cell membrane action

- (10) Models for the modification and selection of crossover probabilities
- (11) Models of mating
- (12) Models of mortality
- (13) The interaction of the above models by means of a flexible computer program.

CHAPTER 4

SOME SPECIFIC ASPECTS OF THE SYSTEM

In the previous two chapters, we presented a description of the system in fairly broad terms and concluded by stressing its generality. As the title of this chapter suggests, we will be concerned with certain aspects of the overall system to the extent of pointing out difficulties which arise individually and in combination with one another. The present chapter will serve as introduction to the specific experiments to be discussed in the following chapters, but first we will point out some features of the integrated operation of this biological system.

As a background to this discussion we will briefly survey some of the work done in the simulation of genetic systems. Prominent names are Fraser [1,2], Martin and Cockerham [10], and Gill [11]. The fundamental difference between these simulations and this one is the method used to determine the phenotype. For example, in Fraser's simulations mentioned briefly in section 3.7, the value of the phenotype is a weighted sum of the values of the individual alleles. There are several components to the phenotype, a strictly additive one which in its simplest form is the sum of all the 1-alleles, perhaps modified by dominance considerations and those due to epistatic interaction which may be of quadratic, cubic or higher order. What is most important is that these are preset and cannot be altered during the course of an experiment.

To a large extent this fact limits the scope of the experiments which can be performed. There is no interaction between the population and the environment if indeed there is an environment. A new generation is formed by producing a large number of offspring, ordering them accord-

ing to their phenotypes, then selecting a subset of these on the basis of an arbitrary fitness level. The very determination of the phenotype is artificial with preset values for the alleles, of which there are in most cases only two per gene, and preset epistatic effects. The main question studied is the rate at which the fitness of the population changes with respect to the values of single parameters or sets of parameters.

But in a very important sense, the population and its evolution is very artificial. It does not correspond to the simplest type of natural population in that there is no time dependent activity, no actual environment, and no interaction between members of the population and the environment. Thus the populations defined by Fraser's system are one-dimensional and constant. Each member has a characteristic value which serves only to determine whether or not it will be a parent. In fact if its value is within the proper range it will be a parent. For such populations there is no reason to have the effect of continuity due to overlapping generations; there is no possibility of investigating such questions as the evolution of a cohesive gene complex and the evolution of linkage groups. Further it is not possible to study simple ecological situations involving the utilization of certain environmental resources by the population or the manner in which the phenotype varies as a function of the environmental state.

On the other hand, consider the many avenues open for study with a system for which the phenotype is defined in a more natural way. In such a system, the genes define a set of enzymes which control the cell chemistry. The set of chemical concentrations defines a phenotype which varies with time and which also depends on the state of the environment. Selection based on a phenotype which is not merely a transformation on a set

of alleles corresponds more closely to the situation in nature. Epistasis and dominance which are included in a quite artificial fashion in the other simulations arise here directly from the reaction graph which defines the allowable set of chemical reactions. These features are not added to increase complexity but rather are an inherent property of the model.

At the end of the last chapter, there was a list of the main characteristics of the system. One of its main virtues is the ease with which parameters can be varied and entire procedures substituted. Various kinds of reaction graphs can be constructed to stress particular modes of gene interaction. The properties of the cell membrane can be directly or indirectly a function of the genotype and such relations readily explored. One of the most important features is the mechanism for carrying out crossovers which permits their frequencies to be altered by selection pressures. This opens up possibilities for studying the factors contributing to the selection of linkage intensities. Another related question is the effect on selection of the number of crossover positions. The actual number of crossover positions is itself a simple input parameter.

There are other procedures which are easily changed and which offer great flexibility to the model. Included is the mating model which can be random or assortative as discussed in section 3.4. The maintenance of a fixed population size requires maximum ages of the members, and a probability of dying which can depend on age, fitness, chance, or a combination of these.

It seems worthwhile at this point to consider the overall functioning of a population modeled by this system. (We will pursue individual topics in considerably more detail subsequently.) A specific reaction graph is developed and a corresponding set of enzymes constructed. Each

individual enzyme or small subset of enzymes is referenced by a single allele, the complete genotype defining a set of enzymes, not necessarily the complete set. There is great flexibility in the choice of reaction graphs and in the organization of the genotype.

These two form the basic definition of the population and now a large number of additional parameters and procedures must be specified including, for example, the following:

- (a) The mating procedure
- (b) The offspring generating function
- (c) The initial and final population sizes
- (d) The number of biochemical time steps per generation
- (e) The maximum age of a cell and the means for controlling the population bound
- (f) The proportion of the population permitted to mate per generation
- (g) Some parameters associated with the chemistry, namely, the initial concentrations within the cells, the subset of chemicals allowed to permeate and their permeation constants
- (h) The procedure for assigning the initial concentrations to offspring

These points can be better appreciated if we outline a typical simulation over a period of one generation.

Many of the simulations carried out here begin with a population of about two hundred cells whose genetics have been generated randomly. This is also the upper bound in many cases. The reason for beginning with a full population is to provide the maximum amount of genetic variation

from the outset. It is intuitively clear that a major problem associated with small, closed populations is the relatively rapid loss of alleles due to the effects of random sampling. This loss would certainly be accentuated in a very small population. Since at this time we are not concerned with the growth of a population, the benefits of starting with the maximum allowable size are clear. Each cell has the same set of chemical concentrations, the same set of chemicals which can permeate and for the early simulations the same fixed set of permeation constants. In these early simulations permeation is carried out on the basis of equation 3.3-1. In a single biochemical time step each cell in turn undergoes the chemical changes brought about by the enzymes defined by its genotype. The operations carried out in each time step have been described in Chapter 2. After each cell has undergone the prescribed chemical reactions, the changes due to the permeation of the chemicals are computed. When these computations have been completed for every cell, the new distribution of chemicals for the environment is calculated.

After a preset number of biochemical time steps has been carried out, mating takes place. This period of biochemical activity serves the two purposes of providing interaction between the cell and the environment and allowing sufficient time for the chemical concentrations in the cell to reflect the genetics of the cell. At the outset each cell has the same chemical constitution and it will take a number of biochemical time steps to produce a set of chemical concentrations which is representative of that cell's individual genetics. Thus the actual number of such time steps is linked to the chemical dynamics and the rate at which significant changes occur in the chemical concentrations (see Section 4.3).

The number of matings which occur are proportional to the number of

cells in the population at the time of mating. A preset parameter called the mating fraction is used to multiply the number of cells in the population and yield the number of matings to be carried out. For a value of this parameter of .1 and 200 cells, the number of matings will be 20 involving 40 cells. We have previously discussed various types of mating rules and regardless of which is used a pair of parents eventually is determined. On the basis of the combined fitness of both the parents, a number of offspring is assigned to them. This is a very important point in that it is not generally the case in mathematical genetic analysis that the production of offspring is dependent on the fitness of two individuals. Each individual has a fitness which is a function of its reproductive and mortality properties and future state of the population is determined on the basis of the fitnesses of the individual members. It is a very important feature of this simulation that the offspring at any generation are determined on the basis of the pairs of cells which have been selected for mating. And we point out once again that the fitnesses of the parents themselves are phenotypically based and are determined through the time dependent chemical reactions controlled by their genotypes.

The genetics of the offspring are assigned by means of the usual mechanism, one gamete being obtained from each parent. The structure of the gamete itself will depend on whether or not the individual genes on the parents' chromosomes are strongly linked. Whatever these conditions are, the production of gametes is a straightforward process. The initial chemical constitution of the offspring must now be specified. The form of this chemistry, as we have stated previously, is open to experimentation. Later in this chapter we will discuss several such forms and outline some of the difficulties which arise.

Before the offspring can become members of the population, other cells must be removed to make place for them at those times for which the population is at its full size. The basis for removal of cells has been mentioned previously. After all the matings have been completed, each cell is aged by one unit except for the newly formed offspring which start at age zero. Any cells which exceed the maximum age are removed at this point. The population now undergoes the next stage of biochemical development.

The organization of the simulation into successive periods of mating and what we will call development parallels in a certain fashion the activity of many animal populations. Mating takes place at fixed intervals in the year and otherwise the animals are involved in the process of growth and survival. In this system offspring are not permitted to mate in the same mating period in which they are produced, although if this is desired it can be readily implemented. However, they are allowed to mate at the next earliest opportunity. Another restriction is that cells can mate only once during any single mating period. This is not really severe for even in a population of 100 cells, the probability of choosing the same cell twice is only one in ten thousand. If a cell survives it has the opportunity to mate in subsequent generations.

The use of the term generation in this context means that period consisting of development and mating. Similar to most natural populations, the generations overlap and many cells of age greater than zero coexist with new offspring. In many discrete time analyses, overlapping generations are not allowed which means that all the cells of the previous generation are replaced by offspring. The preservation of a measure of continuity between generations is another important feature of the present model.

We will now discuss in some detail a few of the specific issues raised.

4.1 REACTION GRAPHS AND THE GENOTYPE

In section 3.2, the term reaction graph was defined and an example given. In this chapter we wish to examine some properties of reaction graphs and their relationship to the structure of the genotype. The earlier discussion suggested that reaction graphs could be constructed by building up words from a small alphabet. For example, given an alphabet of three letters, the total number of words from length one to length five is fifty-five. Each forward chemical reaction corresponds to the formation of a word from two shorter ones and there are 180 such constructions.

For instructive purposes we will use this reaction graph in an experiment for a population whose genetics consists of ten genes and seven alleles per gene. There will be fifty-five chemicals with 180 forward reactions and 180 corresponding reverse reactions. Among the seventy different alleles which can be used in the genotype all the 360 forward and reverse reactions are represented. Of course any particular genotype contains only twenty alleles which can define on the average about one hundred reactions. The purpose of this experiment is to study the range of phenotypes produced with such a reaction graph and such a genotype organization.

The phenotype for this experiment is defined by a function f of the form described in section 3.4. In this case it consists of eighteen chemicals and the sum of the preset concentrations in the function is about .5, i.e., one-third of the chemicals is required to have about one-half of the chemical concentration. The initial population consists of twenty-five cells, each having a uniform chemical concentration distribution.

After forty biochemical time steps during which time no mating occurs, the cells are examined with respect to their chemical state.

It is found that the chemical concentrations of all the cells are very much alike. This results in the values of f being very similar across all the cells. Therefore, selection based on the positive assortative scheme discussed previously will not have sufficient phenotypic differentiation to be operatively effective. Some reasons for this situation based on the form of the reaction graph will now be discussed.

Given a word of the maximum length five, there are at least four different ways to produce it. This suggests that there is probably too much 'parallelism' or redundancy in the reaction matrix. By this it is meant that there are so many alternate paths connecting the various chemicals that many sets of enzymes defined by different genetics will produce very similar concentration distributions especially if the enzyme rate constants are very much alike.

There are at least two remedies for this situation. The first is to design a reaction graph with the parallelism greatly reduced. This is simply accomplished by removing a large number of reactions. For example, allow a five-letter word to be produced in only one or two ways instead of a maximum of eight. Such reaction pathways would be more directed and much more similar to actual metabolic pathways. The second remedy is to use a highly parallel reaction graph but specify through the genotype only a small subset of the allowable reactions. We will discuss an experiment to test this later in this section but first some further comments on the construction of the genotype.

In terms of this model, the genotype is simply the set of genes for a cell. The term set is quite appropriate, for the genotype is literally

a collection of genes, and their relative linear position is arbitrary. This is not the case in nature where the genotype in any living organism is the product of a long history of evolution, resulting in a cohesive well-integrated complex. In this artificial system, the functions of the alleles are arbitrarily specified with little a priori knowledge as to how they will interact. It is hoped that some measure of cohesion will arise through selection pressures on the population.

Two alternatives have been considered in the choice of alleles at a given locus. To specify a large number of reactions by a genotype with few genes it is necessary for each allele to define a set of reactions which differs from that defined by the other alleles. In those cases for which the number of reactions is relatively small it is possible for all alleles at a locus to define the same set of reactions, the alleles differing only in the rate constants associated with these reactions. Later experiments will involve models in which the genotype is based on a compromise of these schemes. Finally, it should be noted that it will usually be the case that an allele must define a set of reactions rather than a single one in order to limit the number of genes in the genotype.

In this model, the organization of the genotype is a very crucial element and much flexibility is available in the selection of genotypes to carry out specific tasks. However, in all cases there are complex interactions present among the functions of individual alleles as manifested through the reaction graph. It is very unlikely that optimum genotypes can be selected a priori on the basis of a knowledge of the reaction graph and the fitness function.

Now let us examine the possibility of increasing chemical differentiation among the cells by reducing 'parallelism' in the reaction graph.

The reaction graph being used is a subgraph of the one described in Section 4.1. It consists of all the reactions involved in constructing words of length two and three from the alphabet of three letters. This results in a chemical system of nineteen chemicals with twenty-four forward and twenty-four reverse reactions. The reaction graph and reaction matrix are shown in Fig. 4.1-1 and 4.1-2.

Three different models for the genotype will be tested. These will differ primarily in the average number of chemical reactions which are defined. That is, for each model, the reaction graph will be the same but a typical genotype will define a substantially different number of chemical reactions among the set of forty-eight. The table below sums up the distinctions among the models.

MODEL	Number of Genes	Number of Alleles	Mean Number of Reactions
1	5	5	33.6
2	3	5	14.6
3	5	8	11.5

Table 4.1-1
MODELS 1, 2, & 3

Chemicals 4, 5 and 6 are allowed to permeate with each permeation constant fixed at .05. Only chemical 14 is used in the f-function which has the form

$$f = (.074 - x_{14})^2 .$$

Each model is tested with a population of twenty cells with different genetics for each population. The main point of comparison is the distribution in the values of f for each model after thirty biochemical time steps have been carried out, during which no mating is allowed. The

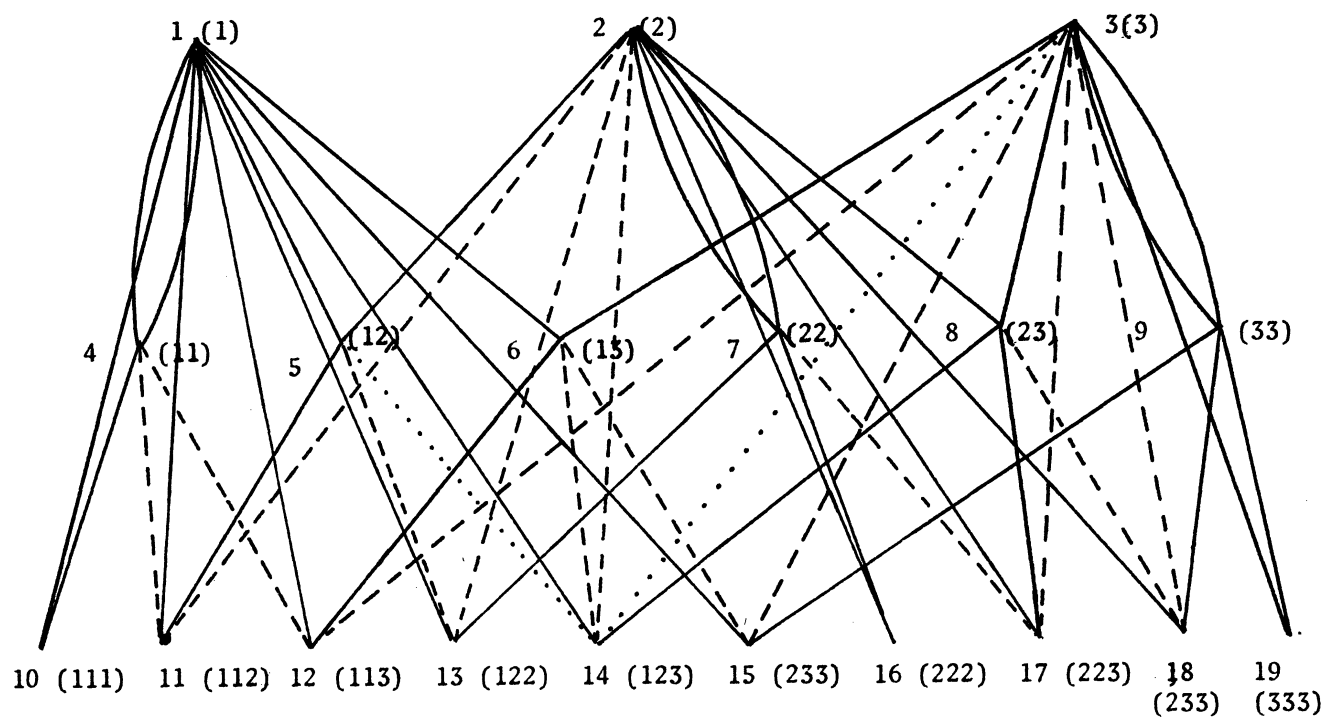


Fig. 4.1-1 Reaction Graph.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	4	5	6	10	11	12	13	14	15										
2		7	8	11	13	14	16	17	18										
3			9	12	14	15	17	18	19										
4				0															
5					0														
6						0													
7							0												
8								0											
9									0										
10										0									
11											0								
12												0							
13													0						
14														0					
15															0				
16																0			
17																	0		
18																		0	
19																			0

Fig. 4.1-2 Reaction Matrix

results are shown in Table 4.1-2. It is evident that the variation in values of f increases markedly from model 1 to model 3. Because the genetics are different in each population, cells of the same number across the models should not be compared. Instead, one should compare the range of values across all the cells of each model.

Cell No.	MODELS		
	1	2	3
	$\times 10^{-3}$	$\times 10^{-3}$	$\times 10^{-3}$
1	.4273	.3744	2.7210
2	.3962	.4190	.6080
3	.3977	.4103	.1836
4	.3843	.3793	.4057
5	.1713	.3723	1.0850
6	.2945	.0406	.2320
7	.5248	.3644	.3910
8	.3795	.0941	.6060
9	.1729	.2309	.0032
10	.3990	.3649	.3740
11	.4258	.3955	.1700
12	.1776	.1639	.4000
13	.3984	.3853	.4170
14	.3970	.0487	1.2190
15	.4012	.3894	.1700
16	.3656	.1657	.0040
17	.3986	.1782	2.2740
18	.4084	.3930	.4030
19	.4267	.3720	.3860
20	.4090	.4118	.9540

Table 4.1-2. Comparison in the Values of f for the Three Models

Thus it is possible to induce greater differentiation among the cells by having the genotype define a small subset of the set of all reactions. In some sense, however, this seems to be a rather artificial way of avoiding the problem. It would seem to be more reasonable for each cell to have the possibility of carrying out all the reactions which are allowed. Now this set should contain very little of the parallelism discussed previously.

The major difference among genotypes will be the rate constants which are associated with the enzymes. Thus two genotypes may define the same set of reactions but if the set of rate constants for these differ, the resulting distribution of chemical concentrations will be quite dissimilar. In the next section we will discuss in some detail a model which has been developed along these lines.

It is an important feature of this system that various reaction graphs and genotype organizations can be easily substituted. In fact, both are defined in the system by input parameters and do not require program alterations. Although the reaction graphs discussed in this section have been based on a word-alphabet scheme it is not necessary that this be the case as the model discussed in the next section will illustrate. The alphabet scheme was used mainly for the ease with which it can be described.

4.2 THE DEVELOPMENT OF A MODEL

Based on the discussion of the previous section, a reaction graph is proposed together with a particular genotype organization. These are described in some detail in conjunction with several other issues. For easy reference these are listed below:

(1) A more complete discussion of the chemical reactions themselves especially the constant k which occurs in equation 2.2-3

namely, $\lambda^{t'} = kM^{t'}$

where $\lambda^{t'}$ is the number of collisions in the time interval and $M^{t'}$ is the total number of colliding molecules.

(2) A discussion of the implementation of the crossover mechanism and the means developed for studying it.

(3) A comparison of a few offspring generating functions.

(4) A discussion of some of the difficulties associated with various procedures for assigning the initial chemistry of offspring.

We first turn to an examination of the reaction graph. The chemistry is specified by the reaction graph and matrix shown in Fig. 4.2-1 and Fig. 4.2-2, respectively.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
1	0	7	0	9																					
2		0	8																						
3			0	0	10																				
4				0	0	0	12																		
5					0	11																			
6						0	0	0	0	0	0	0	0	16											
7							0																		
8								0	0	13	0	15													
9									0	0	14														
10										0	0	0	0	17											
11											0	0	0	0	0	0	0	0	0	20					
12												0	0	0	0	0	0	0	0	0	0	0	23		
13													0	0	18										
14														0	0	19	0	22							
15															0	0	0	0	0	0	25				
16																0	0	0	0	24					
17																	21								
18																		0							
19																			0						
20																				0					
21																					0				
22																						0			
23																							0		
24																								0	
25																									0

Fig. 4.2-2 Reaction Matrix

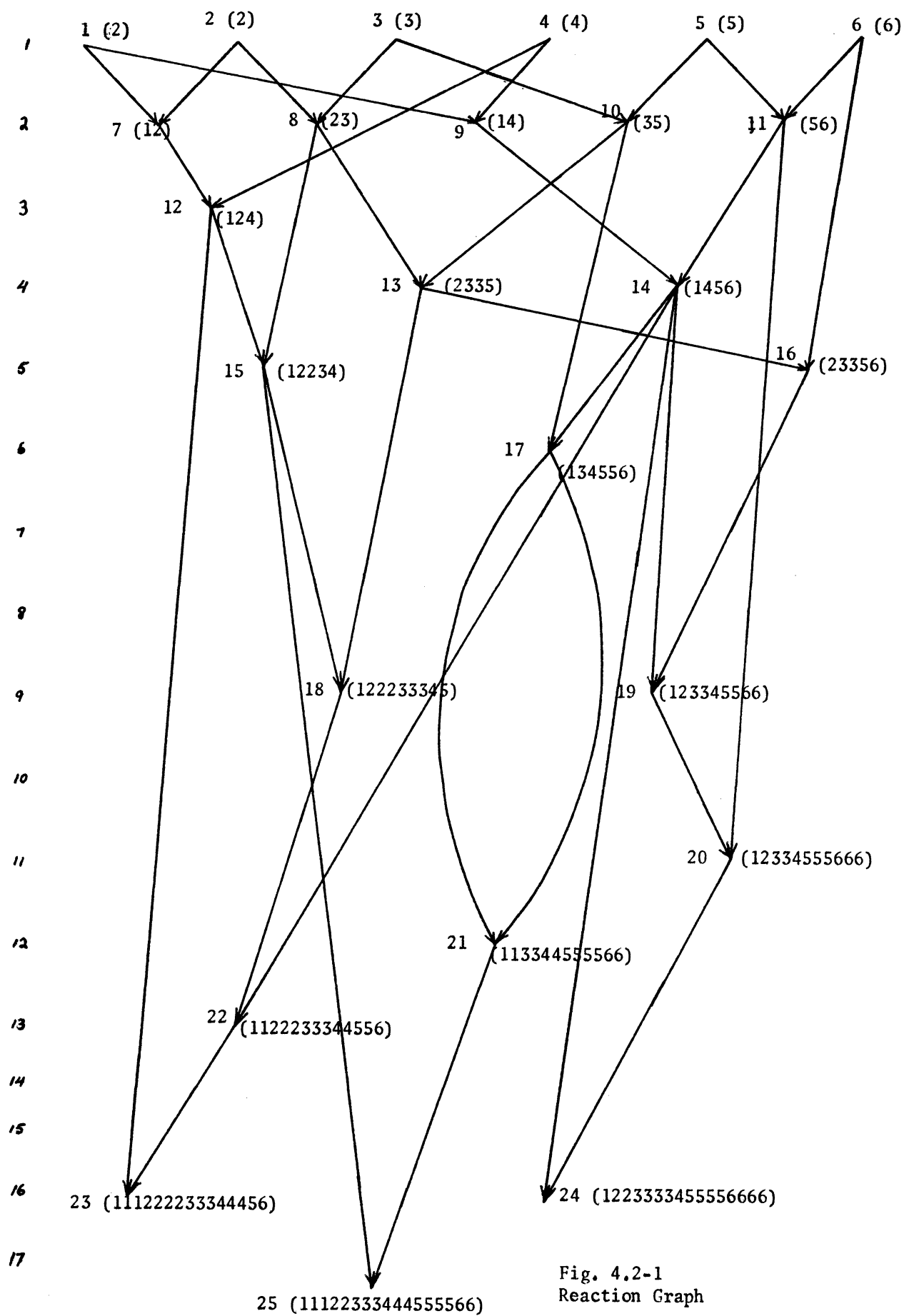


Fig. 4.2-1
Reaction Graph

There are twenty-five chemicals and nineteen reactions among them. A comparison of this graph with the one depicted in Fig. 4.1-1 is revealing. Except for the primary six chemicals of this model, which are not produced by a composition reaction, each chemical is produced in only one way. There are some relatively 'heavy' chemicals, especially 23, 24 and 25 which consist of 16, 16 and 17 primary chemicals respectively.

Although the chemicals in this graph are expressed as strings of primary chemicals, the graph itself was not obtained by constructing all the possible productions of words up to length seventeen and then pruning most of them. This would have been a nearly impossible task considering the very large number of such productions. The actual procedure was to begin with a small number of chemicals, with a bound of approximately twenty-five chemicals overall and with a goal to produce a graph with considerable depth and very little breadth. In this context the terms depth and breadth refer respectively to the number of levels and the number of different chemicals in each level. There are seventeen levels but at most six chemicals per level. The top level consists of the six primary chemicals which are transformed by a set of nineteen reactions to the three heaviest chemicals at the lowest levels. There is a definite sense of directedness in the present reaction graph.

The complete enzyme table is given in Table 4.2-1. The first five genes deal with composition reactions and the remaining four with the decomposition ones. In order to discuss this table, consider gene 1. The alleles are divided into two groups 0 to 3 and 4 to 7. All the alleles in the first group define the same pair of reactions $1 + 2 \rightarrow 7$ and $14 + 16 \rightarrow 19$ but are distinguished by the rate constants. These are shown to be .136, .202, .301 and .448 for alleles 0, 1, 2, and 3

A L L E L E S

G
E
N
E
S

	0	1	2	3	4	5	6	7
1	1+2→7 14+16→19				6+13→16 15+13→18			
	.136	.202	.301	.448	.136	.202	.301	.448
2	1+4→9 16+20→24				5+6→11 15+21→25			
	.136	.202	.301	.448	.136	.202	.301	.448
3	2+3→8 14+18→22				10+14→17			
	.136	.202	.301	.448	.136	.202	.301	.448
4	8+12→15 4+7→12				3+5→10 8+10→13			
	.136	.202	.301	.448	.136	.202	.301	.448
5	12+22→23 17+17→21				9+11→14 11+19→20			
	.136	.202	.301	.448	.136	.202	.301	.448
6	15→8+12 16→6+13 25→15+21				14→9+11 24→16+20			
	.01	.015	.0223	.0362	.01	.015	.0223	.0362
7	7→1+2 8→2+3 18→13+15				10→3+5 11→5+6 21→17+17			
	.01	.015	.0223	.0362	.01	.015	.0223	.0362
8	13→8+10 23→12+22				9→1+4 19→14+16			
	.01	.015	.0223	.0362	.01	.015	.0223	.0362
9	12→4+7 22→14+18				17→10+14 20→11+19			
	.01	.015	.0223	.0362	.01	.015	.0223	.0362

Table 4.2-1 Enzyme Table

respectively. In fact for all the composition alleles, the same rate constants are used. The second group of alleles, 4 to 7, define the reactions 6+13→16 and 15+13→18. A similar description applies for the decomposition reactions but the rate constants are much lower. For this model, a genotype may define the complete set of reactions and differ from another such

genotype in the associated set of rate constants. Homozygosity at any locus implied the inability to carry out a specific set of reactions which may be detrimental in certain situations.

The only chemicals which permeate are the primary ones, namely 1 to 6. Thus the cells are only allowed to receive the lightest chemicals from the environment as raw materials in the production of the heavier ones. As in the previous experiments, the permeation constants are fixed.

This completes the description of the reaction graph and the enzyme table. They have been constructed on the basis of the concepts developed in section 4.1. We turn now to a consideration of the first point in the list given earlier.

In the derivation of the chemical kinetic model, one of the sufficient conditions for the number of molecules to be non-negative is that the constant k in equation 2.2-3 be less than or equal to 1. All the experiments to follow will use a value of k greater than one. The main purpose is to cause the reactions to proceed at a faster rate so that the general trend of chemical concentrations for the cells will be more quickly achieved. This will produce a reduction in the number of biochemical time steps resulting in a significant saving in computer time. The simplest way to determine the value of k to be used is through a trial and error procedure.

Starting with $k = 1$, the values of k are successively increased until the occurrence of negative concentrations. Then a slightly smaller value is tested for a small population over several generations and if no negative concentrations arise this value is used. The occurrence of a negative concentration would cause an immediate halt in the experiment unless special precautions are taken. For most experiments, this limits k to a value no greater than ten and prevents even greater savings in computer time.

However, a method is available which would allow larger values of k .

This method is based on the assumption that the chemical concentrations resulting after one chemical time step using a particular value of k are very much the same as those after two chemical time steps using a value of k which is one-half of the previous value. It is impossible for these two sets of concentrations to be equal because the chemical equations are non-linear and therefore superposition would in general not hold. It is not important that the concentrations are not identical but what is important is that the values of the fitness functions be very similar. Some tests have shown that in most cases these values are in fact very close.

Using a large value of k , about 20, will result in a somewhat higher rate of occurrences of negative concentrations. When a negative concentration occurs in a particular cell, then for that cell k is reduced to 10 and two time steps are carried out in place of the one which produced the negative concentration. Provision is also made for the possibility that a negative concentration occurs for $k = 10$ in which case four time steps are carried out with $k = 5$. A further negative concentration in one of these time steps causes the program to halt.

With $k = 20$, this latter condition did not arise in any of the present set of experiments. Using this high value of k it is now possible to reduce the number of chemical time steps from the usual value of ten to four. That is, four chemical time steps with $k = 20$, produce a change in the chemical concentrations of a cell comparable to that produced during ten time steps with $k = 8$.

The original analysis in Chapter 2, set bounds on parameter values associated with the chemical reactions. These were based on the assumption of a set of conditions which would produce the most extreme result. In any given case such a coincidence is unlikely and this permits certain

parameters to exceed their original bounds. The new limits must be determined experimentally but with the procedure outlined above there is great latitude.

We will now consider the implementation of the crossover mechanism described in section 3.1. Therein the concept of the allele as an ordered pair was introduced. We will refer to x_{ij} as the crossover level associated with the allele at gene j of chromosome i . If at gene i , the alleles (a_1x_{11}) and (a_2x_{21}) are present the probability of crossing over between genes 1 and 2 is a function of x_{11} and x_{21} . In the simplest case, the dependency will be on $x_{11} + x_{21}$, the sum of the levels. Then $p(x_{i,j} + x_{i,j+1})$ is the probability of crossing over between genes j and $j+1$ where p is an arbitrary function.

The crossover levels are generated randomly for each allele by a uniform distribution. For most of the experiments to be discussed there are eight levels running from 0 to 7. Therefore the sum of the levels at any position takes values from 0 to 14. Because the individual levels are each generated by a uniform distribution, the sum will have a triangular distribution with the maximum probability at 7. This will be the initial expected distribution at each crossover position.

We will briefly discuss an experiment to illustrate the manner in which the distribution of crossover probabilities is used. For this experiment, the function p is given by the following table:

Sum of the Crossover Levels	Crossover Probability
0	.000
1	.005
2	.010
3	.025
4	.050
5	.100
6	.125
7	.150
8	.200
9	.225
10	.250
11	.325
12	.375
13	.450
14	.500

Table 4.2-2
Crossover Probabilities

The lowest probability is 0 which means complete linkage, and the highest is .5 which means equal likelihood of crossing over or not crossing over. Crossover is permitted at four positions between genes 2 and 3, 4 and 5, 6 and 7, and 8 and 9. Further detail at this point is not important and so we now consider the four diagrams of Fig. 4.2-3.

The solid line curves represent the crossover distribution at each of the four positions of the initial population. As we stated previously this initial distribution should be triangular, but it is only in (b) that this is nearly the case. The deviation is especially noticeable in (d). Subsequent experiments will involve attempts to affect the direction and evolution of the crossover probability distribution at designated positions. When they are described, additional detail will be presented about the theory and implementation of the crossover mechanism. At present we wish to comment on some of the characteristics of the curves of Fig. 4.2-3.

In Fig. 4.2-3a, there is a marked shift in the distribution towards the lowest probabilities and this is the only clear-cut case of such a response. This suggests that tighter linkage between the gene sections 1-2 and the remainder of the chromosome is occurring although further experiments would be necessary to verify this conjecture. At position 4-5, the initial triangular distribution remains essentially the same and at position 6-7, the initial distribution which is roughly triangular shifts slightly towards the higher probabilities. In Fig. 4.2-3d, the initial distribution, weighted towards the higher probabilities, is very much non-triangular and tends toward a more triangular distribution at generation 15.

In addition to such graphs, statistics on the ratio of actual to potential crossovers for each generation would reflect the tendency at

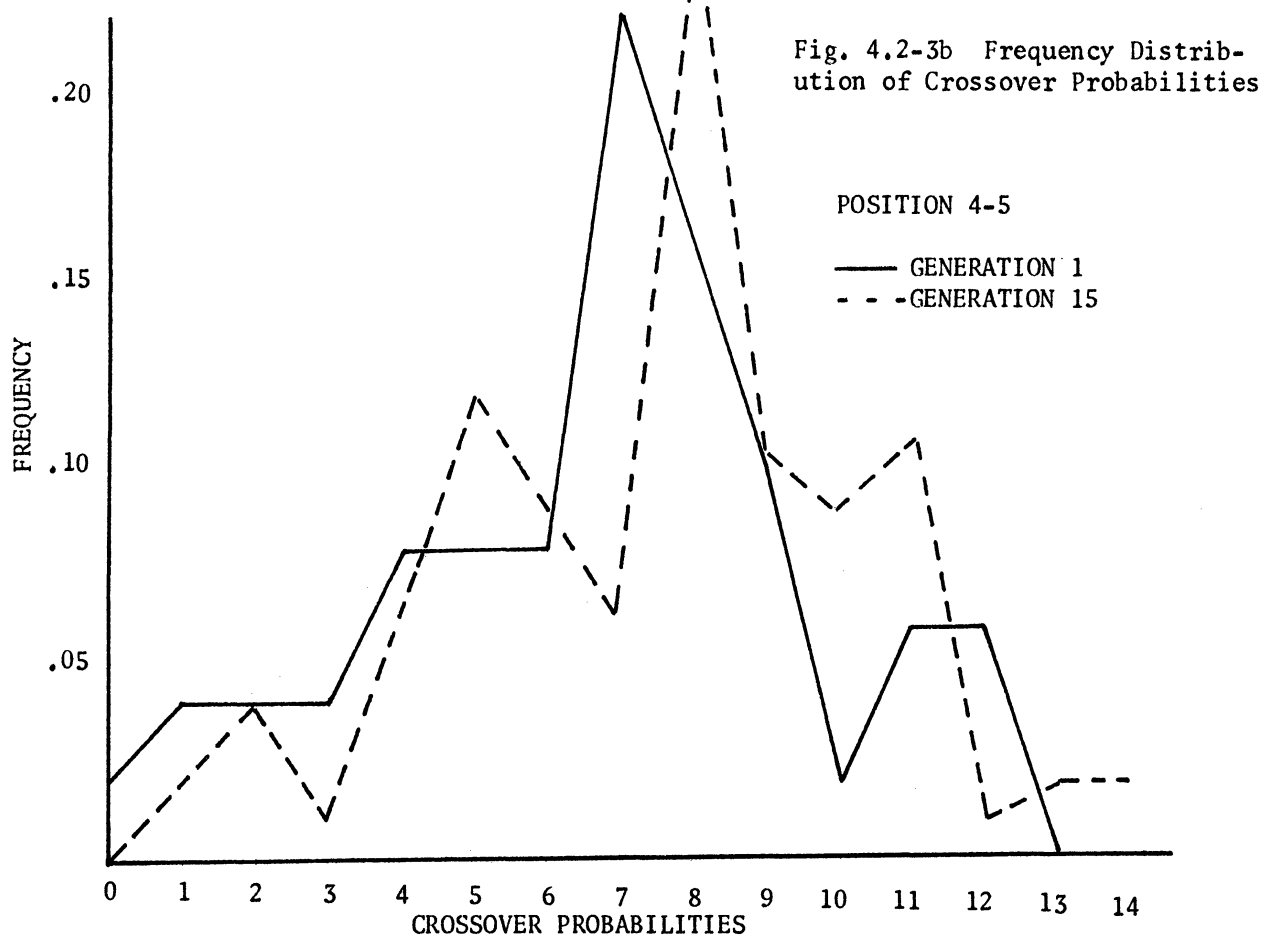
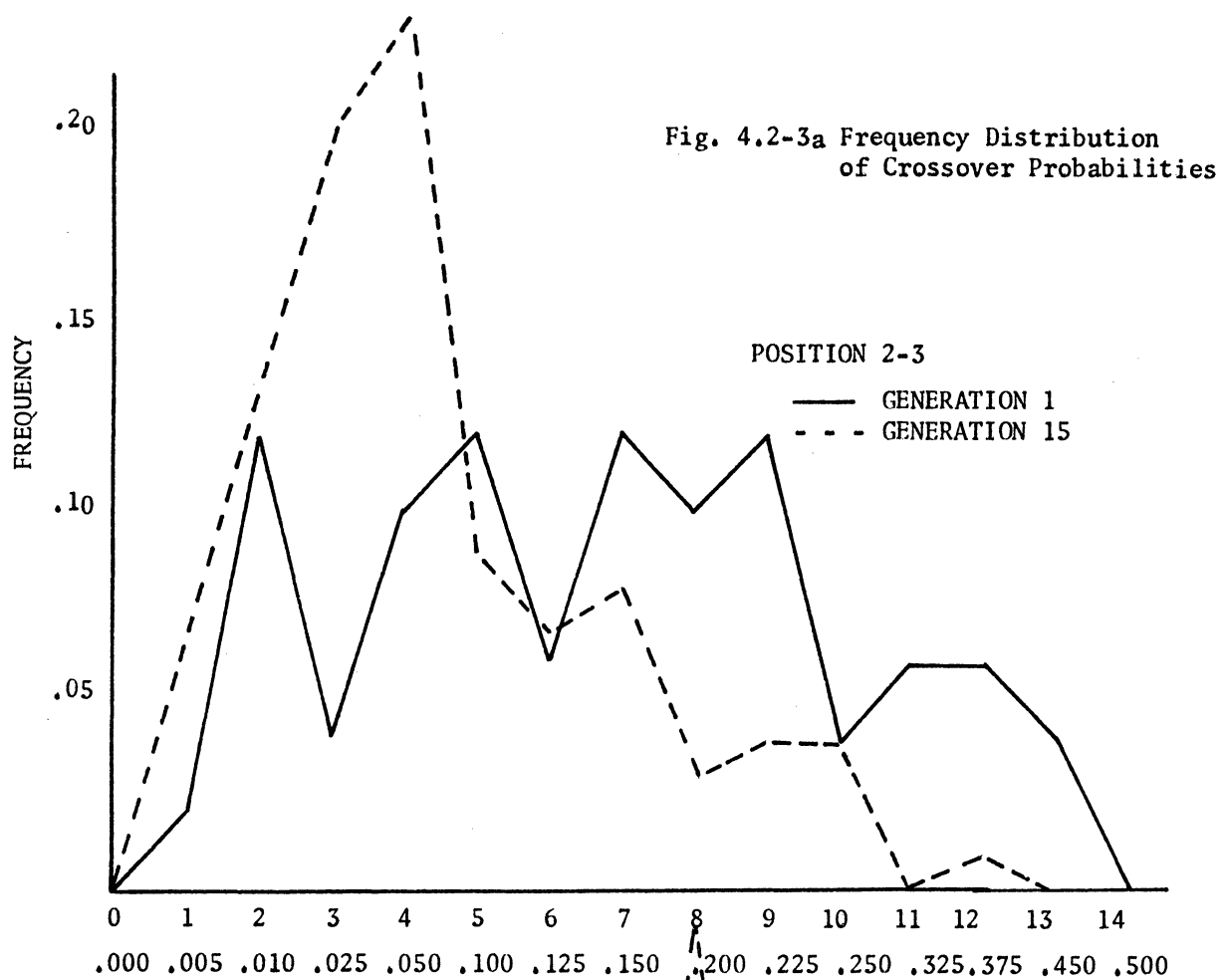


Fig. 4.2-3c Frequency Distribution
of Crossover Probabilities

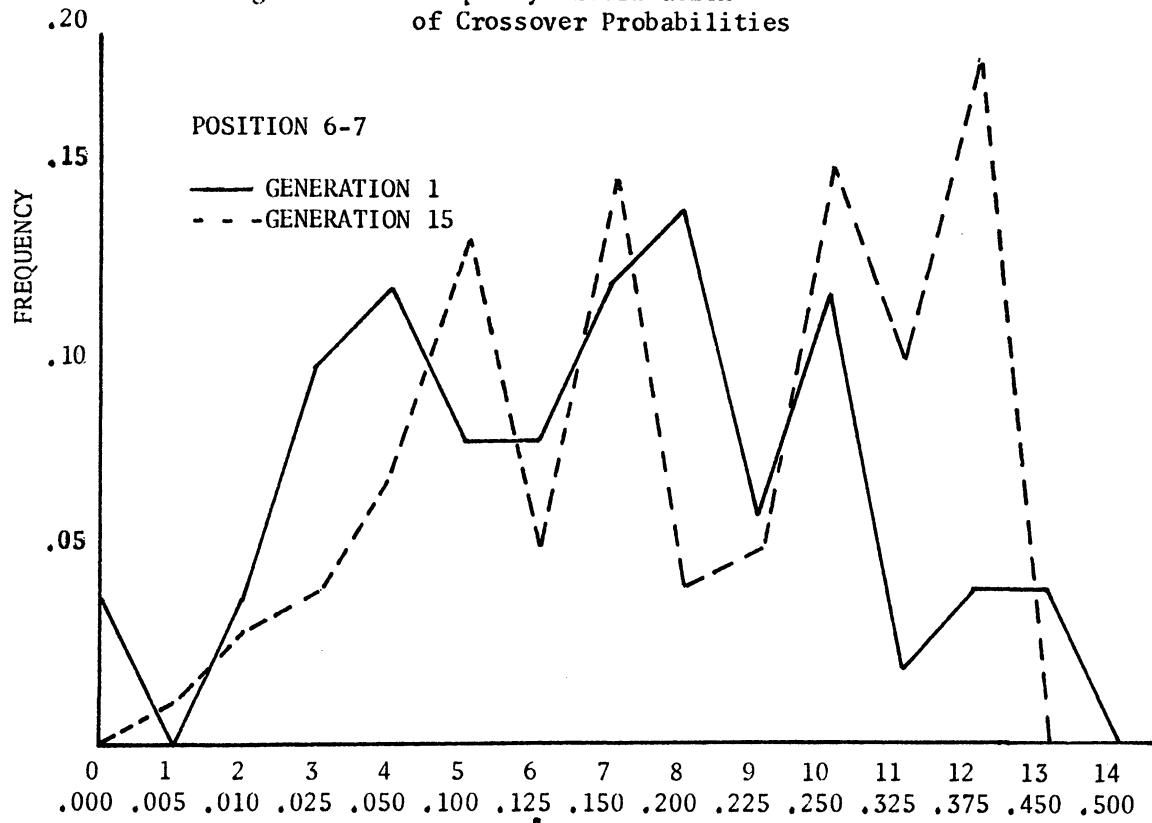
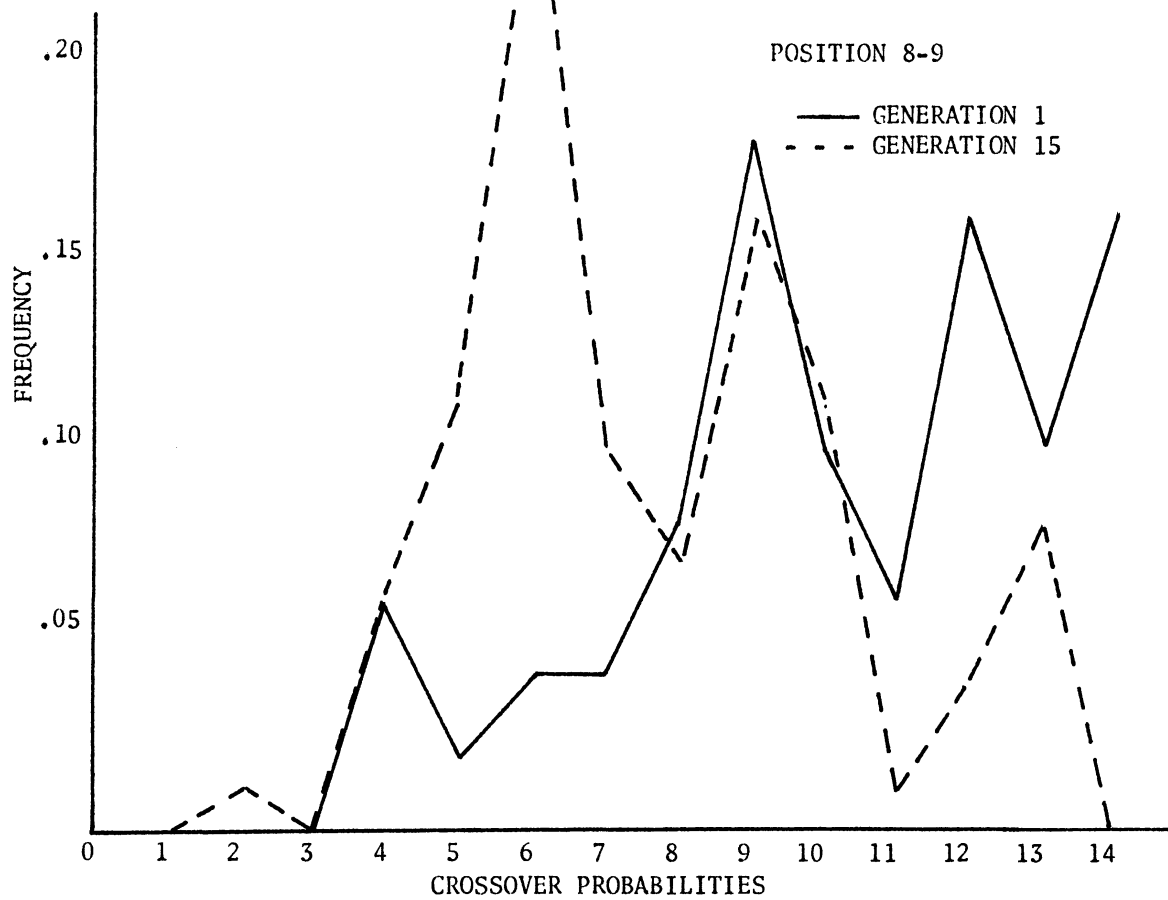


Fig. 4.2-3d Frequency Distribution of
Crossover Probabilities



each position. As we mentioned earlier the possibility of inherited crossover levels and thereby the selection of crossover probabilities introduced into this system an entirely new dimension which is lacking in other simulations.

The third point deals with an examination of various offspring generating functions. The theoretical form of many of these functions was given in section 3.6. The variable s on which the number of offspring to be assigned depends was defined as

$$s = \frac{f_i + f_j}{2\bar{f}}$$

where f_i and f_j are the values of the fitness function and $\bar{f}$ is the mean value over the population. Two important aspects are embodied in the form of s . As we stated earlier, the number of offspring assigned to a pair of parents depends on the fitnesses of both the parents. Also with the inclusion of $\bar{f}$, the relative improvement of the population is taken into account. That is, an improvement in the population as a whole requires potential parents to reflect this improvement if they are to have a large number of offspring.

It is necessary to experiment with the parameters of the offspring function to produce an offspring frequency distribution which has a reasonable form. If it were possible to know the distribution of the fitness values in the population, the offspring function could be more easily constructed but this is not the case. Consider the following set of offspring functions:

	Offspring Function				Number of Offspring
	A	B	C	D	
s ₁	.5	.2	.2	.1	6
s ₂	.7	.35	.5	.3	5
s ₃	.85	.5	.8	.5	4
s ₄	.95	.65	.95	.8	3
s ₅	1.0	.8	1.05	1.1	2
s ₆	1.1	.95	1.15	1.4	1
	otherwise				0

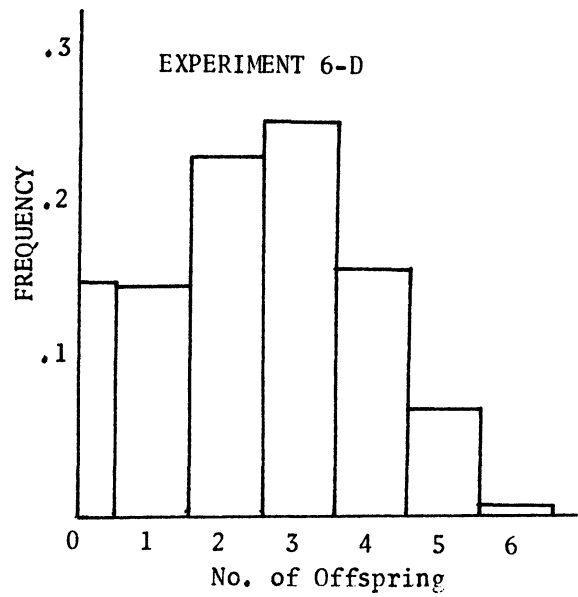
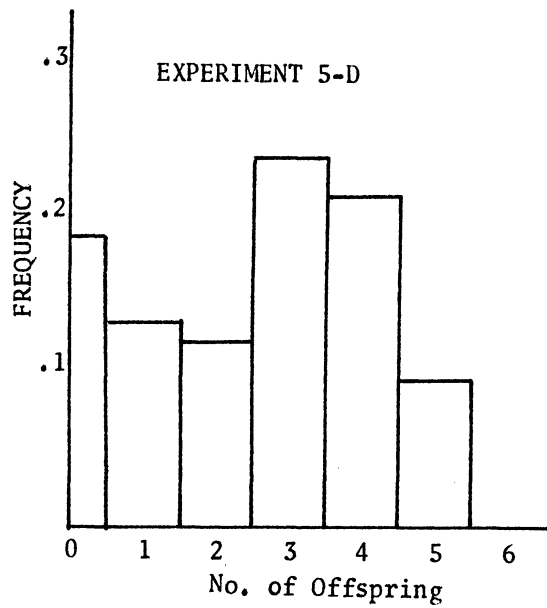
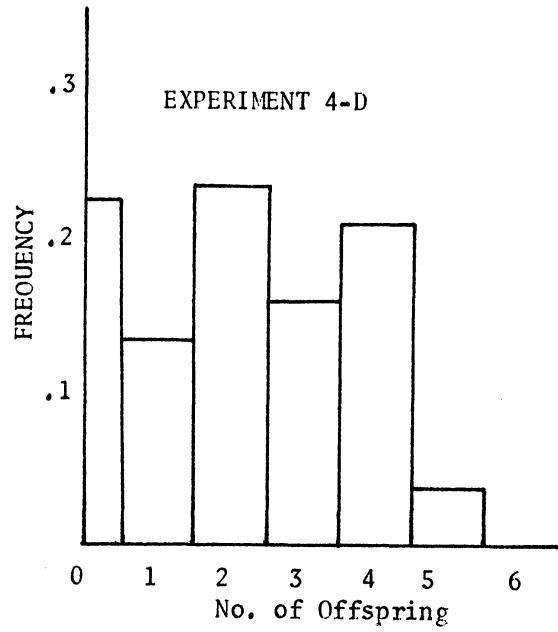
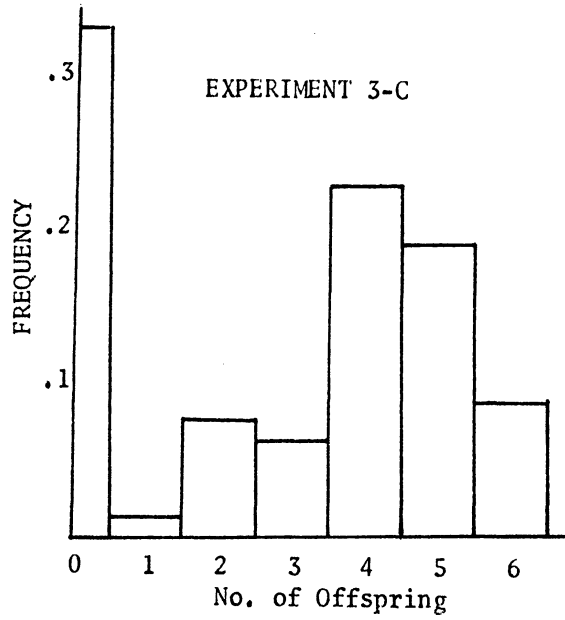
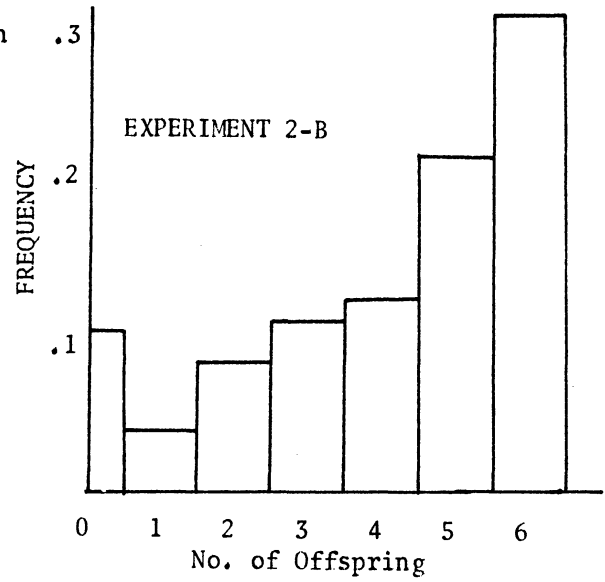
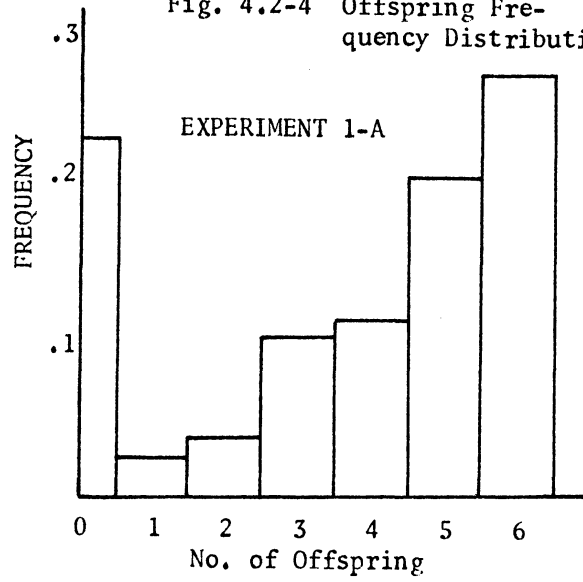
Table 4.2-3 Offspring Generating Functions

The number of offspring can vary between 0 and 6. The variation in parameters between functions A and B will be discussed in terms of the related offspring distributions. The final offspring distributions of a few experiments are shown in Fig. 4.2-4. In experiment 1, in which offspring function A is used the frequency of zero offspring is much too high as is the frequency of six offspring. This suggests function B which has the effect of reducing the frequency of zero offspring but no effect on the frequency of six offspring. Various changes in the experiments require new offspring functions until finally the use of function D with experiment 6 produced a distribution which is weighted towards the middle.

Whenever conditions in an experiment change it may be necessary to use a different offspring function. Care must be taken to determine the proper one for any experiment.

The final point deals with the question of how to assign the initial chemistry of offspring. This is best examined by the discussion of several schemes and their consequences. The subroutine of the program which assigns the initial concentrations is called NEWCON and so we will refer

Fig. 4.2-4 Offspring Frequency Distribution



to these schemes as NEWCON -1, -2, and -3, respectively. NEWCON-1 is constructed in the following way.

Let $\bar{M}$ be the mean number of molecules per cell over the whole population and r and q , integers. The integer q is the cell number of one of the parents chosen in an arbitrary fashion. Then M_i , the number of molecules of chemical i which are contained in each offspring produced by a pair of parents, one of which is q , is

$$M_i = \frac{\bar{M}}{25} \left(\frac{r-1}{r} \right) + M(q) X_i(q) \left(\frac{1}{r} \right)$$

where the number of different chemicals in the cell is 25, and

$\bar{M}(q)$ is the mean number of molecules per cell, i.e., $\bar{M} = \frac{1}{N} \sum_{i=1}^N M(i)$

$M(q)$ is the number of molecules in cell q ,

$X_i(q)$ is the concentration of chemical i in cell q .

The source of these chemicals is the environment and justification for this has been outlined in Section 3.4. The integer r , an input parameter, controls the degree to which the initial offspring chemistry is uniform across all the offspring of any generation as opposed to reflecting the chemical constituency of one of its parents. For a value of $r = 1$, the chemistry of the offspring is an exact replica of the chemistry of one of the parents, whereas for r a large number, it is largely dependent on the mean number of molecules per cell of the parental generation. For example, let $r = 10$. Then the number of molecules of chemical i in the offspring is

$$M_i = \frac{\bar{M}}{25} \left(\frac{9}{10} \right) + M(q) X_i(q) \left(\frac{1}{10} \right) .$$

The main contribution is from the first term and therefore the chemical

concentrations within the offspring will be very similar for all offspring of the same generation.

The fitness function used in a test experiment with NEWCON-1 is

$$f_1 = (.09 - x_{23})^2 + (.10 - x_{24})^2 + (.11 - x_{25})^2$$

where x_{23} , x_{24} , x_{25} are the concentrations of chemicals 23, 24, and 25 respectively.

At the outset, each cell contained one million molecules overall but after thirty-five generations in this experiment, this number was reduced to about eighty-six thousand. This drastic reduction changes the basic nature of the system for the major part of the population mass is gradually being transferred to the environment. The reason for this condition is that NEWCON-1 carries with it the positive feedback situation which will be outlined next.

The chemicals included in the equation for f_1 are the heaviest ones and this requires that the basic metabolic process be the conversion of light molecules to these heavy ones with the resultant reduction in the total number of molecules in the cell. Thus the natural effect of selection is to reduce the mean number of molecules per cell in the population. Now the integer r associated with NEWCON-1 for this experiment is five which means that four-fifths of the total number of molecules of an offspring are determined on the basis of the mean number over the population. Thus selection reduces the mean, and the initial chemistry of offspring is largely tied to the mean resulting in its inevitable gradual reduction over time. To prevent this effect, a modification of NEWCON-1 called NEWCON-2 will be used.

Instead of using the mean number of molecules as a single parameter, a set of parameters comprising the mean number of molecules of each

chemical is used in the following way.

The number of molecules of chemical i , which are contained in each of the offspring is

$$M_i = \frac{\bar{M}_i}{25} \left(\frac{r-1}{r} \right) + M(q) X_i(q) \left(\frac{1}{r} \right)$$

where $\bar{M}_i$ is the mean number of molecules of chemical i calculated by dividing the total number of molecules of chemical i in the population by the number of cells in the population. NEWCON-2 is used in several other experiments and appears to be operating satisfactorily until a problem to be discussed in section 4.5 is discovered. We will resume this discussion at that time.

Much of the present discussion has had the purpose of indicating the various problems which arise in such a simulation. At the same time, the ease with which changes are made has also been stressed. We will continue with a description of some of the statistical methods used to study the response of a population.

4.3 SOME STATISTICAL AIDS

We will digress at this point from the examination of aspects of the system itself to consider briefly a few of the statistical measures used to monitor the response of a population. A fitness function has been defined in Chapter 3 which might be more accurately termed an 'anti-fitness' function in the sense that a decrease in its value reflects an increase in fitness. In most experiments, the goal is to improve the overall fitness of the population under a wide range of conditions. Thus it is clear that the variation of the average fitness of the population over many generations is the main evidence of its response. In every experiment, therefore, the

most vivid indication of the effect of various forces on the population is a graph of average fitness $\bar{F}$ versus number of generations.

This value is computed every generation prior to mating along with the standard deviation. The importance of the latter number is in its indication of the degree of scatter among the fitness values in the population. If the standard deviation is small compared with the average value, the population in most cases will not respond very much further to the selection pressure. If such a relationship holds at the outset of an experiment, it suggests that there will not be enough phenotypic differentiation for selection to be effective. An example of this situation was given in section 4.1 and arose because of the structure of the reaction graph.

Other aids have been discussed previously and these include the frequency distribution of offspring as well as the frequency distribution of crossover probabilities. The average number of molecules per cell and the average weight per cell are also computed every generation as additional indications of the system performance. In the previous section, the drastic decrease in the number of molecules per cell indicated a fault in NEWCON-1 and led to its correction.

The effect of selection on the population is to change its genetic structure as reflected in the increase in frequency of certain alleles and the corresponding decrease of others. A most straightforward way of observing this result is to compute all the allele frequencies at every generation. However, this poses the problem of assimilating a mass of data in order to perceive trends. It suggests the need of a single number to indicate the genetic state which corresponds to the average fitness in reflecting the phenotypic state. This gives rise to the following dis-

cussion which includes an additional reason for such a measure.

It is important in any genetic study to relate the 'fitness' of the population to the genetic variance. Under very special conditions, R. A. Fisher [12] derived what is called 'the fundamental theorem of natural selection,' namely, that the rate of increase of fitness of a species is proportional to the genetic variance. At this point, we do not intend to derive a form of this relation suitable to the present system but we do wish to keep it in mind when studying the change in $\bar{F}$. It is therefore necessary to have some readily obtainable estimate of the genetic variance. The actual calculation for a genetic system with several genes and several alleles per gene is quite lengthy. However, based on the calculations of allele frequencies, a simple alternative is available.

This measure is based on concepts derived in information theory [11]. A few definitions are needed. Consider a transmitter which can function in any one of n states labelled $1, 2, \dots, n$ the states occurring with probability $p_1, p_2, \dots, p_n$, respectively. When in state i , message M_i is transmitted. Then the "informational content" or "entropy" H of this transmitter is

$$H = - \sum_{i=1}^n p_i \log p_i$$

where the base of the logarithm is usually 2. The term entropy arises from its use in physics where it is a measure of the state of organization or disorganization of a system.

It is readily apparent that maximum entropy exists for the situation

$$p_1 = p_2 = \dots = p_n$$

i.e., when all messages are equiprobable or in other words all states are equally likely. If certain states become more probable, H decreases and in some sense the system is more highly organized. The lower H is, the

better estimate one has of the state in which the system is most likely to be. Finally, if only one state is possible, $H = 0$ and we have complete information about the system.

We have been considering a genetic system which is organized as a set of n genes with m alleles per gene. Consider each gene to be a subsystem which can exist in any one of m states. Thus we can define the entropy H_i of gene i , to be

$$H_i = -\sum_j p_{ij} \log p_{ij}$$

where p_{ij} is the probability (frequency) of allele j of gene i . The quantity of information needed to specify the total system is at most as large as the summed quantity of information needed to specify each single subsystem. It may be smaller if certain combinations of subsystem states are impossible or if the probability of certain combinations is not the sum of the individual members because these situations reduce the amount of specification needed. Thus the information or entropy, H , of the total system is

$$H \leq \sum_{i=1}^n H_i .$$

A result similar to this but derived for a different purpose was given by Dancoff and Quastler [12]. However in their discussion, H was given as

$$H \leq \text{no. of genes} \times H(\text{gene})$$

which assumes that the entropy is the same for each gene. The quantity we will use is $\bar{H} = \sum_{i=1}^n H_i$ which is an upper bound to the true entropy. In calculating $\bar{H}$, the p_{ij} are of course the allele frequencies for the whole population. Thus $\bar{H}$ is an estimate of the genetic variance in the sense that equality of all the allele frequencies yields the maximum value of $\bar{H}$ and the increased frequency of certain alleles results in a decrease in $\bar{H}$.

A population under a selection pressure will respond by an increase in the frequency of favorable alleles with a corresponding decrease in the frequency of the unfavorable ones. $\bar{H}$ serves as an approximate measure of the response of the population to selection without conveying any information about the distribution of alleles.

Certain experiments will require special statistical measures and these will be discussed as they arise.

4.4 CHEMICAL DYNAMICS

We wish to discuss some properties of the chemical reactions which affect the response of the population to selection. This can best be done within the framework of the analysis of a pair of experiments. This discussion will also serve to illustrate the use of some of the statistical aids described in section 4.3. Offspring frequency distributions were shown in Fig. 4.2-4 for these two experiments numbered 5 and 6. Previous experiments were briefly referred to.

Both experiments 5 and 6 are carried out under the following conditions:

- (a) Random mating
- (b) Offspring function D
- (c) NEWCON-2
- (d) Same initial population of 50 cells
- (e) $r = 5$ for experiment 5, but $r = 1$ for experiment 6
- (f) $k = 8$
- (g) Maximum cell age of 50
- (h) Mating fraction of .1
- (i) Population bound of 200

Table 4.4-1a shows the allele frequencies for the initial population. Although the genetics are generated randomly using a uniform distribution, there is substantial variation among the allele frequencies. However,

comparing these with the final allele frequencies for the two experiments there does not appear to be any correlation between the initial and final ones.

The graph of $\bar{F}_1$ for experiment 5 is shown in Fig. 4.4-1 along with a plot of the number of offspring produced each generation. The improvement

GENES	A L L E L E S							
	0	1	2	3	4	5	6	7
1	.16	.15	.06	.11	.16	.14	.10	.12
2	.08	.14	.15	.19	.09	.12	.18	.05
3	.16	.15	.13	.12	.13	.09	.14	.08
4	.14	.08	.12	.09	.10	.20	.12	.15
5	.16	.11	.11	.12	.10	.11	.12	.17
6	.10	.09	.12	.14	.13	.18	.11	.13
7	.12	.11	.17	.14	.09	.16	.08	.13
8	.10	.20	.09	.09	.14	.12	.11	.15
9	.14	.15	.13	.14	.15	.09	.08	.12

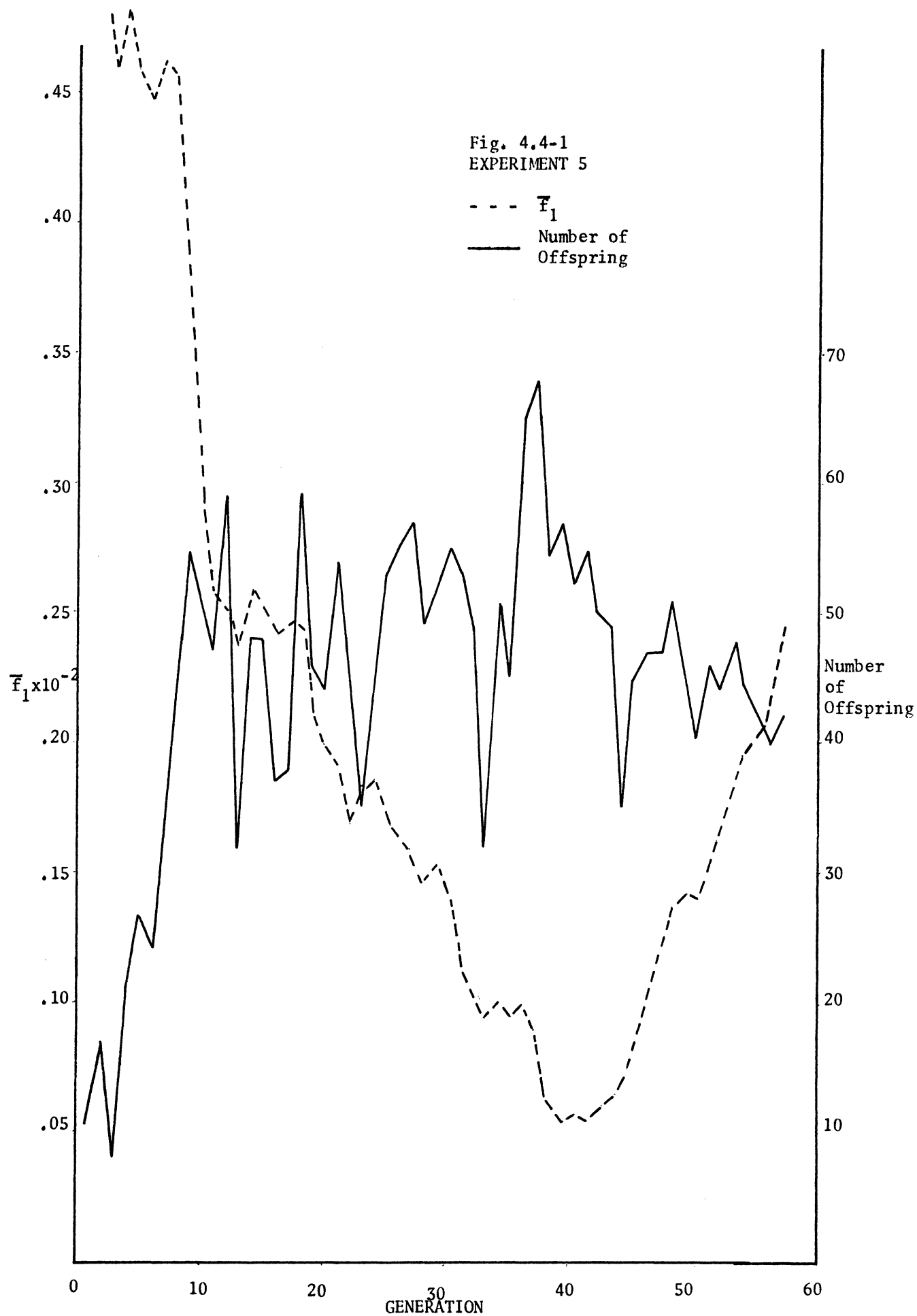
Table 4.4-1a. Allele Frequencies in the Initial Population

GENES	A L L E L E S							
	0	1	2	3	4	5	6	7
1	.0000	.0000	.0000	.0025	.2225	.0000	.7575	.0175
2	.1050	.0175	.0000	.0000	.2225	.0025	.0000	.6525
3	.2400	.0025	.7575	.0000	.0000	.0000	.0000	.0000
4	.0375	.3525	.3375	.0000	.0000	.1175	.1550	.0000
5	.4700	.3375	.0000	.0000	.0000	.0375	.1550	.0000
6	.0375	.3525	.3375	.0000	.1550	.0000	.0000	.1175
7	.0000	.0000	.1200	.0925	.7250	.0000	.0000	.0625
8	.3800	.0925	.0000	.4650	.0000	.0625	.0000	.0000
9	.0925	.2600	.0000	.0625	.5850	.0000	.0000	.0000

Table 4.4-1b. Allele Frequencies at Generation 51 in Experiment 5.

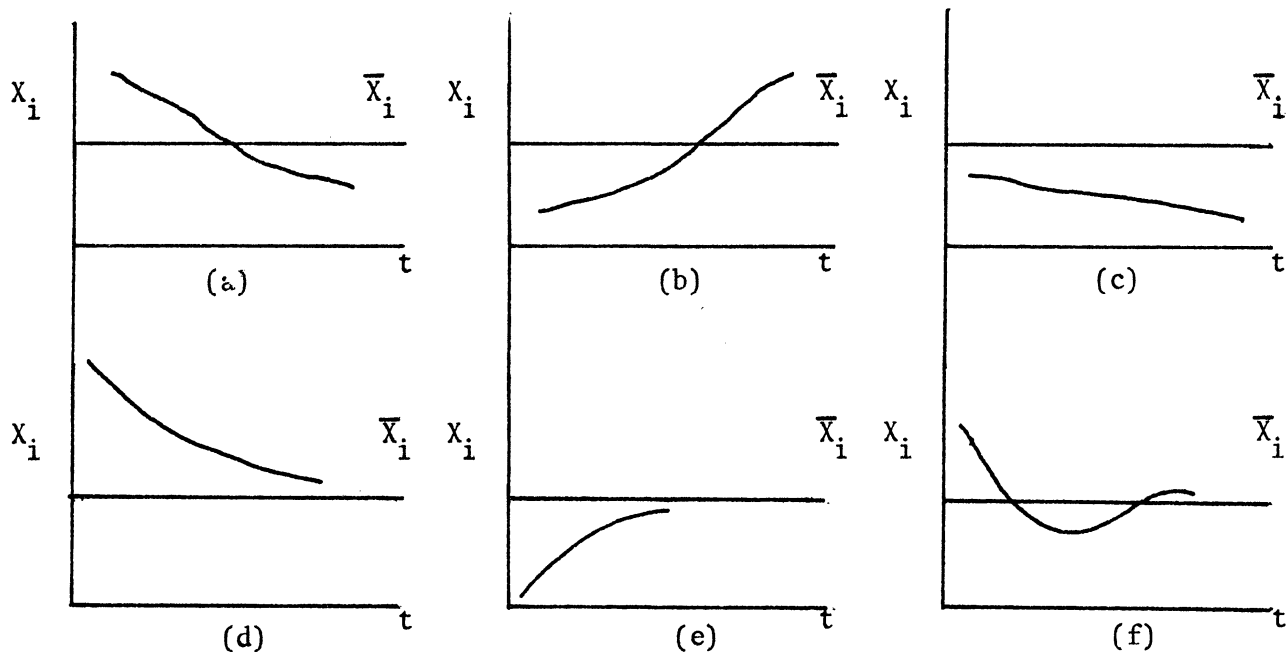
GENES	A L L E L E S							
	0	1	2	3	4	5	6	7
1	.0000	.0000	.0000	.0175	.2525	.1925	.5375	.0000
2	.6625	.0175	.2400	.0000	.0125	.0000	.0325	.0750
3	.0300	.1925	.4625	.0750	.0000	.0000	.2400	.0000
4	.0725	.0000	.0000	.0325	.0000	.2700	.5225	.1025
5	.3025	.5275	.1025	.0000	.0000	.0000	.0675	.0000
6	.0000	.0000	.0000	.0000	.5225	.1050	.0000	.3725
7	.2650	.0000	.0600	.1800	.2250	.0550	.0000	.2150
8	.0525	.0000	.0600	.3525	.2650	.0000	.0550	.2150
9	.4450	.0525	.0000	.2150	.1725	.1150	.0000	.0000

Table 4.4-1c. Allele Frequencies at Generation 51 in Experiment 6.



in the population is very striking for the first forty generations during which time $\bar{f}_1$ is reduced from $.521 \times 10^{-2}$ to $.054 \times 10^{-2}$ or by a factor of about ten. However, from generation forty-one to fifty-seven, $\bar{f}_1$ increases in value to $.242 \times 10^{-2}$ or to about half of its original value. This latter occurrence is difficult to explain but we will suggest some reasons after first discussing the initial behavior.

This discussion is based on two results derived from an experiment we will call 5a. Experiment 5a is the same as 5 except that the population bound is 50 instead of 200 and the number of generations is much fewer. The purpose of this experiment was to study the behavior of cells which survived several generations and to compare the values of f_1 of offspring with those of their parents. The first result is that the value of f_1 for a cell which survives several generations almost invariably increases. The second is that it is also usually the case that the f_1 -value for offspring is less than that of the parents. The following simple diagrams may explain the reason for the first result. In these diagrams, i is one of the chemicals in f_1 , X_i is the concentration of chemical i and $\bar{X}_i$ is the desired concentration.



Cases (a) and (b) illustrate variations in X_1 which produce temporary improvements in f_1 followed by a worsening. In case (c), f_1 will increase from the outset. Cases (d), (e) and (f) represent variations in X_1 which gradually produce improvements in f_1 . From the outset, no attempt has been made to achieve dynamic equilibrium in the chemical system. That is, in general, the chemical concentrations will pass through the desired levels rather than converging towards them. Thus, cases (a), (b) and (c) are much more likely to be representative than are cases (d), (e) and (f). For this reason, the f_1 -value of a cell is likely to increase in the long run although there may be a temporary initial improvement.

The second result is due to the fact that the initial chemistry of offspring is strongly keyed to an 'average' cell defined by the subroutine called NEWCON-2. It is likely that the initial chemistry of offspring will correspond to a lower f_1 than will both their parents, especially in those cases in which only one or two offspring are produced. This is because two randomly selected parents which produce few offspring are likely to have higher f_1 -values than does the 'average' cell. For these situations in which $r = 1$, the f_1 -values of the offspring are the same as one of the parents and thus can be higher or lower than that of the other parent which means that on the average this second result will not occur.

We can now return to an examination of the initial behavior of the curve shown in Fig. 4.4-1. The fifty cells forming the initial population have undergone fifty chemical time steps before mating first occurs in order to allow the chemical concentrations to reflect the cell genetics. For the first few generations (eight) there is a fluctuation in $\bar{f}_1$ but it does decrease slightly. During this phase, it is not necessary to eliminate cells to make room for offspring, since there is space available

while the population grows from fifty to two hundred. Until the upper bound is reached, $\bar{f}_1$ will decrease when the effect of the offspring with their lower f_1 -values overcomes the increase in f_1 -values of the older cells in the population. Conversely when fewer offspring are produced, $\bar{f}_1$ will increase.

Sharp decreases in $\bar{f}_1$ occur as the population size approaches 200, because now the oldest cells (with perhaps the worst f_1 -values) must be eliminated to make room for offspring with their substantially lower f_1 -values. This period of sharp decrease lasts for only a couple of generations until the oldest cells have been removed. Now a period of slower improvement begins.

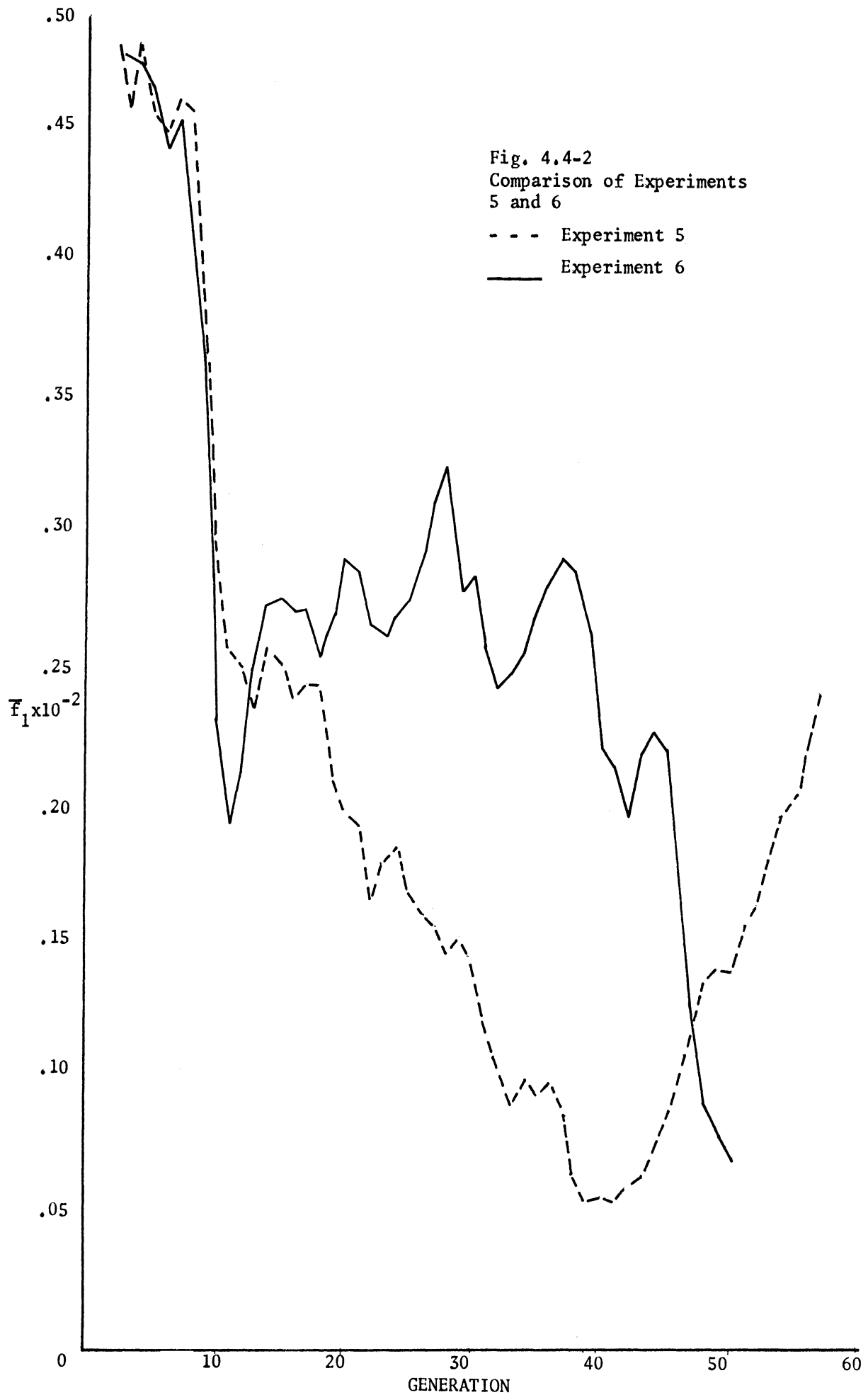
The offspring being produced now will be replacing cells which have not had as long a time to worsen their f_1 -values as the original cells; hence, the improvement in fitness will be much less marked. In fact, due to fluctuations in the number of offspring produced there will be fluctuations in $\bar{f}_1$. Whenever the number of offspring is low, their lowered f_1 -values will be offset by the worsening of the large majority of cells, resulting in small decrease in $\bar{f}_1$, or even an increase. Thus a rough rule is that periods of below normal production of offspring will be accompanied by increasing $\bar{f}_1$ and vice versa. To show this a curve of the number of offspring per generation is plotted together with the $\bar{f}_1$ curve of Fig. 4.4-1.

At generation 42 the upward surge of $\bar{f}_1$ begins. It should be noted that the number of offspring produced over the six preceding generations is unusually high, about 58 per generation. This is followed by a period of very low offspring production which is naturally associated with a worsening population. What actually causes the beginning of the upswing of $\bar{f}_1$ is not clear but the rapid turnover of the population may be the

cause. Once it begins, however, it may be self-propagating because the initial offspring chemistry is tied to the aforementioned 'average cell', and the 'average cell' of course reflects the upturn of $\bar{F}_1$. To test this conjecture, experiment 6 which is the same as 5 except that $r = 1$, is carried out. This means that the initial chemistry of offspring will be identical to that of one of the parents, that one being chosen randomly.

In Fig. 4.4-2, the initial behavior of $\bar{F}_1$ of experiment 6 is very much like that of 5 up to generation 10 and the above explanation for this period is therefore also applicable to experiment 6. After this point, however, there follows a long period of seemingly random fluctuation until generation 44. For the final six generations there is a sharp decrease in $\bar{F}_1$ very much like the first period decrease. It is very clear that after generation 10 these two curves are very dissimilar. It may be that tying the initial chemistry of offspring to the parents instead of the 'average cell' may prevent trends from being self-propagating.

The final major decrease is in direct contrast to the increase of $\bar{F}_1$ of experiment 5. One possible explanation is that the basic stochastic nature of the system allows wide variations in the nature of the $\bar{F}_1$ -curve. But the variation between these two is so wide that doubt arises as to whether or not the selection pressure is strong enough or directed enough. A comparison of Tables 4.4-1b and 4.4-1c reveals that in many instances there are wide disagreements between corresponding allele frequencies. Consider the following examples:



Locations		Experiment	Experiment
GENE	ALLELE	5	6
2	0	.1050	.6625
2	7	.6525	.0750
3	0	.2400	.0300
3	6	.0000	.2400
4	1	.3525	.0000
4	2	.3375	.0000
4	6	.1550	.5225
6	1	.3525	.0000
6	2	.3375	.0000
7	1	.0000	.2650
8	0	.3800	.0525
9	0	.0925	.4450

Table 4.4-2
Specific Allele Frequencies

Some of these differences are so marked that these two populations appear to be headed in different directions as reflected by their respective genetic constitutions. We would expect it to be the case that within reasonable bounds two populations subjected to the same selection pressures and beginning with the same initial population should tend towards approximately the same genetic state. As this is not the case, the difference may be due to the form of the chemistry of the offspring or to some as yet unknown factors.

This rather detailed discussion has pointed up the close relationship between the chemical reactions and the response of the population. Such knowledge is very important in assessing the results of any experiment and this is the reason for the effort devoted to explicating these two experiments.

4.5 BUILT-IN SELECTION

A general question arises with respect to a complex simulation. With so many procedures and parameters how is it possible to prevent circumstances from developing which control or influence the response of the population, unknown to the researcher? Such questions are not restricted

to simulations but rather are accentuated in this type of research because of the very large number of variables under consideration and the even larger number of interactions among them.

One of the only ways of discovering such 'built-in' or hidden factors in a system is to carry out experiments for which the results are fully anticipated. It is hoped that any unrecognized errors will become visible under such conditions by the appearance of results which do not conform to the expected. This technique does not guarantee that all such problems will be uncovered because for all practical purposes there is no such method. To illustrate this matter, some experiments will be presented in which two procedures are shown to influence the direction of evolution. These are the method used to assign the initial offspring chemistry, namely, NEWCON-2 and the method used to eliminate cells. This latter method is called the "oldest-first" procedure and was briefly referred to in section 3.6

Needed places for new offspring are made available by eliminating the oldest cells in the population. For example, if places are needed for three offspring and there are ten cells of age six, the oldest in the population, then three of this age group are randomly selected for elimination. If there are only two in this oldest age group then these two are eliminated immediately and one cell is randomly chosen from the next lower non-empty age group.

In addition to pointing out the actual role of these two procedures, the same experiments will be used to confirm a necessary property of the system. In the fitness function, f_2 used in these experiments

$$f_2 = 10(.15 - x_{23})^2 + (.1 - x_{24})^2 + (.1 - x_{25})^2 ,$$

chemical 23 is heavily weighted over the others. We will expect those

alleles most directly associated with the production of chemical 23 to be strongly selected.

We will now turn to the discussion of the experiments of this section beginning with experiments 7 and 8. The following conditions apply to both experiments:

- (a) Random mating
- (b) Offspring function D for experiment 7
Offspring function E for experiment 8
- (c) NEWCON-2
- (d) Initial population size of 200
- (e) $r = 1$
- (f) $k = 20$
- (g) No maximum cell age
- (h) Mating fraction of .1
- (i) Population bound of 200
- (j) Number of biochemical time steps per generation of 4
- (k) Number of crossover positions is two.

Offspring function E does not depend on the f_2 -values of the cell but is instead a probability distribution, namely the negative binomial. The probability of X offspring, $p(X)$ is given by

$$p(X) = \binom{s+X-1}{X} p^s q^X.$$

With $p = 2/7$, $q = 5/7$, the mean of this distribution, q/p is 2.5. For a value of $s = 3$, and with the values weighted so that $\sum_{X=0}^6 p(X) = 1$, the distribution is

Number of Offspring	Probability
0	.0936
1	.2554
2	.2322
3	.1762
4	.1200
5	.0763
6	.0463

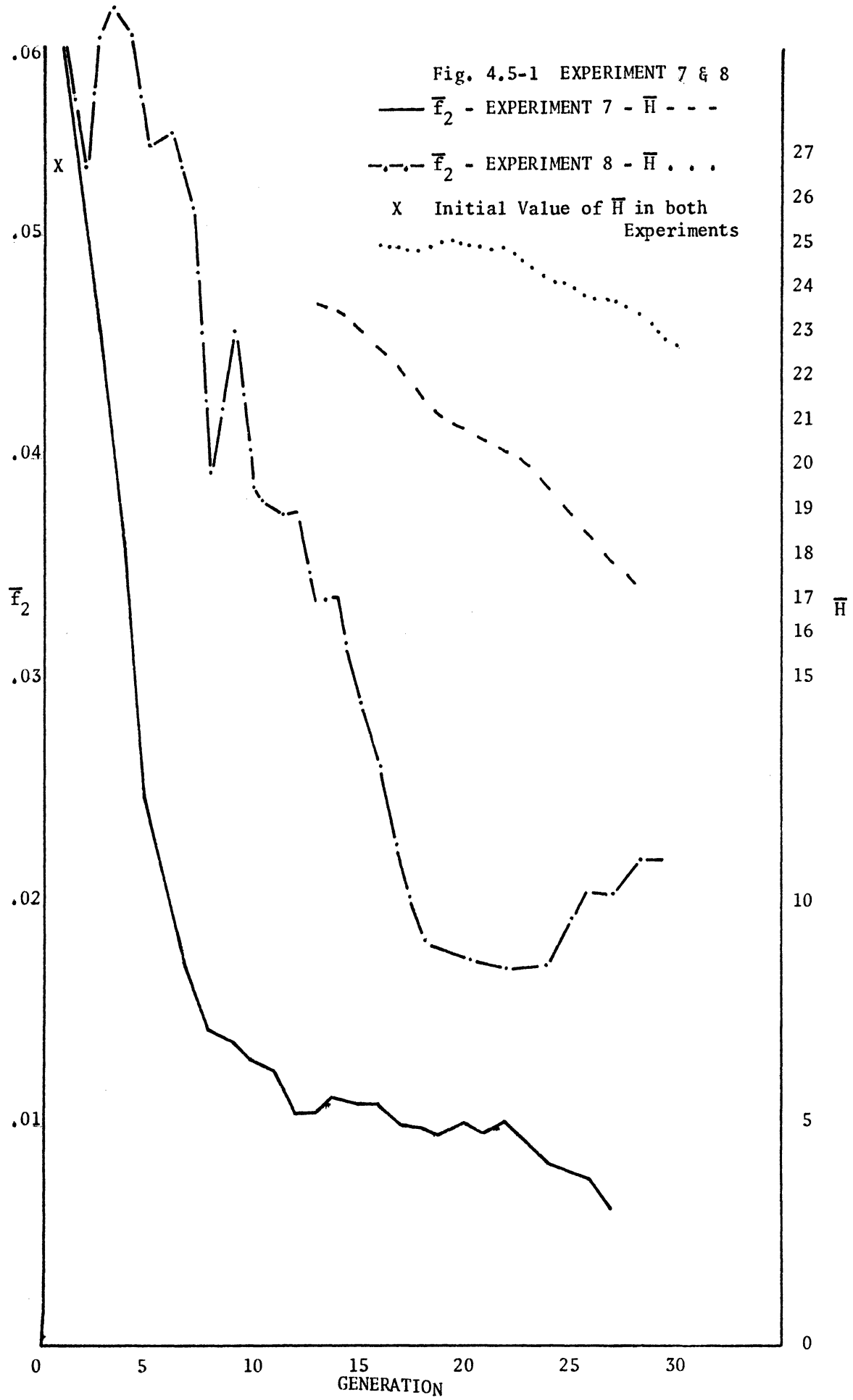
Table 4.5-1 A Random Offspring Function

The $\bar{F}_2$ curves of both experiments are shown in Fig. 4.5-1 and the result is quite surprising. Experiment 7 shows the strong response of a population to selection in which $\bar{F}_2$ is reduced by a factor of almost 10 in 27 generations. Experiment 8 was to be a control experiment in which parents are selected randomly and offspring are assigned randomly, i.e., no selection is operative. Therefore, it is to be expected that the behavior will be quite random. However, the $\bar{F}_2$ graph for experiment 8 seems to indicate that there is selection operative because $\bar{F}_2$ is reduced by a factor of about 3 in 29 generations and this reduction is strongly directed. Thus it appears that there is a hidden or built-in selection force in this system. The next task is to identify it.

Before doing this, however, it seems worthwhile at this point to comment on the $\bar{H}$ curves. These do not begin from generation 1 because the program which calculates $\bar{H}$ was not in operation at that time. Both of these experiments were carried out in two stages and it was during the second part that $\bar{H}$ was calculated. At generation 28, the value of $\bar{H}$ in experiment 8 is about 17 whereas in experiment 9 it is 23.5. This indicates a greater degree of genetic fixation in the former experiment which is to be expected since its population is under a clear selection pressure.

The maximum value of $\bar{H}$ is 27 as the following calculation shows. There are 9 genes and 8 alleles per gene in each cell. The maximum value of $\bar{H}$ occurs for the condition of equal frequencies of alleles at each gene. In this case with the same number of alleles per gene, this common frequency is one-eighth. Thus $\bar{H}$ is given by

$$\begin{aligned}\bar{H} &= 9 \left(- \sum_{i=1}^8 \frac{1}{8} \log_2 \frac{1}{8} \right) \\ &= 9 \log_2 8 \\ &= 27.\end{aligned}$$

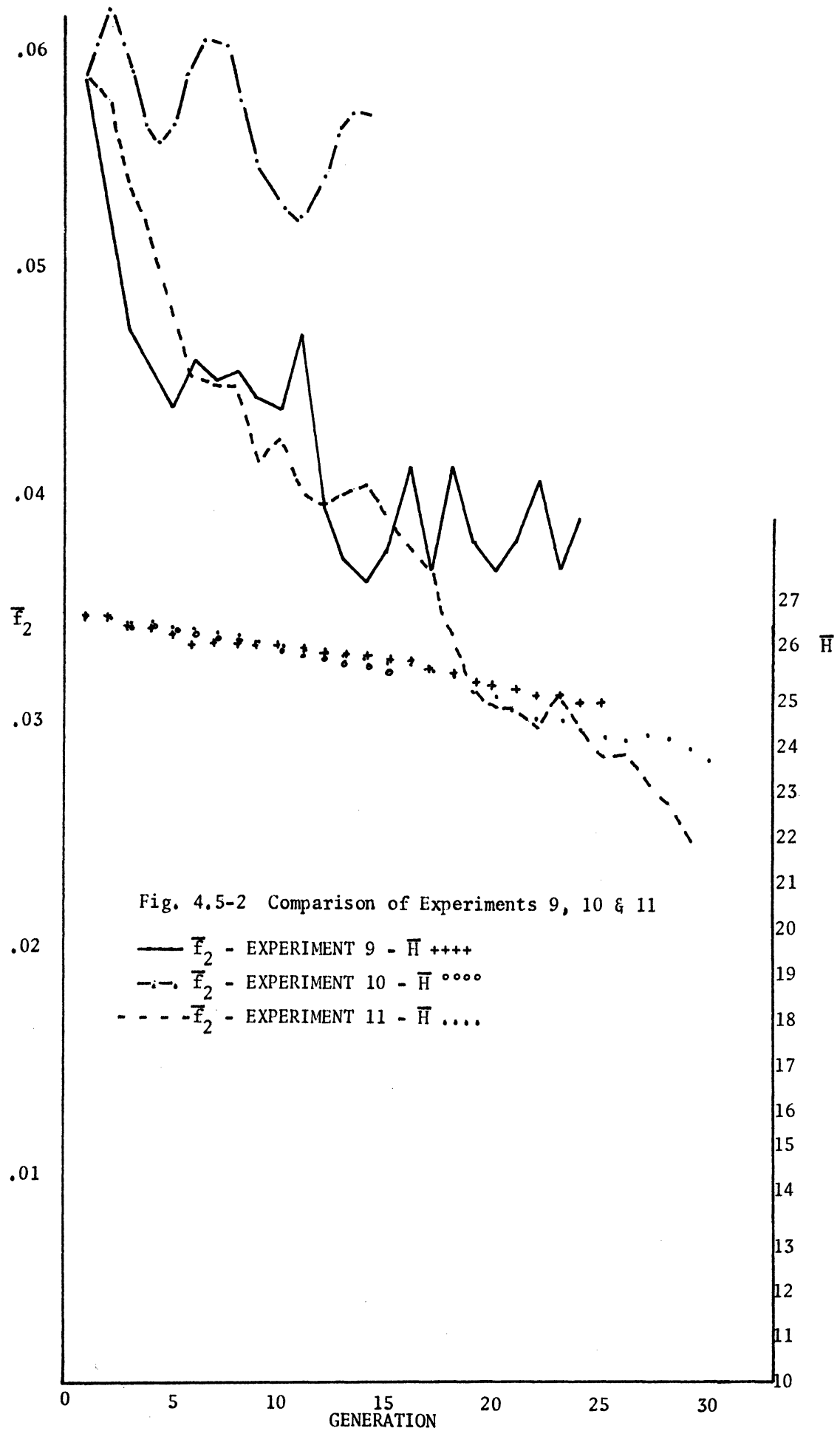


Because the genetics for the initial population is generated randomly, the value of $\bar{H}$ at generation 1 is very close to 27. Therefore in experiment 7, $\bar{H}$ has decreased by 37% by generation 28 whereas in experiment 8, it has decreased by 17%. This is a clear indication of the effect of the known selection pressure.

We will now discuss those experiments which uncovered at least two instances of built-in selection. The first procedure to be tested is the elimination of cells to provide space for offspring. In all the experiments up to this point the oldest cells are removed first when the need arises (see Section 3.6). In a previous discussion (Section 4.4), it was pointed out that the oldest cells have the poorest f -values and therefore their elimination will tend to improve the population's fitness independent of selection.

The simplest alternative is to eliminate cells randomly without regard to age. Experiment 9 is the same as experiment 8 except for this change in the method of eliminating cells. The resulting graph of $\bar{f}_2$ is shown in Fig. 4.5-2. In comparison with the graph of $\bar{f}_2$ for experiment 8, there is much less selection evident. However, $\bar{f}_2$ does drop from about .06 to about .036 before the oscillatory behavior begins. This initial drop may be a stochastic effect. Although the number of offspring are being assigned randomly, it is possible that enough good offspring appear initially to set up a short-lived trend. The fact that this improvement lasts 14 generations indicates that there may be another source of hidden selection.

A possible candidate is NEWCON-2, the subroutine which specifies the initial chemistry for the offspring. In experiments 7, 8 and 9, the parameter $r = 1$ is used which means that the initial chemistry of the

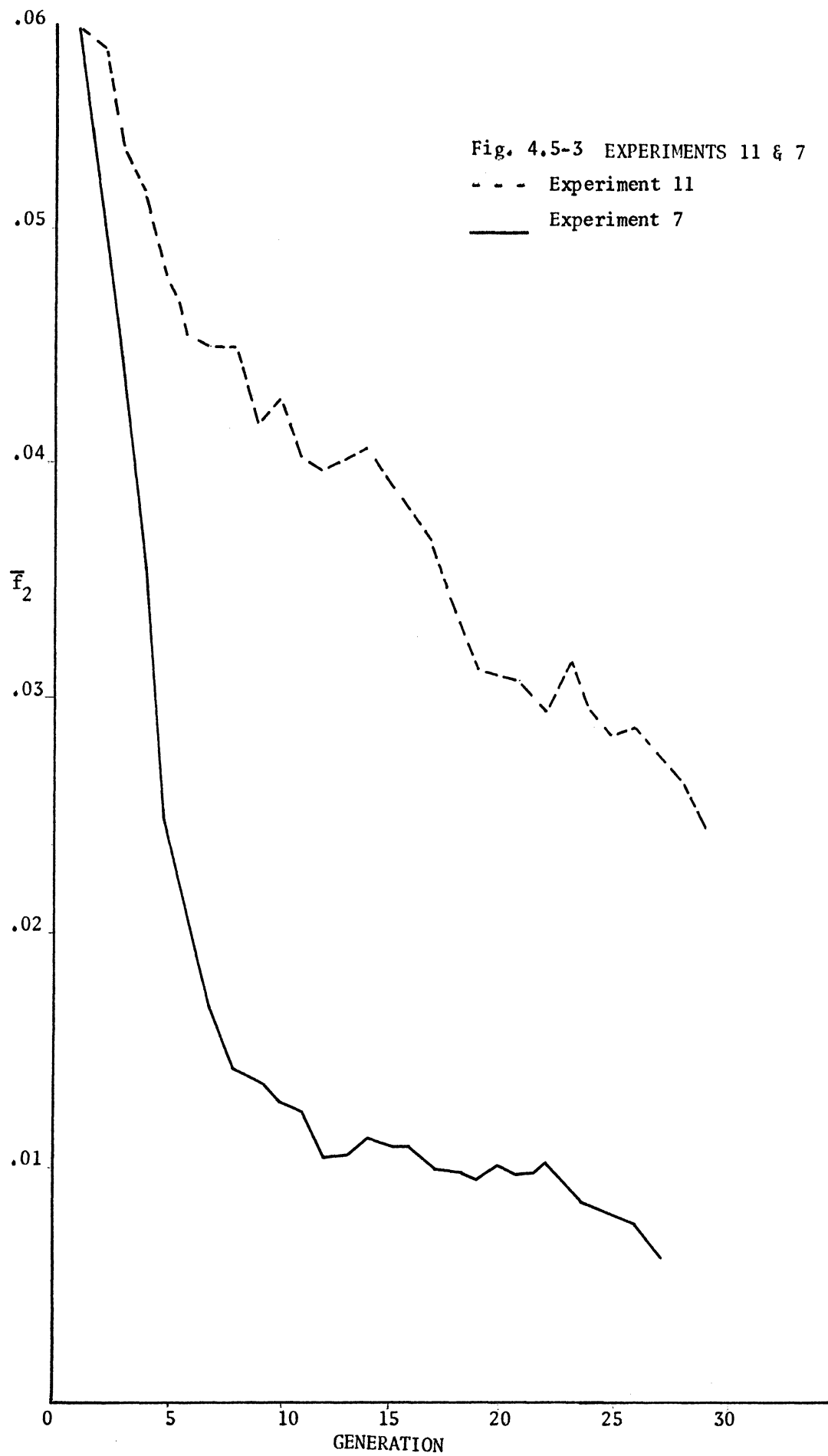


offspring is exactly that of one of its parents. If a trend is initiated it is then favored by the assignment of an initial chemistry which reflects this trend. The obvious extreme alternative is a chemistry totally independent of the state of the population.

Such an alternative denoted as NEWCON-3 is the assignment of an equal number of molecules of each type to each offspring. With 25 chemicals, this number was chosen to be 40,000, resulting in a total number of one million molecules per cell. Differing from experiment 9 only in the NEWCON subroutine and from experiment 8 in the procedure for eliminating cells as well, experiment 10 was carried out and the graph of $\bar{F}_2$ is shown in Fig. 4.5-2. Its oscillatory shape is that expected of a population which is undergoing no selection. In fact it appears now that there is no hidden selection for the population is behaving in a more or less random fashion.

As a final experiment, selection is returned to the population in the nature of the offspring function D. Both changes, however, are incorporated and it is in this respect that experiment 11 differs from experiment 7. In comparison with experiment 10, experiment 11 is quite striking. The value of $\bar{F}_2$ is reduced from .06 to about .025 in 29 generations and appears capable of further reduction. There is a very direct response to the selection pressure.

Experiments 7 and 11 are compared in Fig. 4.5-3, where it is seen how much faster the population of experiment 7 has improved. The removal of the selective effects due to the form of the initial offspring chemistry and the means of eliminating cells, has definitely reduced the rate of evolution of the population. This is also reflected in the values of $\bar{H}$ for the two populations. In experiment 7, $\bar{H}$ has a value of 17.3 at gener-



ation 28, whereas in experiment 11, $\bar{H}$ has a value of 21.8 at generation 29. This is a reduction of 36% in the former and 23% in the latter.

Through the use of a control experiment, the hidden contributions of two schemes have been clarified. In the following chapter, new versions of these will be formulated based on the information derived from the experiments of this section. Finally, the effect of the form of f_2 on the selection of certain alleles is noted.

The form of f_2 put a premium on matching the preset concentration of chemical 23, because the associated term was heavily weighted over the other two terms. An examination of Table 4.3-1 shows that alleles 0 to 3 of gene 5 are concerned with the production of chemical 23. The required concentration of .15 is high when compared to the uniform concentration value of .04. Thus among the alleles 0 to 3 of gene 5, those with the larger rate constants should be favored. Consider the following table:

Gene 5 Alleles	Allele Frequencies		
	Experiment 7 Generation 28	Experiment 11 Generation 30	Both Experiments Generation 1
0	.2350	.0500	.1075
1	.0800	.1900	.1425
2	.5925	.4575	.1150
3	.0000	.1725	.1500
4	.0000	.0000	.1175
5	.0575	.0650	.1250
6	.0000	.0250	.1175
7	.0350	.0400	.1250

Table 4.5-2
ALLELE FREQUENCIES FOR EXPERIMENTS 7 & 11

It is very clear that allele 2 has been selected very strongly in both experiments and this suggests that both populations are moving in the same direction but that the one from experiment 11 is moving much more slowly.

At this point, the system appears to be performing in an adequate

manner and more specific experiments can be conducted.

4.6 SUMMARY

The main theme of this chapter has been a demonstration of the generality and flexibility of the system. It is from the definition of phenotype that much of the importance of this model arises. The phenotype is dynamically determined through gene action which is expressible in terms of enzymatic control as well as interaction between the cell and the environment. With a phenotype defined in this way, questions can be considered which are not appropriate for other simulations. Some of these have been discussed in the opening portions of this chapter, and others will be dealt with in some detail in the following two chapters.

We have stressed the problems inherent in a complex system with many parameters and have investigated the unrecognized contributions of two procedures. There is no way to be sure that the model is free of such unknown controls and so a basic requirement of further research is the inclusion of control experiments wherever possible. These can be of the type meant to evoke random responses or else to produce results which are very predictable on the basis of present knowledge. This is necessary to increase the weight of evidence for treating results produced by this system as reasonable.

Before concluding, we make note of one of the problems involved in a stochastic simulation. None of the experiments will produce exactly repeatable results with respect to the behavior of the average value of the fitness function. In general, a stochastic simulation requires many repetitions of each experiment to produce an average experiment. We do not have sufficient computer time at this point to carry out such a

large number.

Therefore instead of trying to prove anything conclusively, we wish to strongly suggest that certain conjectures may be true. It is possible for a misleading conclusion to arise from a single experiment in a stochastic process. However, when a set of experiments varying in a single parameter, are performed which produce a consistent result, it leads one to place some reliability on this result. Certain characteristics are so dominant that even with stochastic variations they stand out.

Thus it is our goal to make conjectures and to carry out experiments which tend to support them. The key word is tend because we do not have the resources to attempt a more complete and fuller confirmation and in such a complex model no result can ever be totally valid. With respect to problems in the real world, of course, the latter statement is trivially true. When working with a model, it needs to be stressed.

CHAPTER 5

SELECTION AND THE NUMBER OF CROSSOVER POSITIONS

In the last chapter, the experiments on the model described in section 4.2 were carried out on a population for which crossover was allowed at only two positions between genes 3 and 4 and genes 6 and 7. The results of one experiment described there dealt with a population with four crossover positions but the purpose of that discussion was to introduce new terms and not to explore the relation between the number of crossover positions and the response to selection. It is the purpose of the present chapter to explore this relation.

Two other areas for investigation will arise in the course of the above study. The first, a matter of necessity, appears because of the radical changes that the environment is undergoing as a result of the selection pressures on the population. The second, a matter of choice, is the desire to compare the effects of different numbers of offspring per generation on the evolution of the population. Neither these nor the experiments on crossover positions will produce conclusive results. Rather, they will lay the foundation for a future and more extensive research effort.

5.1 PRELIMINARIES

In section 4.5, it was pointed out that the methods comprising the solution to the problem of built-in selection would not become the standard techniques for future experiments. The alternative to the 'oldest-first' method of eliminating cells was a method for eliminating them

randomly. This latter procedure while effective does not correspond too well to the situation in nature. It is certainly not the case that all members of a population have an equal chance of dying independently of age, fitness or any other relevant parameter.

For these experiments, we will use a method which depends on the age of the cells. The maximum age any cell may achieve is six and it is automatically eliminated before its age is increased to seven. There are two stages involved in determining which cell to eliminate to make room for an offspring. First, one of the age groups from zero to six is chosen through the use of an age-differential mortality distribution. Within the age group, the cell to be eliminated is chosen randomly.

This method allows for the possibility that a cell of any age may be eliminated but weights the probability in favor of older cells being chosen. The actual mortality distribution used was constructed on the basis of an exponential function which was arbitrarily chosen. The distribution given in tabular form below,

Age	Probability of Dying	Cumulative Distribution
0	.050	.050
1	.083	.133
2	.114	.247
3	.144	.391
4	.175	.566
5	.204	.770
6	.230	1.000

Table 5.1-1
Age Dependent Mortality Distribution

is derived from the equation,

$$\text{probability of dying at age } t = 1.05 - e^{-.033t}.$$

As we stated, the actual form of this distribution has been chosen arbitrarily but with the restriction that it be 'reasonable.' This form

could itself be an object of study at some future time.

To be more precise about the process of eliminating cells in these experiments, not every cell can be removed. During a mating period, offspring produced earlier in the period cannot be eliminated to provide room for those produced later. It is also the case that offspring cannot become parents in the same mating period in which they are produced. As we stated earlier, other conditions can be readily implemented.

The description of NEWCON-3 as given in Section 4.7 is not quite correct. Instead of assigning 40,000 molecules of each chemical to each offspring there should be a mechanism similar to that in the other initial concentration functions to allow some influence of the parental chemistry. The actual form of NEWCON-3 used in these experiments will operate in the following fashion. The number of molecules of chemical i , M_i , assigned to any offspring is

$$M_i = 40,000 \left(\frac{r-1}{r} \right) + M(q) X_i(q) \frac{1}{r}$$

where the variables have been defined previously.

The justification for this type of assignment of the initial chemistry of offspring, has a basis in nature. It is generally the case that in a given species, the chemicals of the egg, which provide almost all the initial chemicals for the offspring, are the same for all offspring. Any deviation in the chemistry of the egg would reflect in some way the particular chemistry of the female which produced it.

The degree to which the chemistry of the egg is uniform depends, in our system, on the parameter r . Thus for a high value of r , the initial chemistry of offspring will be largely similar except for the perturbation due to the chemistry of one of its parents. The following

diagram illustrates this point:

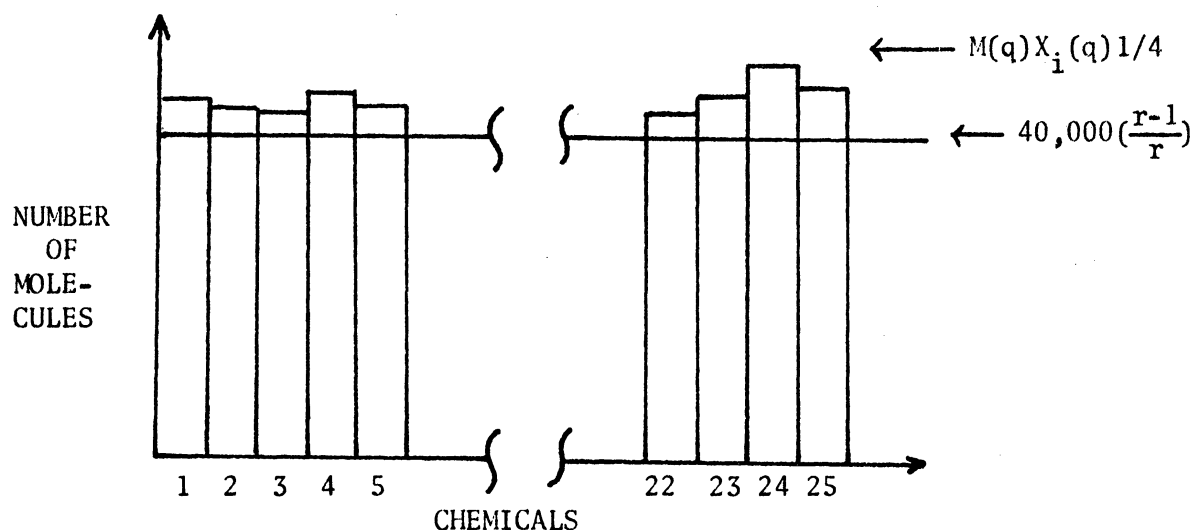


Fig. 5.1-1 Initial Chemistry for Offspring

An additional point is that with r large, there is much less likelihood of a trend being propagated through the inheritance of a 'good' chemistry. Because all offspring have basically the same initial chemical constitution, they will distinguish themselves mainly on the basis of their genetics and to a much lesser degree on any inherited chemical factors.

The experiments to be discussed in this section will serve to illustrate this point. In some sense they are similar to the experiments of section 4.5 in again pointing out features which may lead to a response to selection but which have not been so recognized previously. Before beginning the discussion, we will list those features which are common to all the experiments of this chapter. They are as follows:

- (1) The initial population size is 200 and the initial genetics is the same for all experiments.
- (2) The upper bound is 200.
- (3) The number of biochemical time steps per generation is 4.
- (4) $k=20$.
- (5) The maximum age of a cell is 6 and the mortality distribution (as described in this section) is used.
- (6) $f_2 = 10(.15 - x_{23})^2 + (.1 - x_{24})^2 + (.1 - x_{25})^2$.

(7) NEWCON-3 (as described in this section) is used.

(8) This modification of Table 4.2-2. (See Table 5.1-2)

Two new offspring generating functions are used in these experiments and are given below. (See Table 5.1-3). With the changes made in the population, offspring function D was no longer suitable. Again $s = (f^i + f^j)/2\bar{f}$ where f^i and f^j are the values of f_2 for cells i and j , respectively (the parents).

The specific conditions for experiment 1 are as follows:

- (1) Two crossover positions
- (2) $r = 5$
- (3) Random mating
- (4) Offspring function J
- (5) Mating fraction of .1

Sum of the Crossover Levels	Crossover Probability
0	.000
1	.005
2	.025
3	.050
4	.100
5	.150
6	.200
7	.250
8	.275
9	.300
10	.325
11	.350
12	.400
13	.450
14	.500

Table 5.1-2
CROSSOVER PROBABILITIES

	Offspring Function		Number of Offspring
	J	K	
$s \leq .1$	.1	.2	6
$s \leq .3$	.3	.4	5
$s \leq .5$	.5	.6	4
$s \leq .8$	.8	.8	3
$s \leq 1.2$	1.2	1.1	2
$s \leq 1.5$	1.5	1.7	1
	Otherwise		0

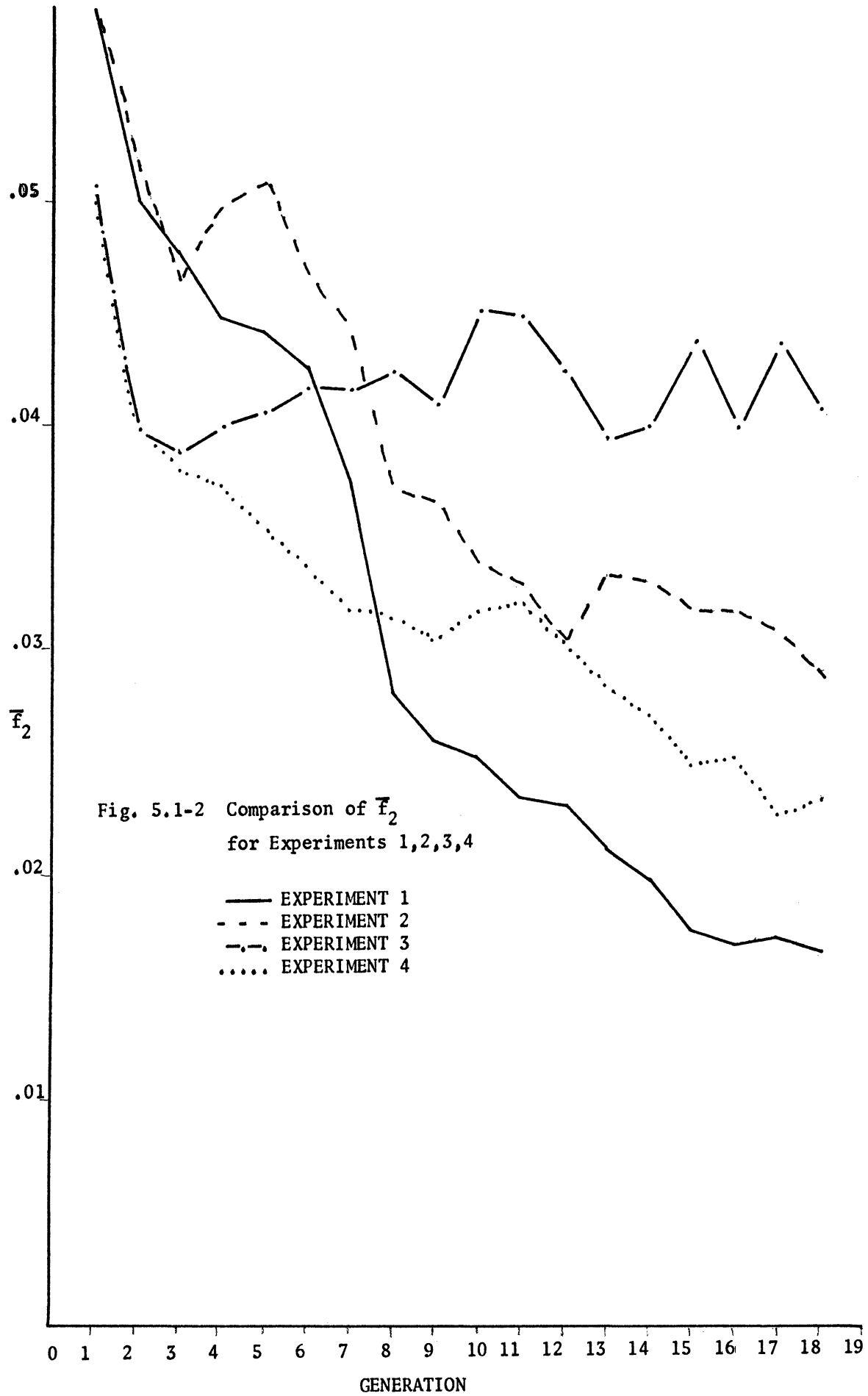
Table 5.1-3
OFFSPRING GENERATING FUNCTIONS

The initial population has been run for thirty biochemical time steps prior to generation one. The graph of $\bar{F}_2$ is shown in Fig. 5.1-2 for this experiment as well as for the others discussed in this section.

With the changes made, it seems necessary, once again, to carry out a control experiment of the type discussed in section 4.5

Experiment 2 differs from experiment 1 only in the use of offspring function E, the negative binomial distribution described in section 4.5. From a comparison of the graphs of $\bar{F}_2$ for these two experiments it appears that the population of experiment 2 which is not under selection pressure is acting as if it were, although not to the same degree as that of experiment 1. The value of $\bar{F}_2$ is reduced by about a factor of 3.5 in experiment 1 and 2 in experiment 2.

Two changes will be made to attempt to remedy this situation. First, r will be increased from 5 to 10 in light of previous discussion. Thus a parent will have about a 10% effect on the initial chemical constitution of the offspring. Because the offspring will begin with a more or less uniform set of concentrations, perhaps the original population should begin under similar conditions. That is, instead of a preliminary set of biochemical time steps, the population, beginning with a uniform



chemistry for each cell, will run for the usual four time steps before mating begins.

There is of course no reason at this point to choose the uniform concentration distribution above any other distribution. The main point is that the distribution of concentrations be largely independent of the state of the population. As with many of the procedures devised in this present simulation, we will choose a particular one and not treat it as an experimental parameter if it gives satisfactory results. Of course, we always have the prerogative of treating any specific method as a subject of experimental study.

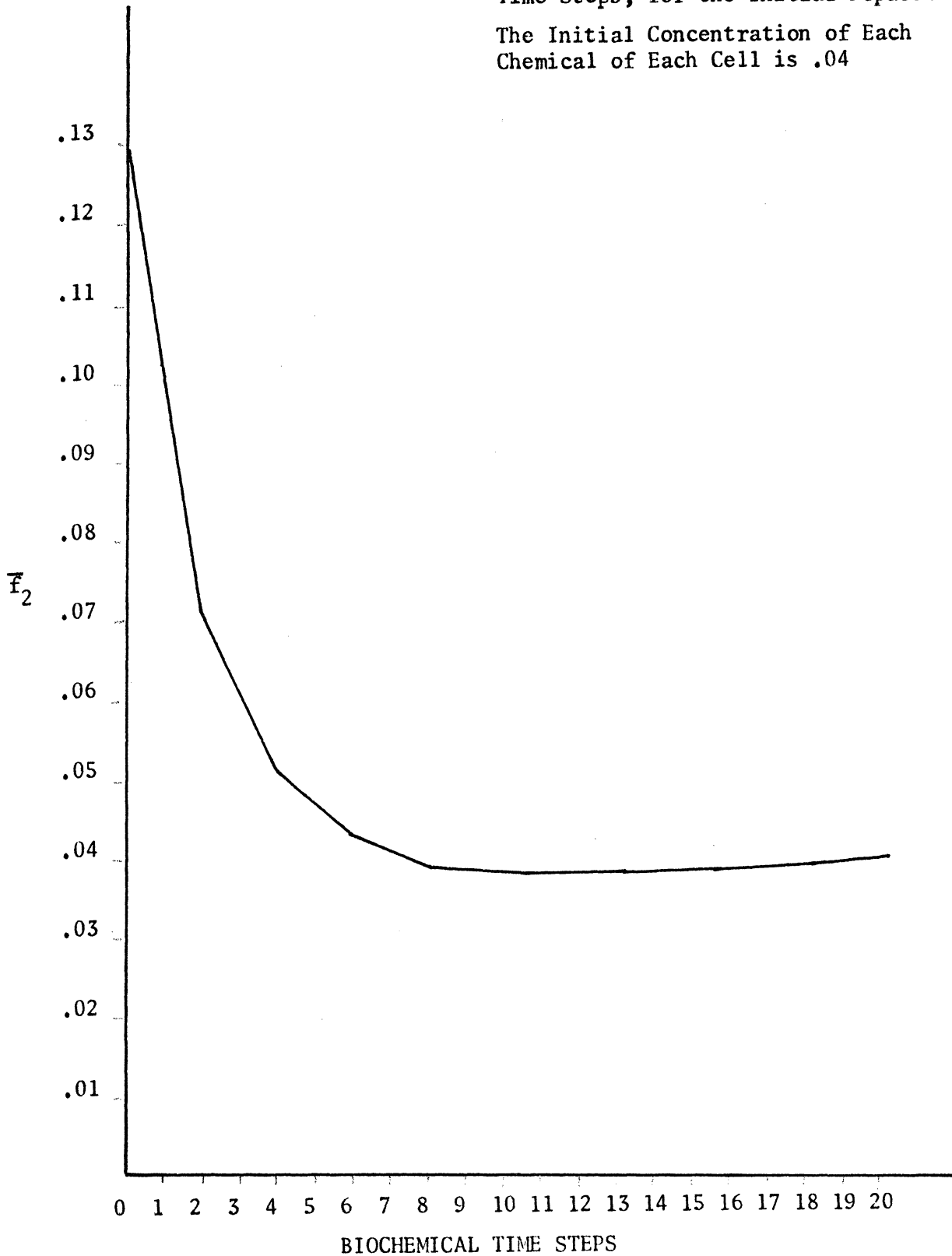
Experiment 3 incorporates these two changes along with offspring function E. The resulting curve shown in Fig. 5.1-2 appears to exhibit a random behavior. Experiment 4, which is the same as experiment 3 except for the replacement of offspring function E by J (i.e., the introduction of a selection force) evidences a clear response to the selection pressure.

The initial sharp drop in $\bar{F}_2$ between generations 1 and 2 in both experiments can be readily explained. Fig. 5.1-3 shows the action of the initial population with uniform chemical concentrations over a period of twenty biochemical time steps. At time step 4, the value of $\bar{F}_2$ is .0508, which is of course the value of $\bar{F}_2$ at generation 1. At time step 8, it is .0383. In both experiments there are about 50 offspring produced in the first generation. Thus 150 cells of the original population are undisturbed. The value of $\bar{F}_2$ at generation 2 is about .0405 in both populations which is very close to the value which the original two hundred cells would have reached in eight time steps. The

Fig. 5.1-3

The Variation in $\bar{f}_2$ Over 20
Time Steps, for the Initial Population

The Initial Concentration of Each
Chemical of Each Cell is .04



initial drop. therefore, is just a manifestation of the natural direction of the system.

If generation 2 is taken as the actual beginning of these experiments, the contrasts in their behavior is all the more striking. From this point on, all the experiments of this chapter will begin with uniform chemical concentrations and with the value of $r = 10$.

We will now continue with experiments for which the number of crossover positions is the key parameter.

5.2 CROSSOVER POSITIONS AND SELECTION

We wish to explore the relationship between the number of positions at which crossover is possible and the response of the population to selection. In the earliest experiments in which there was no crossover, there was a rapid loss in the genetic variation of the population. The reason for this is quite clear for with no possibility of crossover, whole chromosomes can be lost through the effects of random sampling. In this context, the term sampling applies to the production of gametes, one from each parent, which combine to form the genetics of the offspring.

When there is no crossover, a gamete is simply one of the chromosomes of the parent and since the chromosome is selected randomly it is very common for chromosomes to be lost from the population. When the possibility of crossing-over between chromosomes is allowed, then even if a chromosome is not chosen initially some of its genes may contribute to the gamete if crossing-over does occur. Thus crossover plays the very important role of preventing the rapid decrease of genetic variation in such a small closed population.

But this is not its primary role, which is to provide the mechanism to explore the genetic space. Each gene can be viewed as a single dimension or axis of a multi-dimensional space.

The alleles are the allowable positions along the axis and therefore the coordinates of any point in this space are an ordered set of alleles. The fact that each genotype is a pair of chromosomes is a complication but one way of treating the genotype is to describe it by a pair of points in this space. The phenotype is a non-linear transformation of this pair of points in the genetic space to a point in some 'chemical concentration' space.

The process of evolution consists in exploring this genetic space to determine genotypes which produce the 'best' phenotypes. Exploration is directed through the ranking of phenotypes on a 'fitness' scale. In terms of this discussion the role of crossing-over can be seen more clearly. Suppose no crossing-over is allowed.

Then the genotype of an offspring is simply the pair of points in the genetic space determined by taking one point from one parent and one point from the other. It is obvious that the number of possible genotypes determined this way is quite limited. For example, even with two hundred cells initially, there is a maximum of four hundred different chromosomes. This maximum depends on the number of genes and number of alleles per gene. Suppose there are n genes and m alleles per gene. Then the number of different chromosomes is m^n . In the present model there are nine genes and eight alleles per gene. This results in a number of chromosomes equal to 8^9 or over 128×10^6 . However, in the initial population there can be at most four hundred different chromosomes. With no crossing-over the maximum number of distinct genotypes

which can ever be produced in this population is $400 \times 400 = 160,000$. This is a very small fraction of the total possible number.

With crossover allowed between every pair of genes the number of distinct genotypes which could be produced is comparable to the total possible number. We wish to make one point very strongly. Given a population of two hundred cells and a limited number of generations, only a relatively small number of genotypes can ever be produced. What then is the importance of crossing-over if the actual number of genotypes produced is very low, certainly much less than 160,000? The importance is not the number of distinct genotypes but the regions of the genetic space in which they are found.

With no crossover allowed, genotypes can be produced in a region of the genetic space which is quite restricted. In mathematical terms, if the coordinates of a genotype are (i,j) let I be the set of distinct points represented by the genetics of the initial population. Then the region of the space represented by genotypes which could be produced is

$$G = \{(i,j) \text{ for which } i,j \in I\} .$$

This is a limited region of the total space and without crossover it is impossible to explore any region but this one. With crossover it is possible to send out probes to any part of the space in an attempt to discover the best genotypes. The region of the genetic space open for exploration is now $G' = \{(i,j) \text{ for which } i \text{ and } j \text{ could be produced}$

through crossover from $I\}$.

Depending on the number of crossover positions and the variation in I , G' could be as large as the entire genetic space. Again it is not the possibility of producing all genotypes which is important for there is

not sufficient time to do this. The importance of crossover lies in the fact that all regions of the space are open for exploration.

In this section, three experiments will be carried out which are the same in all respects except for the number of positions at which crossover can occur. The common conditions for these three experiments are as follows:

- (1) $r = 10$
- (2) Random mating
- (3) Offspring function J
- (4) Mating fraction of .1

The description of the crossover positions for the three experiments is given in the table below.

Experiment	Number of Crossover Positions	Crossover Positions (in terms of genes)
5	2	3-4, 6-7
6	4	2-3, 4-5, 6-7, 8-9
7	8	between every pair of genes

Table 5.2-1

CROSSOVER POSITIONS FOR
EXPERIMENTS 5, 6, AND 7

There are some things we can say about populations with few crossover positions and those with many such positions. In the former case, on the basis of our previous discussion, less of the total genetic space is open for exploration. Thus the conjecture arises that the populations with fewer crossover positions will respond more directly in the initial phase of their evolution. That is, there will be much less exploration of the genetic space in the first stages as evidenced in the behavior of $\bar{F}_2$. As the loss in genetic variation is more rapid in such populations, their eventual genetic constitution is more quickly achieved.

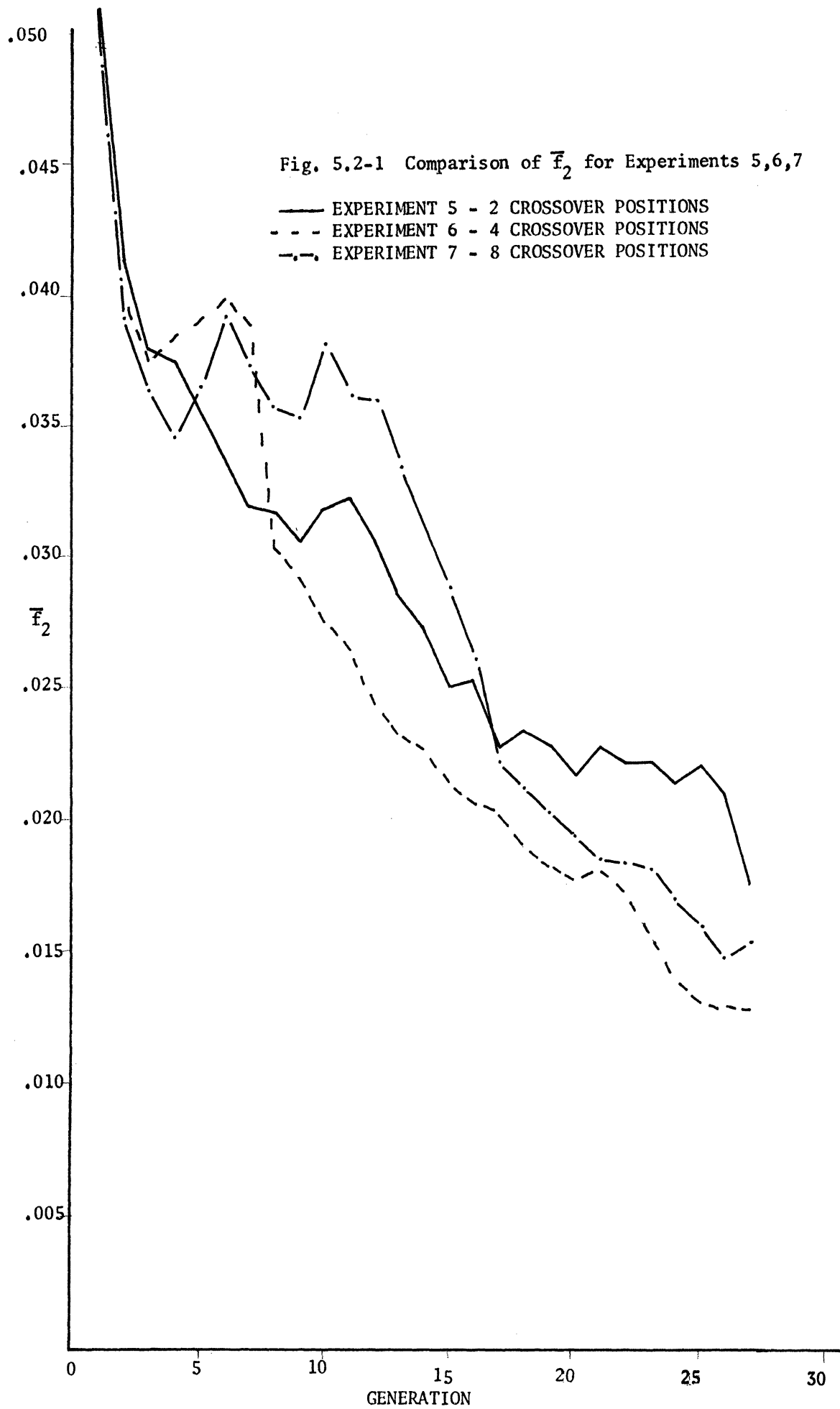
It is also very likely that they will not do as well in the long run because of their limited ability in exploring the genetic space. In contrast a population with many crossover positions may not do well during the first part of its lifetime which is devoted to some exploration of the genetic space. However, once success is achieved it will be exploited very rapidly and such a population should overtake one which is limited in its crossover potential.

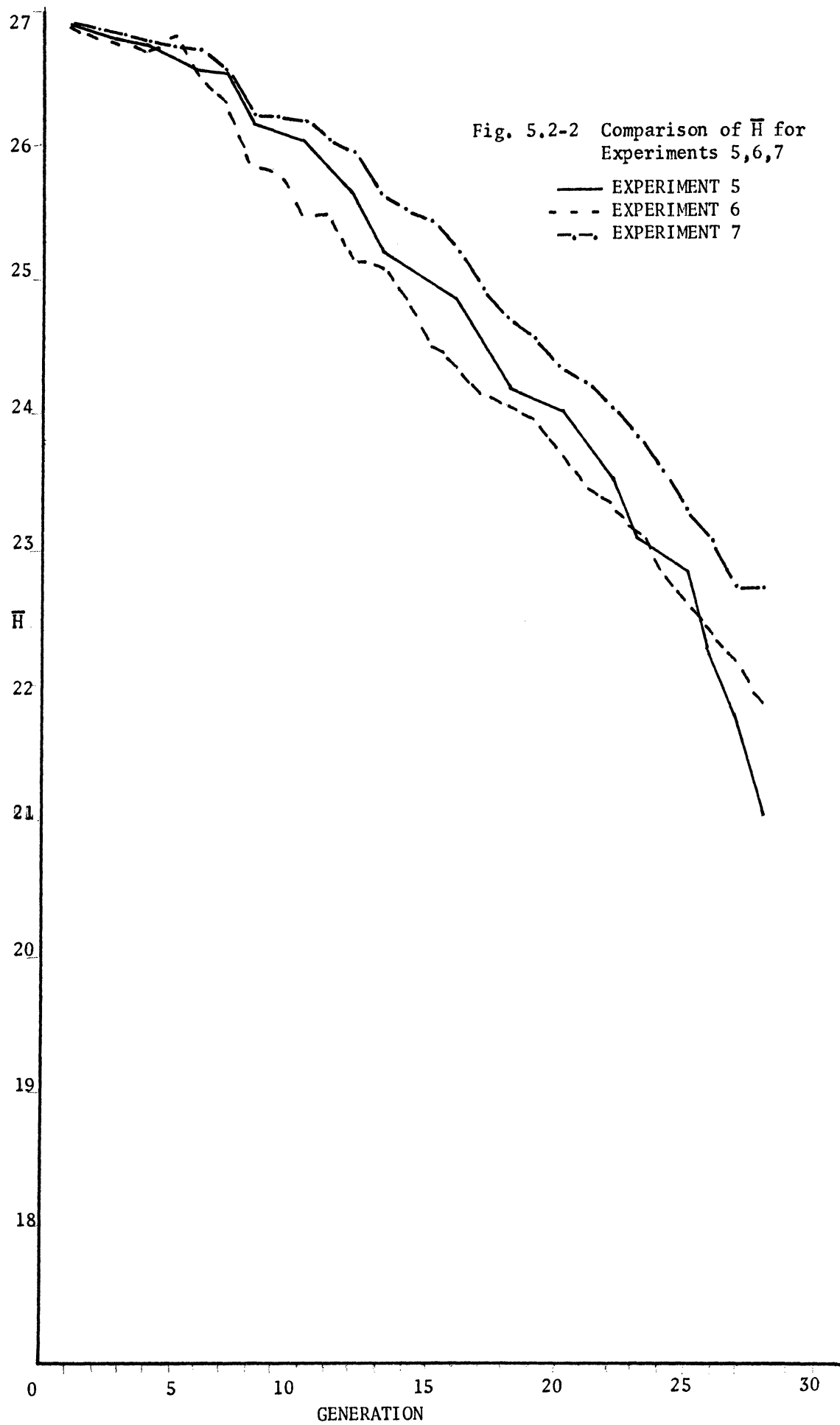
For brevity, we will refer to populations with a limited number of crossover positions as narrow populations and to those with many crossover positions as wide populations. With respect to the evolution of a population, the rate of evolution and the eventual evolutionary state are very important. Our previous discussion suggests the following conjecture:

The initial evolutionary rate of a narrow population is greater than that of a wide one but its eventual evolutionary state is less fit.

The present experiments suggest that this conjecture may hold but there is a difficulty in determining more precisely the actual number of crossover positions which distinguish narrow from wide populations. In the present model with nine genes we cannot say if under the fitness criterion being used, four crossover positions represents a narrow or a wide population.

Another question which we will not be able to answer is whether or not for such a model as we are using there is some optimum number of crossover positions which will result in a high rate of evolution and an eventually high population fitness. We will now turn to the experiments themselves for which the results are shown in Fig. 5.2-1 and Fig. 5.2-2.





The results of Fig. 5.2-1 seem to bear out the hypothesis concerning rates of evolution and the direction of the initial segment. In both experiments 6 and 7, the $\bar{F}_2$ curve overtook that of experiment 5 which had the advantage initially. Except for generation 28, which shows a strong downturn for $\bar{F}_2$ in experiment 5, it appeared that this population had essentially stopped improving. This is probably not the case for $\bar{H}$ has decreased from about 27 to 21, a decrease of about 22% which means there is still much genetic variation left in this population.

One very interesting result of these experiments is the degree to which each population has responded to the selection pressure as manifested in the tendency towards fixation of alleles 0-3 of gene 5. It will be recalled that these alleles are responsible for the production of chemical 23 which is the chemical toward which $\bar{F}_2$ is heavily weighted. Since the alleles are distinguished by an increase in the rate constant from 0 to 3 it is to be expected that the higher numbered alleles will be more favored and this is indeed the case as the following table of allele frequencies at generation 28 shows:

Experiment Number	Frequency of Alleles				Entropy of Gene 5, H_5	Average Entropy Per Gene, $\bar{H}/9$	Ratio of H_5 to $\bar{H}/9$
	0	1	2	3			
5	.0325	.0950	.1725	.5725	1.9144	2.3993	.80
6	.0850	.1875	.2625	.3950	2.0751	2.4281	.85
7	.0900	.0575	.0325	.6925	1.5953	2.5289	.63

Table 5.2-2 ALLELE FREQUENCIES AND ENTROPY FOR EXPERIMENTS 5, 6, AND 7

Although the average entropy per gene in experiment 7 is the highest, the entropy at gene 5 is the lowest. Thus while maintaining more overall genetic variation than in the other experiments, experiment 7 has achieved

a greater degree of fixation at the most crucial gene. It appears then that the effect of an increased number of crossover positions is to maintain genetic variation ($\bar{H}$ has only decreased by about 16% in 28 generations) while allowing for flexibility in the exploration of the genetic space.

The fixation of alleles 2 and 3 of gene 5 represents the most direct response of the system since the simplest way to decrease $\bar{F}_2$ is to increase the concentration of chemical 23. The concentrations of the other two chemicals will not become important until little further improvement can be achieved with respect to chemical 23. In some sense then the population responds to effect the coarsest (and also major) improvement immediately with further improvement expected to be much less dramatic because of the lesser importance of chemicals 24 and 25 in the fitness function.

Alleles 0 to 3 of gene 2 are responsible for the production of chemical 24 and alleles 4 to 7 of the same gene are responsible for the production of chemical 25. Consider the following table, again of allele frequencies at generation 28.

Experiment	Frequency of Alleles at Gene 2								Sum of Alleles 0-3	Sum of Alleles 4-7
	0	1	2	3	4	5	6	7		
5	.220	.035	.235	.0325	.0025	.2875	.1125	.0750	.5228	.4775
6	.040	.0825	.040	.3050	.145	.155	.195	.0375	.4675	.5325
7	.080	.090	.0775	.2150	.145	.270	.060	.0625	.4625	.5375

Table 5.2-3

ALLELE FREQUENCIES AT GENE 2 FOR EXPERIMENTS 5, 6, AND 7

The figures in the final two columns of this table are of some interest. Since the contributions of chemicals 24 and 25 to the fitness

function are equal we would expect the alleles producing them to occur with equal frequency and this is very nearly the case. We might wonder, however, why the alleles with the highest rate constant do not have the highest frequency. Among alleles 0 to 3, allele 3 does have by far the highest frequency in experiments 6 and 7 but for alleles 4 to 7, the frequency of allele 7 is very low in all experiments whereas the frequency of alleles 4 and 5 is generally high.

As a final comment, curves of $\bar{H}$ are generally much smoother than those of $\bar{F}_2$. They seem less subject to temporary fluctuations. In the long run, they are also an accurate reflection of the state of the population with respect to the average fitness and present a smoothed version of the overall response to the selection force.

5.3 STRUCTURE OF THE ENVIRONMENT

Up to now we have not discussed the structure of the environment, other than to comment on the necessity of its being homogeneous as well as the usefulness of considering it to have a grid-like organization (see section 3.3). This latter form was found to be convenient for characterizing the permeation of chemicals through the cell membrane. In the computational process, the state of the environment is held constant until the permeation reactions have been carried out for all the cells after which the new state of the environment is calculated.

For all the experiments being discussed, the environment serves two main purposes. It provides the source of the initial chemicals for offspring (see section 4.3) and serves the usual function of being a food source. In this model, the food consists of chemicals 1 to 6, the

only chemicals allowed to permeate. Because the environment is being used to supply the chemicals for the offspring it is necessary that it contain these chemicals in sufficient quantity. For the set of experiments beginning with experiment 3 of this chapter, the environment begins with 40,000,000 molecules of each chemical. With a grid size of 1000 this results in each environment square having 40,000 molecules of each chemical.

The mechanism of providing the initial chemicals for offspring by taking them in a single 'lump' from the environment is itself unrealistic. But the interpretation given in section 3.4, that the purpose is to bypass the period of growth to maturity by supplying sufficient chemicals which ordinarily would be derived gradually from the environment, does appear to be reasonable. Upon the death of cells their chemicals are added directly to the corresponding chemicals of the environment.

Upon a cursory examination of the situation, it might be expected that these two complementary processes would achieve a balance of sorts. But a more careful consideration reveals the fact that due to the selection pressure on the population, the environment will be driven in an unsatisfactory manner. Certain chemicals will be rapidly depleted whereas others will become very abundant. Such an unbalance will occur because of the form of f_2 . The following detailed description will explain the reason for this situation.

As an aid to the discussion, we will employ the following diagram.

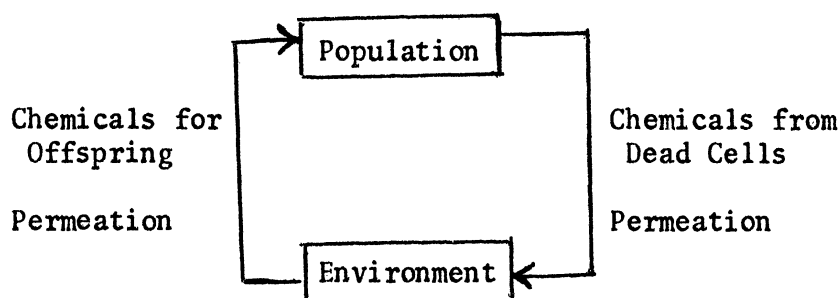


Fig. 5.3-1 CHEMICAL FLOW IN THE SYSTEM

This diagram depicts the nature of the chemical flow in the population-environment system. It is reasonable to assume that the effect of permeation on the state of the environment is minimal because it acts on the basis of a simple gradient for each chemical and therefore would by itself tend to maintain a balance. The major disruptive force arises from the difference between the flow of chemicals from the environment for offspring and the flow of chemicals to the environment from dead cells. For a full population, one cell must be killed to make room for each new offspring.

The chemical composition of offspring is roughly the same for all offspring over all generations, namely, 40,000 molecules per chemical. There is no such uniformity in the chemical constitution of a typical cell in the population of age greater than zero. The function $f_2 = 10(.15 - x_{23})^2 + (.1 - x_{24})^2 + (.1 - x_{25})^2$ favors the situation in which the three heaviest chemicals comprise 35% of the total concentration of a cell. Thus the population may be likened to a factory in which the raw materials are the primary chemicals, 1 to 6, and the finished products are the heaviest chemicals 23, 24, and 25.

Now the disequilibrium in the environment can be understood. Offspring begin life with an approximately uniform concentration distribution but under the chemical processes favored by the selection force, this distribution will shift towards the heaviest molecules. When such a cell is eliminated, the environment receives a disproportionately large number of the heaviest molecules with a correspondingly small number of the lightest ones. Since the offspring receives nearly equivalent numbers of all molecules, the net result is a buildup in molecules

23, 24, 25 and a depletion of molecules 1 to 6 in the environment. In terms of the factory analogy, the population is converting light molecules to heavy ones and then dumping these heavy molecules into the environment. This situation can be seen in the state of the environment after twenty-eight generations of experiments 5, 6 and 7. Consider the table below.

Chemical	Experiments		
	5	6	7
1	11,044,325	10,482,312	13,797,534
2	6,689,254	4,443,737	4,621,791
3	5,923,628	4,369,423	6,032,943
4	4,212,072	6,280,104	4,192,996
5	8,295,335	6,876,330	7,248,413
6	7,210,216	9,813,438	5,833,516
7	33,057,648	35,688,501	30,303,500
8	9,960,745	9,196,623	8,878,563
9	36,714,616	37,982,343	36,254,028
10	10,485,096	13,176,209	14,996,083
11	26,301,918	32,653,823	35,620,678
12	9,442,077	10,638,505	9,599,529
13	33,421,435	40,268,080	33,197,936
14	5,132,032	6,035,336	4,124,592
15	17,080,378	18,053,244	16,177,391
16	24,661,764	23,964,215	29,479,181
17	11,857,622	11,255,982	10,668,425
18	34,097,909	31,213,137	32,846,148
19	37,154,590	38,842,093	41,691,188
20	27,900,942	23,998,657	31,193,488
21	39,728,245	38,955,444	31,441,422
22	26,164,701	25,547,753	27,230,929
23	76,738,570	77,736,304	75,531,932
24	64,567,030	64,933,976	57,664,341
25	63,345,685	62,924,734	70,629,180

Table 5.3-1 DISTRIBUTION OF CHEMICALS IN THE ENVIRONMENT OF EXPERIMENTS
5, 6, AND 7

In all these experiments, the environment at the outset consisted of 40,000,000 molecules of each chemical. Since the distribution of molecules is quite consistent across the three experiments we can speak of any one of them as being typical. Two observations are immediate. Chemicals 2 to 6 and 14 are well below 10,000,000 and in some cases

below 5,000,000 molecules whereas chemicals 23, 24, and 25 are of the order of 70,000,000 molecules. If this trend continues, these lighter molecules will disappear from the environment, bringing the entire system to a halt. Now that we understand this process, how do we remedy it? The first thought is to begin with an environment distribution which has a larger number of light molecules and a smaller number of heavier ones. At best this would provide a temporary cure because it is an essentially static solution to a dynamic problem.

The solution which we have derived proves to be very instructive. It requires the existence of a process in the environment to decompose the heavy molecules. This immediately suggests the analogous situation in nature in which an individual receives food from the environment and converts it into complex organic molecules. When the individual dies, it is the turn of various forces in the environment including scavengers, lower order animals and bacteria, to decompose the organism. This then is one interpretation we can give to the process of reducing heavy molecules to light ones.

An alternative interpretation is to view the total system as a simple two-member ecological system. The population is the more complex part of this system and is undergoing selection pressure. It is necessary to construct a very simple process in the environment to complement the action of the population. Thus the population produces heavy molecules and the undefined 'population' in the environment decomposes them. Future research in this direction calls for a structure to be put on the environment 'population' and its role to be more carefully described.

This analysis has resulted in a reappraisal of the relationship between the population and the environment. The initial implicit assumption that the environment could be passive for these experiments proved to be false. To complement a dynamic population under a selection pressure it is necessary to have a dynamic environment.

The actual implementation of the process in the environment is done in an interesting way. The key point is to treat the environment as a special member of the population. It is given a fixed set of genetics which allow only decomposition reactions to occur. In the computation process, after the biochemical reactions have been carried out for each of the cells, the decomposition reactions are carried out for the environment.

The nature of the chemical flow depicted in Fig. 5.3-1 can now be specified in more detail.

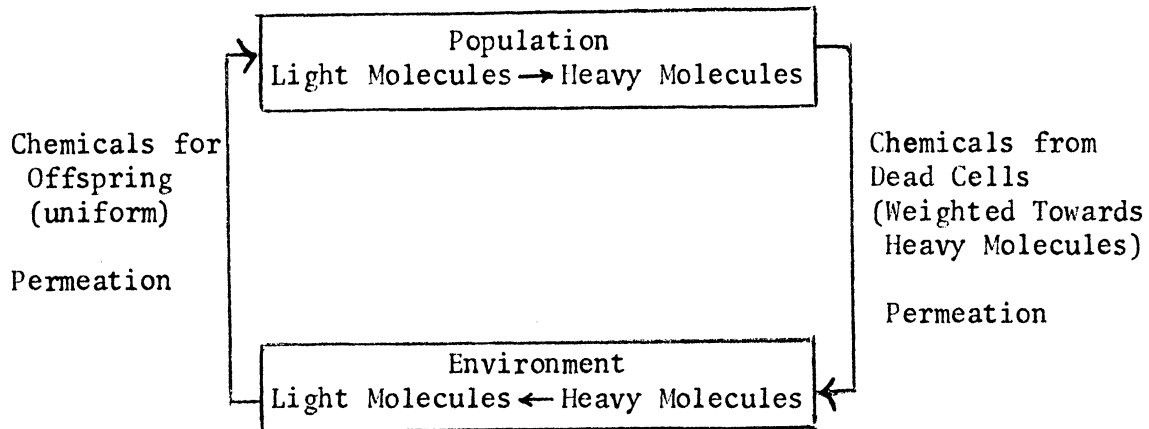


Fig. 5.3-2 REVISED CHEMICAL FLOW IN THE SYSTEM

All experiments to be performed from this point on, will have what is called 'environment genetics.' Before concluding this section, we will mention one experiment which has this feature.

Experiment 8 uses offspring function J and has eight crossover positions. To appreciate the effect of environment genetics, consider

Table 5.3-2 which presents the state of the environment of experiment 8 at generation 38. The state of the environment represented by Table 5.3-1 occurred at generation 28

Chemical	Number of Molecules
1	44,279,748
2	23,637,878
3	16,725,022
4	26,881,021
5	33,780,761
6	33,557,721
7	22,454,430
8	22,568,639
9	38,462,220
10	24,693,603
11	51,026,148
12	22,418,571
13	44,222,319
14	29,531,600
15	24,700,989
16	59,223,251
17	21,560,327
18	25,030,859
19	28,539,571
20	21,428,742
21	40,829,078
22	37,938,437
23	70,849,013
24	49,545,843
25	47,081,084

Table 5.3-2 DISTRIBUTION OF CHEMICALS IN THE ENVIRONMENT OF EXPERIMENT 8

There is only one chemical with fewer than 20,000,000 molecules and that is chemical 3 which has 16,725,022. The method of choosing a proper set of genetics for the environment is similar to the method for choosing a proper offspring function. It is essentially a trial and error method which requires that there be modification of the parameters until the outcome is acceptable.

The above table suggests that chemical 23 is not being decomposed quickly enough and that also the rate of the decomposition reaction which produces chemical 3 should be increased. On the whole, however, there does not seem to be the danger of exhausting any chemical in the near future.

5.4 INCREASED MATING FRACTION

Up to now we have used a mating percentage of .1 in all the experiments discussed in this chapter. In a population of 200 cells this results in 20 matings per generation. With the offspring functions that have been used, the average number of offspring per generation lies between 45 and 50. This means that there is almost a complete turnover in the population every fifth generation. The question naturally arises about the effect of the degree of population turnover per generation on the evolution of the population.

With respect to the previous discussion about sampling rate, the question was how to make the sampling more efficient, where efficiency was defined on the basis of the average number of 'good' offspring per mating. Might not a less efficient sampling rate with a greater number of samples be equivalent to a more efficient rate with fewer samples? In other words, in a few generations the production of a large number of offspring per generation could have an effect on the evolution of the population which is comparable to the effect achieved by the production of fewer offspring over a greater number of generations.

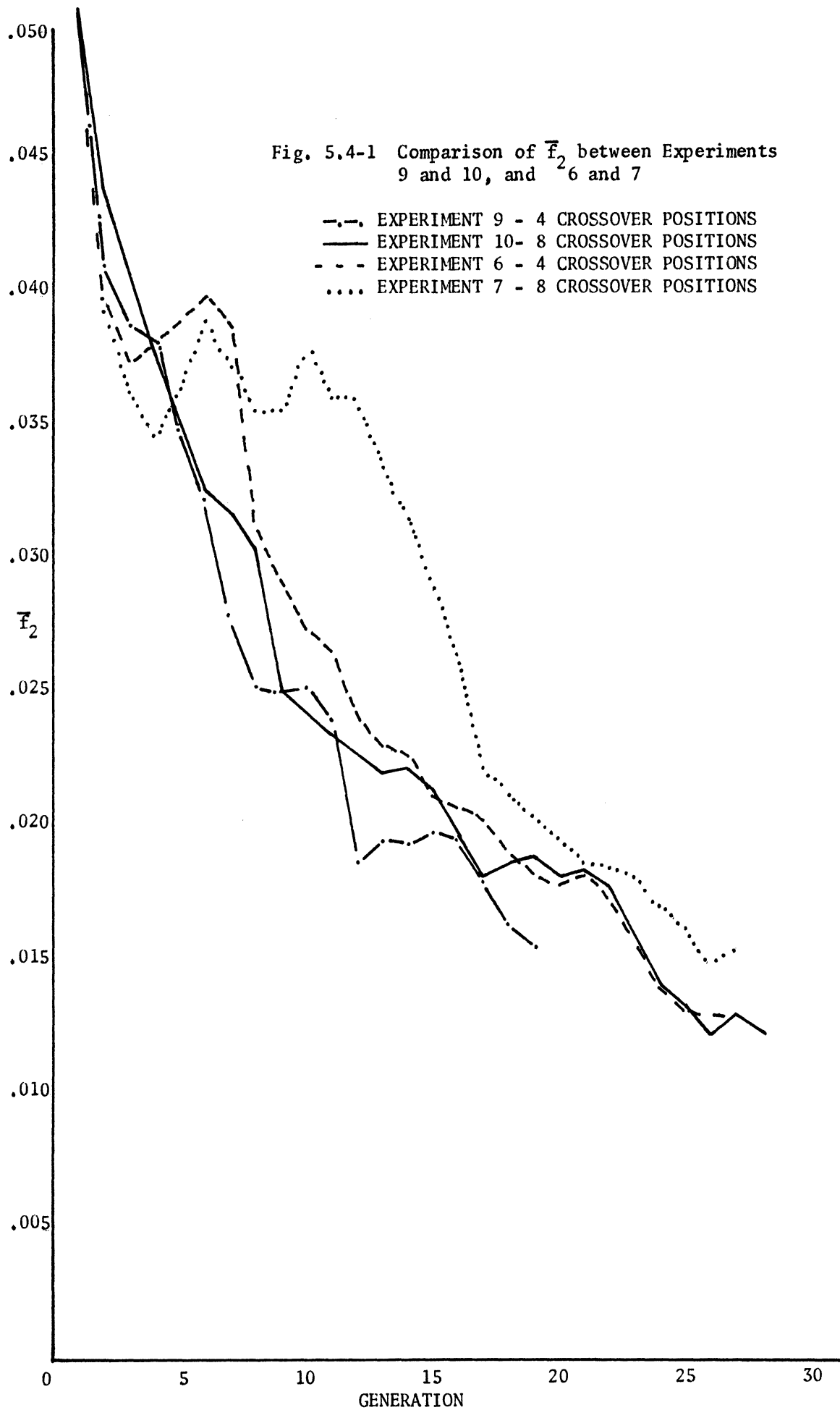
There is the problem in the former method that equally large numbers of cells must be eliminated to make room for the offspring. Thus a cell will have its number of opportunities for mating drastically reduced

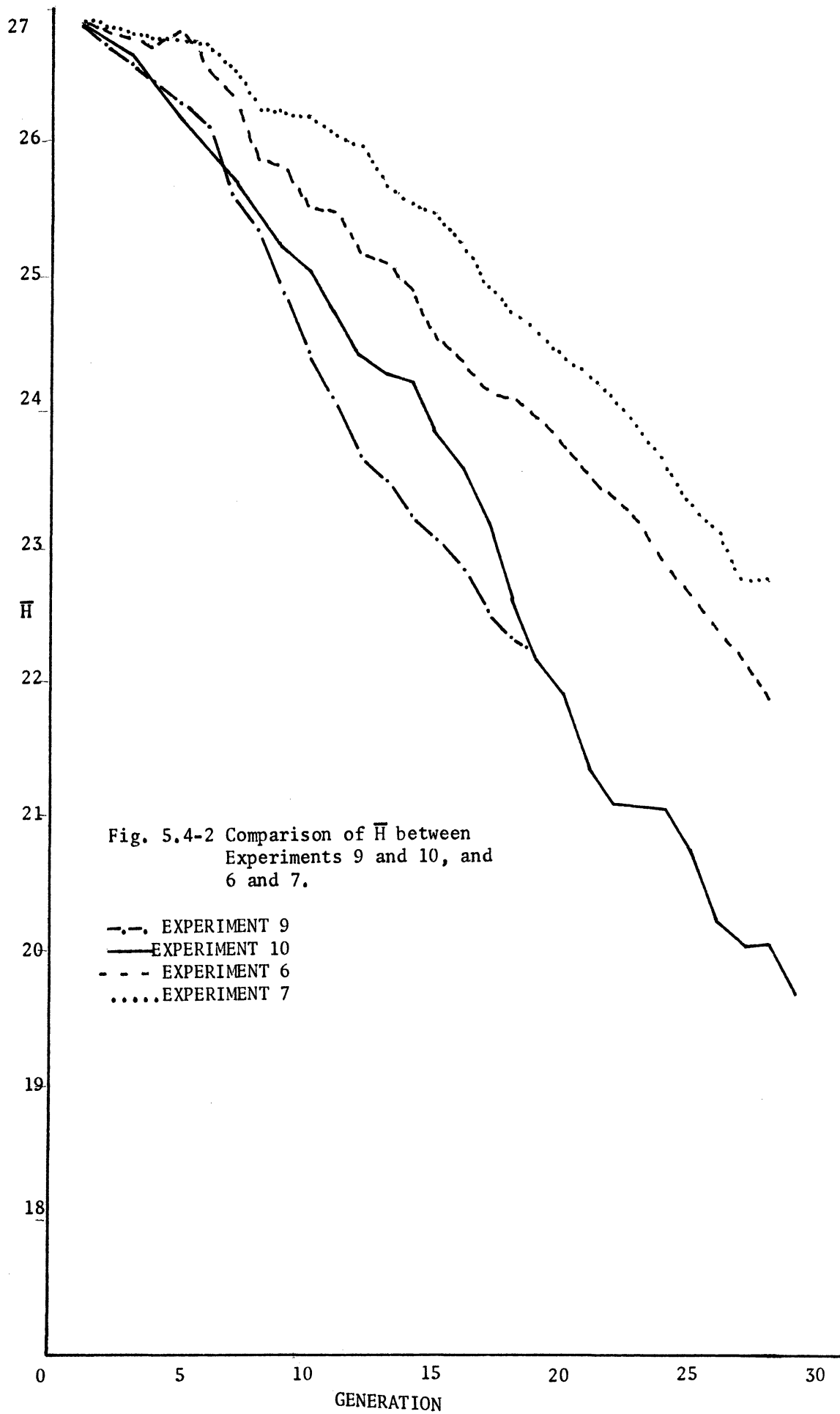
and therefore a good cell will not be able to contribute as much to the future population as it could under conditions of lower population turnover.

With a large population turnover, a cell will undergo a reduced number of biochemical time steps or equivalently a reduced life period. Part of the importance of this simulation is the weight attached to the interaction between the population and the environment over an extended period of time. Any severe reduction in the length of the overall biochemical cycle would impair the usefulness of the simulation. Nevertheless, it is interesting to see if there is any significant effect produced by increasing the mating fraction but not by too much.

In both experiments 9 and 10, the mating fraction is .2 (double the previous value) and the offspring function is K. Experiment 9 will have four crossover positions and experiment 10 will have eight. The best way to study these experiments is to compare them with comparable experiments done earlier. We have chosen experiments 6 and 7 which have four and eight crossover positions respectively, but differ in three ways: offspring function J instead of K, no environment genetics, and what is most important, a mating fraction of .1 instead of .2. The results are shown in Fig. 5.4-1 and Fig. 5.4-2.

Both experiments 9 and 10 have avoided the period of search between generations 1 and 10 which distinguished experiments 6 and 7. In experiment 10 the average number of offspring per generation is over 90 and scarcely any offspring can expect to live past the age of 2. Because of the heavy selection pressure, the populations of experiment 9 and 10 respond very rapidly but by generation 20, they are over-taken (but not passed) by the populations of the earlier experiments.





It is unfortunate that these experiments cannot be extended beyond generation 28, because of the problem with the environments of experiments 6 and 7. The comparison of the $\bar{H}$ curves of experiments 7 and 10 demonstrates most clearly the effect of the increased mating fraction. In experiment 7, $\bar{H}$ has decreased by nearly 16% but in experiment 10, it has decreased by nearly 26%. With respect to their values of $\bar{F}_2$, there has been a decrease of nearly 70% for experiment 7 and 74% for experiment 11. Thus, while there have been very comparable reductions in $\bar{F}_2$ over 28 generations, there has been a considerably greater loss in the genetic variation of the population subject to a greater population turnover.

This suggests the possibility that an increase in mating fraction from .1 to .2 overdrives the population initially and perhaps limits its future development. Because of reasons given previously, such a hypothesis cannot be tested on these experiments but future experiments are called for.

5.5 SUMMARY

Although the title of this chapter refers to crossover and selection, several other issues were discussed. This chapter continued the study of built-in selection begun in the previous chapter. The form of NEWCON-3 was amended and a description was given of a mortality distribution for the age groups of the population. We noted once again that it was important for the initial chemistry of the offspring to be largely independent of the state of the population.

In anticipation of the experiments on the number of crossover positions we introduced the notions of a genetic space and the exploration

of such a space. The experiments themselves verified to some extent the predictions about how the number of crossover positions would affect the behavior of the populations. A large number of crossover positions results in an initial period of search but the population eventually performs well with respect to populations with many fewer crossover positions. The question may be partially answered in the next chapter which is concerned with the selection of crossover probabilities. With a mechanism which allows the selection of crossover probabilities, the best strategy would be to allow the maximum number of crossover positions at the outset. Then through the process of selection the probability of crossover at certain positions would be greatly diminished and favorable groups of genes would arise. Other positions would still be 'open' to allow continued testing of new configurations. We postpone further discussion until the next chapter.

The nature of the relationship between a population under selection pressure and its environment was clarified. The implicit assumption of a passive environment existing harmoniously with a dynamic population was shown to be totally inadequate. There must be complementary processes in the population and the environment if a chemical balance is to be maintained. It was found necessary to provide the environment with a process for decomposing these chemicals which are being accumulated by the population responding to the selection force. Analogies were readily made with similar processes in nature for reducing complex organisms to simpler structures. A beginning was thereby made in extending the scope of the simulation to include simple ecological systems.

The chapter concluded with a discussion of the role of the degree

of population turnover on the evolutionary process as well as two experiments directed to this issue. Results were tentative but for the purposes of this simulation it was felt that a large population turnover each generation would be detrimental.

CHAPTER 6

SELECTION OF CROSSOVER PROBABILITIES

In the summary of the previous chapter, it was noted that a possible answer to the question of how many crossover positions should be allowed for the 'best' response of the population would depend on whether or not selection could operate on the crossover probabilities as well as the genes. In section 3.1, the concept of an ordered pair (a_i, x_i) for an allele was introduced. The first member of this pair, a_i , defines an enzyme (or enzymes) in the usual fashion and the second member, x_i , defines a "linkage" factor. A pair of linkage factors at a particular gene together define a crossover probability. Additional detail on the crossover probability and associated distribution curves was given in section 4.2.

It is the goal in this chapter to demonstrate the selection of crossover probabilities in the following sense. Along with the genetics, the crossover levels (x_i 's) are generated randomly for the initial population. Depending on the form of the functional relationship between the crossover levels and the crossover probability, there will be a characteristic shape for the initial distribution of the crossover probabilities. For the scheme used in previous experiments, this shape is triangular (see section 4.2). The appearance after many generations of a distribution curve which is weighted towards the lowest crossover probabilities would be evidence of the action of selection on them.

Given a chromosome of which two genes work well together, or produce a greater fitness than do any two other genes against the same genetic background, it is to the benefit of the population that these two genes remain together. To achieve this, it is necessary for the probability of crossing-over between the two genes to be reduced. For this reason we

have developed a mechanism which allows the linkage factor of a gene to be an object of selection. The fact that alleles $a_{ij}=(a_i, x_j)$ and $b_{jk}=(b_j, x_k)$ on some chromosome are very compatible means that the enzymes defined by a_i and b_j are compatible and it is possible to select values of x_i or x_j which are low enough to prevent this favorable combination from being lost.

Independent selection of the allele function and the linkage tendency is therefore possible. It is true that the linkage tendency exists only between neighbouring genes and therefore if alleles a_{ij} and b_{jk} are separated by several genes, it is necessary for all these crossover levels to be reduced if the a_{ij} - b_{jk} pair is to be preserved. As a first step, we will construct models in which the two genes of interest are in fact adjacent and thereby study the process of reducing the probability of crossing-over between a single pair of genes. The extension to larger numbers of genes should follow directly.

Computer simulations of genetic systems with linkage have been carried out by Fraser [1, 2] and Martin and Cockerham [10]. However, these differ significantly from our experiments in the following ways. Because the term recombination fraction is used instead of our term crossover probability when referring to their experiments, we will use the former. In the Martin and Cockerham experiments, the recombination fraction is the same between every pair of consecutive loci and is fixed at the beginning of an experiment. It is used as a parameter and its effect on selection is determined in the natural way by carrying out a series of experiments each for a different value of the recombination fraction. Fraser's experiments differ only in the fact that different recombination fractions can be assigned between each pair of adjacent loci. There is no possibility for exploring the process of selection of the recombination fraction.

Such a possibility is one of the significant features of the present simulation.

The first model presented in this chapter is used to show that it is not possible to achieve the desired goal by constructing a system based solely on the cell chemistry. In other words, it is necessary for there to be a relationship between the chemistry and the genetics which has a firm base in genetic theory. Such is the case with the second model which is derived from principles of theoretical genetics. In the development of this model, the action of the genes in terms of the enzymes they produce is carefully related to the theoretical description of their interaction.

Early problems which arise seem to be caused by the limited number of alleles at each position which results in sharp discontinuities between parents and offspring because of the very small number of distinct phenotypes. This number is readily increased by having eight alleles at each gene instead of two. Two other points are considered at this time. To aid the selective process the criterion for elimination of cells is extended to include fitness as well as age. An age group is chosen initially, but within this group the cells are eliminated on the basis of their fitness. The method for computing the crossover probability from the crossover levels is altered to weight more heavily the lower level of the pair. It is thought that this will help the selection of crossover probabilities.

These changes produce a much better response to the selection pressure but no response at all in the distribution of crossover probabilities. An analysis of this experiment brings to light another example of the unrecognized contribution of the initial chemical concentrations within offspring. The assignment of uniform concentrations relieves the selection pressure directed towards heterozygosity. This is an instance of a

non-genetic inheritance causing a cell to appear to be more fit than its genetics would suggest. Only the primary chemicals are assigned in any quantity to the offspring.

With this and other changes, an experiment is carried out which suggests that a theoretical analysis might be helpful. Indeed this analysis reveals that the system will not respond to selection until the mortality scheme is more effective. Also necessary is an increased range of offspring which can be assigned to any mating. All this is brought about through an awareness of the role of coupling and repulsion heterozygotes in the population, especially with respect to the effect of the magnitudes of their crossover probabilities.

The selection of crossover probabilities is achieved together with a better knowledge of the dynamics of this system. Other important factors involved are the effect of the form of the f -function on the nature of the selection, the relation of the range and distribution of crossover probabilities to the eventual improvement of the population, and the effects of large numbers of offspring on this improvement. All of these points will be discussed in more detail subsequently.

6.1 THE FIRST MODEL

This model was developed to illustrate the necessity of a firm theoretical base in such a study. The construction of a model based only on some features of the reaction graph is shown to be inadequate. To be more specific, consider the reaction graph shown in Fig. 6.1-1a. There are two distinct parts of the graph, each part producing one chemical which is used in the fitness function given below.

$$f_1 = (.30 - x_{10})^2 + (.2 - x_{11})^2 \quad .$$

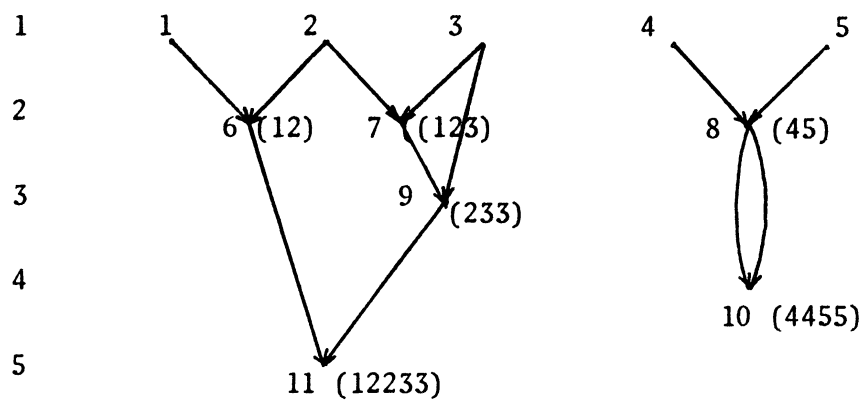


Fig. 6.1-1a REACTION GRAPH A

	1	2	3	4	5	6	7	8	9	10	11
1	0	6									
2		0	7								
3			0	0	0	0	9				
4				0	8						
5					0						
6						0	0	0	11		
7							0				
8								10			
9									0		
10										0	
11											0

Fig. 6.1-1b REACTION MATRIX A

		ALLELES							
GENES	1	0	1	2	3	4	5	6	7
	1	1+2→6, 2+3→7				3+7→9, 6+9→11			
	1	.0362	.0610	.111	.165	.0362	.0610	.111	.165
2	2	6→1+2, 7→2+3				9→3+7, 11→6+9			
	2	.0100	.0183	.0362	.0740	.0100	.0183	.0362	.0740
3	3	4+5→8, 8+8→10				8→4+5, 10→8+8			
	3	.0362	.0610	.111	.165	.0100	.0183	.0362	.0740

Fig. 6.1-1c ENZYME TABLE A-1

The genetics consist of 3 genes and 8 alleles per gene whose functions are shown in Fig. 6.1-1c.

The first two genes define all the reactions in the larger part of the reaction matrix and are responsible for the production of chemical 11. Gene 3 is responsible for the production of chemical 10. These processes are independent of each other. Genes 1 and 2 are so intimately related through the cell chemistry and at the same time unrelated to the activity of gene 3 that they should become strongly linked under selection pressure.

Two experiments are carried out on this model which differ only in the enzyme table. For the second experiment, the enzyme table labeled A-2 differs from A-1 in that alleles 4-7 of gene 1 are interchanged with alleles 0-3 of gene 2. The purpose is to make the genes uniform in the sense that each gene will have four alleles defining forward reactions and four alleles defining the corresponding reverse ones. Except for this difference in the enzyme table, both experiments are carried out under the following conditions:

- (a) Random mating
- (b) Offspring function D (see section 4.2)
- (c) The initial population size is 160 and the initial genetics is the same for both experiments. The population bound is 200.
- (d) The number of biochemical time steps per generation is 4.
- (e) $k = 12.0$
- (f) The maximum age of a cell is 6 and the mortality distribution described in section 5.1 is used.
- (g) Two crossover positions are used and the crossover probabilities are those of Table 5.1-2.
- (h) $r = 10$

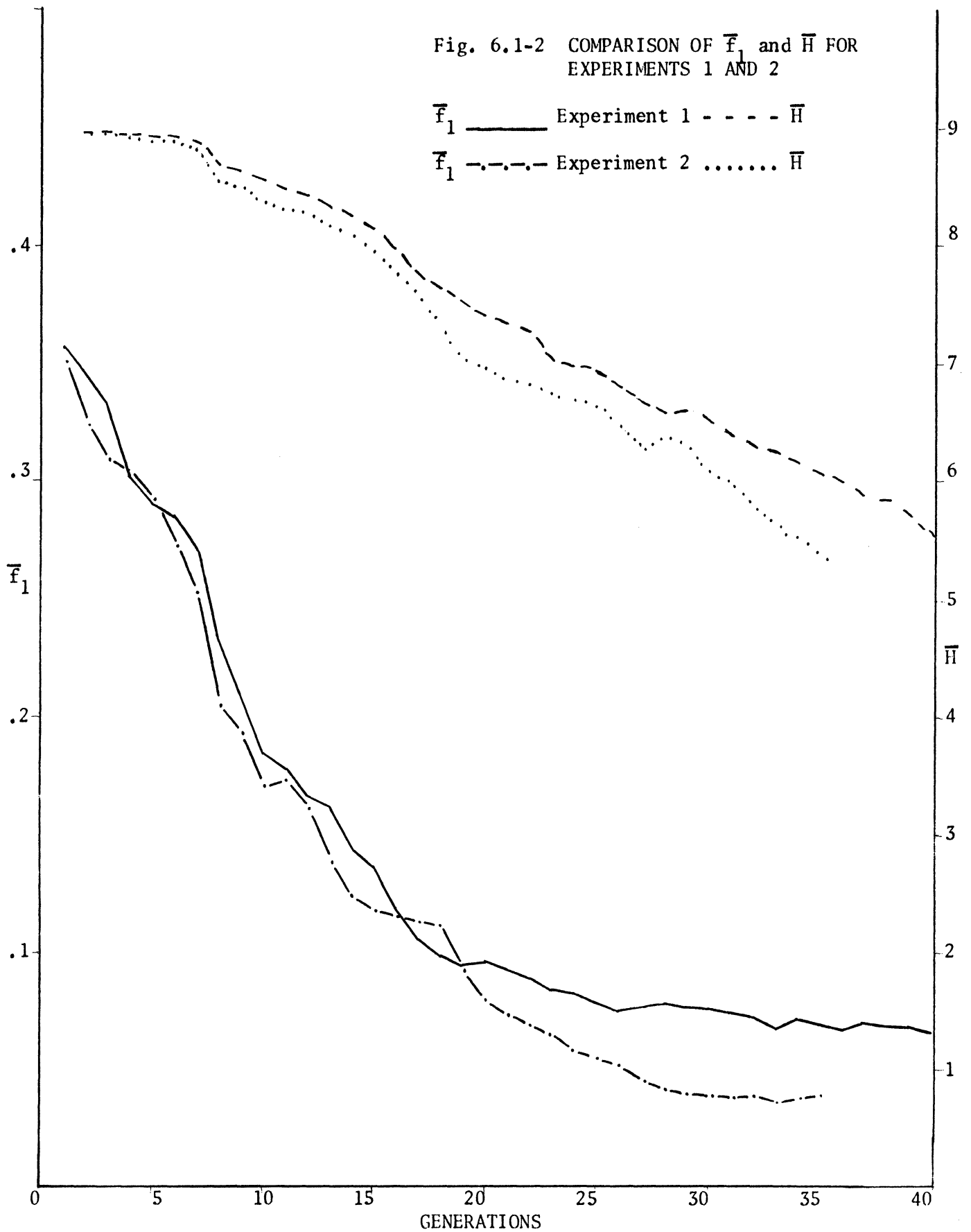
- (i) Mating fraction of .1
- (j) Chemicals 1 to 5 are allowed to permeate with fixed permeation constants of .0025.
- (k) Each cell has 1,000,000 molecules initially and concentrations of .091 for chemicals 1 to 10 and .090 for chemical 11.
- (l) The form of NEWCON used is similar to that described in section 5.1, except that instead of $40,000\frac{(r-1)}{r}$ molecules for each chemical of the offspring, $91,000\frac{(r-1)}{r}$ molecules are assigned for chemicals 1 to 10 and $90,000\frac{(r-1)}{r}$ for chemical 11.
- (m) Experiment 1 has enzyme table A-1 and experiment 2 has enzyme table A-2.

The $\bar{F}_1$ and $\bar{H}$ curves of these experiments are shown in Fig. 6.1-2. Both experiments show a very good response to selection pressure, especially experiment 2, where $\bar{F}_1$ is reduced from .35 to .04, nearly 90% in 35 generations. However, the important question is the nature of the distribution of crossover probabilities. Figs. 6.1-3a and 6.1-3b present the evolution of the distribution of crossover probabilities for both experiments. In experiment 1, this distribution is still very nearly triangular after 41 generations with a small shift of the peak from .250 to .200. In experiment 2, it is rather irregular but with some shift towards the center and low end.

However, in neither instance is there any indication of a pronounced shift towards low crossover probabilities. The type of reaction graph we have constructed is inadequate to produce the required result.

A search of the relevant literature uncovered a paper which proposed the model to be discussed in the next section.

Fig. 6.1-2 COMPARISON OF $\bar{F}_1$ and $\bar{H}$ FOR
EXPERIMENTS 1 AND 2



DISTRIBUTION OF CROSSOVER PROBABILITIES
FOR EXPERIMENTS 1 AND 2

Fig. 6.1-3a EXPERIMENT 1
POSITION 1-2

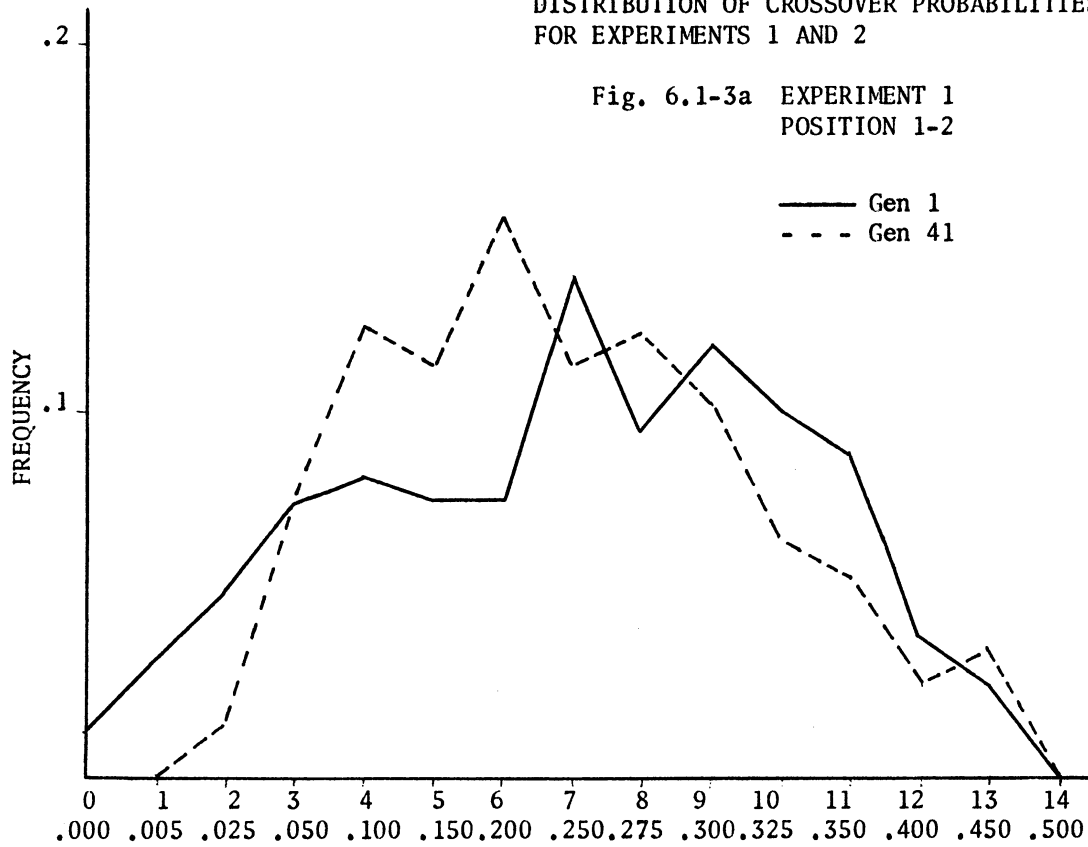
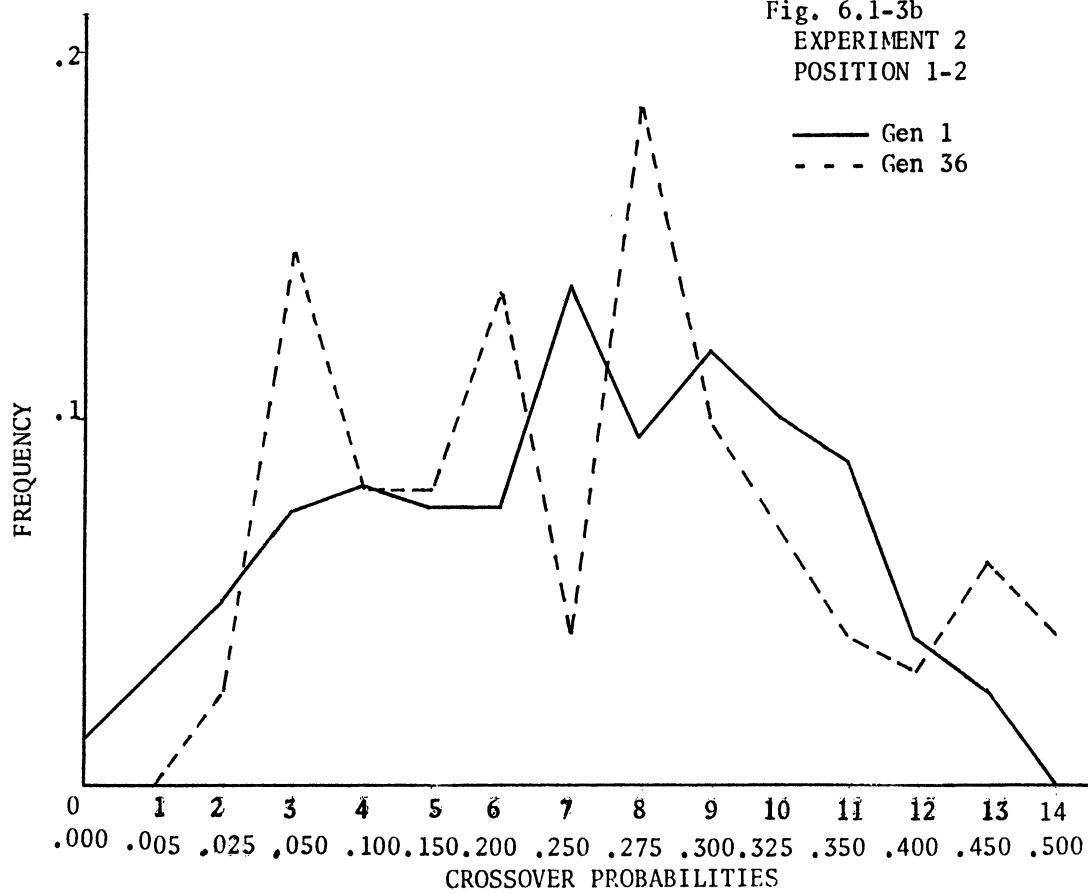


Fig. 6.1-3b
EXPERIMENT 2
POSITION 1-2



6.2 THE SECOND MODEL

Kimura in his paper [14] presents a discussion of some of the history of research into the modification of linkage intensity by natural selection. He reports that Fisher [10], without going into detail, states that "...the presence of pairs of factors in the same chromosome, the selective advantage of each of which reverses that of the other, will always tend to diminish recombination, and therefore tend to increase the intensity of linkage in the chromosomes of that species."

No discussion is given on the ways in which the two loci are maintained in a polymorphic state but only the consequences with respect to selection for decreased recombination of such a state. Kimura next discusses a specific model proposed by Dr. P. M. Sheppard at the 1955 Cold Spring Harbor Symposium (XX: Population Genetics) in which a stable polymorphism is maintained. Sheppard's model is outlined as follows:[14]

Assume a locus with a pair of alleles, say A_1 and A_2 , kept in balanced polymorphism by heterozygote superiority in fitness. Another pair of alleles, B_1 and B_2 are at another locus on the same chromosome, and interact with the genes in the first locus in such a way that A_1 is advantageous in combination with B_1 but is disadvantageous in combination with B_2 , while the situation is reversed for the gene A_2 . Then the second locus will remain polymorphic if linkage between the two loci is sufficiently close. Under this model, close linkage would maintain a larger fraction of the fitter genotypes than loose linkage and crossover reducing mechanisms will be favored by selection.

Kimura carries out a mathematical analysis of this model but it is based on the usual assumption of a large, randomly mating population. Since we are interested in the evolution of a small, closed population, the precise results derived by Kimura are not directly applicable. However, the model itself seems to be a very worthwhile object for experimentation, but first it must be recast in terms of a reaction graph and an enzyme table.

Consider the reaction graph of Fig. 6.2-1a. Like the previous model, it is relatively simple and consists of a few chemicals, in this case ten. The apparent simplicity of this model should not be misleading. In studying the selection of linkage intensity, we are beginning with a single pair of adjacent genes. That part of the cell biochemistry controlled by these two genes need not be very complex and in fact reaction graph B could be thought of as representing a small portion of the total chemistry.

The fitness function f_2 is of the following form:

$$f_2 = (.25 - x_9)^2 + (.25 - x_{10})^2 .$$

We are now able to discuss the enzyme table shown in Fig. 6.2-1c, with respect to the Sheppard model. Alleles 0,1 of gene 1 correspond to alleles A_1, A_2 and alleles 0,1 of gene 2 correspond to alleles B_1, B_2 . It is clear that gene 1 will be kept in balanced polymorphism for the following reason. Allele 1 is necessary to produce chemical 10 which is one of the chemicals in the fitness function. But the production of chemical 10 requires chemical 6, whose production is controlled by allele 0, and chemical 8, whose production depends on chemical 7 which again is controlled by allele 0. Thus both alleles are required to produce chemical 10.

For ease in discussing the inter-relationship between the alleles of gene 1 and those of gene 2 we will use subscripts on the alleles. Thus allele 1 of gene 2 is designated by 1_2 . Allele 0_1 is advantageous with allele 0_2 for two reasons. First there is no competition for chemicals between them and second the production of chemical 9 is necessary for the cell fitness. However, there is competition between alleles 0_1 and 1_2 over chemical 2, and hence they are disadvantageous in combination.

Alleles 1_1 and 1_2 are not in competition for any chemical and as the production of chemical 5 by allele 1_2 is necessary to produce chemical 9,

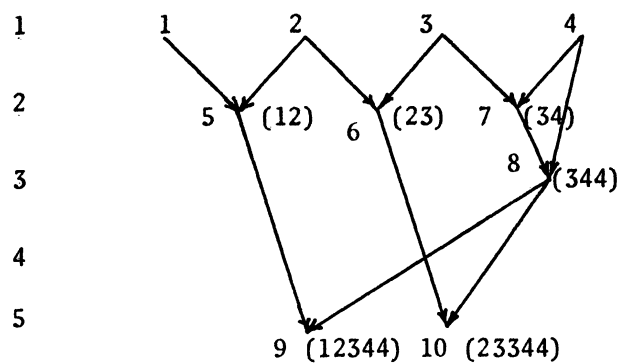


Fig. 6.2-1a REACTION GRAPH B

	1	2	3	4	5	6	7	8	9	10
1	0	5								
2		0	6							
3			0	7						
4				0	0	0	8			
5					0	0	0	9		
6						0	0	10		
7							0			
8								0		
9									0	
10										0

Fig. 6.2-1b REACTION MATRIX B

ALLELES

G E N E S	0		1	
	1	2+3→6, 3+4→7 .111	1	4+7→8, 6+8→10 .111
	2	5+8→9 .111	1	1+2→5 .111
	3	6→2+3, 7→3+4 .0362	1	6→2+3, 7→3+4 .0362
	4	8→4+7, 10→6+8 .0362	1	8→4+7, 10→6+8 .0362
	5	9→5+8 .0362	1	9→5+8 .0362
	6	5→1+2 .0362	1	5→1+2 .0362

Fig. 6.2-1c ENZYME TABLE B-1

which improves the fitness, they are advantageous together. But there is competition between alleles 1_1 and 0_2 over chemical 8 which is needed in the production of chemicals 10 and 9 respectively, which makes these two alleles disadvantageous together. At each of the remaining genes the alleles are the same and thereby provide each cell with the same set of reverse reactions. The cells only differ, therefore, in the genetics of the first two genes.

We will briefly discuss three experiments based on this model. The experiments are carried out under the following common conditions:

- (a) Random mating
- (b) Maximum cell age is 6 and the mortality distribution described in section 5.1 is used.
- (c) There is one crossover position between genes 1 and 2 and Table 5.1-2 is used.
- (d) $r = 10$
- (e) Mating fraction is .1
- (f) Chemicals 1 to 4 are allowed to permeate with fixed permeation constants of .005.
- (g) Each cell has 100,000 molecules initially.

The differences are based on the use of a new offspring function and two new initial chemistry procedures which are defined below.

Offspring Function L	Number of Offspring
$s \leq .3$	6
$s \leq .5$	5
$s \leq .6$	4
$s \leq .7$	3
$s \leq .9$	2
$s \leq 1.2$	1
Otherwise	0

Table 6.2-1
OFFSPRING
FUNCTION L

Chemicals	NEWCON-4	NEWCON-5
1 - 4	$16,600 \left(\frac{r-1}{r}\right) + M(q) X_i(q) \frac{1}{r}$	$13,000 \left(\frac{r-1}{r}\right) + M(q) X_i(q) \frac{1}{r}$
5 - 7	$8,400 \left(\frac{r-1}{r}\right) + M(q) X_i(q) \frac{1}{r}$	$9,000 \left(\frac{r-1}{r}\right) + M(q) X_i(q) \frac{1}{r}$
8	$4,200 \left(\frac{r-1}{r}\right) + M(q) X_i(q) \frac{1}{r}$	$8,000 \left(\frac{r-1}{r}\right) + M(q) X_i(q) \frac{1}{r}$
9 - 10	$2,100 \left(\frac{r-1}{r}\right) + M(q) X_i(q) \frac{1}{r}$	$6,500 \left(\frac{r-1}{r}\right) + M(q) X_i(q) \frac{1}{r}$

Table 6.2-2 NEWCON-4 and NEWCON-5

Unlike the previous NEWCON's, the number of molecules varies among the chemicals. Therefore, the index i has the range indicated in the column labeled Chemicals. It was thought that this distribution would be better than the strictly uniform distribution we have been using.

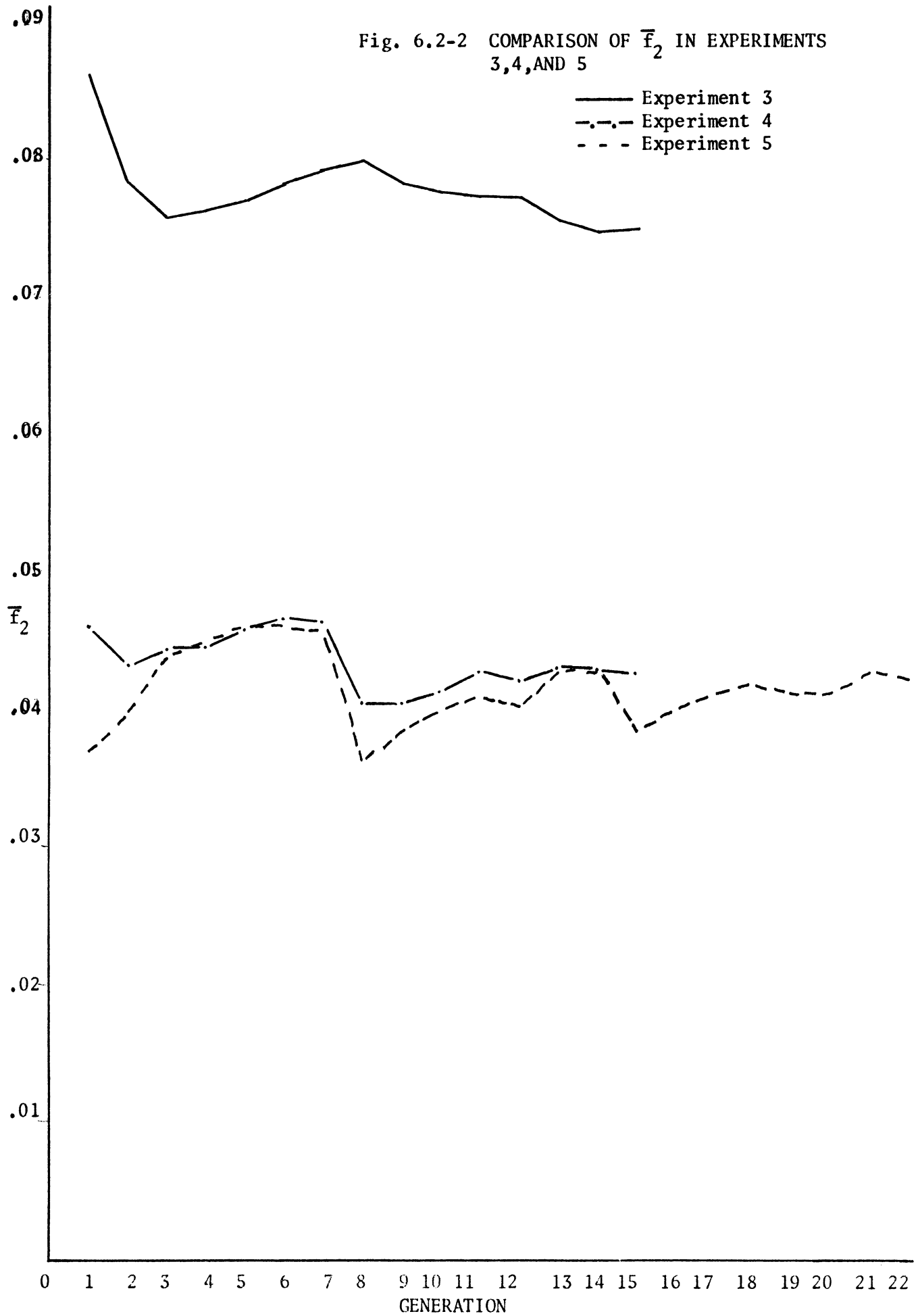
The three experiments differ in the following respects:

EXPERIMENT	OFFSPRING FUNCTION	NEWCON	NUMBER OF BIOCHEMICAL TIME STEPS	k	INITIAL POPULATION SIZE
3	D(Sec.4.2)	4	4	12.0	160
4	L	5	4	15.0	160
5	L	5	8	20.0	200

Table 6.2-3 DIFFERENCES IN EXPERIMENTS 3, 4, AND 5

The graphs of $\bar{f}_2$ for these three experiments is shown in Fig. 6.2-2. In all cases, the populations are not responding to any appreciable degree to the selection pressure. The offspring function used in experiment 3 produced a highly skewed offspring distribution with almost 90% of the matings producing either one or two offspring. This led to the use of offspring function L in the other two experiments along with an increased value of k in an attempt to produce greater phenotypic differentiation.

Fig. 6.2-2 COMPARISON OF $\bar{F}_2$ IN EXPERIMENTS
3,4,AND 5



Also included in this attempt was an increase in the number of biochemical time steps in experiment 5. While this was achieved, the populations still were not responding to the selection pressure. Before a discussion of the possible reasons, it is worthwhile to compare the phenotypes with their corresponding genotypes. For a two-gene system there are nine different genotypes. In the table below the genotypes are ordered according to their fitnesses (recall that a lower value of f_2 corresponds to a higher fitness). These results are taken from a typical set of values in experiment 6.

GENOTYPE	PHENOTYPE (f_2 -value)
$\frac{00}{11}$ or $\frac{01}{10}$	.0089
$\frac{00}{10}$	.0326
$\frac{11}{10}$	.0424
$\frac{11}{01}$	.0631
$\frac{10}{10}$	.0732
$\frac{01}{00}$	.0901
$\frac{11}{11}$	.0908
$\frac{00}{00}$	.1096
$\frac{01}{01}$	.1235

Table 6.2-4 ORDERING OF GENOTYPES BY
PHENOTYPIC VALUES

As expected, the double heterozygote has the best fitness by a factor of about 3.7. It is followed by three single heterozygotes and then by a double homozygote. Perhaps the only one out of place in this

arrangement is the genotype $\frac{01}{00}$ which might have been expected to precede the genotype $\frac{10}{10}$. In any case there is a strong correspondence between the expected order of the genotypes based on the structure of the reaction graph and the enzyme table, and their actual order.

The basic difficulty in the system appears to be due to the limited number of distinct phenotypes. A mating between 'good' cells is likely to produce several offspring which have poor fitnesses. With a larger number of phenotypes, there should not be such large discontinuities between parents and offspring. This suggests a new enzyme table with an increased number of alleles per gene to produce a larger number of genotypes. The following enzyme table will be used in the next experiment.

		ALLELES							
		0	1	2	3	4	5	6	7
G E N E S	1	2+3→6, 3+4→7				4+7→8, 6+8→10			
		.074	.111	.165	.247	.074	.111	.165	.247
	2	5+8→9				1+2→5			
		.074	.111	.165	.247	.074	.111	.165	.247
	3	6→2+3, 7→3+4				8→4+7, 10→6+8			
		.0183	.0272	.0362	.0510	.0183	.0272	.0362	.0510
	4	9→5+8				5→1+2			
		.0183	.0272	.0362	.0510	.0183	.0272	.0362	.0510

Table 6.2-5 ENZYME TABLE B-2

Although the number of genes has been reduced, the number of effectively different alleles has been increased, thereby increasing the number of genotypes. The experiment to be discussed next will also include a few changes, two of them major, to improve the response of the population. The first major change is a modification of the procedure for eliminating

cells. At present, cells are eliminated on the basis of the mortality distribution given in Table 5.1-1. Thus the only criterion for eliminating a cell is its age. To improve the response of the population, the fitness of the cell will also be made a factor in its survival in the following way.

In section 3.4, a method for positive assortative mating was discussed. The same principle is employed for the present purpose. That is, within every age group a cell is chosen on the basis of its fitness. Each cell is represented by a segment on the interval 0 to 1 whose length is inversely proportional to the fitness of the cell. A random number, uniformly generated, determines the cell to be eliminated. This method of elimination will be called the age-fitness mortality scheme. Such a scheme constitutes an additional site of selection pressure, for cells with higher fitnesses will have a better chance to survive and hence further opportunity to mate.

The second major change is related to the method for computing the crossover probability. The method used up to now was described in section 4.2. The probability of crossing over was made a function of the sum of the crossover levels. Thus two levels which sum to the same value correspond to the same crossover probability. For example, levels 1 and 8 are equivalent to levels 4 and 5. The scheme to be introduced will tend to weight the lower value of the pair of levels more heavily in computing the overall value. The crossover probability is based on the following function t of the two levels x_i and x_j

$$t(x_i, x_j) = 2 \frac{x_i x_j}{x_i + x_j}$$

For example, for the two pairs of crossover levels given above

$$t(1,8) = 16/9 = 1.778$$

and

$$t(4,5) = 40/9 = 4.444.$$

The effect of the one low crossover level in the pair (1,8) is to reduce significantly the overall value of the pair. Such a scheme, by heavily weighting the low crossover levels, should aid in the selection of low crossover probabilities.

With this procedure, each pair of crossover levels corresponds to a unique value. Thus to produce fifteen different probabilities as is done in the previous method, only five levels are necessary. The following table gives the relationship between the crossover levels and the corresponding crossover probability.

CROSSOVER LEVELS	$t(x_i, x_j)$	LINEAR ORDER	CROSSOVER PROBABILITY
1,1	1.00	0	.000
1,2	1.33	1	.067
1,3	1.50	2	.100
1,4	1.60	3	.120
1,5	1.67	4	.133
2,2	2.00	5	.200
2,3	2.40	6	.240
2,4	2.67	7	.267
2,5	2.86	8	.286
3,3	3.00	9	.300
3,4	3.43	10	.343
3,5	3.75	11	.375
4,4	4.00	12	.400
4,5	4.44	13	.444
5,5	5.00	14	.500

Table 6.2-6 CROSSOVER PROBABILITIES

The determination of the probabilities is based on the second column of this table.

There will be a return to the uniform concentration distribution for offspring. This has the same form as NEWCON-3, except that the number of molecules of each chemical M_i assigned to the offspring is

$$M_i = 10,000\left(\frac{r-1}{r}\right) + M(q)x_i(q)\frac{1}{r}.$$

The use of a uniform concentration distribution in conjunction with the present reaction graph and enzyme table results in a situation which is paralleled in nature. This will be discussed subsequently.

The last change is in the offspring function. As pointed out earlier, it is necessary to experiment continually with the offspring function in order to find one which produces a satisfactory offspring distribution. This is the reason for introducing so many different functions. For this experiment, offspring function N is used.

OFFSPRING FUNCTION N	NUMBER OF OFFSPRING
$s \leq .08$	6
$s \leq .2$	5
$s \leq .4$	4
$s \leq .8$	3
$s \leq 1.1$	2
$s \leq 1.8$	1
otherwise	0

Table 6.2-7 OFFSPRING FUNCTION N

Compared to offspring function L, function N reflects the great range of phenotypes in the population. The bound for six offspring in L

is .3 but it is .08 in N. However, the bound for one offspring is 1.2 in L and 1.8 in N.

The other conditions for this experiment are

- (a) $k = 12.0$
- (b) number of biochemical time steps = 4
- (c) initial population size is 200

and all the conditions common to experiments 3, 4 and 5, except for the number of crossover positions. In this experiment there are three crossover positions.

The graphs of $\bar{F}_2$ and $\bar{H}$ for experiment 6 are shown in Fig. 6.2-3. There is a significant improvement in the response of this population. In twenty generations, $\bar{F}_2$ has been reduced from .02 to .004, a factor of 5. This population is responding very well and shows the possibility of continued improvement. In fact, the standard deviation at generation 20 of .0039 is greater than .0037, the value of $\bar{F}_1$. The changes most responsible for this improvement are the new enzyme table and the age-fitness mortality scheme.

The distribution of crossover probabilities for generations 1 and 21 at position 1-2 is shown in Fig. 6.2-4. The expected initial distribution for crossover probabilities is no longer triangular. Each of the crossover levels from 1 to 5 is generated by a uniform distribution. Then the probability of each pair of levels x_i, x_j is $1/25$ for $x_i = x_j$ and $2/25$ for $x_i \neq x_j$. The present distribution at generation 1 corresponds fairly closely to the expected distribution. It is clear that at generation 21, the distribution has not shifted towards the lower probabilities. In fact, there appears to be a shift towards the highest probabilities. This experiment has not achieved the desired goal and reasons for this

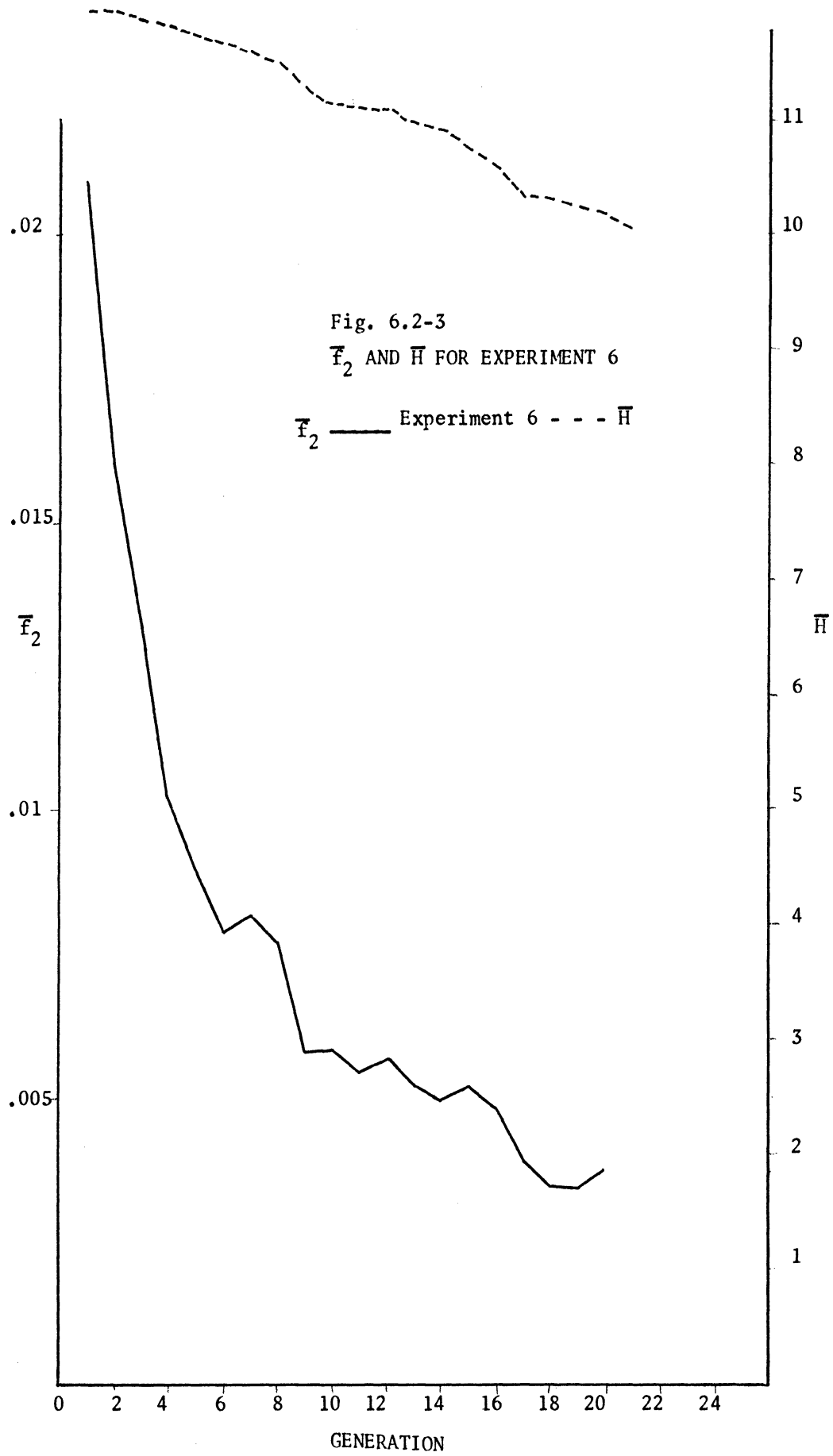
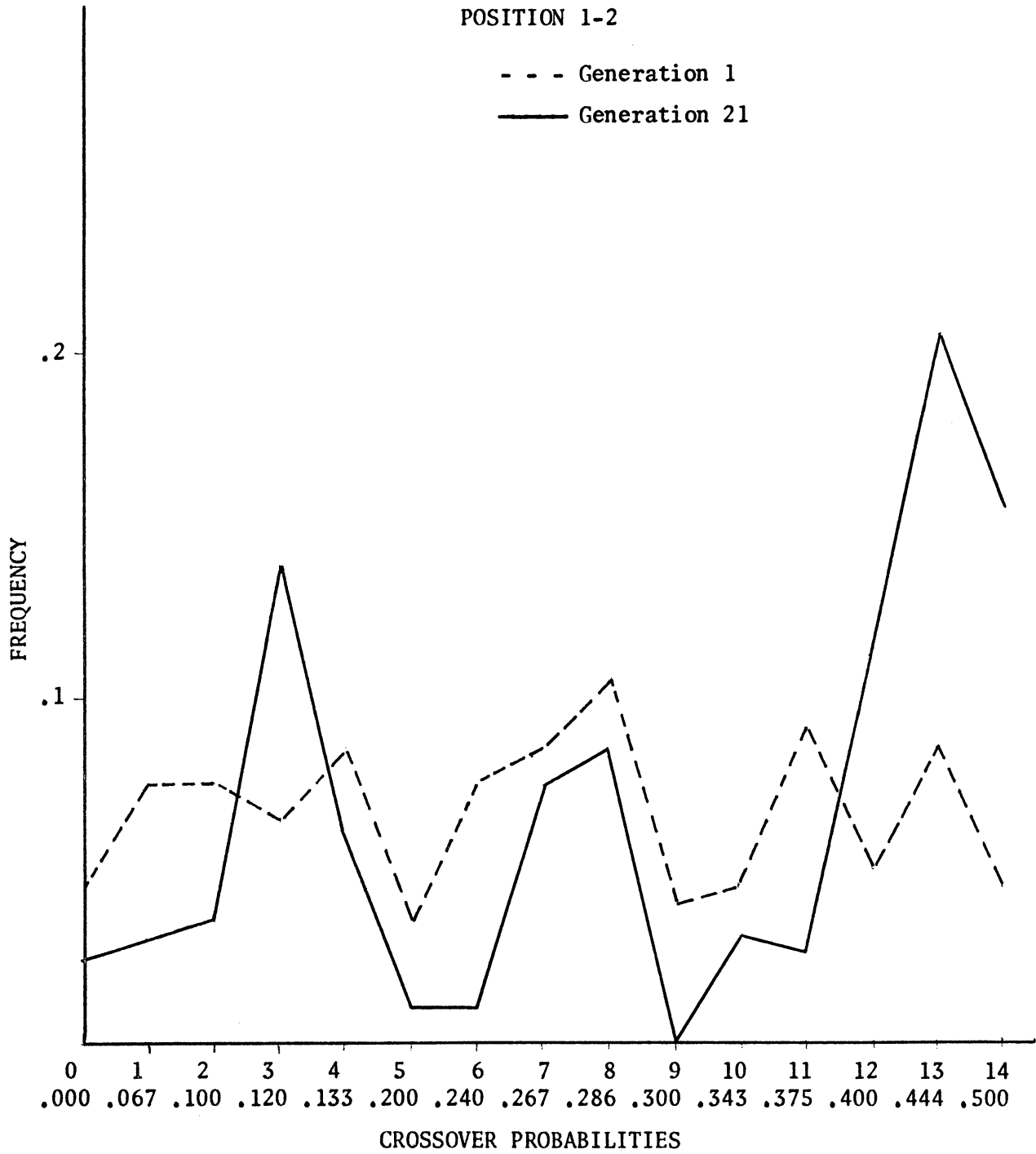


Fig. 6.2-4 DISTRIBUTION OF CROSSOVER PROBABILITIES FOR EXPERIMENT 6

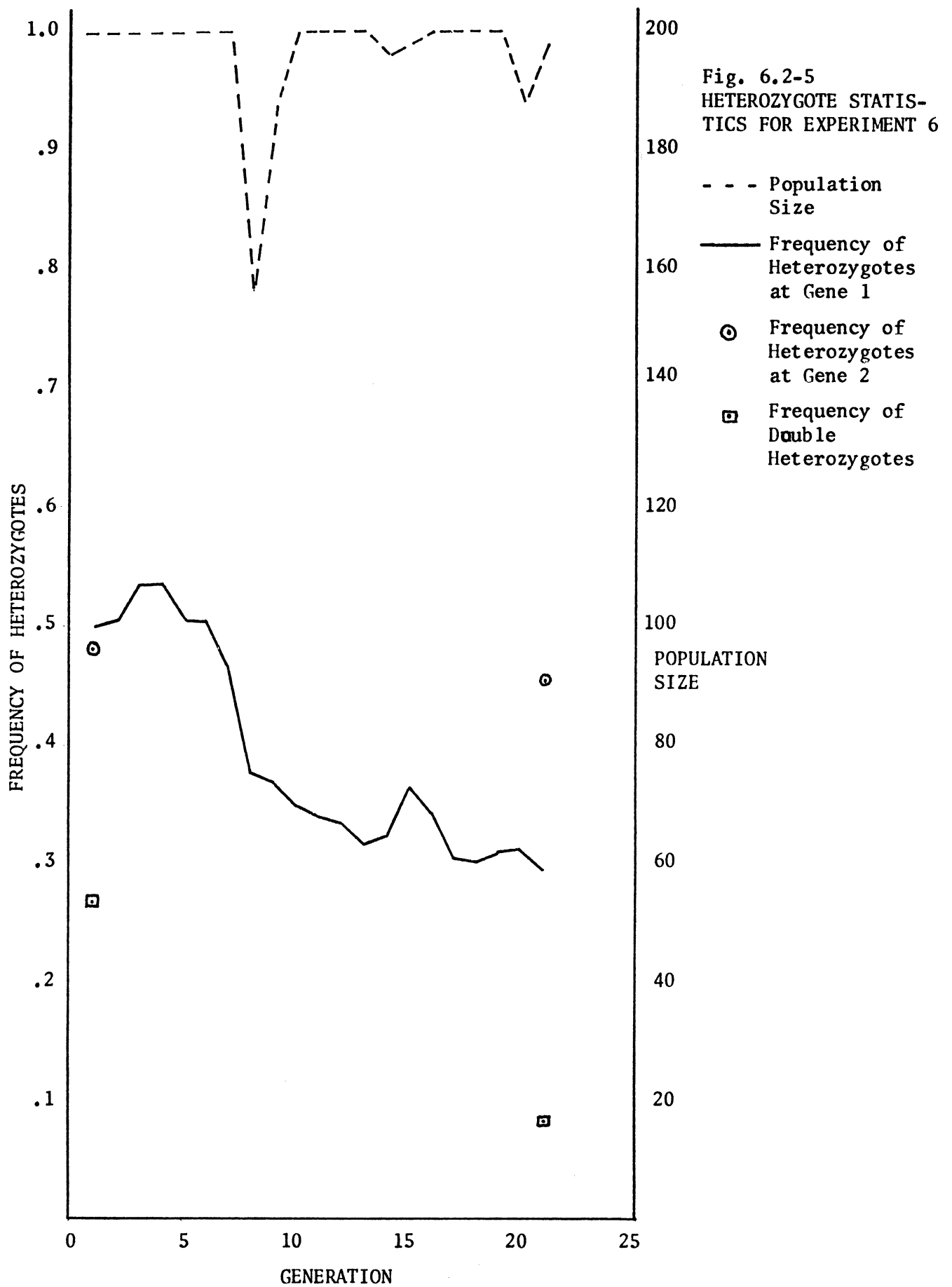


must now be sought.

An additional statistical aid has been developed to indicate an important aspect of the response of the population. The Sheppard model requires that the two loci in question be heterozygous in order for the crossover probability between them to be reduced. Thus it is important to monitor the degree of heterozygosity at each locus as well as the combined heterozygosity at both. For this experiment, only the first locus is so studied but the particular values at the second as well as the combined values are given for generations 1 and 21. These are shown in Fig. 6.2-5.

Because the genetics are generated randomly at the outset, it is to be expected that the frequency of heterozygotes at each locus will be .5 and the combined heterozygosity will be .25. These are very close to the actual values. However, by generation 21, the frequency of heterozygotes at gene 1 has fallen to .3 and the combined heterozygosity has fallen to .085. It is only at gene 2 that the frequency has remained relatively constant. Thus the population is able to improve as evidenced by the shape of the fitness graph but without having to increase or even maintain heterozygosity at the two relevant loci. This is contrary to the requirements of the Sheppard model. How then is the population able to improve with a decreasing number of double heterozygotes?

The answer is an example of several features in the system combining to produce a result which otherwise would not arise. In this case it is the combination of the initial concentration function and the structure of the enzyme table. For this experiment, a uniform concentration distribution is assigned to the offspring. The enzyme table was constructed so that it would be necessary for both genes 1 and 2 to be heterozygous if



a good fitness value were to be achieved. For the population to perform well without a significant number of double heterozygotes means that the other genotypes are achieving better fitnesses than they should.

This is indeed the case as a simple example will show. The worst cell is a double homozygote which has alleles in the range 0 to 3 at gene 1 and 4 to 7 at gene 2. It is the worst because such a cell does not have the capacity to produce either chemical 9 or 10. However, the initial chemistry function assigns concentrations of approximately .1 to both these chemicals. Whatever the decomposition reactions are, they do not reduce these values very much because of their low rate constants. The fitness value for this cell is

$$f_1 = (.25-.10)^2 + (.25-.10)^2 = .045.$$

It should be noted that at generation 20, the value of $\bar{f}_1$ is nearly .040. Therefore, because of the initially high concentrations even a cell which should have a low fitness because of its genotype, does not.

Cells are being supplied with chemicals which they should have to produce themselves in order to achieve high fitnesses. Therefore, there is little reason under these circumstances for double heterozygotes to have much of a selective advantage. This situation is somewhat similar to a natural phenomenon called Dauermodifikationen discussed by Sonneborn [15] , p. 308, reporting Jollos:

"By the term Dauermodifikationen, Jollos means temporarily inherited alterations producible by environmental means, the degree of effect or the duration of its persistence often being increased by repeated or long-lasting environmental treatment. The effects persist for hundreds or even thousands of fissions, but sooner or later they invariably disappear, the disappearance being hastened by changes of environment and by fertilization. The physical basis of Dauermodifikationen is presumed to be in the cytoplasm or in the macronucleus or in both."

Cells in the present experiment are able to survive and reproduce,

because the chemicals assigned to them as offspring are adequate for their needs independent of their specific genetics. In some sense this is comparable to Dauermodifikationen in which inherited cytoplasm in combination with special environments prolong the life of certain effects in paramecia.

To achieve the desired goal with respect to the Sheppard model, it is necessary therefore to assign a different distribution of chemicals to the offspring. We will discuss this procedure, as well as an experiment incorporating this change and others, in the next section.

This section has involved the construction and development of a new model which will demonstrate the selection of crossover probabilities. Several experiments were discussed which pointed out problems associated with the functioning of the model. A highly interesting development was the occurrence of a situation which has a strong parallel in nature. The relationship to this phenomenon called Dauermodifikationen once again illustrates the fact that the present system incorporates many features found in natural populations.

6.3 A THEORETICAL ANALYSIS

In the previous section, the procedure used to assign the initial offspring chemistry was found to have an influence on the evolution of the population which overrode the genetic limitations. That is, because of the high concentration of the chemicals assigned, the cells had relatively high fitnesses and were able to survive and reproduce beyond the point to which their genetics would otherwise have limited them. To remedy this situation, offspring will be provided in quantity with only the primary chemicals, 1, 2, 3 and 4. The concentration function

NEWCON-6 is defined below

CHEMICALS	NUMBER OF MOLECULES OF CHEMICAL i
1 - 4	$10,000(\frac{r-1}{r}) + M(9) x_i(9) \frac{1}{r}$
5 - 10	$M(9) x_i(9) \frac{1}{r}$

Table 6.3-1 NEWCON-6

With this function, all cells will receive nearly equivalent amounts of chemicals 1 to 4. Those offspring of good parents will receive a small amount of chemicals 9 and 10. In a certain sense, therefore, cells are able to inherit extra-chromosomal properties from their parents and the offspring of good parents are able to benefit from this. Of course with a value of $r = 10$, the quantity of chemicals assigned to offspring on the basis of their parent's chemistry is quite limited.

In this chapter, experiment 7 will be described which will serve as a source of statistics in an analysis of the performance of the population. Because the population is basically a two-gene two-allele system, it is possible to predict a great deal about its behavior, especially possible reasons for its lack of success. But first we present additional background for this experiment.

The enzyme table given below has been constructed to focus attention more clearly on genes 1 and 2.

It differs from enzyme table B-2 in that the production of chemicals 7 and 8 has been made common for all cells. Thus each allele of genes 1 and 2 references only one reaction and the interaction among these

A L L E L E S

		0	1	2	3	4	5	6	7
G E N E S	1	2+3→6				6+8→10			
		.074	.111	.165	.247	.074	.111	.165	.247
	2	5+8→9				1+2→5			
		.074	.111	.165	.247	.074	.111	.165	.247
	3	6→2+3 7→3+4				8→4+7 10→t+8			
		.183	.0272	.0362	.0510	.0183	.0272	.0362	.0510
S	4	9→5+8				5→1+2			
		.0183	.0272	.0362	.0510	.0183	.0272	.0362	.0510
	5	3+4→7				3+4→7			
		.111	.111	.111	.111	.111	.111	.111	.111
	6	4+7→8				4+7→8			
		.202	.202	.202	.202	.202	.202	.202	.202

Table 6.3-2 ENZYME TABLE B-3

reactions is sharper. Because the initial chemistry of offspring consists mainly of chemicals 1 to 4, and all cells require the primary chemicals 3 and 4, the permeation constants are increased. The present values are .01, .03, .03 and .04 for chemicals 1 to 4 respectively, compared with the previous values of .005 for each. The differences are related to the number of times each chemical can appear in a reaction and the magnitude of the rate constants associated with these reactions.

Once again a new offspring function is used but on this occasion it is of a different form as well. The offspring function variable s is now defined as

$$s = f_2^i + f_2^j$$

where f_2^i and f_2^j are the f_2 -values of the parents i and j . No longer does s also depend on $\bar{F}$ as in previous functions. The reason for this is the limited range of possible phenotypes. Observation of previous experiments

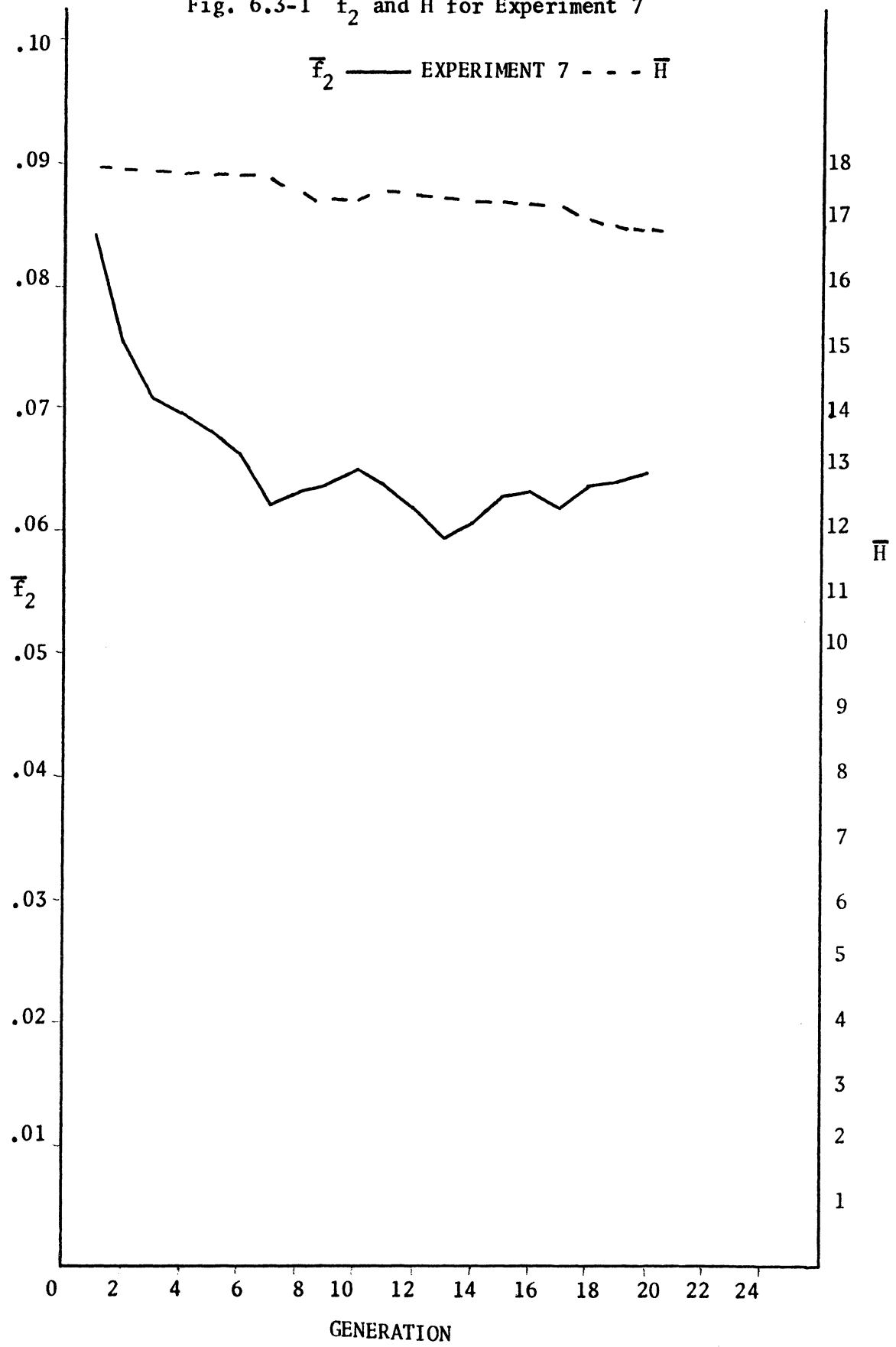
OFFSPRING FUNCTION 0			NUMBER OF OFFSPRING
s	$\leq$	.05	6
s	$\leq$	.07	5
s	$\leq$	.10	4
s	$\leq$	.14	3
s	$\leq$	.18	2
s	$\leq$	.21	1
otherwise			0

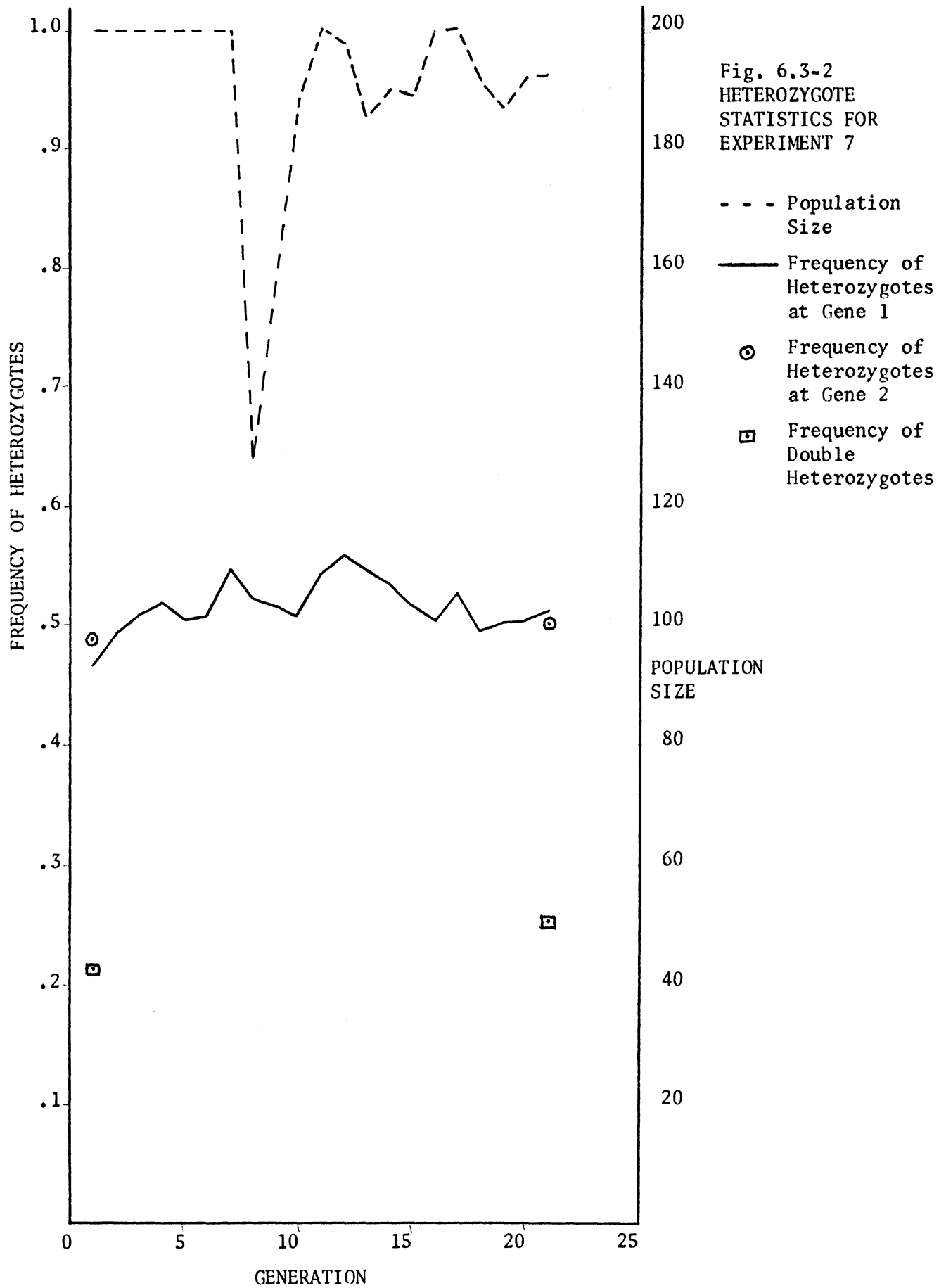
Table 6.3-3 OFFSPRING FUNCTION 0

indicates that there are three apparent groupings of fitness values, namely (.01-.03), (.06-.08), and (.10-.125). This suggests that the assignment of offspring should not be dependent on the overall fitness of the population because cells of these types will always be present. Thus the range of good cells, for example, will always be the same and hence such cells should always be awarded with the same number of offspring.

Other changes are an increase in the value of k to 15.0 and the number of biochemical time steps from 4 to 6. The purpose is to increase the rate at which cells approach their ultimate fitnesses. The number of molecules assigned to cells initially is 40,000 with 10,000 each to chemicals 1 to 4. This is in line with the form of NEWCON-6. It is clearly the case now that cells will have to be heterozygous at both genes 1 and 2 in order to succeed.

The results of experiment 7 are shown in Fig. 6.3-1 and Fig. 6.3-2. Both the $\bar{F}_2$ and the $\bar{H}$ curves indicate very little response in this

Fig. 6.3-1 $\bar{F}_2$ and $\bar{H}$ for Experiment 7



population. In fact, $\bar{H}$ has only decreased by 5.6% in twenty generations. Another indication of this is the percentage of double heterozygotes in the population. In twenty generations this percentage has only increased from 20.5% to 24.9%. Unless there is a substantial increase in their number, the population will not improve. It seems necessary at this point to carry out a theoretical analysis of the population in order to understand its behavior.

As mentioned previously genotypes can be divided into three classes on the basis of the phenotypes. In this classification, only genes 1 and 2 are used to determine the genotype because the major contribution to the phenotype derives from them and not genes 3 and 4. Ignoring differences in phenotype due to differences in age, these are the three groupings of the cells: (.01-.03), (.06-.08), and (.10-.125). They are written in terms of fitness values. The corresponding division of the genotypes is given in the table below.

		PHENOTYPES		
		.11 - .125	.06 - .08	.01 - .03
G E N O T Y P E S		1	2	3
	a	$\frac{00}{00}$	$\frac{10}{00}$	$\frac{00}{11}$
	b	$\frac{11}{11}$	$\frac{11}{01}$	$\frac{01}{10}$
	c	$\frac{01}{01}$	$\frac{01}{00}$	
	d	$\frac{10}{10}$	$\frac{11}{10}$	

Table 6.3-4 GENOTYPE CLASSES

The four double homozygotes are in class 1 which corresponds to fitness in the range (.10-.125). The four single heterozygotes are in class 2 which corresponds to fitnesses in the range (.06-.08) and the two double heterozygotes are in class 3, which has the lowest fitness range, (.01-.03). Any specific genotype can be referred to by the column-row label; for example, the genotype $\frac{01}{00}$ is labeled 2c.

We will now carry out all possible matings among cells from these three classes to determine the distribution of offspring. The cases can be listed in the following way:

MATINGS	NUMBER OF SUBCASES
1 x 1	10
1 x 2	16
1 x 3	8
2 x 2	10
2 x 3	8
3 x 3	3

Table 6.3-5 MATING CLASSES

Assume that the probability of crossing over between genes 1 and 2 is p . Then the following table lists the proportions of each gamete which can

G E N O T Y P E S										
	$\frac{00}{00}$	$\frac{11}{11}$	$\frac{01}{01}$	$\frac{10}{10}$	$\frac{10}{00}$	$\frac{11}{01}$	$\frac{01}{00}$	$\frac{11}{10}$	$\frac{00}{11}$	$\frac{01}{10}$
	1a	1b	1c	1d	2a	2b	2c	2d	3a	3b
GAMETES 00	1				1/2		1/2		1/2(1-p)	1/2p
01			1			1/2	1/2		1/2p	1/2(1-p)
10				1	1/2			1/2	1/2p	1/2(1-p)
11		1				1/2		1/2	1/2(1-p)	1/2p

Table 6.3-6 GAMETE PRODUCTION

be produced by each genotype. It is only in the gamete production of the double heterozygotes that the crossover probability p plays a role.

All the possible matings are shown in Figs. 6.3-3 to 6.3-8 inclusive. The final column in these figures gives the proportion of each class produced and is calculated in the following way. Since the initial population is generated randomly, the distribution of genotypes is shown in the table below together with the actual initial distribution in experiment 7.

THEORETICAL					EXPERIMENT 7	
CLASSES	GENOTYPES	FREQUENCY	EXPECTED NUMBER	EXPECTED NUMBER IN EACH CLASS	ACTUAL NUMBER OF EACH GENOTYPE	ACTUAL NUMBER IN EACH CLASS
1	$\frac{00}{00}$	1/16	12.5	50	10	54
	$\frac{11}{11}$	1/16	12.5		14	
	$\frac{01}{01}$	1/16	12.5		14	
	$\frac{10}{10}$	1/16	12.5		16	
2	$\frac{10}{00}$	2/16	25	100	22	103
	$\frac{11}{01}$	2/16	25		28	
	$\frac{01}{00}$	2/16	25		34	
	$\frac{11}{10}$	2/16	25		19	
3	$\frac{00}{11}$	2/16	25	50	16	43
	$\frac{01}{10}$	2/16	25		17	

Table 6.3-7 DISTRIBUTION OF GENOTYPES IN EXPERIMENT 7 COMPARED WITH RANDOM DISTRIBUTION

		1 X 1										Propor. e.c.prod.
		1ax1a	1ax1b	1ax1c	1ax1d	1bx1b	1bx1c	1bx1d	1cx1c	1cx1d	1dx1d	
1	$\frac{00}{00}$	1										
	$\frac{11}{11}$					1						
	$\frac{01}{01}$								1			1/4
	$\frac{10}{10}$										1	
	$\frac{10}{00}$				1							
2	$\frac{11}{01}$						1					
	$\frac{01}{00}$			1								1/2
	$\frac{11}{10}$							1				
	$\frac{00}{11}$		1									
	$\frac{01}{10}$											
3	$\frac{00}{11}$											
	$\frac{01}{10}$											1/4
	$\frac{10}{10}$								1			

Fig. 6.3-4 Matings within Class 1

		2 X 2										Propor. ea.class prod.
		2ax2a	2ax2b	2ax2c	2ax2d	2bx2b	2bx2c	2bx2d	2cx2c	2cx2d	2dx2d	
1	$\frac{00}{00}$	1/4		1/4					1/4			
	$\frac{11}{11}$					1/4		1/4			1/4	
	$\frac{01}{01}$					1/4	1/4		1/4			1/4
	$\frac{10}{10}$	1/4			1/4						1/4	
	$\frac{10}{00}$	1/2		1/4	1/4					1/4		
2	$\frac{11}{01}$					1/2	1/4	1/4		1/4		
	$\frac{01}{00}$		1/4	1/4			1/4		1/2			1/2
	$\frac{11}{10}$		1/4		1/4			1/4			1/2	
	$\frac{00}{11}$		1/4		1/4		1/4			1/4		
	$\frac{01}{10}$		1/4	1/4				1/4		1/4		1/4

Fig. 6.3-5 Matings within Class 2

1 X 3

		1ax3a	1bx3a	1cx3a	1dx3a	1ax3b	1bx3b	1cx3b	1dx3b	Proportion of Each Class Produced
1	$\frac{00}{00}$	$\frac{1}{2}(1-p)$				$\frac{1}{2} p$				1/4
	$\frac{11}{11}$		$\frac{1}{2}(1-p)$				$\frac{1}{2} p$			
	$\frac{01}{01}$			$\frac{1}{2} p$				$\frac{1}{2}(1-p)$		
	$\frac{10}{10}$				$\frac{1}{2} p$				$\frac{1}{2}(1-p)$	
	$\frac{10}{00}$	$\frac{1}{2} p$			$\frac{1}{2}(1-p)$	$\frac{1}{2}(1-p)$			$\frac{1}{2} p$	
2	$\frac{11}{01}$		$\frac{1}{2} p$	$\frac{1}{2}(1-p)$			$\frac{1}{2}(1-p)$	$\frac{1}{2} p$		1/2
	$\frac{01}{00}$	$\frac{1}{2} p$		$\frac{1}{2}(1-p)$		$\frac{1}{2}(1-p)$		$\frac{1}{2} p$		
	$\frac{11}{10}$		$\frac{1}{2} p$		$\frac{1}{2}(1-p)$		$\frac{1}{2}(1-p)$		$\frac{1}{2} p$	
	$\frac{00}{11}$	$\frac{1}{2}(1-p)$	$\frac{1}{2}(1-p)$			$\frac{1}{2} p$	$\frac{1}{2} p$			
	$\frac{01}{10}$			$\frac{1}{2} p$	$\frac{1}{2} p$			$\frac{1}{2}(1-p)$	$\frac{1}{2}(1-p)$	

Fig. 6.3-6 MATINGS BETWEEN CLASSES 1 AND 3

2 X 3

		2ax3a	2bx3a	2cx3a	2dx3a	2ax3b	2bx3b	2cx3b	2dx3b	Proportion of Each Class Produced
1	$\frac{00}{00}$	$\frac{1}{4}(1-p)$		$\frac{1}{4}(1-p)$		$\frac{1}{4} p$		$\frac{1}{4} p$		1/4
	$\frac{11}{11}$		$\frac{1}{4}(1-p)$		$\frac{1}{4}(1-p)$		$\frac{1}{4} p$		$\frac{1}{4} p$	
	$\frac{01}{01}$		$\frac{1}{4} p$	$\frac{1}{4} p$			$\frac{1}{4}(1-p)$	$\frac{1}{4}(1-p)$		
	$\frac{10}{10}$	$\frac{1}{4} p$			$\frac{1}{4} p$	$\frac{1}{4}(1-p)$			$\frac{1}{4}(1-p)$	
	$\frac{10}{00}$	1/4		$\frac{1}{4} p$	$\frac{1}{4}(1-p)$	1/4		$\frac{1}{4}(1-p)$	$\frac{1}{4} p$	
2	$\frac{11}{01}$		1/4	$\frac{1}{4}(1-p)$	$\frac{1}{4} p$		1/4	$\frac{1}{4} p$	$\frac{1}{4}(1-p)$	1/2
	$\frac{01}{00}$	$\frac{1}{4} p$	$\frac{1}{4}(1-p)$	1/4		$\frac{1}{4}(1-p)$	$\frac{1}{4} p$	1/4		
	$\frac{11}{10}$	$\frac{1}{4}(1-p)$	$\frac{1}{4} p$		1/4	$\frac{1}{4} p$	$\frac{1}{4}(1-p)$		1/4	
	$\frac{00}{11}$	$\frac{1}{4}(1-p)$	$\frac{1}{4}(1-p)$	$\frac{1}{4}(1-p)$	$\frac{1}{4}(1-p)$	$\frac{1}{4} p$	$\frac{1}{4} p$	$\frac{1}{4} p$	$\frac{1}{4} p$	
	$\frac{01}{10}$	$\frac{1}{4} p$	$\frac{1}{4} p$	$\frac{1}{4} p$	$\frac{1}{4} p$	$\frac{1}{4}(1-p)$	$\frac{1}{4}(1-p)$	$\frac{1}{4}(1-p)$	$\frac{1}{4}(1-p)$	

Fig. 6.3-7 MATINGS BETWEEN CLASSES 2 AND 3

3 X 3				
	3ax3a	3ax3b	3bx3b	Proportion of Each Class Produced
1	$\frac{00}{00}$	$1/4(1-p)^2$	$1/4p(1-p)$	$1/4$
	$\frac{11}{11}$	$1/4(1-p)^2$	$1/4p^2$	
	$\frac{01}{01}$	$1/4p^2$	$1/4p(1-p)$	
	$\frac{10}{10}$	$1/4p^2$	$1/4(1-p)^2$	
2	$\frac{10}{00}$	$1/2p(1-p)$	$1/4(1-p)^2 + 1/4p^2$	$1/2$
	$\frac{11}{01}$	$1/2p(1-p)$	$1/4(1-p)^2 + 1/4p^2$	
	$\frac{01}{00}$	$1/2p(1-p)$	$1/4(1-p)^2 + 1/4p^2$	
	$\frac{11}{10}$	$1/2p(1-p)$	$1/4(1-p)^2 + 1/4p^2$	
3	$\frac{00}{11}$	$1/2(1-p)^2$	$1/2p(1-p)$	$1/4$
	$\frac{01}{10}$	$1/2p^2$	$1/2(1-p)^2$	

Fig. 6.3-8 MATINGS WITHIN CLASS 3

Within each class the numbers of each genotype are the same. Fig. 6.3-7 will be used to carry out a sample calculation. There are 4 genotypes in class 2 and 2 genotypes in class 3; hence the number of different mating genotype pairs is eight. Each column contains the distribution of offspring resulting from the designated mating. Therefore the sum of the terms in each column is 1. The proportion of matings of each pair is one-eighth the total because the pair-wise products are all equal. Thus to determine the overall distribution among the classes, the terms in each column are divided by eight and all the columns are summed. The resulting values are grouped into classes, yielding the final distribution which is the same for all the matings.

The overall distribution of offspring is $1/4$, $1/2$, $1/4$ with respect to the classes 1, 2, and 3, respectively. This distribution is independent of the crossover probability p and, what is most important, independent of selection. That is, even if some matings have a greater number of offspring assigned, the distribution will remain the same. In experiment 7, the initial distribution among the classes is 54, 103 and 43 respectively. After twenty generations, the distribution is 43, 101 and 49 respectively. The number of active cells is 183 and the proportions are .237, .552 and .268 respectively. There has been very little shift from the theoretically expected distribution.

An increase in the number of double heterozygotes is to be expected because of the age-fitness mortality scheme which favors the elimination of less fit cells. But the effect of this scheme is not as strong as it might be because of the low number of offspring produced per generation and the rather large numbers of cells eliminated because of old age. However, a more effective mortality scheme will be an important selection

force.

The foregoing analysis suggests that as presently constituted the population will never respond to selection pressure unless several changes are introduced. These depend on a more careful examination of the mating classes, especially matings within class 3. It is only within this class that crossover can play a significant role. The two types of double heterozygotes $\frac{00}{11}$ and $\frac{01}{10}$ are commonly termed coupling and repulsion, respectively, and will be referred to here as C and R. Because an improvement in the population can come about only by an increase in double heterozygotes it is necessary to investigate mechanisms which can effect this result. Part of Fig. 6.3-8 is reproduced below in a slightly altered form.

MATINGS	DOUBLE HETEROZYGOTE OFFSPRING DISTRIBUTION								
	C			R			C+R		
	PROPOR- TION OF C	p=0	p=1/2	PROPOR- TION OF R	p=0	p=1/2	PROPORTION OF C+R	p=0	p=1/2
C x C	$1/2(1-p)^2$	1/2	1/8	$1/2p^2$	0	1/8	$\frac{1}{2}(1-2p+2p^2)$	1/2	1/4
C x R	$1/2p(1-p)$	0	1/8	$1/2p(1-p)$	0	1/8	$p(1-p)$	0	1/4
R x R	$1/2p^2$	0	1/8	$1/2(1-p)^2$	1/2	1/8	$\frac{1}{2}(1-2p+2p^2)$	1/2	1/4

Table 6.3-8 DISTRIBUTION OF DOUBLE HETEROZYGOTE OFFSPRING
FROM DOUBLE HETEROZYGOTE PARENTS

In the C x C and R x R matings, the proportion of double heterozygote offspring increases from 1/4 to 1/2 as p, the crossover probability, decreases from 1/2 to 0. However, such a decrease in p for the C x R matings results in a decrease of this proportion from 1/4 to 0. There is a clear disadvantage to a reduction in the crossover probability for C x R matings, which are the most probable matings if the number of C's and R's is equivalent. Even with a p of .5, these matings produce only 1/4 double

heterozygotes. Thus an increase in double heterozygotes can occur only through C x C or R x R matings, and these matings will become more frequent if the number of either C's or R's increases. This suggests the following scheme for increasing the number of double heterozygotes and for selecting low crossover probabilities.

The number of C's and R's is the same at the outset. Through a more effective age-fitness mortality procedure the number of double heterozygotes will increase. Due to random fluctuations, there will be the possibility of increases in C over R or vice versa. For the moment assume that the number of C's exceeds the number of R's. This will increase the probability of C x C matings. Those C x C matings among C's with low crossover probabilities will produce a high frequency of C offspring. These C offspring will again have low crossover probabilities and will lead to an increased number of C's. It is by such a mechanism that a temporary increase in one of the double heterozygote types over the other can lead to a more substantial increase with an associated decrease in the crossover probabilities. This is the crucial point in the following experiments.

6.4 FIRST SUCCESSFUL RESULTS

A number of changes whose purpose is to favor conditions necessary to produce the selection of crossover probabilities will be presented. The most important of these, a more effective mortality procedure was mentioned in the last section. The new procedure was developed with natural populations in mind and involves mortality at two different times.

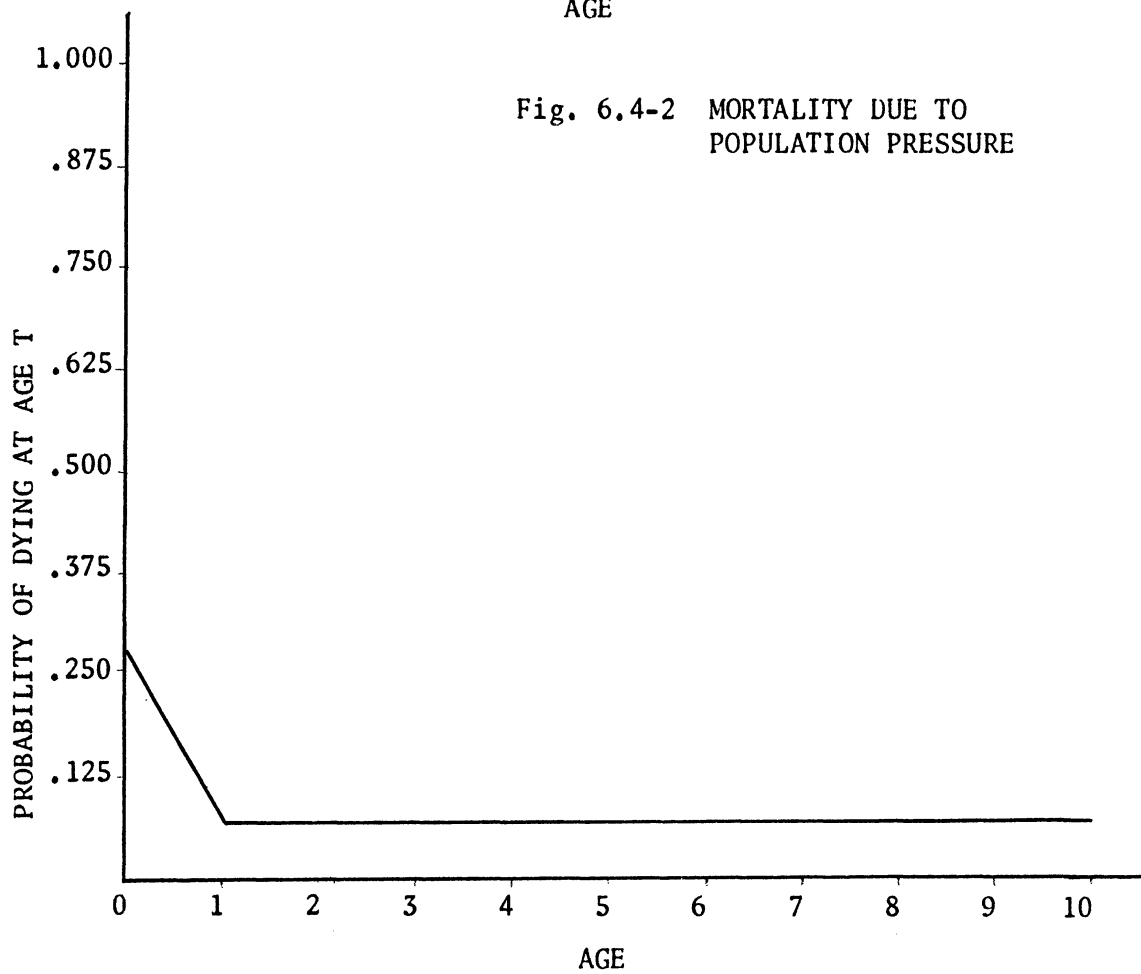
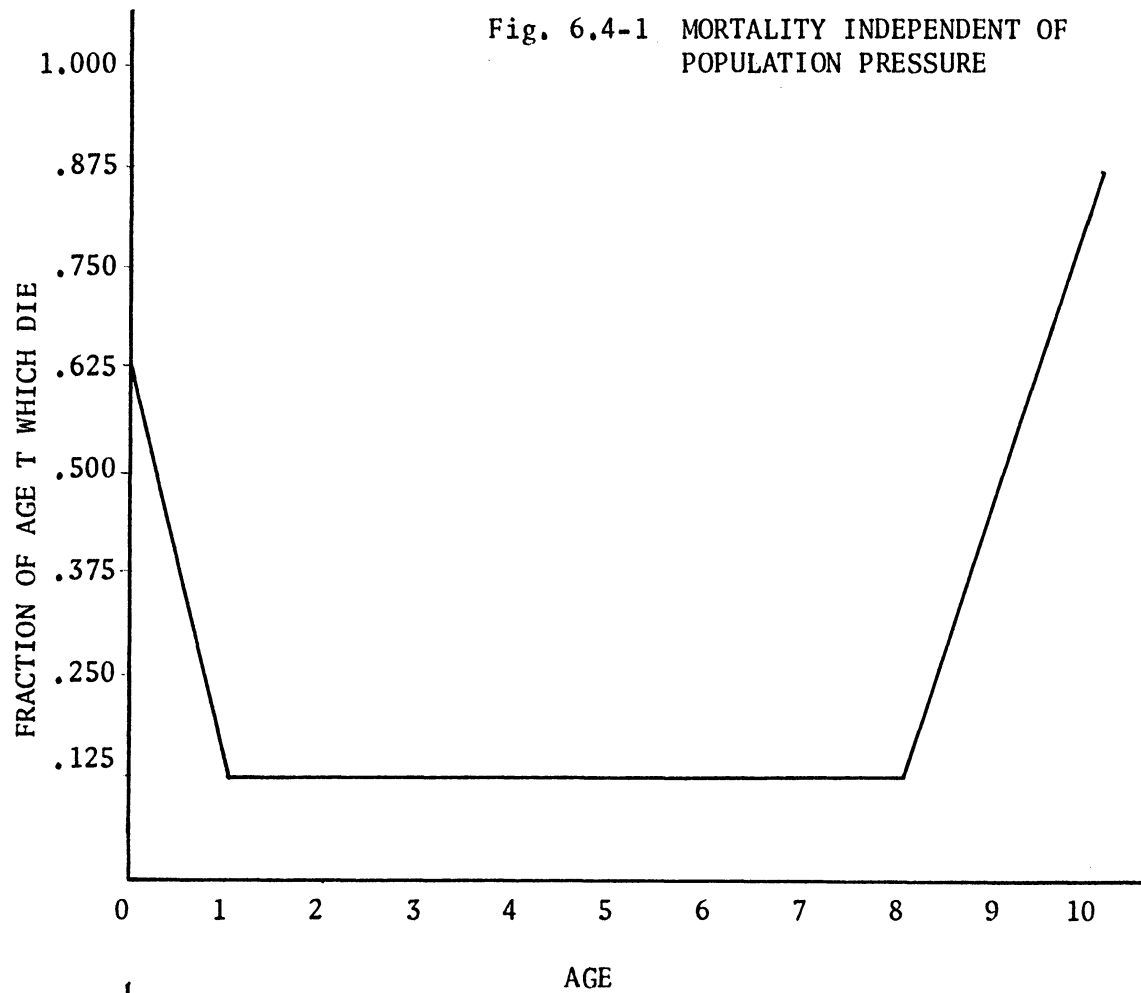
Just before mating occurs, the population is subjected to an elimination of cells for causes which can be grouped under the heading of losses

independent of population pressure. In nature, some of these might be due to lethal genotypes, predation and aging. Thus the form of this mortality is weighted towards the low and high age groups. A set fraction of cells of each age group is removed on the basis of fitness. Fig. 6.4-1 shows the actual distribution used for most of the experiments. This mortality is especially severe on the cells of age zero with five-eighths being eliminated. Thus for one hundred offspring only thirty-seven will have the opportunity to mate and most of these will be double heterozygotes. After this initial elimination of cells, the size of the mating population will be reduced to about 135 cells from a full population of 200.

It should be noted that the number of age groups has been increased from 6 to 10. The purpose is to allow cells to have a longer life span, thereby increasing the number of chances of mating. This is important in increasing the opportunity of double heterozygotes to influence the state of the population, because it is only this kind of genotype which is likely to survive many generations.

With this type of mortality it is necessary that the population not begin with all its members at age zero. In such a situation only three-eighths of the population would survive for the first period of mating. To avoid this, ages are assigned randomly to the cells on the basis of an arbitrary distribution. This distribution, given in the table below, is designed so that only 40 to 50 cells are eliminated initially.

The application of the first mortality distribution results in a number of places available for offspring. If the number of offspring produced exceeds this available space, additional places are provided on the basis of an age-fitness mortality. In a certain sense, this action is dependent on the pressures of population size and can be related to



AGE	PROPORTION OF CELLS ASSIGNED THIS AGE
0	.25
1	.125
2	.100
3	.100
4	.100
5	.075
6	.075
7	.075
8	.050
9	.035
10	.015

Table 6.4-1 INITIAL AGE DISTRIBUTION

such factors as migration and infant mortality due to the lack of nursing care for large numbers of offspring. The probability distribution used in the experiments to follow is given in Fig. 4.2-2. The combined action of both mortality distributions constitutes a significant selection pressure.

Two other changes have been made. The first is a slight change in the enzyme table B-3 shown in Table 6.3-2. Because chemical 1 is only used in one reaction, its increased quantity tends to cause the reaction 1+2→5 to proceed at a faster rate than the other reactions defined by genes 1 and 2. To bring the system of equations into better balance the rate constants for alleles 4 to 7 of gene 2 are reduced to .061, .091, .136 and .202, respectively. The second change is an increase in the range of offspring from six to eight. This increase should have the purpose of aiding in the production of double heterozygotes. Matings

between double heterozygotes produce the maximum number of offspring.

An increased number of offspring will provide a better opportunity for slight advantages in the production of C's or R's, due to decreased cross-over probabilities, to be translated into actual increases.

In addition to the changes outlined above, experiment 8 differs from experiment 7 in two points. The mating percentage is increased from .1 to .15 because of the reduced size of the mating population. A different offspring function labelled Q is given below. As before the offspring variable s is given as $s = f_2^i + f_2^j$.

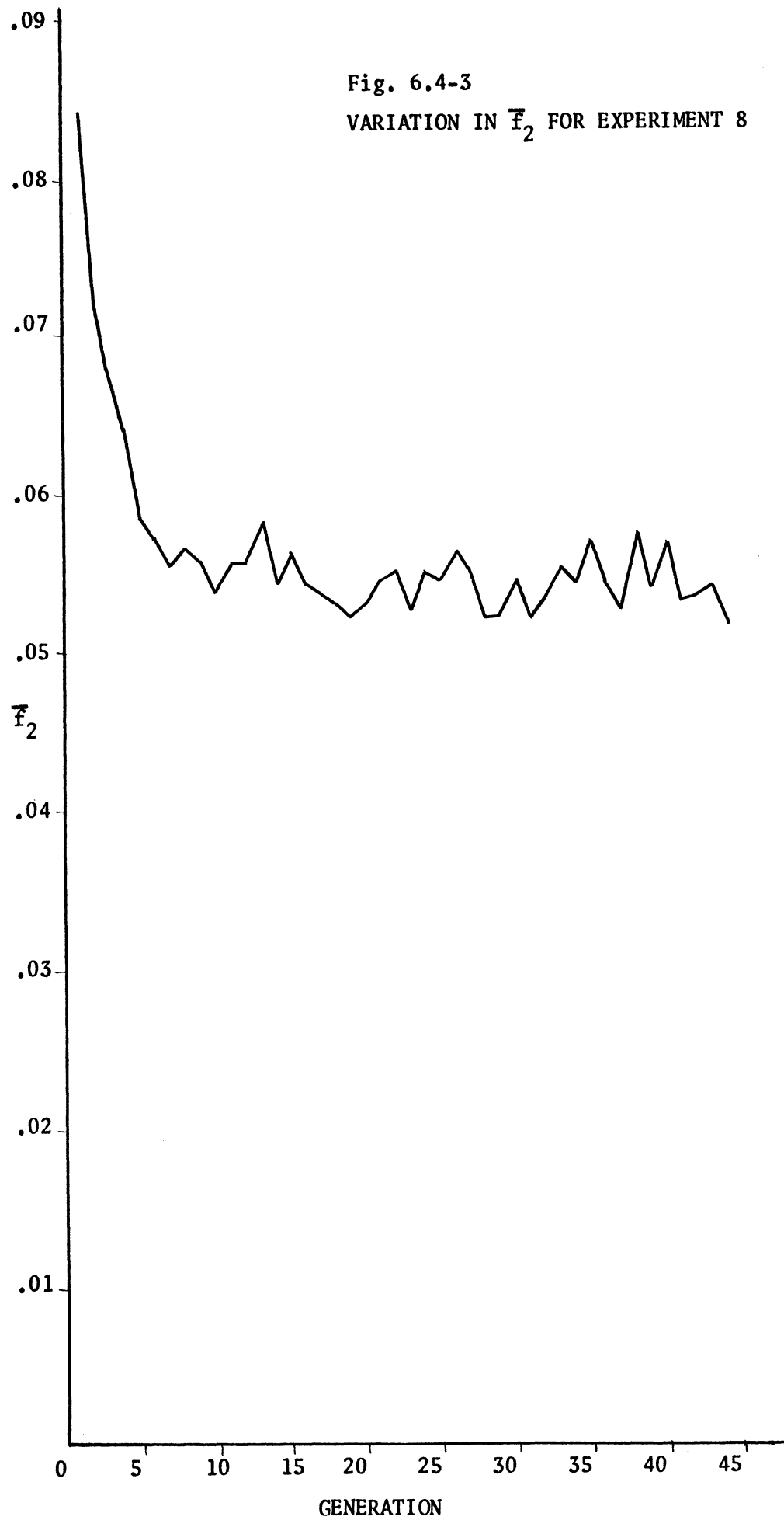
OFFSPRING FUNCTION Q	NUMBER OF OFFSPRING
$s \leq .05$	8
$s \leq .06$	7
$s \leq .10$	4
$s \leq .12$	3
$s \leq .15$	2
$s \leq .21$	1
otherwise	0

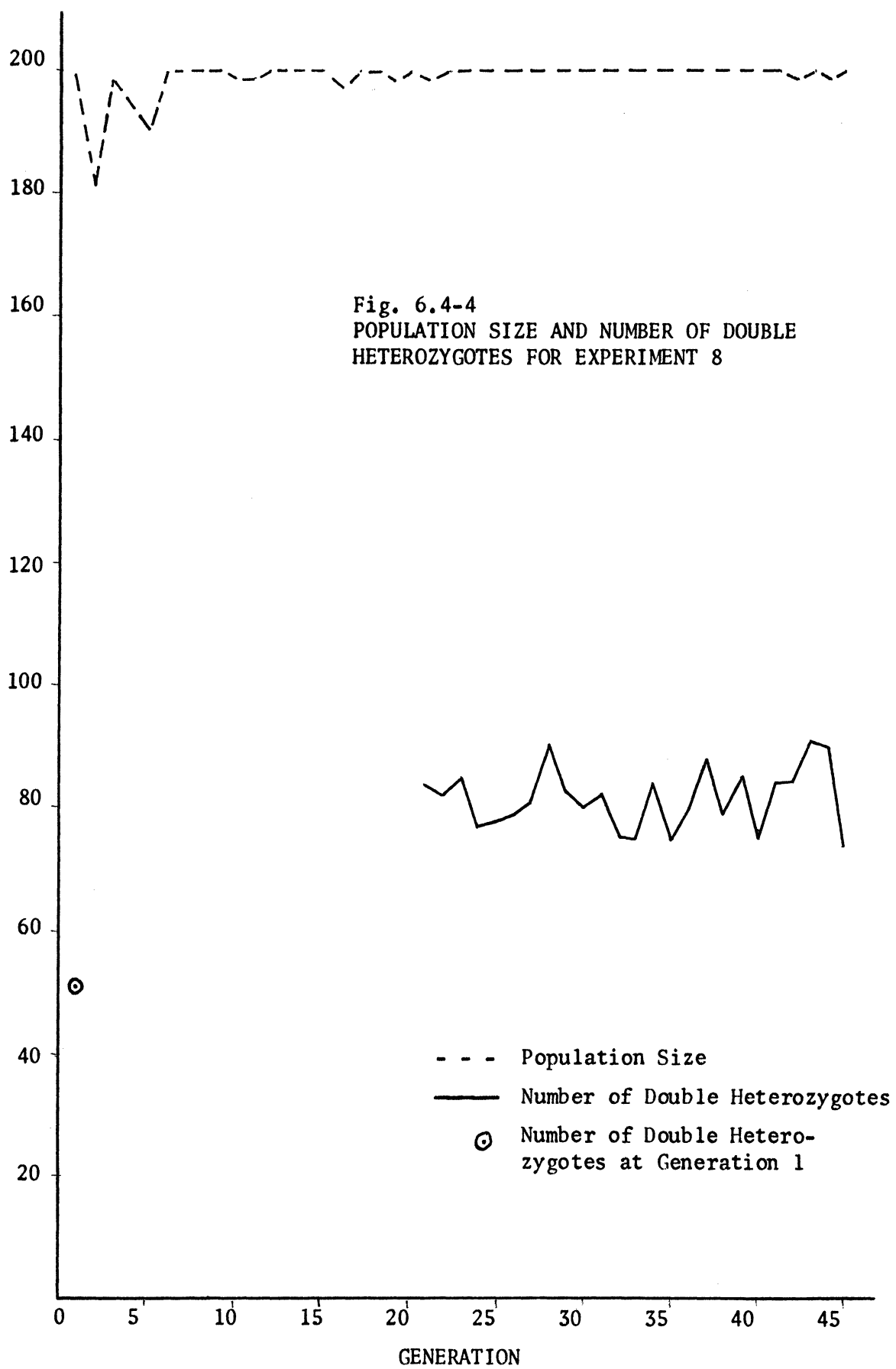
Table 6.4-2 OFFSPRING FUNCTION Q

No provision is made for the production of 5 or 6 offspring. It should be the case that matings of type 3 x 3 will produce 7 or 8 offspring and those of type 2 x 3 or even 2 x 2 will produce 4 or 3 offspring. The purpose is to have a large separation between the maximum number of offspring and the next highest number. Note that the level associated with 8 offspring is one-half the value of that associated with 4 offspring. There is a linear relationship between the number of offspring and the sum of the parental fitnesses.

The results of experiment 8 are shown in Fig. 6.4-3 and Fig. 6.4-4.

Fig. 6.4-3

VARIATION IN $\bar{F}_2$ FOR EXPERIMENT 8



It can be readily seen from the graph of $\bar{F}_2$ in Fig. 6.4-3 that from generation 7 on, the population shows very little improvement. The number of double heterozygotes never exceeds ninety and this limits the possibilities of improvement in $\bar{F}_2$. There is an increase in double heterozygotes from 51 to a maximum of 90 and this indicates that the mortality procedures are effective. The lack of a further increase suggests that the C's are not predominant enough over the R's or vice-versa. At generation 1 there were 19 C's and 32 R's. By generation 21, there were 52 C's and 32 R's — a complete reversal. This difference should have been enough to start a substantial increase in C's but this did not happen. In fact at generation 45 there were 36 C's and 38 R's.

The reason that the expected increase in C's did not occur is directly attributable to the lack of low crossover levels in the population. Because it required 20 generations to produce the first substantial spread between C's and R's, there was sufficient time for a loss of alleles with low crossover levels. At generation 25, the sum of the frequencies of crossover probabilities from .0 to .200 was only .105. This means that there is no mechanism available which will favor C x C matings and therefore the population drifts towards equal numbers of C's and R's and correspondingly higher crossover probabilities. Indeed, at generation 25, the sum of the frequencies of crossover probabilities from .375 to .5 is .455. We will return to this discussion in the next section.

The main difficulty seems to be that a difference in the number of C's and R's is not capitalized upon more quickly. An important parameter in this respect is the average number of like double heterozygotes that a given double heterozygote produces in its lifetime. This parameter, sometimes termed the reproductive factor, is indicative of the degree

to which a class of double heterozygotes, either C or R, is self-sustaining. Among matings, except for C x C or R x R, the frequency of double heterozygotes has an expected value of 1/4. There will be an increase in the number of C's say, only if a given C contributes on the average more than one C to the population in its lifetime. An estimate of the reproductive factor for C's can be determined in the following way.

At any generation the C's produced are listed and their life history is traced until they have all died. This study of their history includes all their matings and all the offspring they produce. For example, in generation 21 of experiment 8 there were 52 C's and 32 R's and among the offspring 16 C's were produced. Over a period of ten generations there were 27 matings involving this group of C's in which 8 other C's also took part. The total number of C offspring produced was 34. Thus the average number of C offspring produced per C of this original group is $34/(27+8) = .971$. The C's are therefore not reproducing in sufficient quantity to increase their number.

To correct this situation, the range of offspring which can be produced in any mating will be increased from 8 to 16. This should allow advantages in C x C or R x R matings to take effect more rapidly. It was for the same purpose that the previous increase to 8 was made but this increase was insufficient. This leads to experiment 9 which differs from experiment 8 only in the mating percentage, which is reduced to .1 because of the increased offspring range. A new offspring function R is given below. Because of the range of fitness values, the linear relationship between the number of offspring and the sum of the parental fitnesses does not hold in this case. However, changes will be made later to recover

OFFSPRING FUNCTION R	NUMBER OF OFFSPRING
$s \leq .04$	16
$s \leq .05$	14
$s \leq .10$	4
$s \leq .12$	3
$s \leq .15$	2
$s \leq .21$	1
otherwise	0

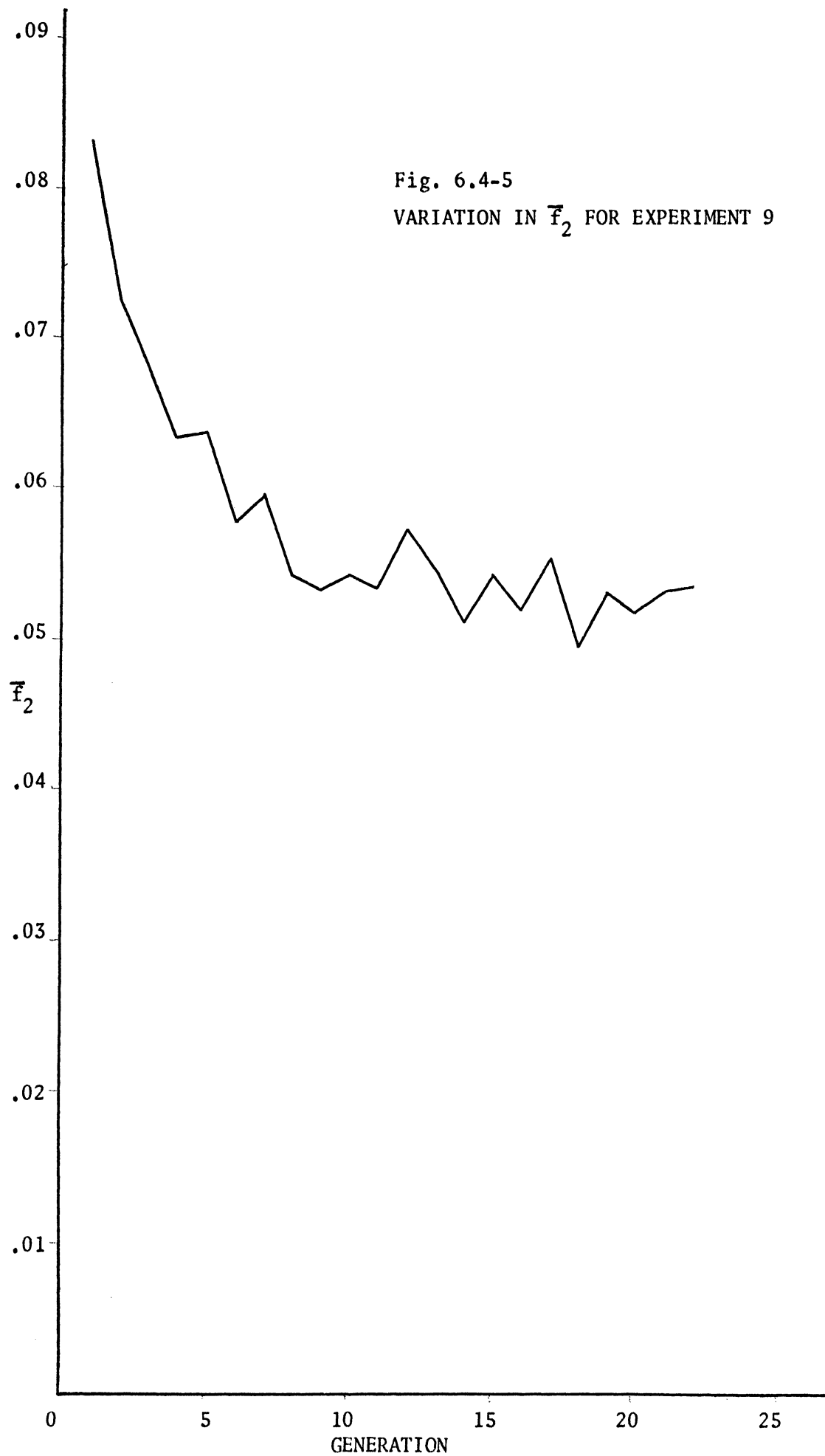
Table 6.4-3 OFFSPRING FUNCTION R

this condition.

Results of this experiment are shown in Figs. 6.4-5, 6.4-6, and 6.4-7. The graph of $\bar{f}_2$ for experiment 9 is not very much unlike that of experiment 8. It does appear to range slightly lower, however, and this is because of the fact that the number of double heterozygotes averages about 97 from generation 14 onwards whereas in experiment 8, the number of double heterozygotes averages about 82 from generation 20 onwards. The question of whether the $\bar{f}_2$ graphs can show a greater degree of separation will be discussed shortly, but first an examination of Fig. 6.4-7 is in order.

The distribution of crossover probabilities at position 1-2 shows a strong shift towards the lowest probabilities and compared with the distributions at the other possible crossover positions, indicates that this has been a successful experiment. In fact the sum of the frequencies of the crossover probabilities .0 to .2 is .715 as compared with .290 at generation 1. A more detailed examination of this experiment proves to be interesting.

At generation 23 there were 88 C's and 18 R's, a ratio of nearly



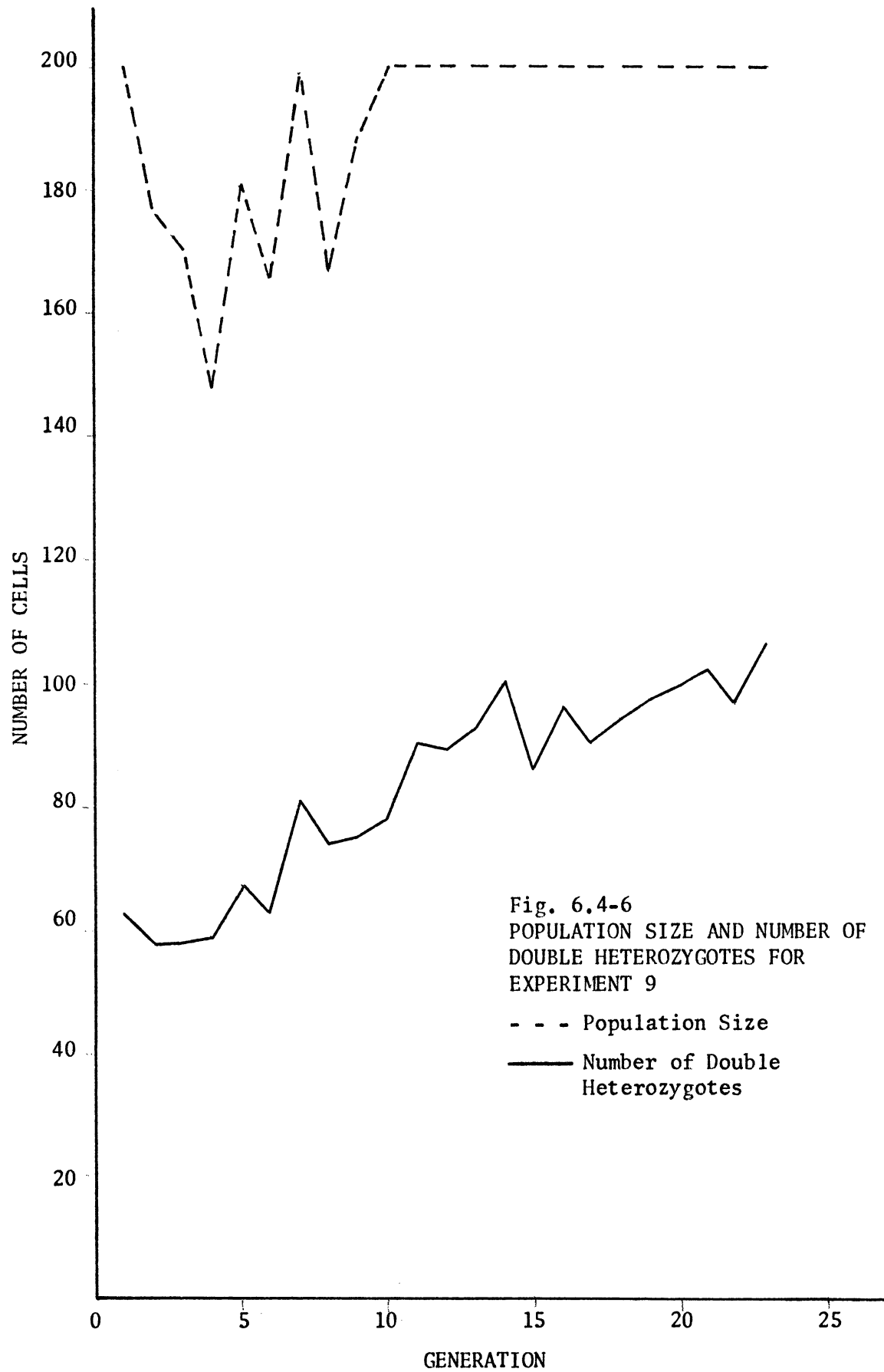
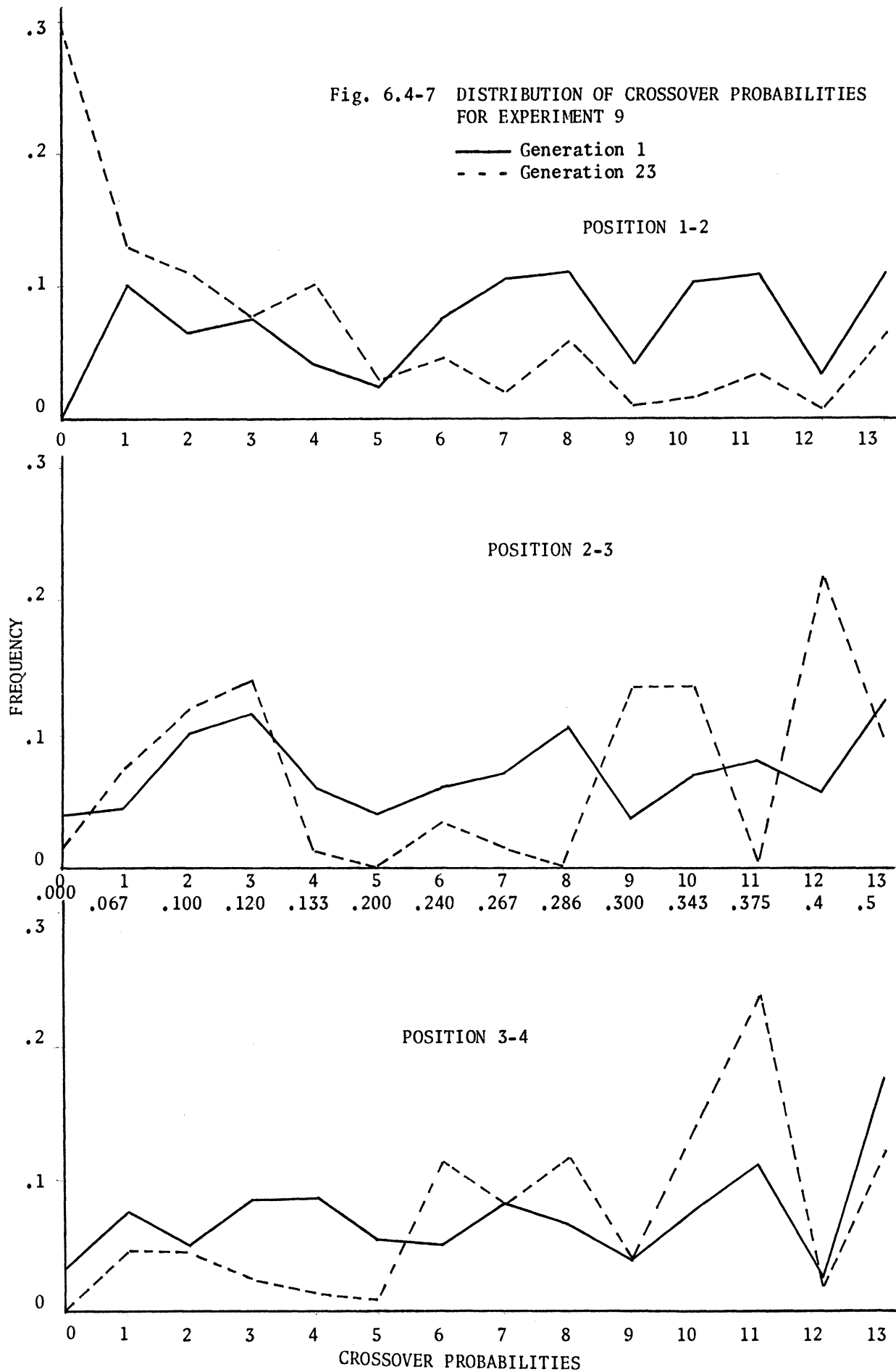


Fig. 6.4-7 DISTRIBUTION OF CROSSOVER PROBABILITIES
FOR EXPERIMENT 9



5 to 1 as compared with 38 C's and 25 R's at generation 1. The reduced crossover probability at position 1-2 and the high ratio of C's to R's are very much in agreement with the theory outlined earlier. An additional aid in analyzing the behavior of the population is the ratio of the actual number of crossovers to the potential number. In Fig. 6.4-8 this value is plotted for each crossover position. Up to generation 16, the three graphs are almost indistinguishable but after this point the graph for position 1-2 diverges sharply and is subsequently distinct. This parallels the shift towards low crossover probabilities. In generation 16, a few fortuitous matings between cells with relatively low crossover probabilities results in the production of 23 C's which have crossover probabilities of less than .2 at position 1-2.

By generation 22, 14 of these are dead and in the interval the 23 C's together with 9 other C's have produced 65 C offspring. With 9 C's still remaining, the average C of the group produced in generation 16 has produced $65/(23+9) = 2.03$ C's. This is a sufficiently high reproductive factor to produce the heavy weighting of C's over R's at generation 23.

The shift towards the low crossover probabilities first occurs at generation 16 as can be clearly seen in Fig. 6.4-9, which shows the time history of the frequencies of the first five crossover probabilities beginning at generation 14. These frequencies moved up sharply at generation 17. From generation 21 to 22 there is a very sudden increase in the frequency of the crossover probability .0 from .17 to .29. This is brought about by the production of a large number of offspring with zero crossover probabilities by some of the cells originally produced in generation 16.

This discussion suggests that some of the success of this experiment

Fig. 6.4-8 RATIO OF ACTUAL NUMBER OF CROSS-OVERS TO POTENTIAL NUMBER IN EXPERIMENT 9

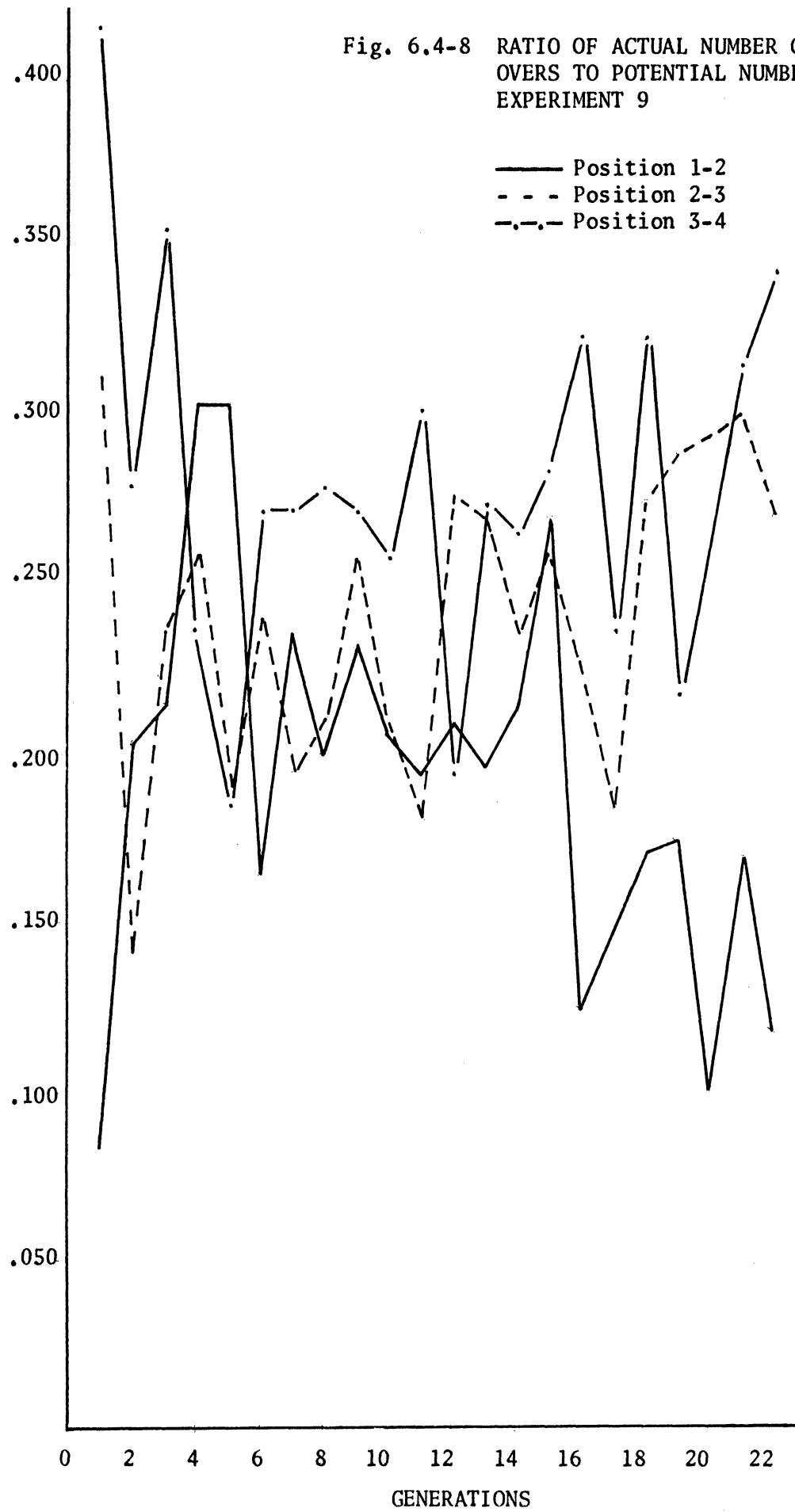
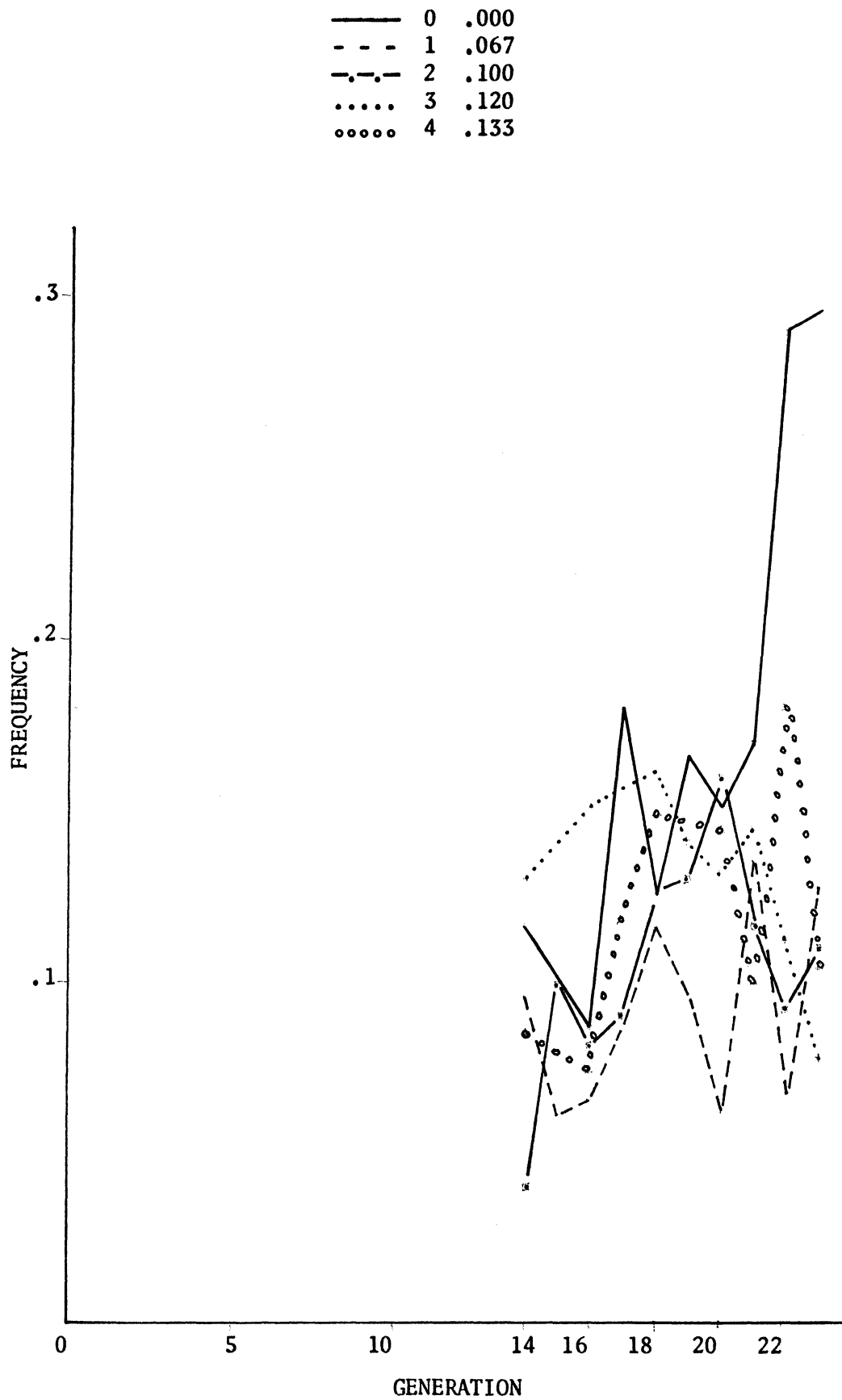


Fig. 6.4-9 DISTRIBUTION OF CROSSOVER PROBABILITIES
AT POSITION 1-2 IN EXPERIMENT 9



is due to good fortune. The necessary conditions for capitalizing on such an occurrence have been made available by the series of improvements incorporated into the system. The important event in this experiment was the occurrence of a few matings between cells with low crossover probabilities and the consequent production of a sizeable number of similar offspring. Within a very few generations these offspring were responsible for altering the state of the population. It is important to note that the success of the population is largely due to the fact that sufficient individuals were available with low crossover probabilities. In the next section we will discuss a method for increasing the likelihood of this condition.

With the present form of f_2 , it is very unlikely that there will be much separation between the graphs of $\bar{F}_2$ of two populations which differ in their number of double heterozygotes by less than 25. For example, in experiment 8 the value of $\bar{F}_2$ at generation 43 is .542 and this population has 92 double heterozygotes. The population of experiment 9 which has 102 double heterozygotes at generation 21 has a value of $\bar{F}_2$ of .533. This suggests that a different form of f is necessary to provide a sharper distinction between different populations. The form of this new f will be discussed in the next section.

6.5 FINAL EXPERIMENTS AND CONCLUSIONS

For the experiments of this section, a new f is defined which has the following form

$$f_3 = 10[(.25-x_9)^2 + (.25-x_{10})^2]^2$$

or in other words $f_3 = 10f_2^2$. The obvious effect of f_3 is to square differences which previously existed among the phenotypes. The table

below shows these differences,

CLASS	AVERAGE VALUE OF f_2	AVERAGE VALUE OF f_3
1	.11	.12
2	.07	.049
3	.02	.004

Table 6.5-1 COMPARISON OF f_2 AND f_3

The double heterozygotes (class 3) are now very much superior in fitness to the other genotypes and differences will be more apparent. Indeed, the mortality selective pressure will be significantly increased.

The offspring function given in the table below has the same property as offspring function Q. That is, 16 offspring are assigned for s less than .02 but only 4 for s less than .08.

OFFSPRING FUNCTION s	NUMBER OF OFFSPRING
$s \leq .02$	16
$s \leq .025$	14
$s \leq .08$	4
$s \leq .10$	3
$s \leq .15$	2
$s \leq .20$	1
otherwise	0

Table 6.5-2 OFFSPRING FUNCTION s

It is important that there be a sufficient number of alleles in the population with low crossover probabilities if the population is to have an opportunity of responding to selection pressures. If the number of

C's temporarily exceeds the number of R's but there are few C's with low crossover probabilities, this advantage will not be maintained and there will be a balance between C's and R's. To maintain a sufficient quantity of low crossover probabilities in the population, two alterations are made in Table 6.2-6. First, the maximum probability is reduced from .5 to .3 and second, a greater number of low probabilities are included. An examination of Table 6.2-6 reveals that there are only two probabilities less than .1 whereas there are nine greater than .2. This means that if a large number of alleles with crossover level 1 are lost, there will be very few crossover probabilities less than .2. The new crossover probabilities are listed in Table 6.5-3. In this format there are seven probabilities

LINEAR ORDER OF CROSSOVER LEVEL PAIRS	CROSSOVER PROBABILITY
0	.0
1	.015
2	.030
3	.045
4	.060
5	.075
6	.090
7	.11
8	.13
9	.15
10	.17
11	.20
12	.23
13	.26
14	.30

Table 6.5-3 CROSSOVER PROBABILITIES

below .1 and eight above .1.

Two experiments are carried on the basis of these changes. Both experiments have a reduced mating fraction of .07. This is necessary to control the number of offspring produced per generation. Some test experiments indicate that if the number of offspring produced greatly exceeds the space made available by the application of the first period of mortality, the additional cells to be eliminated are mainly double heterozygotes. This is so because the double heterozygotes constitute a large percentage of the population above the age of zero. It is this part of the population which is subjected more than 70% of the time to the second period of mortality.

Experiment 10 includes the crossover probabilities of Table 6.5-3. Experiment 11 serves as a control experiment, for all its crossover probabilities are .3. It is to be expected that the relative numbers of C and R will be approximately equal because there is no opportunity for selecting reduced crossover probabilities.

The results of these experiments are shown in Figs. 6.5-1, 6.5-2, 6.5-3, and 6.5-4. The graphs of $\bar{F}_3$ are very nearly the same for both experiments in the early stages, but there is improvement in the last three generations of experiment 10. The significant differences between the two populations are revealed by their respective numbers of double heterozygotes. In experiment 10, there are 114 R's and only 8 C's at generation 23 whereas the numbers in experiment 11 are 46 R and 59 C. This bears out the prediction that the numbers of C and R will be approximately equal when there is no mechanism available for reducing the crossover probabilities. On the other hand, where such a mechanism does exist, there will eventually be a large preponderance of one type over

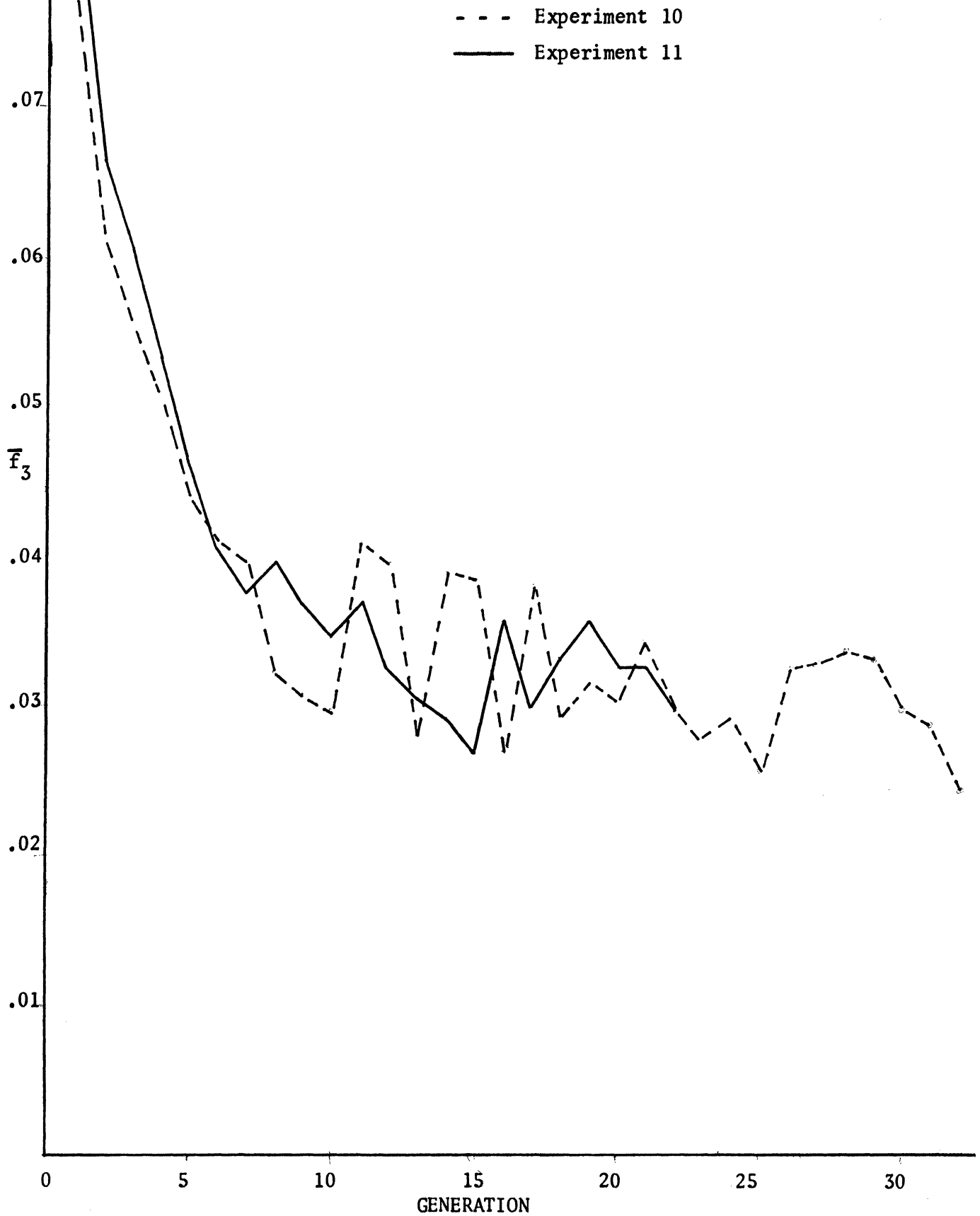
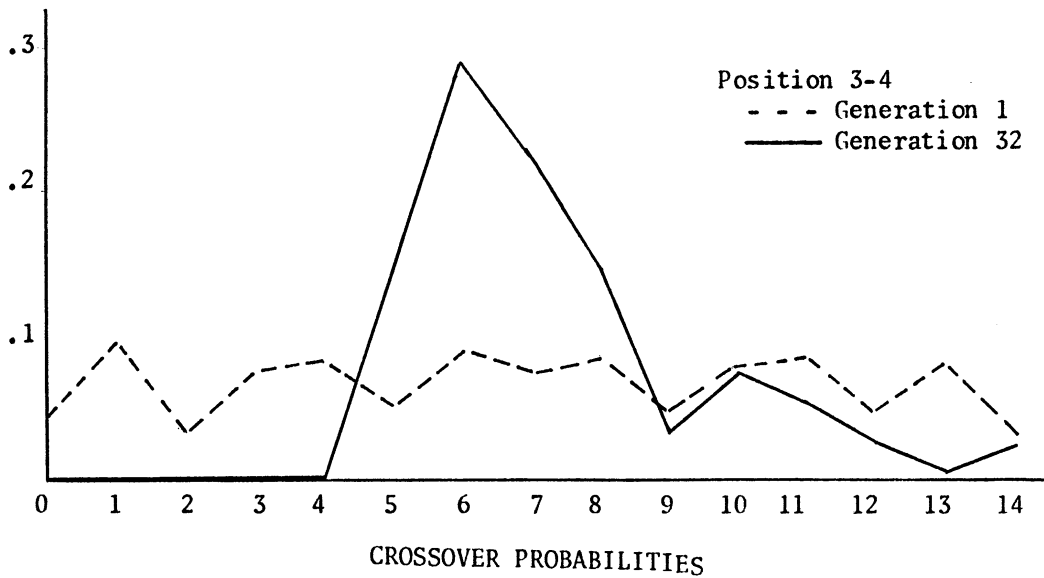
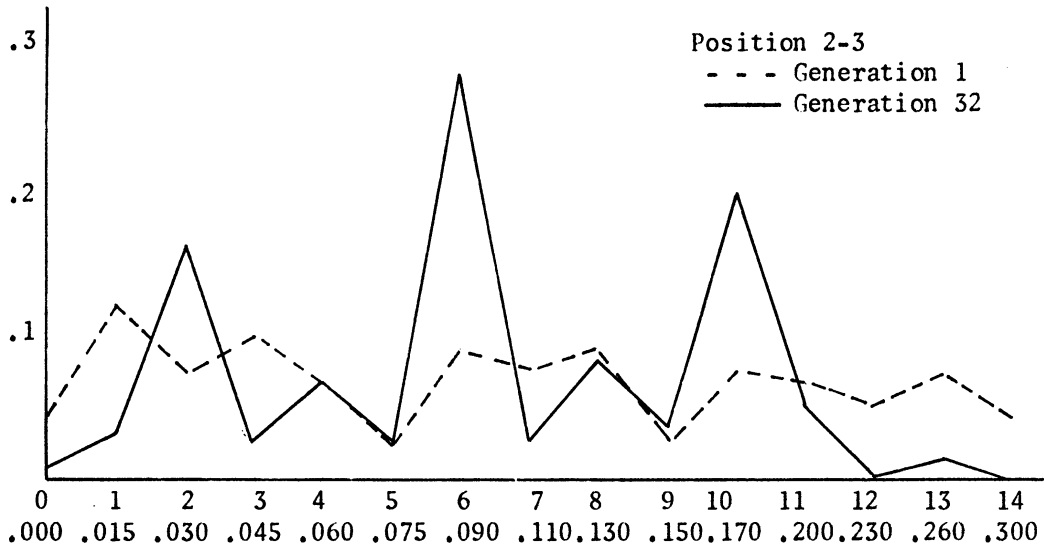
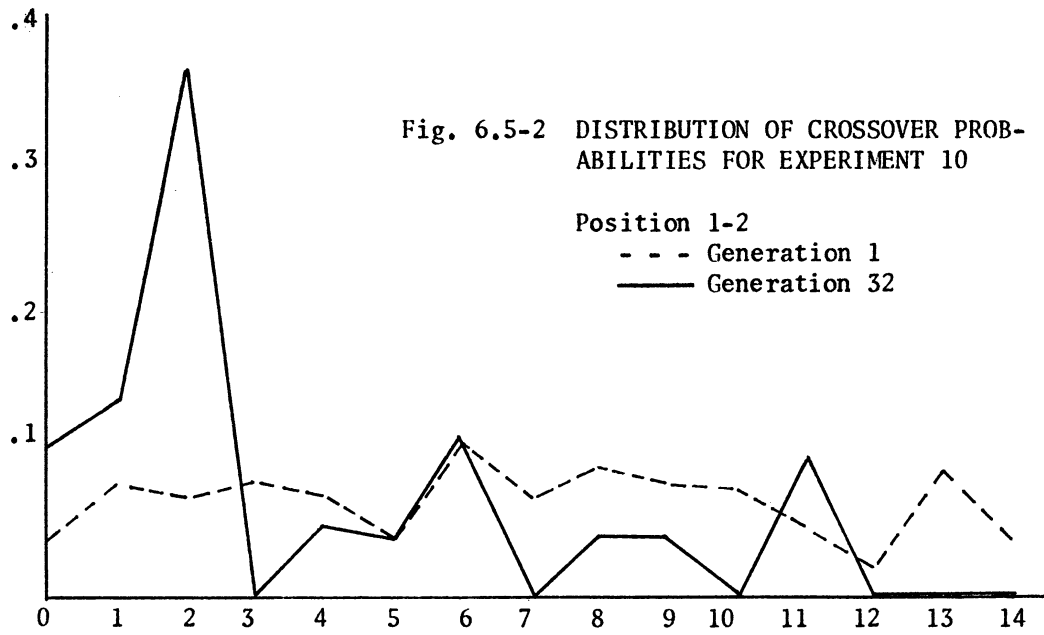
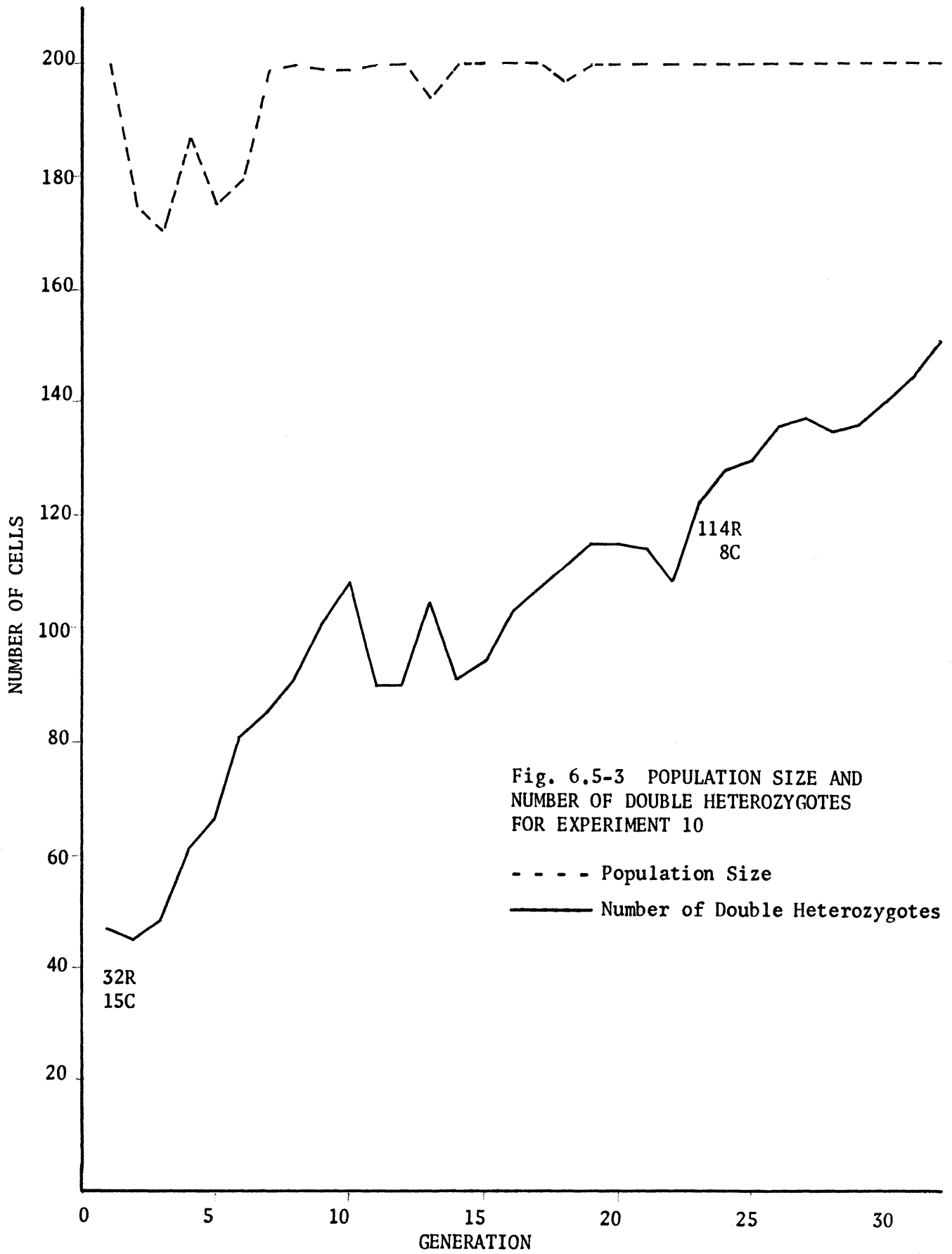
Fig. 6.5-1 VARIATION IN $\bar{F}_3$ FOR EXPERIMENTS 10 and 11

Fig. 6.5-2 DISTRIBUTION OF CROSSOVER PROBABILITIES FOR EXPERIMENT 10





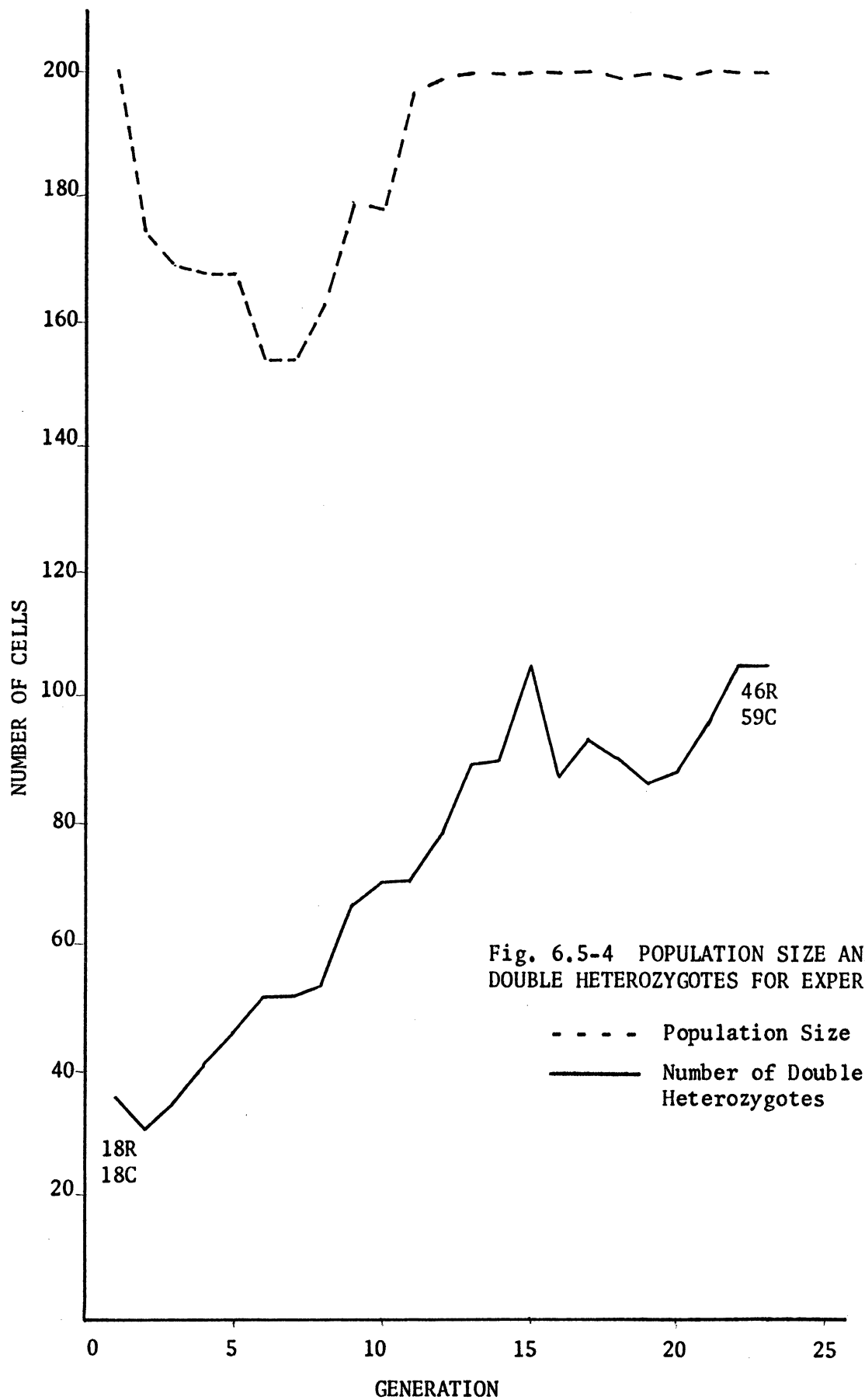


Fig. 6.5-4 POPULATION SIZE AND NUMBER OF DOUBLE HETEROZYGOTES FOR EXPERIMENT 11

- - - - Population Size
— Number of Double Heterozygotes

the other. This is indeed the case in experiment 10 where the ratio of R to C is greater than 14 to 1.

An examination of Fig. 6.5-2 reveals a strong selection of the lowest crossover probabilities with the three lowest having a frequency of .606; that is, 60% of the population have a probability of crossing over at position 1-2 which is less than .03. In the control experiment, of course, all the crossover probabilities are .3.

It should be noted that the reductions in the number of double heterozygotes at generations 11, 14, and 22 of experiment 10 are due to sudden increases in the number of offspring. This is because of the reason stated earlier for decreasing the mating fraction. The present occurrences suggest further reductions will be useful, and from generation 23 onwards the mating fraction is .05.

Another kind of control experiment would involve the random assignment of offspring. How would such a population compare with a population under selection pressure? In this context, random assignment will mean equal numbers of offspring, say 10, to all matings. With 10 matings per generation, the number of double heterozygotes will build up more rapidly than in a selected population. This is due to the fact that of 100 offspring, 25 can be expected to be double heterozygotes. Along with the differential mortality due to fitness, there will be a rapid increase in this number. In a selected population, the number of offspring is relatively low in the early generations because the proportion of double heterozygotes in the mating population is low. Thus there is a more gradual build-up in the number of double heterozygotes.

An expected distribution of genotypes in a population in which offspring are assigned uniformly to all matings is readily derived.

The distribution of offspring can be expected to be 25, 50, 25 for classes 1, 2 and 3 respectively. By means of both mortality processes, 100 cells are removed to make room for these offspring. In the most favorable case for double heterozygotes, the remaining distribution is 0, 0, 100. The addition of the offspring results in the distribution 25, 50, 125 among the three classes. Some experiments not presented here showed results along this line.

As an indication of how well a population might do under selection, consider a population of 200 C's with no possibility of crossover allowed. Suppose 100 offspring are produced each generation. Since there is no crossover, all matings will yield either C's or genotypes of class 1 in approximately equal numbers. Thus the first mating period will produce 50 C's and 50 members of class 1, resulting in an overall distribution of 50, 0, 150. Because the members of class 1 have values of f_3 thirty times larger than those of class 3, it will probably be the case that all 50 will be eliminated in order to provide room for new offspring. Thus it is likely that the expected distribution is 50, 0, 150.

It was found that in most cases the value of f_3 for a double heterozygote decreases by a factor of about 3 as the age increases from 0 to 6. This suggests the real reason for the lack of a greater separation among $\bar{F}_3$ curves is the fact that many of the newly produced double heterozygotes have not reached their lowest values. The remedy would be to increase the number of biochemical time steps, but this would result in substantial increases in computer time. In this case, the genetic state of the population is the best measure of its overall fitness.

The selection of crossover probabilities has been demonstrated for what is essentially a two-gene system. Extension to a multi-gene system

appears to be straightforward. The steps along the way have revealed much about the nature of the system. It was found necessary to carry out a limited theoretical analysis of a simple two-gene system in order to discover the source of experimental difficulties. This is a very necessary ingredient of any simulation study.

Perhaps the most important feature in this study is the necessity for strong selection forces to combat the disruptive stochastic effects in a small population. Only with the introduction of two separate and powerful mortality distributions together with a wider range in the number of offspring assigned to a mating could the population respond successfully to selection pressures. The mathematical analysis of small populations is generally very difficult and it is the purpose of simulations such as the present one to facilitate such studies. It is clear that strong pressures are necessary to drive a small population.

CHAPTER 7

CONCLUSIONS

A simulation system differing in some important respects from other simulations has been developed and applied to a few specific problems. In the description of the system itself many of its features were discussed in great detail. Particularly stressed were the many similarities to natural populations such as a time-varying phenotype, a dynamic interaction between the population and its environment, and mating rules, mortality rules, and selection rules of wide variety. The definition of the phenotype on the basis of chemical concentrations, whose levels are under genetic control via enzyme action, is the central feature of the model. With a phenotype determined in this way a wide range of useful and interesting research problems can be considered.

Preliminary investigations have been made into several such problems. The relation between the number of crossover positions in the chromosome pair to the rate of evolution of the population was explored. The question of an optimum number of such positions with respect to a particular form of response was raised. With more crossover positions available, a population will have greater scope in exploring its genetic space.

The importance of a dynamic relationship between the population and the environment was stressed. It is clear that a population under selection pressure must communicate with an environment which behaves in complementary fashion. During these studies certain features of the system assumed a growing importance. The form of the initial chemical concentrations for offspring and the method for eliminating cells to make room for offspring were the key features. In the whole series of experiments the form of the initial concentrations was found to have

a considerable effect on the behavior of the population. At first it was based on the characteristics of an 'average' cell but this constituted an instance of built-in selection. In the final set of experiments, it was found necessary to assign offspring equal quantities of only the primary chemicals. The fact that this had not been done earlier negated genetic differences among cells and masked the selection pressure.

An important result was achieved in the final set of experiments in which the selection of crossover probabilities was demonstrated. A model was constructed in which it was to the advantage of the population to reduce the probability of crossing-over between a particular pair of genes. Success in this study involved an appreciation of the magnitude of stochastic variation to which a small population is subjected. Attempts to direct the population require strong selection pressures to overcome this random effect. Of prime importance were a strengthening of the procedures for eliminating cells and an increase in the number of offspring which could be assigned to a mating.

On the latter point, slight advantages in the reproduction of certain genotypes will only become effective when the number of trials (the number of possible offspring for a single mating) is sufficiently large. It is also important to control the total number of offspring per generation, for too many offspring impede the progress by forcing the elimination of some 'good' cells. The range of crossover probabilities and the quantization of this range must be carefully studied. It was found that an increased number of values at the lower end was necessary to guarantee the desired result.

In the rest of this chapter, several areas for further research are presented. Also discussed, is the work of two biologists Mayr [9] and

Waddington [16] as it relates to the present research. Their work, particularly Waddington's, provides a methodological framework against which this system, its aims and potential, should be viewed. In addition, two areas are introduced for the first time, namely morphogenesis and adaptive systems. The present program can be readily modified to be used in the exploration of the growth and development of a single cell and groups of cells. As a final point, the contributions of a genetic simulation like the present one to the theory of adaptive systems are considered. All of these considerations are brief and will serve mainly to broaden the scope of inquiry of this simulation in its future application.

7.1 GENERAL CONSIDERATIONS

This section will include several quotations by Ernst Mayr to provide the supporting framework mentioned previously. The present system is concerned with the development and evolution of a small, closed population which numbers at most two hundred. In a key chapter of his seminal book Animal Species and Evolution, Mayr discusses the role of the local population in evolutionary theory as opposed to the individual. He writes on p. 588, [9],

In actuality it is not the individual but the local population in which selection leads to the increase or decrease of certain genes and gene combinations; it is the local population which becomes conditioned to new habits and which may acquire new ecological preferences; it is the local population in which all intraspecific competition takes place and which in turn is in competition with its sympatric competitors. In short, to repeat, it is the local population which is the key to the solution of all evolutionary problems.

It seems very worthwhile, therefore, to carry out investigations of the interrelation between a dynamic population and its environment. Up to this point, the environment has been, of necessity, structured simply.

The most direct extensions will be towards increasing the complexity of both the environment and the population, even to the inclusion of other competing populations, in order to carry out fruitful explorations of the evolution of a small population. The nature of the complexities to be introduced into the environment should involve an increase in the number of niches. At present the entire environment is really a single niche for a single population. Even if a single population is maintained it would be beneficial to make several niches available to it. This would allow a much more realistic interaction between the population and its multi-dimensional environment.

In a further comment on this inter-relationship Mayr writes on p. 239, [9],

Increased selection pressure and reduced population size unquestionably lead to an increased rate of selective elimination of genes from a population. Conversely, temporary relaxation of selection pressure will permit survival of otherwise inferior genotypes, facilitating rare genetic recombinations that would be impossible during periods of severely adverse selection. Certain of these "impossible" genotypes may be retained in the population when the period of higher selection intensity is reestablished.

And on page 251, he notes: " ... diversity of the environment favors the genetic diversity of populations and thus counteracts the depletion of genetic variation." [9] All of these comments reinforce the importance of a richer environment in which to study the development of a small population. They also suggest the usefulness of including a time-varying selection pressure as part of the overall study of this situation.

In Section 4.1, a preliminary discussion was presented with respect to the nature of the organization of the genotype. At that time, it was stated that initially the genotype was merely a set of genes with very little cohesiveness and that the goal of the evolutionary process was to develop such cohesiveness. The importance of this concept is stressed by

Mayr in a discussion of misconceptions held by some individuals not familiar with modern genetics. The third point in this discussion on page 613 follows.

The concept that genes can be classified into those that are superior and will automatically be incorporated into the gene pool and those that are inferior and will inexorably be eliminated is an oversimplification that is actually quite misleading. The selective value of a gene is not absolute, but in an individual case is determined to a large extent by the external environment and the epigenetic system in which it operates. Natural selection consequently is not a problem of simple arithmetic. It must, furthermore, be assumed that the genes of a gene pool, being members of a successful team, will make on the average a greater contribution to fitness than randomly added new genes.

... The genotype is not a beanbag of unconnected genes. Since most actions of genes are interactions during development, the genotype is an epigenetic system, or, as Waddington [16] called it, an epigenotype.

It is for such reasons that the present simulation was conceived in its present form. The necessity of viewing the genotype as a highly organized structure required its manifestation through a set of controlled biochemical reactions. This is of course a first step and suggestions for further study will be proposed subsequently. We will now consider a few comments of C. H. Waddington which have a bearing on this system.

The very title of a recent book by Waddington, The Strategy of the Genes, suggests the long range and complex nature of the gene action. Waddington is particularly concerned with the developmental approach to the theory of evolution. That is, an approach which directs itself to questions about the pathways between the genes and the phenotype, the relation of individual genes to the total genetic background as well as the complex concepts of genetic homeostasis and canalization viewed within a cybernetic framework. In discussing the many interactions among chemical reactions in terms of variations of rate constants, he notes on

p. 24 with respect to open systems, [16],

... Genes, in fact, can be introduced into our open system picture in two different ways. If, to take the first of these, one concentrates attention on the gene-activities which steer the developmental processes, then the genes can be considered as catalysts which determine the rates at which the various reactions will occur.

This, in fact, is almost precisely the fashion in which the genes act in our simulation. In another context, Waddington speaks of 'canalized' or 'buffered' paths (p. 19), [16],

... Secondly, the whole course of development from the initial stage on up to the final adult condition is a 'most-favored path'; that is to say if the mass of material is developing along one such path and is at some time during the course of development forced out of it by some experimental means, it will exhibit 'regulative behavior' and tend to return to the normal path. To express this character, I have spoken of such paths as being 'canalized' or 'buffered'.

And on p. 38, he writes, [16],

... We still have very little information about the frequency with which the canalization of a developmental system is specific to one particular stress as opposed to providing a general buffering against a wide variety of different stresses. It seems probable, however, that the canalization evolved in populations under natural conditions usually involves a rather general buffering, since populations must have to cope with variation in many environmental conditions rather than extreme stresses of any one kind, such as tend to be used in laboratory experiments.

It will be recalled that in Section 4.1 a description of some problems arising from a particular type of reaction graph was given. The difficulty was seen to be in a lack of differentiation among the phenotypes of individuals with widely varying genetics and was attributed to a factor called 'parallelism' in the reaction graph. The remedy was a reaction graph with few multiple paths. In light of the above quotations, much more detailed investigations are called for into the very complex interactions among the genes as manifested through the reaction graph. Indeed much further study is required into the relationship between the nature of the reaction graph

and the evolution of the population as a whole.

As a final quotation from Waddington, consider a comment on page 65 about selection [16],

... Selection usually does not operate directly on the genes. It depends on the ability of organisms to leave offspring. It is the phenotype throughout its development which is exposed to the rigours of the environment; which falls a prey to disease, the inclemency of natural conditions, or predation; which is successful or unsuccessful in mating; which is fertile or infertile.

This in fact sums up many of the features of and attitudes behind the present model. The primary selective force is the assignment of offspring based on the fitnesses of both parents. As has been mentioned several times these fitnesses are strictly phenotypically determined. There are additional sites of selection, especially in the mortality properties of the individual members of the population. Thus a study of this nature embedded in the present simulation is in the stream of current research.

7.2 SELF-REGULATORY MECHANISMS

The following quotation from T. O. Browning [18] occurs on p. 105.

When ecologists speak of self-regulatory mechanisms in animal populations they are making an analogy with the governor of an engine, which speeds the engine up when it is slowing down and slows it down when it is speeding up, so tending to keep the fluctuations in the speed of the engine within rather narrow limits. They are referring to any situation in nature in which the chances of survival and reproduction of an individual in the population depend primarily on the numbers of other animals of its own kind in the vicinity, in such a way that the chances become less as the numbers increase and vice versa.

He gives several examples in animal populations of this mechanism but states that it is not clearly understood how it operates. The maximum size of the population is fixed because of storage limitations in the computer. How such a control could evolve in a population may be a poss-

ible future research problem. It would be quite simple to tie the rate of reproduction to population size, thereby superimposing a control, but this is too simple a solution.

Mayr points out that in certain species, the frequency of certain genes is correlated with population size. That is, these genes have greater selective value for smaller than for larger population sizes or vice versa. But in these cases, the causal relationship seems to go from population size to the presence or absence of certain genes.

7.3 MORPHOGENESIS

Although the primary purpose of this system is to study the evolutionary development of a small population under various conditions, there is the possibility of using it with limited modifications to investigate some problems in morphogenesis. The term morphogenesis is being used in its broadest sense in this context as pertaining to the production or evolution of the form and structure of animals and plants. Perhaps that part of morphogenesis of most interest at this point is what Waddington calls 'regionalization' and defines in the following way on page 11, [16],

... [Regionalization] is the process by which the various parts of the germ which originally seem alike, come to exhibit clearly apparent differences, one part developing into say an eye, another into the brain, a third into the intestine, a fourth into notochord and so on.

Instead of treating the cells as individual members of a population, they could be viewed as being part of a tissue. Then the problems of differentiation in a relatively small tissue can be studied. Since each cell has the same genetics in this situation, differentiation will depend on the geometry and neighborhood configuration. The role of the semi-permeable membrane assumes prime importance in controlling the flow of chemicals among the individual cells.

7.4 RANDOM FIXATION AND THE FOUNDER PRINCIPLE

Random fixation is a name for the phenomenon whereby a particular allele achieves a very high percentage in a small, isolated population and the corresponding alleles are lost due to accidents in sampling. With respect to random fixation, Sewall Wright [17] has said, "Accidents of sampling ... may be responsible for nonadaptive differentiation of small island populations but are more likely to lead to ultimate extinction ... than to evolutionary advance."

Since the population in the present system is small and readily isolated, it will be possible to observe random fixation. The presence of a strong selection force is of course a major factor in the increased frequency of certain alleles. Others, not perhaps of prime importance, are subject to loss due to the aforementioned random sampling. If the selection pressure changes, the population would have to begin evolving in a new direction. Inability to do so would result inevitably in extinction, a common occurrence among natural populations. Thus a high degree of fixation for many alleles would be a strong indication of the dangers the population would face if the success criteria should suddenly change. This suggests a range of experiments to study the degree to which fixation can be allowed before the population is unable to respond to any change in the selection pressure.

The whole question of genetic load is raised by these considerations. It is obviously of immediate benefit to the population to respond most directly to a selection force by an increase in the frequency of certain alleles. However, it must at the same time retain enough genetic variation to be used in the future. Such alleles are of no immediate use to the population and hence constitute a genetic load. The obvious questions are how much of a load can be maintained and what is the mechanism or

mechanisms for maintaining it. These questions can certainly be considered within the scope of the present system.

The founder principle is a term chosen by Mayr to designate the establishment of a new population by a few original founders which carry only a small fraction of the total genetic variation of the parental population. In effect, beginning a simulation with a population of a few individuals is a situation falling well within the category describable by the term founder principle. In most simulations the initial population will have as much genetic variation as possible at the outset. However, an interesting set of experiments would involve determining the relationship between the initial state of the population and its state after many generations. This would require a number of different sets of the initial genetics.

7.5 ADAPTIVE SYSTEMS THEORY

The only reason for including this part is to summarize a few results with respect to their wider applications in a theory of adaptive systems. The study of adaptation in nature is really a study of evolution and it seems reasonable that the development of a theory of adaptive systems should include an evolutionary aspect. The following is a possible model, only briefly developed, based on certain biological analogues.

As a structural analogue to chromosomes and genes, the individual adaptive system could consist of a set of subroutines each performing a specific task. This individual system is part of a larger population embedded in a complex environment. As in nature, the population will be subjected to a selection pressure related to the goal or goals of the system. The set of subroutines can be re-organized by the usual genetic means

such as crossover and mutation.

If modelled on a computer, serial operation will be necessary, but in general it will be better to allow parts of the system to behave in a sequential fashion while others are behaving in a parallel one. Information should be permitted to enter the system from many points simultaneously as well as leaving the system by such means.

One point should be clear from this cursory proposal: a model for the study of adaptive systems is the technique of evolutionary change exemplified in nature.

7.6 FINAL REMARKS

These will be particular brief for many points have already been considered. What has been outlined is a broad, flexible system for the study of evolutionary problems related to a closed small population. The clear distinction between genotype and phenotype, with the dynamic determination of the phenotype, the complexity of the environment, and the inclusion of many genetic characteristics are some of the main features of the model. It is hoped that the discussion both in the previous chapters and in the present one has indicated the potential of this as a useful and important research tool.

APPENDIX A

PROGRAM DESCRIPTION

The purpose of this appendix is to discuss programming details of the present system. No listing of the program is included because many features are tightly linked to a particular computer and language, namely the IBM 7090 and the MAD language, respectively. For this reason, it was thought that a brief discussion of the structure of the program and the functioning of each of the parts would provide sufficient information. This discussion should be read in conjunction with the block diagrams of Fig. A-1, Fig. A-2, and Fig. A-3.

Before beginning, the list of input parameters to the program will be given and briefly described. The main reason is to indicate the flexibility of the program as well as the many points which must be specified. The list is organized into two sets of data, A and B, and into two classes of parameters, program control and structural. The sets A and B refer to two different times at which data is read. The program control parameters specify features such as length of experiment in terms of biochemical time steps, and printing and punching instructions. It is the structural parameters which define the model and many of the conditions under which it is being studied. Other conditions are specified through the use of interchangeable subroutines. The input data is given as follows:

Data Set A

Control Parameters

- TIMAX - the total number of biochemical steps allowed for the run
- GENS - generation number at the start of the experiment

Fig. A-1

MAIN

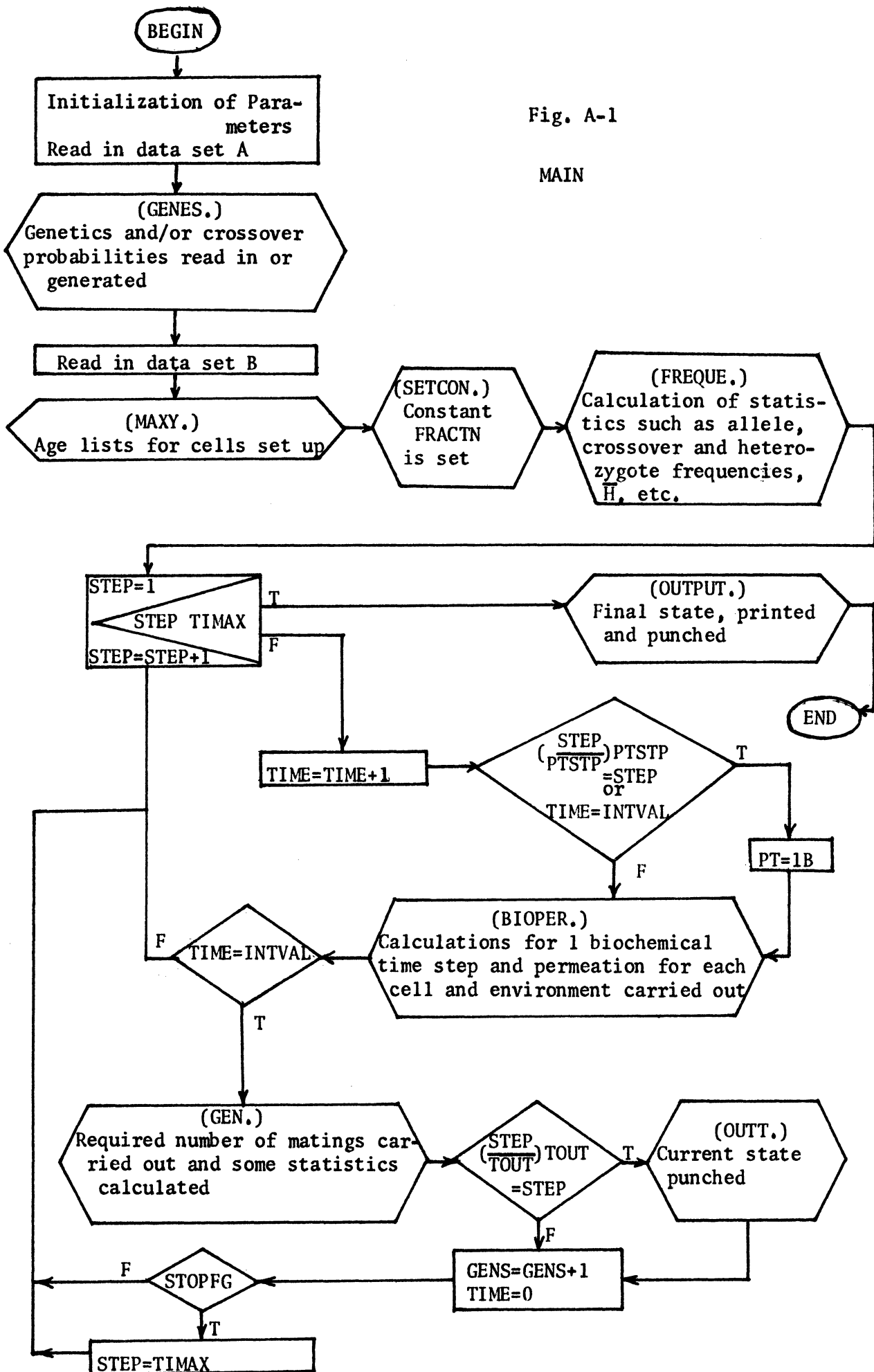


Fig. A-2

BIOPER

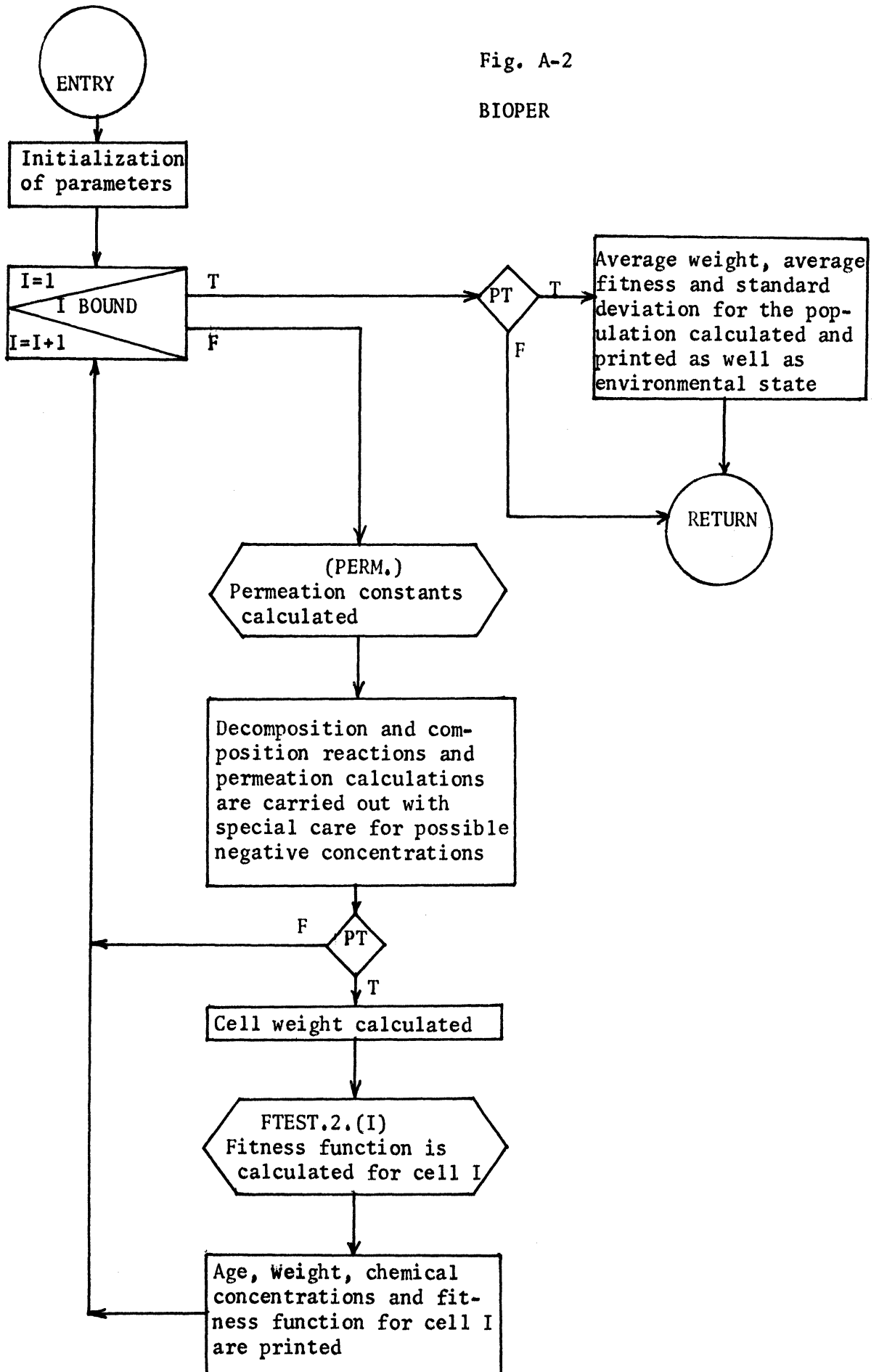
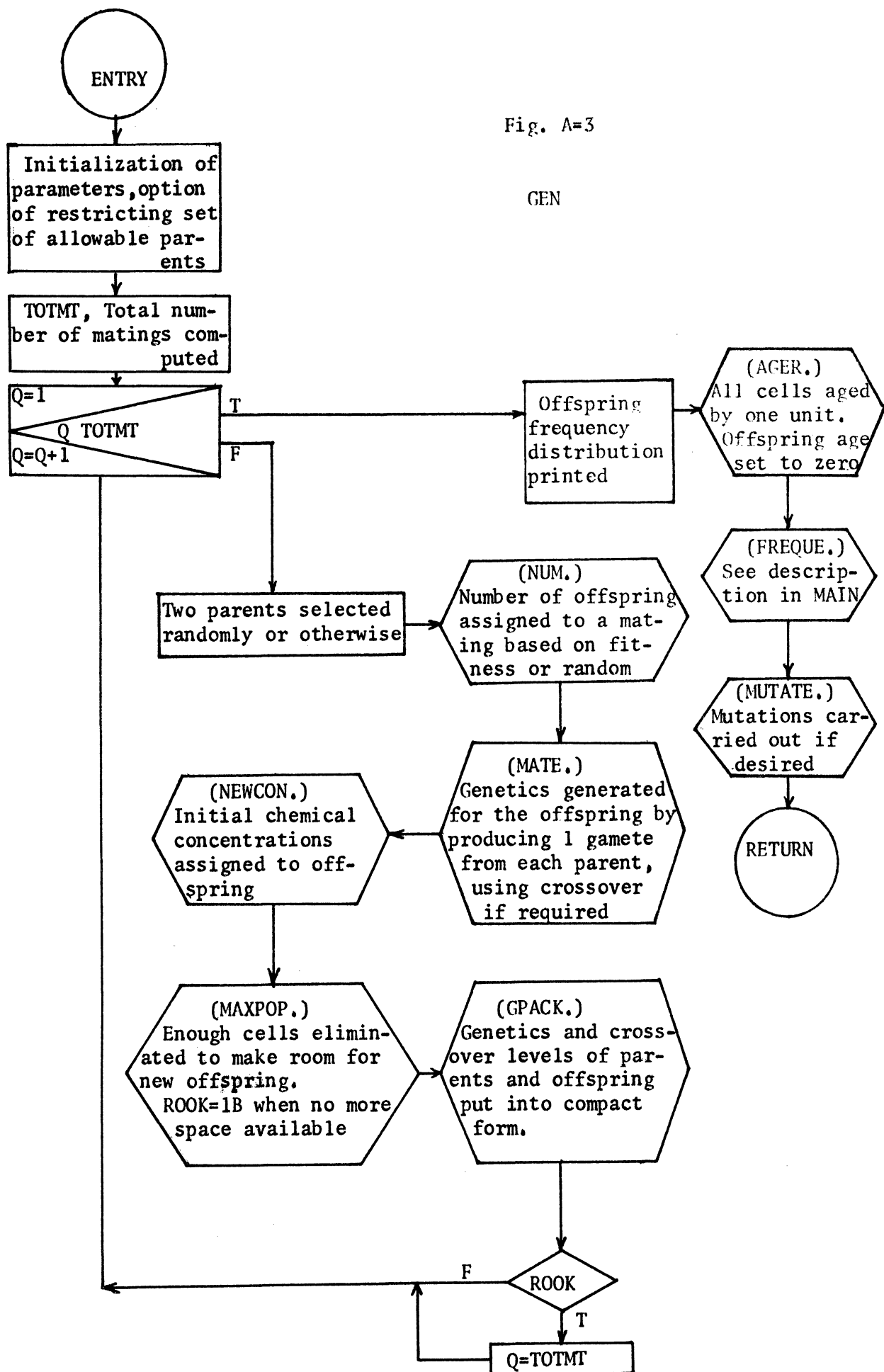


Fig. A=3



- TIME - counter for the biochemical time steps; usually started at zero
- PTSTP - printing will take place in BIOPER every PTSTPth time step
- IMAX - the highest numbered active cell at the outset
- PCHSTP - the complete state will be punched every PCHSTPth generation
- BIN - equals 1 if the concentrations for the cells 1 to IMAX are to be read in binary form; 0 otherwise
- ITERAT - If BIN = 1 this is ignored, otherwise it is 1 if the same concentrations are to be read from a single set of data for all the cells, and 0 if a different set is to be read for each cell from 1 to IMAX in decimal form
- AB - is 1 if the subroutine GENES is to generate the genetics for cells 1 to IMAX and 0 if the genetics are to be read
- BC - is 1 if there are any genetic overrides, 0 otherwise
- DE - is 1 if the subroutine GENES is to generate the crossover levels for cells 1 to IMAX and 0 if the genetics are to be read
- EF - is 1 if there are any crossover level overrides, 0 otherwise

Structural Parameters

- AGE(1)=x,y,... - the ages of the cells; needed if the genetics are not generated by GENES
- INTVAL - number of biochemical time steps per generation
- BOUND - maximum number of cells allowed
- FRACTN - corresponds to the integer r defined in section 4.2
- KLAM - corresponds to the number k of equation 2.2-3
- MTPCT - the mating fraction
- KILLAG - the maximum age any cell can achieve after which it is killed

CROP(1)=-,-,-,...	- crossover is allowed at positions corresponding to 1's in this list
CROSB=y,y,z,...	- crossover probabilities, e.g. Table 4.2-2
XLl=n,x,y,z,...	- n is the number of chemicals used in the fitness function; x,y,z,... are the chemicals
FNl(1)=x,y,z,...	- preset concentrations of chemicals x,y,z,... in the fitness function
CHEMAX	- the highest numbered chemical in the system
GENMAX	- number of genes
MAXAL	- alleles are numbered from 0 to MAXAL
MAXCRO	- crossover levels are numbered from 0 to MAXCRO
EMAX	- number of chemicals which permeate
KMAX	- highest numbered reaction-rate constant level
GRID	- number of squares in environment structure
MOLE(1)=x,y,z,...	- 'molecular' weights of chemicals x,y,z,...

Data Set B

Structural Parameters

Genetics and Crossover levels - present if AB = 0 and DE = 0

Environment genetics - to carry out decomposition reactions in the environment

Reaction matrix description in terms of vectors

List of chemicals which permeate

Chemical concentrations of the cells

List of permeation constants

Number of molecules of each type in the environment

Enzyme table

List of reaction rate constants

Total number of molecules in each cell

This completes the list of input parameters.

The program is organized as a small main part with a number of subroutines two of which carry out the major operations. The MAIN program is shown in Fig. A-1 and will be discussed first in terms of the subroutines it calls.

GENES

The control parameters AB and DE determine the function of the subroutine GENES. The former controls the generation of the genetics and the latter, the crossover levels. Both the genetics and the crossover levels are determined by the use of a uniform random distribution whose range is specified by the parameters MAXAL and MAXCRO. When the genetics are generated, ages are assigned to the cells, either all zero or according to some specified distribution. The genetics and crossover levels are punched in a very concise binary form.

MAXY

This is the name of one entry point to a subroutine whose other entry points are MAXPOP and AGER. At this call, the cells are arranged in a threaded list according to their ages. The other functions will be described as they arise.

FREQUE

This subroutine has the function of computing certain statistics and printing them every generation. Among these are the allele fre-

quencies at every gene, the entropy per gene, and the overall entropy, $\bar{H}$. The crossover probability distribution at each position is also presented. Other statistics are the percentage of heterozygotes at specified loci as well as the combined heterozygosity at several loci. The number of actual crossovers occurring as a percentage of the possible number is also computed. Additional functions can be included if desired.

SETCON

This is one entry point into the subroutine NEWCON. At this call, the parameter FRACTN is incorporated into the procedure for assigning the initial chemistry to offspring.

OUTPUT

The final state of the experiment is printed and punched. The state consists of the following items:

- (1) Genetics and crossover levels
- (2) Chemical concentrations of each of the cells
- (3) Total number of molecules in each cell
- (4) Age of each cell
- (5) Number of molecules of each chemical in the environment

Another entry to this subroutine is named OUTT, and it is called in order to punch the state but not to print this information. It would be called during a long experiment to produce a permanent and convenient record of the state of the experiment. An experiment can be resumed easily at any point for which the state has been preserved on punched cards.

The operations of the subroutines BIOPER and GEN are outlined in Fig. A-2 and Fig. A-3 respectively. As its name indicates it is in BIOPER that the biochemical reactions and the permeation exchanges are carried out for each cell.

PERM

This subroutine has limited use up to now but it can be used to calculate the permeation constants when they have a functional form as described in section 3.3.

FTEST2

The form of the fitness function is specified in this subroutine partly by the vectors XL1 and FN1. The sum of the fitness functions (for the population) as well as the sum of the squares are also computed.

Depending on certain control parameters the concentrations, ages, weights and numbers of molecules in the cells are printed out at regular intervals. The average fitness, standard deviation and environmental state are printed every generation prior to mating.

GEN

It is through this subroutine that the matings are carried out. The number of allowable matings is computed on the basis of the number of active cells and the mating fraction, MTPCT. Two parents are selected randomly or on the basis of some other scheme.

NUM

The number of offspring for these two parents is determined either

on the basis of their combined fitnesses or in some other fashion by this subroutine. In the course of the experiments there were more versions of this subroutine than any other.

MATE

The genetics for the offspring are determined in this subroutine. One gamete is obtained from each parent in the following manner. One of the chromosomes is chosen randomly and a random walk is begun with the possibility of crossing over at allowed points. The method for combining the crossover levels to determine the probability is specified in this subroutine.

NEWCON

The initial chemistry for the offspring is assigned in this subroutine. This can depend on the chemistry of the parents as well as preset parameters. Various procedures have been discussed in Chapters 3, 4, 5, and 6.

MAXPOP

This subroutine is called if the number of offspring produced in a given mating exceeds the number of available places. In this situation cells must be eliminated and the procedure for this is specified in MAXPOP. The basis for the elimination can be age, fitness, chance or some combination of these. It is possible for a condition to arise for which continued elimination of cells is not allowed. In this case, a control variable ROOK is set to stop the matings at this point.

GPACK

The genetics and crossover levels are usually stored in a compact form. After the offspring have been produced and places made available for them, both their genetics and those of the parents are put back into the compact form.

AGER

This subroutine is called after all the matings have been completed to set the age of the offspring to zero and to increase the age of all other cells by one. Those cells whose age now exceeds KILLAG are removed from the population.

MUTATE

If mutation is desired, it is carried out in this subroutine. None of the experiments discussed involved mutation and so this subroutine had no function. It is included in the program to provide generality.

As a final function, the frequency distribution of offspring is printed in GEN. Both the one for the current generation as well as the cumulative distribution over all the previous generations including the current one is printed. Also printed is the age distribution of the population in terms of the number of cells of each age as well as the mean age. For each mating such information as the genetics of the parents, offspring, and their concentrations is made available.

This completes the description of the program and a few additional remarks will conclude the appendix. For a population of two hundred

cells and with the model of section 4.2, which consists of 25 chemicals, 38 reactions, 9 genes, and 8 alleles, one biochemical time step takes 13 seconds. Twenty matings which produce about 50 offspring take 14 seconds. A typical experiment of 20 generations for two hundred cells will have the following timings for the parts:

Assume there are four biochemical time steps per generation and in the fourth, means are printed but not the concentrations.

	Time (sec)
Reading of 200 cells, input parameters and a call to FREQUE	12.2
Estimate for one generation:	
3 biochemical time steps (no printing)	36
1 biochemical time step (means printed)	12.9
20 matings	14
1 call to FREQUE	2.5
Total for one generation	65.4
Printing and punching state of 200 cells	55
Time for 20 generations = $12.2 + 20(65.4) + 55$ = 22.9 minutes	

This is a comparatively long period of time for the IBM 7090, but a great deal of computation is being carried out.

REFERENCES (BIBLIOGRAPHY)

1. Fraser, A.S., "Simulation of Genetic Systems by Automatic Digital Computers. 5-Linkage, Dominance and Epistasis," in Biometrical Genetics, edited by Oscar Kempthorne, Pergamon Press, New York, 1960.
2. Fraser, A.S., "Simulation of Genetic Systems by Automatic Digital Computers. II Effects of Linkage on Rates of Advance Under Selection," Australian Journal of Scientific Research, Series B: Biological Sciences 10, 1957, pp. 492-499.
3. Watson, James D., The Molecular Biology of the Gene, W.A. Benjamin, Inc., New York, 1965.
4. Garfinkel, David, "Simulation of Biochemical Systems," Chapter 6 in Computers in Biomedical Research, Vol. I, edited by Ralph W. Stacy and Bruce D. Wassman, Academic Press, New York, 1965.
5. Jacob, Francois and Jacques Monod, "Genetic Regulatory Mechanisms in the Synthesis of Proteins," J. Mol. Biol. (1961) 3, pp. 318-356.
6. Kempthorne, Oscar, Introduction to Genetic Statistics, John Wiley, New York, 1957.
7. Falconer, D.S., Introduction to Quantitative Genetics, The Ronald Press Company, New York, 1961.
8. Kojima, K., "The Evolutionary Dynamics of Two-Gene Systems," Chapter 9 in Computers in Biomedical Research, Vol. I, (see 4 above.)
9. Mayr, Ernst, Animal Species and Evolution, The Belknap Press of Harvard University Press, Cambridge, 1965.
10. Martin, Jr., Frank G. and C. Clark Cockerham, "High Speed Selection Studies," in Biometrical Genetics (see 1 above).
11. Gill, J.L., "Effects of Finite Size on Selection Advance in Simulated Genetic Populations," Aust. J. Biol. Sci., 1965, 18, 599-617.
12. Fisher, Ronald A., The Genetic Theory of Natural Selection, 2nd revised edition, Dover Publications, Inc., New York, 1958.
13. Fano, Robert M., Transmission of Information, M.I.T. Press and John Wiley & Sons, Inc., New York, 1961.
14. Kimura, Motoo, "A model of a Genetic System Which Leads to Closer Linkage by Natural Selection," Evolution 10:278-287, September, 1956.
15. Sonneborn, T.M., "Recent Advances in the Genetics of Paramecium and Euplotes," in Advances in Genetics, Vol. I, edited by M. Demerec, Academic Press, New York, 1947.

16. Waddington, C.H., The Strategy of the Genes, George Allen and Unwin Ltd., London, 1957.
17. Wright, Sewall, "Adaptation and Selection," in Genetics, Paleontology, and Evolution, G.L. Jepsen, E. Mayr, and G.G. Simpson, editors, Princeton University Press, Princeton, 1949.
18. Browning, T.O., Animal Populations, Science Today Series, Harper and Row, New York, 1963.

Errata and Changes

<u>Page</u>	<u>Line</u>	
29	3	"dnoted" should read "denoted"
31	17	should read "yield 4 cells each with n chromosomes."
35	first equation	The first element of the summation in the numerator should be " $i = 1$ "
47	16, 17, 18	This sentence should be removed. The actual probability is not derived this simply.
62	17	"1-2" should read "3-4"
79	19	"offsprng" should read "offspring"
103	4	"contrasts" should read "contrast"
104	14	"seem" should read "seen"
146	15	" $\bar{f}_1$ " should read " $\bar{f}_2$ "
151	4 to 15	This argument should be changed. The main reason poor cells can do well is the existence of precursors in sufficiently high concentrations as well as the high concentrations of chemicals 9 and 10 themselves. " f_1 " should read " f_2 "
152	15	"Dauermodifationen" should read "Dauermodifikationen"
153	Table 6.3-1	All the "9's" should read "q's"
154	Table 6.3-2	In gene 3, alleles 4-7, the reaction should read " $10 \cdot 6 + 8$ "
200	2	"Waddinton" should read "Waddington"
	19	"Browing" should read "Browning"

UNIVERSITY OF MICHIGAN



3 9015 03695 4587