A MODEL OF NATURAL SELECTION BASED ON A MATHEMATICAL THEORY OF GUESSING

by

Frederick E. Warburton

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Department of Genetics
McGill University,
Montreal, Canada

June 1963

ACKNOWLEDGMENTS

I wish to thank Miss Christine Scott, who prepared the manuscript of this thesis. My wife, Dorothy, helped to prepare the tables and graphs, suggested ways of clarifying the exposition, and was indispensable in many other ways.

About two years ago I reported on a preliminary study foreshadowing this thesis, at an informal seminar in the Department of Zoology at Yale University, and wish to thank Professor G.E. Hutchinson for encouraging me to continue the investigation.

Originally I began my doctoral research as a student of Professor F.C. Fraser, on a subject unrelated to the present one. I am grateful to Professor Fraser for graciously releasing me from my obligations to him when the opportunity to work on this topic arose, and for help and encouragement in many other ways.

Theses in the Department of Genetics have traditionally been reports of experimental investigations, and I owe special thanks to Professor J.W. Boyes, Chairman of the Department,for granting me permission to undertake a purely theoretical "armchair" project.

I am particularly grateful to Professor A.F. Naylor, without whom this thesis would never have been written. When Professor Naylor first became acquainted with my ideas on evolution, I had reached the early part of what I describe as

"phase 2" in the introduction, and had no hope that the work would ever be more than a private, spare time project. Professor Naylor brought my ideas to the attention of the Department, and advocated their acceptance as a thesis topic. Since then his advice has largely molded the thesis into its present form.

Contents

I Introduction

The purpose of this thesis is to introduce a new mathematical approach to the study of evolution by natural selection. Practically all of contemporary mathematical evolution theory is encompassed by the field of population genetics, which takes the empirical facts of Mendelian genetics as given and infers their evolutionary consequences. The writer has approached the subject from the opposite direction. This thesis takes the occurrence of evolution by natural selection as a given fact, and emphasizes that mere chance could not have produced the highly adapted organisms we see around us within the limitations of time and space to which natural selection has been restricted. It then attempts to infer the properties of a genetic system, and other conditions, which are necessary and sufficient to allow natural selection to have yielded products so different from those which would have been yielded by chance alone.

At the level of broad analogy, evolution by natural selection can be compared to a guessing game. Each new mutant or recombinant is a guess at a better adapted genotype, and the environment acts as a referee to classify each guess as better or worse than the one before. The analogy can be made more precise and illuminating by a mathematical analysis of the transfer of information from referee to player in an abstract guessing game. A mathematical model of Darwinian evolution as a kind of guessing game can then be developed.

The validity of the model is confirmed by the fact that it allows many of the important empirical properties of genetic systems to be deduced as necessary conditions of evolution by natural selection. The model also offers a way of attacking various problems which cannot be analyzed in terms of population genetics.

Mathematical Approaches to Evolution Theory

Both Charles Darwin and A.R. Wallace credited Malthus' semi-mathematical "Essay on the Principle of Population" as their inspiration for the concept of natural selection (Francis Darwin 1887; Marchant 1916). Although The Origin of Species (Darwin 1859) contains nothing more mathematical than a few arithmetical examples of potential population growth, the theory of natural selection is basically quantitative, and has been the object of mathematical attack and support since its formulation.

In 1870, for example, Bennett published some calculations intended to show that many consecutive variations in the same direction would have to occur before improvements large enough for selection to act would appear. Wallace (1870) rebutted the argument simply by making different assumptions about the sizes of variations, their mode of inheritance, and the sensitivity of selection. As long as little was known about the mechanism of heredity, nothing conclusive could come from such debate. A more pertinent argument was presented by Fleeming Jenkin (1867) who showed quite clearly that selection could not operate on occasional large variations transmitted by blending inheritance. This argument led Darwin to the speculations on changed conditions of life as a cause of widespread variability which are so prominent and disturbing in his later works. Although Jenkin intended to prove the impossibility of effective natural selection, Darwin used Jenkin's argument and the assumption that natural selection was in fact operative to make the first mathematically founded inferences about the mechanism of heredity.

Although Darwin's work solidly established the fact that evolution had occurred, argument about the mechanism continued for decades. None of the hypotheses brought forward to challenge the hypothesis of natural selection appear to have been supported by serious mathematical analysis, however, if one ignores repetitious and irrelevant calculations proving that "chance alone" cannot account for the results of evolution. Neo-Lamarckianism, the Mutation Theory ("which postulated that large mutations, and not small 'continuous variations', were the raw material of evolution, and actually determined most of its course, selection being relegated to a wholly subordinate position" (Huxley 1942)), and the numerous metaphysical theories of directed evolution, rested on qualitative arguments. Mathematical studies of evolution were devoted almost exclusively to working out the consequences of selection.

The validity of such studies of course rested on the correctness of their assumptions concerning the mechanism of heredity. Unfortunately, the rediscovery of Mendel's laws did not immediately reform mathematical evolution theory. In the early decades of the 20th Century, in the words of Julian Huxley (1942), evolution theory "...had reached its biometric phase. Tracing its origin to Galton, biometry blossomed under the guidance of Karl Pearson and Weldon. Unfortunately this, the first thorough application of mathematics to evolution, ... was for a considerable time rendered sterile by its refusal to acknowledge the genetic facts discovered by the Mendelians. Both sides, indeed, were to blame. The biometricians stuck to hypothetical modes of inheritance and genetic variation on which to exercise their mathematical skill; the Mendelians refused to acknowledge that continuous variation could be genetic, ... or that a mathematical theory of selection could be of any real service to the evolutionary biologists."

An essential first step in developing the modern mathematical theory of evolution, which is a synthesis of Mendelian genetics and the theory of natural selection, was taken by Hardy (1908) and Weinberg (1908; cited by Stern 1943). Their publications established the concept of a <u>gene frequency</u> in a population, and demonstrated that the frequencies of genes and the characters they influence will not spontaneously change in the absence of mutation, selection, or migration; in particular, they corrected the prevalent misconception that a recessive character tends to decrease in frequency merely because it is recessive.

The mathematical theory of evolution in almost its modern form then appeared with the publications of Fisher (1930), Sewall Wright (1931), and J.B.S. Haldane (1932). (These are chosen as "classical" references; the work of Wright and of Haldane, in particular, is contained in numerous papers). These publications laid an astonishingly extensive foundation, although of course much has been added since. The subject is adequately summarized by Li (1955) and by Falconer (1960).

Non-genetic approaches to the mathematical study of natural selection have of course been tried. Those which deal only with attempts to establish a quantitative way of measuring evolutionary phenomena, such as rates of morphological change in paleontological series, will be omitted from discussion.

Thermodynamics, which affords a basis for deducing the direction which physical and chemical processes will take and the terminal states they will achieve, has attracted attention from evolutionists. In the form of chemical kinetics, it has been useful in permitting inferences about the conditions required on the surface of the earth for the formation of those chemical substances and complexes we associate with life (Oparin 1938; Blum 1955; Florkin 1960), but except for Blum's claim to have shown that organic evolution and the laws of thermodynamics are not incompatible, its usefulness in discussing the evolutionary process as a whole has not been demonstrated.

Lotka wrote a number of papers trying to relate natural selection to the overall energetics of the earth, his basic principle (Lotka 1922) being that since free energy is the fundamental object of competition in ecological systems, "... natural selection tends to make the energy flux through the system a maximum". This principle, interpreted to mean that an unused or inefficiently used source of energy constitutes an unoccupied ecological niche, has become important in attempts to explain the trophic structure of communities, but it has never become incorporated in evolutionary theory in a form amenable to mathematical analysis.

Population dynamics will perhaps contribute to population genetics by clarifying the genetic changes to be expected in expanding, shrinking, or fluctuating populations. Gause (1934) however intended his experimental and mathematical studies of interspecific competition and predator-prey relations to contribute to the theory of natural selection in a more fundamental way. The principle which bears his name (two or more forms with identical ecological requirements cannot coexist indefinitely in the same environment) laid a foundation for rigorous discussion of the notion of an ecological niche. The subject has been discussed in a mathematical way by Hutchinson (1957), MacArthur (1957) and Slobodkin (1961), in an attempt to answer the question "Why are there so many kinds of animals?" (Hutchinson

1959), in other words to give a precise but general statement of the conditions required to permit an additional species to evolve.

Lewontin (1961) pointed out that "population genetics is not genetics of populations but genetics in populations", and is not suitable to explain how populations achieve an adequate degree of outbreeding, rate of recombination, and dispersal to allow them to compete successfully with other evolving populations. He suggested that analysis by Game Theory might assist in defining the best "strategy" for such situations. His publication was offered chiefly as a suggestion, not worked out in enough detail to allow a firm judgement of its value.

Meanwhile a new body of applied mathematics, known as Information Theory, was growing rapidly. Brillouin (1956) attributes the concept on which it is based to a 1929 paper by Leo Szilard pointing out that the thermodynamic entropy of a system corresponds to the quantity of information about the molecular configuration of the system which is not accessible to instrumental determination. The early history of the subject is also reviewed by Quastler (1958). The mathematical basis of information theory was largely developed by Wiener (1948) and Shannon (Shannon and Weaver, 1949), chiefly in its application to automatic machinery and to communication systems, although Wiener and Weaver also discussed its applicability to biological problems. In a qualitative way, its potential value in genetics was pointed out by Kalmus (1950).

Attempts to discuss evolution in terms of information theory have been few in number, if one excludes those whose authors have merely adopted the terminology of information theory without any attempt at mathematical analysis. It is noteworthy, for example, that neither of two major symposia on Information Theory in biology contains a paper on evolution (Quastler

1953; Yockey, Platzman and Quastler 1958). The first serious discussion of evolution in terms of information is by Ashby (1952) who dealt, as an analogy, with an imaginary chess-playing machine which could "learn" to outplay its designer. Ashby concluded that natural selection was a process for transforming random "noise" into information. Warburton (1958), in a very brief communication, concluded that natural selection is a mechanism for transferring pre-existing information from the environment to the genetic material of organisms. Kimura (1961), in a mathematically sophisticated paper, concluded that natural selection creates information (presumably out of nothing) which is then accumulated by the genetic material of populations, and estimated its rate of accumulation. These papers will be reviewed in a later section of this thesis.

The rarity of such publications suggests that great difficulties exist, some practical and some semantic, in applying the concept of a quantity of information to evolutionary problems. Some of these difficulties will be discussed in Section II of this thesis.

The Need for a Supplementary Mathematical Theory of Evolution.

Almost the whole of modern mathematical evolution theory is therefore included within the discipline of population genetics, and deals exclusively with rates of change of gene frequencies under the influence of mutation, migration, random sampling, deviations from panmixia, and selection, and with rates of change of quantitative characters controlled by polygenes. The theory is essentially a mathematical model of the empirical facts of genetics. This has well served its original purpose of reconciling Mendelian genetics and Darwinian evolution, but it is far from being a complete mathematical theory of evolution. It requires to be supplemented by a fresh mathematical approach meeting the following requirements:

First, the supplementary theory should be easily interpreted in terms of information theory, since the concept of a quantity of information is becoming increasingly important in molecular genetics and as a basis for interdisciplinary discussions intended to relate different fields of science to each other. The question of the source of the large quantities of information carried by the genetic material, for example, must be clearly resolved before organic evolution can be correctly related to other aspects of the cosmological process.

Secondly, the theory should permit a discussion of the entire evolutionary process, from the origin of life to the present day, including the period before the complex genetic systems of modern organisms arose. If possible it should be sufficiently general to serve as a guide for speculation about evolution on other worlds. If the theory is general enough for such purposes, it will permit evolutionary phenomena on earth to be discussed either in terms of the classical particulate gene, a mode of description to which natural history and population genetics will perhaps long be confined by practical difficulties, or in terms of nucleic acid structure and protein synthesis, a mode of description which will become increasingly important if evolution theory is to be related to modern concepts of gene structure and function.

To meet the above requirement, the theory must be freed from too rigid dependence on the empirical facts of genetics as a material basis for a mathematical model. One might hope, indeed, that the theory would permit some of the empirical facts of genetics to be deduced from its axioms. What is needed is not a system for deducing the evolutionary consequences of the facts of genetics -- population genetics already meets this need -- but a system for deducing the necessary conditions, some of which will be genetic conditions, for evolution.

Ideally, the theory should also be useful in discussing evolutionary interactions between populations which do not exchange genes, i.e. for discussing the evolutionary development of taxonomic categories above the species level, and the evolution of genetic systems. It should also permit an evaluation of the propriety of various analogies which have been drawn between organic evolution and other processes, such as learning (Pringle 1956), and the development of serological immunity (Burnet 1958).

Development of the Present Thesis.

The work leading to the theory presented in this thesis evolved through three main phases. Initially the writer believed that an adequate supplement to classical population genetics could be developed simply by restating its principles in terms of information theory. He therefore attempted to write expressions for such things as the quantity of information carried in the genome of an organism, the changes in this quantity resulting from mutation, etc. At that time he was unaware that a similar attempt had been made by J.B.S. Haldane (mentioned in Haldane 1953) who presumably found the approach unproductive, as did the writer, since he makes no further mention of it in later papers. The reasons for the difficulties encountered in such an attempt are made sufficiently clear in Section II of this thesis.

The next phase involved inventing and trying to develop a number of "mechanical" analogies to organic evolution. The first of these was frankly patterned after Ashby's (1952) hypothetical chess-playing machine. A number of authors have published such analogies, which will be mentioned here to indicate the kind with which the writer worked. Moore (1956) wrote a fanciful but fascinating discussion of "artifical living plants", floating factories carrying their own blueprints stored on tape, and utilizing solar power to build duplicates of themselves from minerals extracted from the sea, which could be harvested as a supply of metals. He mentions the possibility that evolutionary potentialities could be built into such machines, but does not discuss the matter except to point out that if would be safer not to do so. To design floating factories which would evolve, but could be guaranteed not to become dangerous predators on peaceful shipping, would indeed tax the theoretical understanding of contemporary evolutionists.

Friedman (1959) described an imaginary "selective feedback computer" for designing electrical circuits. The machine would construct a circuit from a variable combination of components, test it against predetermined criteria, and by restricting further trials to small random changes in circuits which had passed the easier early tests would rather quickly "...converge upon a type which passes the greatest number of tests". He pointed out that the number of trials required to find the best circuit by this procedure is of the order of the logarithm of the number required to find it by testing random combinations of components, unrelated to those already tested.

Rapoport (1961) analyzed the behaviour of a "perfect learner" which must learn, by trial and error, to give the correct response to each of a large set of stimuli. The learner is assumed to make trials in a systematic way, and to have a perfect memory. More instructive to the evolutionist would be a "learner" with a very imperfect memory, which learns to give the correct response (a well adapted phenotype) to a single extremely complex stimulus (the environment).

Michie (1961) pointed out the analogy between evolution and the process by which a very simple machine (MENACE; Matchbox Educable Noughts And Crosses

Engine) rapidly "learns" to become unbeatable at the game of noughts and crosses. Gardner (1962) described a similar but even simpler machine, HIM (Hexapawn Instructable Matchboxes).

A study of these analogies, and of others invented by the writer, was instructive in several ways. For example, they made it clear that to produce an extremely improbable end result one must accumulate slightly improbable events over a period of time. Muller (1929, 1958) had already made this abundantly clear in a purely biological context. Although it lay within his grasp, Muller did not state the fact that such a cumulative process can yield a highly improbable result in roughly the logarithm ofthe time required for it to occur by chance, and Friedman, although he stated this conclusion, did not demonstrate it.

Although suggestive, all these analogies required too many assumptions or restrictions peculiar to themselves to be of truly general biological applicability. The writer gradually became aware, however, that all the models with which he was working either wwre guessing games, like Twenty Questions, or could be readily transformed into guessing games. This realization led him to the third phase, the development of a "Theory of Guessing". No such formal theory appears to have existed before. It reveals principles which may be useful in considering any system in which highly improbable end results must be produced by the accumulation of slightly improbable "improvements" each of which is the result of a sort of trial and error process.

The notion of a quantity of information is central to the theory of guessing, and a section of this thesis is therefore devoted to introducing that concept. The next section discusses guessing games, and is developed to the point at which a particular kind of guessing game, a "Darwinian game", can be defined which serves as a useful model of evolution by natural selection. Then a section is devoted to fairly detailed analysis of a Darwinian game which is deliberately and explicitly oversimplified, but which leads to conclusions relevant to evolution theory. This is followed by a section consisting of examples which illustrate the results of substituting more realistic though more complex assumptions for the oversimplified assumptions used in the model which has been analyzed in detail. Finally, a discussion of the source of genetic information is presented.

II The Concept of Information

The qualitative notion that genetic material serves as a carrier of information is fundamental in modern biology. The most elementary discussion of molecular genetics mentions the hypothesis that information is coded as a sequence of base pairs in DNA and is transformed into a sequence of amino acids in protein molecules. The exact quantity of information carried by a genome and transformed during the development of an organism is difficult to compute, but it is universally recognized to be large. One of the major problems of evolutionary theory is therefore to discern the source of the large quantities of information carried by the genetic material of modern organisms, and to describe the process by which it has become incorporated in that material. This cannot be done without a clear understanding of the concept of a quantity of information. However, there is no need (in this thesis) to carry the mathematical development of the subject beyond an elementary level.

A good starting point is Ashby's (1957) suggestion of defining information as a measure of reduction of variety -- a suggestion which simplifies and clarifies exposition of the subject, but which Ashby unfortunately did not develop beyond a qualitative level.

Variety and Information.

With any set one may associate a number which measures the variety of the set. For example, an obvious measure of variety is the number of kinds of objects in the set. Of particular importance in information theory is the variety of a set of possibilities. The quantity of information carried by anything (such as a message or signal) is a measure of the amount by which it reduces the variety of some set of possibilities.

For example, imagine that one card from an ordinary deck of 52 has been chosen and hidden from me, and I wish to know what it is. The set of possibilities has 52 members. A message informing me that the card is the Queen of Hearts reduces the number of possibilities to one, and might therefore be said to carry 52 - 1 = 51 "arithmetic units" of information.

Measuring variety by the number of members in a set has certain undesirable mathematical consequences, illustrated by the examples below.

1. Let the message identifying a card be divided into 2 parts, i.e.

    A. The card is a Queen

    B. The card is a Heart

Part A reduces the number of possibilities from 52 to 4, and therefore carries 48 arithmetic units of information. Part B reduces the number of possibilities from 4 to 1, therefore carrying 3 units. However, if part B is received first, it reduces the number of possibilities from 52 to 13, i.e. it carries 39 arithmetic units, and part A then carries 12. In either case, A and B together jointly carry 51 units.

It is desirable to have a convention for measuring variety which makes the information carried by each of 2 or more independent messages (i.e. messages in which nothing about message B can be deduced from message A, and <u>vice versa</u>) independent of the order in which they are received or considered. This can be done in the present case by using the logarithm of the number of members in a set as a measure of its variety. The variety of possibilities is then $\log_{10} 52 = 1.7160$ <u>decimal units</u>, which is reduced by $\log_{10} 13 = 1.1139$ units by the message "The card is a Queen" and by $\log_{10} 4 = 0.6021$ units by the message "The card is a Heart", regardless of the order of the two messages.

2. Let the message identifying a card take the form "The probability

is 0.99999 that the card is the Queen of Hearts". This message clearly carries information in the everyday sense of the word, but it does not reduce the number of possibilities at all; its sole effect is to distribute the probability of being the card in question unequally over the members of the set of possibilities. The difficulty of measuring information in such a situation can be overcome, while retaining the advantages of a logarithmic measure, by using a quantity called <u>entropy</u> to measure variety. The entropy, $\underline{H}$, of a set of N possibilities is given by

$$H = - \sum_{i=1}^{N} p_i \log_b p_i \quad ,$$

where p_i is the probability of the i'th possibility, and $\underline{b}$ is the base to which logarithms are taken.

Before the message is received, the probability of being the chosen card is equally distributed over the set of 52 possibilities, and the entropy of the set of possibilities is therefore

$$H_1 = - 52 \left(\frac{1}{52} \log_{10} \frac{1}{52}\right) = 1.7160 \text{ decimal units.}$$

The message redistributes the probabilities so that

$$H_2 = - 51 \left(\frac{.00001}{51} \log_{10} \frac{.00001}{51}\right) - 0.99999 \log_{10} 0.99999 = 0.0001$$

decimal units.

Hence the message carries $H_1 - H_2 = 1.7159$ decimal units, or practically as much information as a message identifying the card with certainty.

Anything which reduces the entropy of a set of possibilities by $\underline{I}$ units may therefore be said to carry $\underline{I}$ units of information. This definition will be adhered to throughout the present thesis.

The units in which entropy, and therefore information, are measured depend on the base chosen for logarithms. If logs are taken to the base 10, entropy and information are measured in decimal units, or, more usually, decimal digits (since the entropy of the set of all possible sequences of n + 1 digits is 1 unit greater than that of the set of sequences of n digits). If natural logarithms are used, the units may be called nepits (a contraction of Naperian digits; Goldman 1953), or natural units. In the literature of information theory it is conventional to use the base 2 for logarithms; the units are then called bits, a contraction of binary digits.

A most useful relationship is

$$H = - \sum_{i=1}^{N} p_i \log_b p_i = \log_b N,$$

and is a maximum, if p_i is constant for all i. The entropy of a set of possibilities is a maximum if all possibilities are equally probable, and is equal to the logarithm of the number of possibilities.

If a carrier is to carry I bits of information, the set of possible carriers must have an entropy of at least I bits. If a receiver is to receive I bits of information per minute, the set of possible sequences of states which it can assume in one minute must have an entropy equal to or exceeding I bits. In communications engineering, therefore, a basic principle is that the information capacity of a channel, a transmitter, or a storage device is at most equal to its entropy. A communication system must be able to carry any one of a set of possible messages, and the engineering problem is to ensure that an adequate variety of states of the system is in fact possible. In communication theory therefore, one often reads of the "entropy of a message", which is numerically equal to the information content of the message, when what in fact is meant is the entropy of the set of possible messages. This is a frequent source of confusion, leading readers to conclude that information is entropy, or entropy is information. The term "negative entropy" or its contraction, "negentropy", have also become current as synonyms for information. Information is a difference between two entropies, and is not entropy any more than a difference between two temperatures is a temperature, although it is measured in the same units. The confusion can be avoided by keeping in mind that entropy is a characteristic only of sets of possibilities, not of real objects.

For example, how much information is carried by the message ABA? This message has no entropy; it is what it is, and there is no probability that it is anything else. To estimate its information content, one must know the entropy of the set of possibilities from which it was chosen. If the set of possible messages is all sequences of 3 capital letters, ABA carries $\log_2 (26^3)$ bits; if lower case letter could be used as well, ABC carries $\log_2 (52^3)$ bits; if longer and shorter messages are included in the set of possibilities, the information carried by ABA is increased.

Biological Relevance.

Most discussions of information theory discuss situations in which the set of possibilities is easily defined. There is no difficulty in reaching agreement that the identity of a playing card is one of 52 equally probably possibilities, though one may hesitate momentarily about whether or not the deck contains a Joker. Elementary expositions of communications theory tend to deal with Morse alphabets -- dot, dash, letter space, word space -- or with sequences composed exclusively of capital letters and spaces. Even when communications theorists deal with frequency-modulated electromagnetic waves, in which the variety of possible frequencies would appear to be infinite since frequency is a continuous variable, their first step is to establish a "band width" which imposes finite limits on the entropy of possible sequences of frequencies which can occur in a finite period.

The major difficulty in applying information theory to biological problems lies in the fact that no such clearly delimited sets of possibilities seem to be available, with a few exceptions.

For example, consider an attempt to compute the information content of a DNA molecule which serves as the genophore of a micro-organism. Let the molecule be L nucleotide pairs long. At some stage in the life of the organism, this molecule will replicate itself. The set of possible DNA molecules of length L has 4^L members, and if each is considered equally probable the set has an entropy of 2L bits. The parent molecule therefore supplies almost 2L bits of information to the daughter molecule -- almost, but not quite, since errors in replication may occur; the set of possible daughter molecules has a large number of members, most of which are of very small probability.

The DNA molecule is also instrumental in synthesizing a large number of protein molecules, forming what Yčas (1958) calls the "protein text". Assuming that three nucleotide pairs are required to determine an amino acid, the total length of the protein text is L/3 amino acids. If 20 different amino acids occur, the set of possible protein texts has $20^{L/3}$ members, for a maximum entropy of 1.441 L bits. Therefore, in this context, the DNA molecule transfers at most 1.441 L bits of information to the protein complement of the organism.

If one takes into consideration the fact that the 20 free amino acids exist in very different concentrations in the cell, one must consider that not all protein texts have equal a priori probabilities. This reduces the entropy of the set of possible protein texts, and therefore reduces the estimated quantity of information supplied by the DNA.

Part of the genophore may constitute regulator and operator genes, controlling the temporal sequence of protein synthesis. The set of possibilities, among which the genophore determines the one which will be realized, is then the set of all possible temporal sequences of protein synthesis under particular external conditions. It is difficult, if not impossible, to estimate the entropy of such a set of possibilities in any acceptable way.

Just as the information transferred from DNA to protein was less than the maximum which could be carried by the DNA, the information transferred from enzymes to the products of metabolism may be further decreased, due to the possibility that enzymes differing somewhat in their amino acid composition may catalyse the synthesis of, for example, identical carbohydrates. The end product of all such information transfers is that set of properties which is collectively called the "phenotype" of the organism. A basic question, of course, is how much information is transferred from the genetic material to the phenotype. Any meaningful estimate of this quantity would require an estimate of the entropy of the set of all possible phenotypes -- not, it must be emphasized, all possible viable phenotypes, or all possible phenotypes of a given species, but all possible phenotypes. It is hard to imagine how to attack even the tremendous semantic difficulties which one would need to overcome before one could begin the quantitative study required for such an estimate.

"Information content" is therefore not a property of an object, such as a chromosome, in the way that length or mass is thought of as a property of an object; information is a quantitative relation between two sets of possibilities. The major difficulty in applying information theory to biology lies in achieving a suitable compromise between the choice of a set of possibilities which is easily defined and has an easily computed entropy but which is of little biological interest, such as the set of all possible DNA molecules of a given length, and a set of possibilities which is of fundamental biological importance but cannot be clearly defined in a way which permits its entropy to be computed, such as the set of all possible phenotypes.

Two other basic concepts of information theory must be briefly defined; redundancy and noise. If the entropy of a set of possible messages exceeds the entropy of the set of possible situations of which one is specified by a message, the set of messages is said to be redundant. A measure of redundancy is

$$R = 1 - \frac{H_1}{H_2},$$

where H_2 is the entropy of the set of possible messages, and H_1 is the entropy of the set of possibilities from which one must be specified by a message.

Redundancy may be said to exist in a set of possible messages when the entropy of the set is arbitrarily reduced below its maximum possible value. For example, the set of possible English messages (of given length) has a smaller entropy than the set of possible sequences of letters of the same length, because of the restrictions imposed by the rules of English spelling and grammar. One may think of these rules as assigning unequal probabilities to different sequences of letters, or alternatively, as assigning the same meaning ("nonsense!") to a large number of possible sequences. The former implies a reduction in the entropy of the set of possible messages, the latter a reduction in the entropy of the set of distinguishable situations at the receiving end, but either interpretation implies that redundancy exists in the set of messages.

Redundancy, like information, is a relation between two sets of possibilities. If a triplet of nucleotide pairs is required to specify one amino acid, DNA displays considerable redundancy during protein synthesis, as there are 64 possible triplets but only 20 possible amino acids; but a DNA molecule displays no redundancy when it is thought of as specifying the structure of a descendant DNA molecule.

Various influences may prevent a message from reducing the entropy of a set of possibilities to as low a value as it would do in a hypothetical perfect or ideal communication system. Such influences are collectively called "noise." For example, noise in the everyday sense of the word may prevent me from hearing all of a message identifying an unknown card, so that although I become certain that the card is a Heart, I remain unsure whether it is an Ace or an Eight; the set of possibilities retains an entropy of 1 bit, although the message was ideally capable of reducing the entropy to zero. In biological systems ionizing radiation, traces of toxic substances, and so on, prevent a DNA molecule from specifying a protein text with a probability of 1, and thus constitute noise.

Because of the conceptual and semantic difficulties involved in its use, the writer has tried to be cautious in applying information theory in this thesis. The thesis deals with such sets of possibilities as "the set of all possible living organism", "the genomes of all organisms as well adapted to their environments as those we see around us", etc. It would obviously be extremely dangerous to estimate the entropy of such vaguely defined sets, and no attempt will be made to do so. There are situations, however, which are important in the thesis, in which a quantity of information can be easily and unambiguously defined. For example, if $\underline{r}$ organisms exist of which only one can survive as the parent of the next generation, the selective agency, which determines which shall die and which shall live, can at most supply exactly $\log_2 r$ bits of information. If care is taken to restrict its use to similarly clear-cut cases, information theory can be extremely useful in biology. Often, the exact computation of the redistribution of probabilities brought about by some act of selection or designation is extremely difficult, but all that needs to be known can be summed up by computing the maximum quantity of information which the act can transfer.

III <u>The Analogy between Organic Evolution and a Guessing Game</u>.

A <u>guessing game</u> is a situation involving one or more <u>players</u> and a <u>referee</u>. In order to win, a player must find a <u>right answer</u>, or one of several right answers, within $\underline{T}$ trials. There are $\underline{N}'$ right answers, collectively forming a set $\underline{S}'$, about which the player initially knows nothing except that it is included in a set $\underline{S}$, with $\underline{N}$ members, the <u>initial set of possibilities</u>. That is, the player initially lacks the information required to deduce a right answer, and must seek one by guessing. The referee responds to the player's guesses in a way which conveys information which can be used to reduce the number of trials required to find a right answer below chance expectation.

In some real guessing games, but not all, the referee chooses the right answer, and is considered to have lost the game if a player succeeds in winning it. This is not a general characteristic of guessing games, however, and is not an interpretation which can be applied to the physical, biological or psychological systems which may be analogized to guessing games. In most guessing games the referee cannot be considered to be an opponent of any of the players; his position, as the name indicates, is that of a neutral referee. A guessing game is therefore not a "game" between the referee and the players, in the technical sense assigned to the word "game" in the Game Theory of von Neumann and Morgenstern (1944).

Later in this thesis some discussion will be devoted to competitive guessing games, in which there are several players. In a competitive guessing game the relations between the players may correspond to the relations between players in a von Neumann Game, but in no case is the relation between any player and the referee a competitive one.

Notice that no particular restrictions have been put on the physical nature of systems which can be analogized to guessing games. The player and the referee may be people, constrained in their behaviour by the rules of the game; they may be machines, behaving as they do because of the way they were designed and built; they may be mechanisms, in the philosophical meaning of the word, which must behave the way they do because of the relations between their parts; or they may be imaginary entities which are useful fictions in describing the behaviour of a system. The restrictions the model imposes on the system are almost entirely restrictions on the way information is transmitted from one part of the system to another. Sufficient information to identify a right answer is initially stored by the referee, and is not immediately available to the player. It is transmitted to the player a little at a time, in response to guesses, and the player stores it until he has accumulated enough to specify a right answer.

Let us consider organic evolution as a guessing game. The "player" is an imaginary entity who manufactures living organisms. The original set of possibilities, <u>S</u>, is all those possible organisms which could have descended from the first spontaneously generated organism in an infinite universe, in which selection was inoperative and even the most grotesque and unadapted forms could survive. The set of right answers, <u>S</u>', includes all those organisms as well adapted to life on Earth as those we see around us which are not in immediate danger of extinction.

Each organism manufactured by the player is submitted to the environment, which "passes" some and "rejects" others, thus serving as a referee to inform the player whether each one is a good guess or a bad guess, or at least whether it is better or worse than the previous attempt. Each new mutant or recombinant is indeed a "guess", as the imaginary player does not know whether the change is an improvement or a defect until the organism has been tested and reported upon by the referee. In the Darwinian hypothesis, the referee informs the player, in effect, "this giraffe is better than that giraffe", but does not say what kind of superiority it possesses. According to Lamarckian hypotheses, the referee says, in effect, "this giraffe is unsatisfactory; try one with a longer neck next time". A Special Creator, of course, does not need to refer to a referee for information at all, but knows the right answers before the game begins.

Guessing games fall into several categories. The kind which can best serve as a model of evolution by natural selection may be called a "Darwinian game". Its characteristics will be outlined in the following pages.

Trivial and non-trivial guessing games.

A Darwinian game is a non-trivial guessing game.

A guessing game may be trivial for various reasons. It is trivial, for example, if the referee conveys no information to the player, who must merely continue guessing until he finds a right answer by blind chance. At the other extreme, a guessing game is trivial if the referee "gives away" the right answer, i.e. supplies so much information at an early trial that the player can at once deduce a right answer. Guessing games are of interest only if the referee supplies some information at each trial, but much less than log (N/N!) units, so that the player must accumulate information over a series of trials in order to win. This requirement of a non-trivial game is met by restricting the referee to giving one of a small set of possible responses at each trial. E.g., a binary referee, or a referee of order 2, must give one of only two possible responses, such as "yes" or "no", or "right" or "wrong", at each trial. Such a referee can supply a maximum of one bit of information per trial.

It is standard practice in quantitative biology to decide upon some value, P_c, as a <u>critical probability</u> in testing whether or not a given hypothesis is adequate to explain a set of observations, as in choosing $P_c = 0.01$ as a level of significance when using chi-square to test whether experimental results are adequately explained by a null hypothesis. Now there is always some small probability of winning a guessing game in <u>T</u> random guesses, or by running down the list of the first <u>T</u> members of <u>S</u>. For a given game, if

$$\frac{TN'}{N} \geqslant P_c,$$

it is plausible to postulate that a winning player has won by some such procedure; it is not necessary to assume that the referee has supplied information to the player, or to assume that the player has used the information if it has been supplied; <u>blind chance</u> is an adequate explanation for the player's success. A game which has a probability P_c of being won by blind chance is a trivial game.

In a non-trivial game, the referee must supply information to the player. Usually, at any trial after the first, not all possible guesses are equally valuable to the player in eliciting information from the referee. The player must therefore accumulate information from the referee, and must use the accumulated information to control subsequent guessing. The way he does so may be called his <u>strategy</u>. For example, it is generally useless to repeat a guess already known to be wrong, as this does not give the referee an opportunity to supply additional information. It is usually advantageous for the player to choose each guess so that, as far as he knows, all the possible responses of the referee are equally probable; in this way he obtains the maximum quantity of information per trial which the referee can give.

An <u>adequate strategy</u> for a guessing game is one which provides the player with

$$\frac{\log\,(P_cN/N')}{T}$$ units of information per trial, or more, i.e. which gives him a probability at least equal to the critical probability of finding a right answer in T trials. The optimum strategy, of course, is the strategy which minimizes the number of trials required to win the game, or maximizes the probability of winning it within T trials. One may loosely call a game difficult if no adequate strategy is very different from the optimum strategy, or easy if the player has a rather wide choice of adequate strategies.

Mammalian haploid chromosome sets contain about 4×10^9 nucleotide pairs. Some of these may carry no genetic information which is transmitted to the phenotype, some may "say" things in one way which could be said as well or better in another. Allowing for these possibilities, Muller (1958) guesses that the probability that a random sequence of nucleotide pairs would form a genome as well organized and adaptive as that of a higher organism is roughly $10^{-2,000,000}$. If the set S consists of all possible sequences of 4×10^9 nucleotide pairs, the probability that one taken at random will be a right answer is therefore about 1 in $10^{2,000,000}$. If the Earth consisted entirely of nucleotides, which rearranged themselves in double chains millions of times per second for billions of years, only about 10^{50} or 10^{60} different sequences would have an opportunity to occur. The chance of a right answer turning up by chance in this number of trials is negligibly small. The evolutionary "game" therefore cannot be won by blind chance, and must be non-trivial.

The important point is that we know a game of this sort has been played and won, and could not have been won by blind chance. Our task is to elucidate the necessary characteristics of an adequate strategy.

An example of a trivial guessing game is the children's game "I Spy". The referee begins the game by announcing "I spy with my little eye something that begins with (for example) 'C'". The player then names objects until he discovers the right one or gives up; i.e., T depends on the player's patience. At each guess, the referee responds with "right" or "wrong".

The only strategy available to the player is to maintain a record of past trials and thus avoid repeating wrong guesses; i.e. to run through the list of possibilities one at a time. I Spy can therefore only be won if it is trivial. The referee's response to each guess enables the player to eliminate that guess, and only that guess, from the set of possibilities. As each guess is much more likely to be wrong than to be right, the referee, who is a binary referee, supplies much less than one bit of information per trial.

To win a non-trivial game the player must be able to eliminate more than one member of S at each trial (or so redistribute the probability of being a right answer among the members of S that the reduction in entropy is equal to that following elimination of several members from the set). Clearly, the player cannot merely eliminate some random collection of possibilites at each trial; the ones he eliminates must be related in some way to the ones he has actually guessed. The set S must therefore be relevantly classifiable, i.e. divisible into subsets of related members in a way which is relevant to the probability of each member being a right answer.

In any guessing game with a binary referee, therefore, the player's procedure can be analyzed into a cyclic repetition of three steps:

1. Divide the set of possibilities, S, into two classes.
2. Obtain from the referee a statement of which of the two classes is

better, or information from which such a statement can be inferred. If p_i is the probability that a member taken at random from the class C_i is a right answer, class C_1 is better than C_2 if p_1 exceeds p_2. I.e., class C_1 is better if the frequency of right answers among its members exceeds the frequency in C_2.

3. Store the result of the comparison, and repeat the cycle.

In order to win the game, within T trials the player must be able to delimit a class C' for which p' exceeds P_c, where C' is the logical product of the better of each pair of classes compared.

If the referee is of order greater than 2, let us say of order r, the player may divide the set of possibilities into r classes at each repetition of the cycle.

Question Games and Trial-and-Error Games.

Guessing games fall into two major categories, question games and trial-and-error games. A Darwinian game is a trial-and-error game.

A question game is a guessing game in which each guess constitutes a question about a whole class of possibilities, and the referee's responses explicitly apply to whole classes. A familiar example of a question game is Twenty Questions. In this game, the referee announces whether the right answer is animal, vegetable, or mineral. The player then asks such questions as "Is the subject a real animal?...a living animal?...a human being?...a Canadian statesman?", etc., to which the referee answers either "yes" or "no".

The game rests on the fact that the set of possibilities can be completely classified in a dichotomous way, and the right answer must belong either to the class named or its complementary class. The player's strategy is to try to choose each question in such a way that "yes" and "no" are equally probable, thus eliciting the maximum of one bit of information from the binary referee, so that at each trial he can eliminate a whole class, containing about half the remaining possibilities. If $T > \log_2 (P_c N)$, this strategy is adequate.

The important point is that in a question game, _any_ classification of the members of _S_ is a relevant classification. The player is therefore perfectly free to classify the set of possibilities as he likes, and should have no difficulty in meeting the strategically desirable aim of comparing two nearly equal classes at each trial. Question games therefore tend to lack variety and interest from the mathematical point of view.

Trial-and-error games are those in which the player is restricted to naming (or producing) individual members of the set _S_, and the referee's responses apply explicitly only to the individuals named. In trial-and-error games the player must be able to draw inferences from the referee's statements about individuals to larger classes which these individuals represent. It is the player's inability to make such inferences validly which renders I Spy a trivial game.

Evolution by natural selection must be analogized to a trial-and-error game. The referee does not report on entire classes of possible organisms presented to it in the abstract; it merely passes or rejects those particular individuals which actually come into existence and are exposed to the environment. A Darwinian game, therefore, is a trial-and-error game.

Primary Referees (Selectors) and Secondary Referees.

In a trial-and-error game, the referee's response to each guess constitutes a comparison between individuals, from which the player infers a comparison between classes. A binary referee, which is restricted to giving one of only two possible different responses, can only give a series of responses which constitute a series of comparisons between individuals if the individuals, which are members of the set S, can be ordered in some way.

An ordering relation is any 2-termed, asymmetric, transitive, non-reflexive relation (Langer, 1953), such as "...exceeds...", "...is north of...", "...is closer to x than...". If the members of the set S are connected by an ordering relation, a rank can be assigned to each member of S. If there are as many ranks as S has members, and only one member per rank, S is a completely ordered set. If there are fewer ranks than members, so that some members have the same rank, S is an incompletely ordered set. The integers can be completely ordered by the relation "...exceeds..."; words can be completely ordered by the relation "...alphabetically precedes..."; points on a plane can be incompletely ordered by the relation "... is farther north than...".

In any guessing game, the set S can be at least incompletely ordered by some such relation as "...is better than...". There must be some way of ranking each possibility as "better" or "worse" than other possibilities, with the right answers having the highest rank. This relation will serve to divide S into at least two ranks, as the set of possibilities must contain both right answers and wrong answers, and the right answers are obviously "better than" the wrong answers. This relation (or similar ones, such as "...is better adapted than..." in biology) may be called the primary ordering relation of the set S, which serves to assign a primary rank to each member. The player of course must not know the primary rank of any member of S, except insofar as he can infer it from information received from the referee.

The members of S may or may not also be connected by secondary ordering relations, which can be known to the player without giving away the right answer. If the set of possibilities consists of all the integers from 1 to 10^{10}, for example, the player can obviously assign a secondary rank to each possibility, since he knows what numbers are bigger than others, but he cannot assign a primary rank to each number until he learns from the referee that certain numbers are better than others.

The referee in a trial-and-error game must therefore perform one of four possible kinds of comparison:

1. He might compare the secondary ranks of two guesses ("x_1 is larger than x_2"). This information, however, is already available to the player and its reiteration by the referee is useless.

2. He might compare the primary rank of each guess with the primary rank of the right answer. This is equivalent to a simple statement that each guess is either right or wrong, as in I Spy, and cannot supply enough information to permit the player to win a non-trivial game.

3. He might compare the secondary rank of each guess with the secondary rank of the right answer ("the number you have guessed is too large"). Responses of this kind can supply useful information to the player. For example, let S be the set of integers from 1 to N, inclusive. If the player's first guess is N/2, and the referee's response is "too large", the player can infer with certainty that the right answer lies between 1 and N/2, i.e. he can exclude half the possibilities after a single trial.

A referee who compares the secondary rank of each guess with the secondary rank of the right answer may be called a secondary referee.

4. He might compare the primary ranks of two guesses ("this guess is better than the last one"). Such a referee could be called a primary referee but will be called a selector in this thesis, for biological reasons.

The set of all possible organisms probably cannot be secondarily ordered in any complete way, although its members could be partially ordered by size, for example. The set of all possible DNA molecules could be secondarily ordered in an "alphabetical" arrangement by adopting a suitable convention for listing the four possible nucleotide pairs in a fixed order. Nevertheless, there is no reason to consider the possibility that the referee in the evolutionary "game" conveys any information to the player about the relative secondary ranks of each guess and the right answer. The environment would seem not to supply information of the form "this giraffe's neck is too short", or "this dinosaur is too big", and it certainly does not supply information of the form "this genophore alphabetically precedes the fittest possible genophore". The environment supplies information only in the form "this organism (or population) is fitter than that organism", where the relation"...is fitter than...", or "...is better adapted than..." is a primary ordering relation. The referee in a Darwinian game is a selector.

Relevant and Irrelevant Classifications.

At his first trial, the player must divide the set S into two classes (if the referee is a binary referee, or ideally into r classes if the referee is of order r), which can be compared by the referee. At the next trial, he must divide the better of these two classes into two classes, which can be compared by the referee, and so on. The members of S must therefore be classifiable in a hierarchy. The members of S can be completely classified in a hierarchy of L levels, in which the entire set S forms the class at level 0, if they can be so distributed among classes that:

1. There are more classes at level i + 1 than at level i.

2. The members of any given class at level i + 1 are either all included in or all excluded from any given class at level i.

3. The members of any given class at level i + 1 are all included in one class at level i; and any given class at level i includes all the members of at least one class at level i + 1.

4. Each class at level L contains only 1 member.

If any class at level L contains more than 1 member, the hierarchy serves to classify the set incompletely.

E.G. the integers from 1 to 16 can be completely classified in the hierarchy shown in Figure 1. The classic example of a hierarchy is the taxonomic hierarchy in which organisms are systematically classified; it is an incomplete classification, as each class at the lowest level (subspecies) contains many individuals.

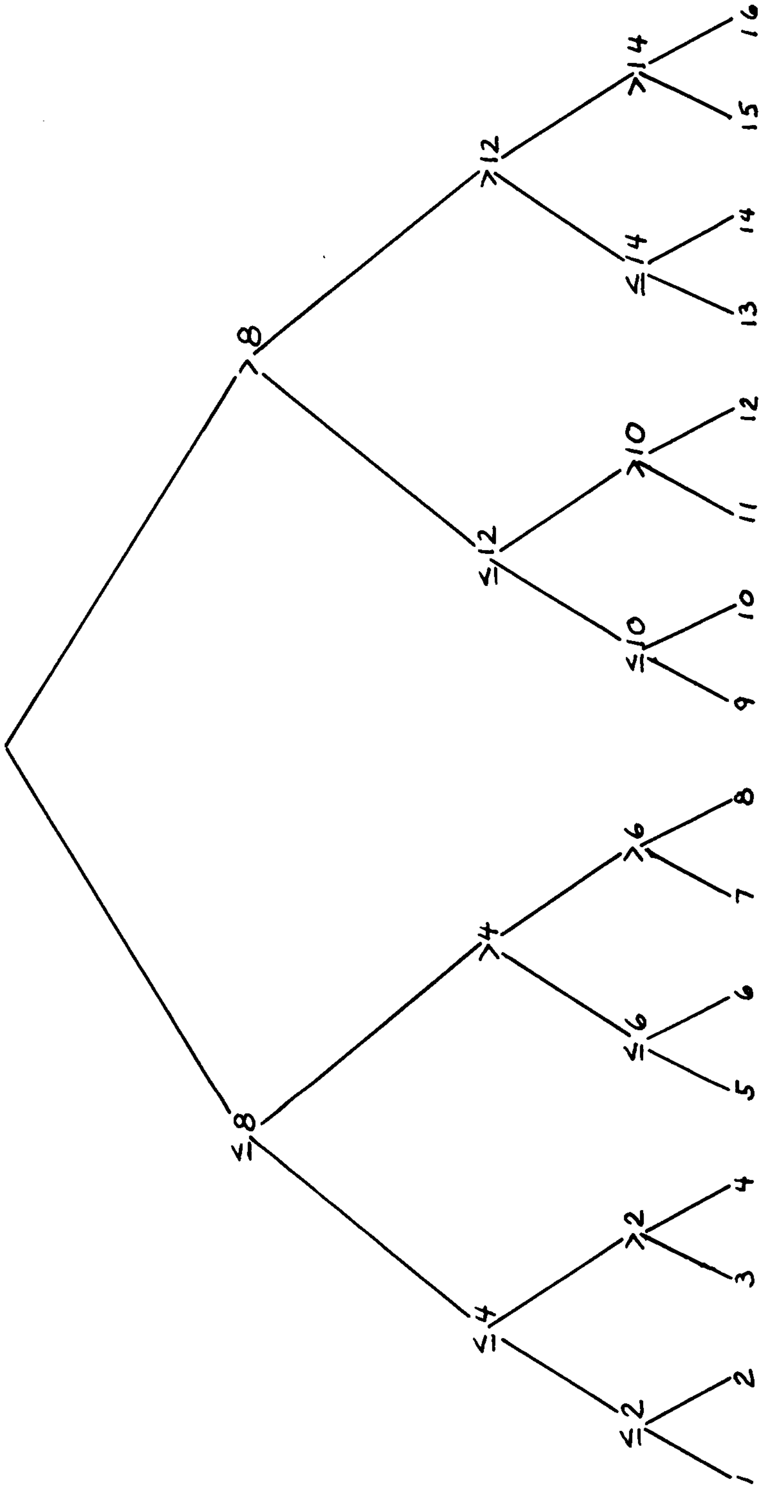
A degenerate hierarchy is a hierarchy in which all classes but one at each level contain only one member. In I Spy the player classifies the set of possibilities in a degenerate hierarchy, as in Figure 2. It is useless for the player to classify S in a degenerate hierarchy unless the game is trivial.

At each trial, therefore, the player knows that the right answer is a member of a particular class, C_{ij}, on the i'th level of the hierarchy (or, more generally, the player knows that the probable frequency of right answers is higher in the class C_{ij} than in any other class at level i). His procedure is therefore to divide the class C_{ij} into 2 (or r) daughter classes at level i + 1, and obtain a comparison of the daughter classes from the referee. The better of the two -- the class $C_{(i+1,x)}$ with the higher probable frequency of right answers -- becomes the next class to be divided.

A hierarchical classification is relevant in a given game if:

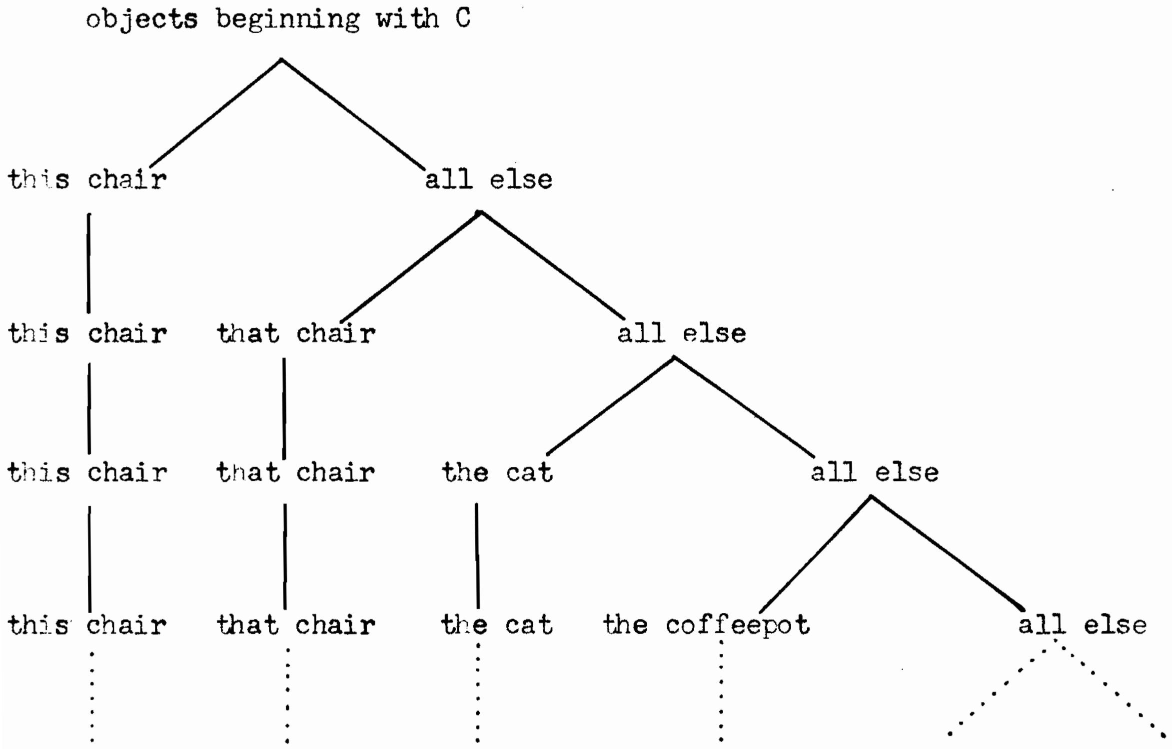
1. Each division (or most divisions), in which any daughter class contains a right answer, divides the right answers unequally between the daughter

Figure 1 A complete classification of the integers from 1 to 16 in a hierarchy of four levels.



≤8
>8
≤4
>4
≤12
>12
≤2
>2
≤6
>6
≤10
>10
≤14
>14
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Figure 2 A partial classification of objects beginning with C, in a degenerate hierarchy, such as might be used in I Spy.


objects beginning with C
this chair
all else
this chair
that chair
all else
this chair
that chair
the cat
all else
this chair
that chair
the cat
the coffeepot
all else

classes.

2. The best of the daughter classes, i.e. the one with the highest frequency of right answers, can be identified with a high probability by the referee's response to the guess.

The first of these requirements will be met by any hierarchical classification; if N is much larger than N', division of S by any criteria will give rise to classes containing no right answers at levels which will be reached within a few trials. In question games, the second requirement is fulfilled by the rules of the game. Hence, as has been stated before, any hierarchical classification is relevant in a question game.

In a trial-and-error game with a secondary referee, the second requirement is met by the properties of ordered sets; having learned, for example, that the guess x_1 is larger than the right answer, the player can infer with absolute certainty that the class "numbers smaller than x_1" is better than the class "numbers as large as or larger than x_1". Any classification of S in which each class is divided into two daughter classes of secondary rank respectively higher and lower than some chosen member of the class is a relevant classification. Because a relevant classification is always obvious, and the optimum strategy is also obvious, trial-and-error games with secondary referees lack variety and are of little mathematical interest.

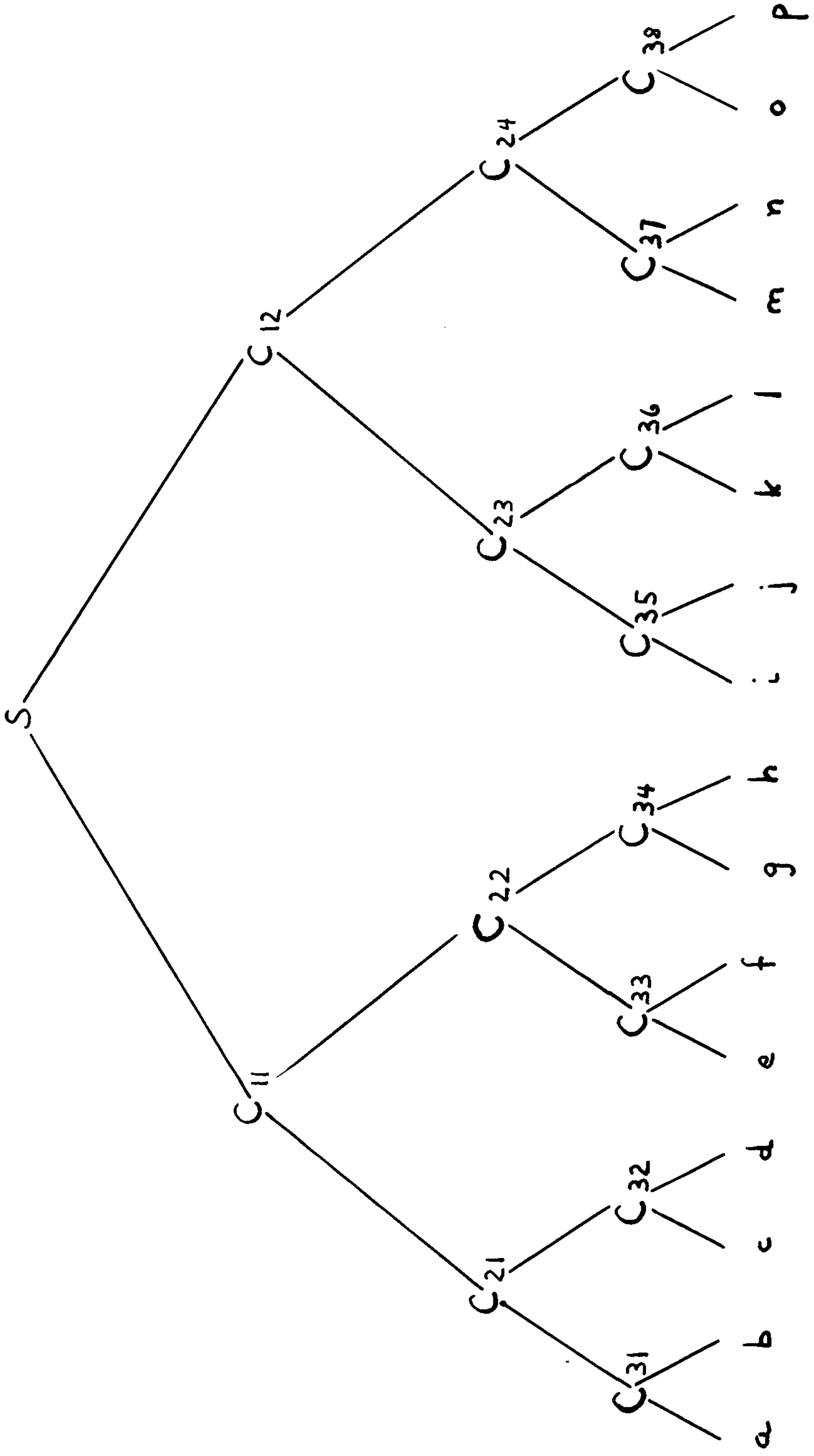
In a trial-and-error game with a selector, the player cannot make logically rigorous inferences about classes from the referee's responses comparing the primary ranks of two guesses, simply because, by definition, the player cannot know what system is being used by the referee to assign primary ranks to the members of S. For example, suppose the player must find an integer, x, between 0 and 10^{10}. The player may assume that since x_1 is better than another guess x_2, x is numerically closer to x_1 than to x_2. If this assumption is correct -- if the referee is in fact ranking guesses by their numerical proximity to x -- the players best next guess is a member of the class containing x_1 and bounded by the number $(x_1 + x_2)/2$. The assumption may be wrong, however; the referee could conceivably be assigning a score to each number, consisting of the sum of the prime factors it shares with x minus the sum of the prime factors it does not share, and ranking the numbers by their scores. If so, the player will be unable to win the game if his classification of S is based on size.

Nevertheless, to win a non-trivial game, the player must be able to classify S in such a way that he can correctly assume that if an arbitrary or random individual from class C_i is of higher primary rank than one from class C_j, the probability exceeds 1/2 that the frequency of right answers among the members of C_i is higher than among the members of C_j.

For example, consider a game in which S consists of sixteen things, labelled abcd...p, which the player can classify as in Figure 3. The best member of the set is p. The player is allowed four trials; i.e. he may name four pairs of members to be compared by the selector. How must the members of S be primarily ranked to yield a 50% probability that the better individual in the fourth trial is the best individual in the set?

If at the fourth trial the player must choose between the two members of C_{38}, he is certain to make the correct choice; if he must choose between the members of any other class at level 3, he is certain to choose an individual which is not the best of the set. Therefore, the player must have a total probability of 1/2 of choosing the correct branch at every fork in the line leading to C_{38}, which means that he must have an average probability considerably larger than 1/2 of making the right choice at each of the three forks. If the members of C_{24} are ranked within the class C_{24} in the order $m = 2$,

Figure 3 A complete classification of 16 objects labelled a,b,c....p, in a hierarchy of four levels. For discussion, see p. 36.



S
C11
C12
C21
C22
C23
C24
C31
C32
C33
C34
C35
C36
C37
C38
a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p

n = 3, o = 1, p = 4, and the player randomly or arbitrarily chooses a member from each of C_{37} and C_{38} for the selector to compare, the probability that the better of the two individuals comes from the better of the two classes is only 1/2. I.e. the probability that the player will make the correct choice among the two subclasses of C_{24} is only 1/2, and to have a total probability of 1/2 of making the correct choice at all three forks, the player would have to be certain of choosing correctly at the two higher forks. This implies that every member of C_{24} must be higher in rank than every member of C_{23}, and every member of C_{12} must be higher in rank than every member of C_{11}. On the other hand, if the members of C_{24} are ranked within that class in the order m = 1, n = 3, o = 2, p = 4, the probability is 3/4 that the better individual comes from the better class, and the restrictions on ranking within the larger classes can be somewhat relaxed. The reader may wish to study Table 1, which shows some primary ranking systems which give the player a 50% of finding the best member of S̲ on the fourth trial, and some which do not.

This example, with the small set of sixteen possibilities, hardly offers a fair illustration of the way the primary ranking scheme must be related to the classification of the set if the game is to be won. After all, the player is permitted to name eight members of the set, and therefore has a 50% probability of naming the right one by blind chance; only those relations between ranking and classification which give a probability greater than 50% are really of interest, Consider, however, a set S̲ containing 2^{1000} members, dichotomously classified in a hierarchy of 1000 levels, with only one right answer. The player can follow a line to an individual member in 1000 trials, naming at most 2000 individuals as he does so; the probability that one of the 2000 will by chance be the right answer is negligibly small. To have a

TABLE 1

THE PROBABILITY, P, OF IDENTIFYING THE BEST OF 16 OBJECTS IN 4 TRIALS

Primary Ranking Systems																
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	P
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	1
1	2	3	4	5	6	7	8	9	10	11	12	13	15	14	16	3/4
1	2	3	4	5	6	7	8	9	10	13	14	11	12	15	16	3/4
1	2	3	4	9	10	11	12	5	6	7	8	13	14	15	16	3/4
1	2	3	4	5	6	7	8	9	10	13	14	11	15	12	16	9/16
1	2	3	4	9	10	11	12	5	6	7	8	13	15	14	16	9/16
1	2	3	4	9	10	11	12	5	6	13	14	7	8	15	16	9/16
1	2	3	4	5	6	7	8	9	10	11	12	14	15	13	16	1/2
1	2	3	4	5	6	7	8	11	12	13	14	9	10	15	16	1/2
5	6	7	8	9	10	11	12	1	2	3	4	13	14	15	16	1/2
5	6	7	8	9	10	11	12	1	2	3	4	13	15	14	16	3/8
5	6	7	8	9	10	11	12	1	2	3	4	14	15	13	16	1/4
8	9	10	11	12	13	14	15	7	6	5	4	3	2	1	16	1/65536
1	1	1	1	1	1	1	1	2	2	2	2	3	3	4	5	1

probability P_c of finding the right answer in 1000 trials, he must classify *S* in such a way that the probability of choosing the correct branch at each fork on the line to the right answer has a geometric mean of at least $\sqrt[1000]{P_c}$. In other words, at every fork on the line, the members of the better class must practically all be higher in rank than the members of the other class.

The situation becomes more complicated if *T* exceeds the number of levels in the hierarchy, so the player can to some extent go back and try again if he makes a mistake; if there are several right answers; or if the player retains the better of the two guesses from one trial for comparison with each new guess at the next trial.

If *T* is large enough, and his first series of *L* trials (where *L* is the number of levels in the hierarchy in which the player has classified *S*) fails to yield a right answer, the player may begin again, using a different system of classification. Nevertheless, in any moderately difficult game, the player cannot repeat the game using more than a few systems of classification. If the player is to win a difficult non-trivial trial-and-error game in which the referee is a selector, therefore, the referee must assign primary ranks to the members of *S* in a way that makes some "natural", "obvious", "simple", or "objective" system of classification relevant. The game cannot be won if the referee merely assigns ranks to the members of *S* in an arbitrary or random way.

The data of paleontology suggest that the referee in the evolutionary game fulfills this requirement. Entire classes, orders, and families of organisms have become extinct, their places being taken by new groups of comparable taxonomic rank which undergo an adaptive radiation. The almost complete replacement of Chondrostean fishes by Holostean fishes, followed by the almost complete replacement of the latter by Teleostei, is a classic example (Romer 1945; Colbert 1955). Such instances suggest that the taxonomic hierarchy used by paleontologists, which is a "natural" classification based almost entirely on the morphology of structures hard enough to become fossils, is a relevant classification. Of course one would hope it would be. If it were irrelevant, one would expect individual species within larger taxa to have become extinct with no particular correlation with taxonomic categories, so that modern organisms would include an apparently random sample of species from all taxa which have existed.

Storage of Data.

In any non-trivial guessing game the player must accumulate information over a series of trials, and utilize it to control subsequent guessing. He must therefore be able to retain some kind of records of past trials, the function of which is to allow him to determine which class should be divided at the next trial. Let us restrict the kind of data he can store to records of two forms:

1. Records of one or more past guesses, with the referee's response to each.

2. A t-record, i.e. a record of the number of previous trials.

A complete record of all past guesses and their outcomes is obviously adequate to permit the player to identify the class next to be divided. More complex situations arise when the player can store records of only a small number of past guesses. In such situations, the t-record may be valuable, as it identifies the level in the hierarchy of classification on which the class to be divided lies.

The only case which must be considered here is a trial-and-error game with a selector. If the player is restricted to keeping a record of only one past guess, this must be the best of the guesses presented to the selector at the last trial; the best past guess is always a member of the class to be divided. A t-record, which will inform the player what level the class to be divided is on, then serves to identify the class to be divided.

In organic evolution, the only "records" available to the imaginary player are the actual organisms which were passed by the selector in the last trial. Rejected organisms or genotypes leave no record; they are lost without trace as far as the player is concerned, although some of them may leave records discernible to human paleontologists. Dead or sterile organisms have no influence on later generations. The imaginary player has no memory; his sole record of the trial just concluded is the actual organisms which passed. The player has no record even of the number of past trials; the genetic material of an organism bears no explicit record of the duration of its history.

In a Darwinian game, the only data stored by the player consists of the best past guess; in particular, he cannot retain a t-record.

Perfect and Imperfect Referees.

Some further possible complications must be mentioned before a mathematical analysis of organic evolution in terms of guessing games can be carried out.

A perfect referee is one who always gives a truthful response to a guess; an imperfect referee is one who sometimes gives a false response. Clearly, if a binary referee is to convey any information at all to the player, he must give a truthful response more than half the time. Operationally, one would recognize an imperfect referee by calling upon him to make the same comparison repeatedly, and observing whether his response was always the same.

The quantity of information supplied per trial by an imperfect binary referee is $1 - (p \log_2 p + (1 - p) \log_2 (1 - p))$ bits, where p is the probability that the referee's response is true. A single response by an imperfect referee does not permit the player to exclude any class of members from the set of possibilities, but does reduce the entropy of the set of possibilities by distributing the probability of being a right answer unequally over its members.

In organic evolution, the selector is an imperfect referee; otherwise every genotype in competition with a superior genotype would have a selection coefficient of 1.

Competitive Guessing Games.

A guessing game may have several players in competition with each other. If each player's aim is merely to be the first to win the game, analysis of the situation involves only an elucidation of the optimum strategy, and therefore adds nothing to what has been said above about non-competitive games. The situation is of interest if, from time to time during the course of the game, some of the players are eliminated before any of them has found a right answer.

The simplest case arises in trial-and-error games with selectors, in which the players to be eliminated are those with the best past guesses of lowest primary rank, i.e. those who are still farthest away from finding a right answer.

One must distinguish carefully between a system which can best be analogized to a competitive guessing game, and one which can best be analogized to a game with a single player who is permitted to make several simultaneous guesses. The distinguishing feature is that competitive players do not receive information in response to each other's guesses, but only in response to their own. A player in a competitive game does not know whether his competitor's best last guess is better than his own until he finds himself eliminated from the game. A single player making multiple guesses, however, may submit as his next group of guesses some sort of mixture or combination of what he believes to be the best features of all the guesses passed by the referee at the last trial.

Both kinds of situation arise when organic evolution is considered as a guessing game. A single Mendelian population can be represented by a single player, making many guesses simultaneously. If the referee's response to these guesses informs him that organisms carrying an allele a" are fitter than those carrying an allele a', he may include a" and exclude a' from all, or most of his future guesses. A competing population, reproductively isolated from this one, represents a competitive player who must discover the superiority of a" over a' for himself, quite independently of its discovery by the first player.

Some aspects of competition between players may lend themselves to analysis by the Theory of Games; the optimum strategy, which leads one to the right answer most rapidly, is not always the same as the strategy which maximizes the probability of remaining one or two jumps ahead of one's competitors. It must be emphasized, however, that the relation between a player and the referee is not a competitive one, and cannot be analyzed by Game Theory.

General.

The preceding discussion has been confined to those aspects of guessing games which are important in setting up a model of organic evolution. During the preliminary stages of the work, however, the writer necessarily examined other aspects of other kinds of guessing games. These suggest that the subject of guessing games in general may be of interest to mathematicians, and perhaps to scientists seeking mathematical models for a variety of physical, biological, psychological, and economic systems. The present thesis, however, is intended to deal with Darwinian systems, and the only kind of guessing game which needs to be considered is a Darwinian game. This section can therefore be adequately recapitulated by the following two definitions:

A Darwinian system is any system which can be usefully analogized to a Darwinian guessing game.

A Darwinian guessing game is a non-trivial trial-and-error game in which the referee is a selector, and in which the data stored by the player can consist only of a record of the guesses to which the referee assigned highest primary rank at the last trial; in particular, the player cannot keep a t-record.

To permit simple cases to be analyzed first, certain options will be retained in this definition: a Darwinian game may have either a perfect or an imperfect referee, and it may be a competitive or a non-competitive game.

Permutable and Non-permutable Hierarchies

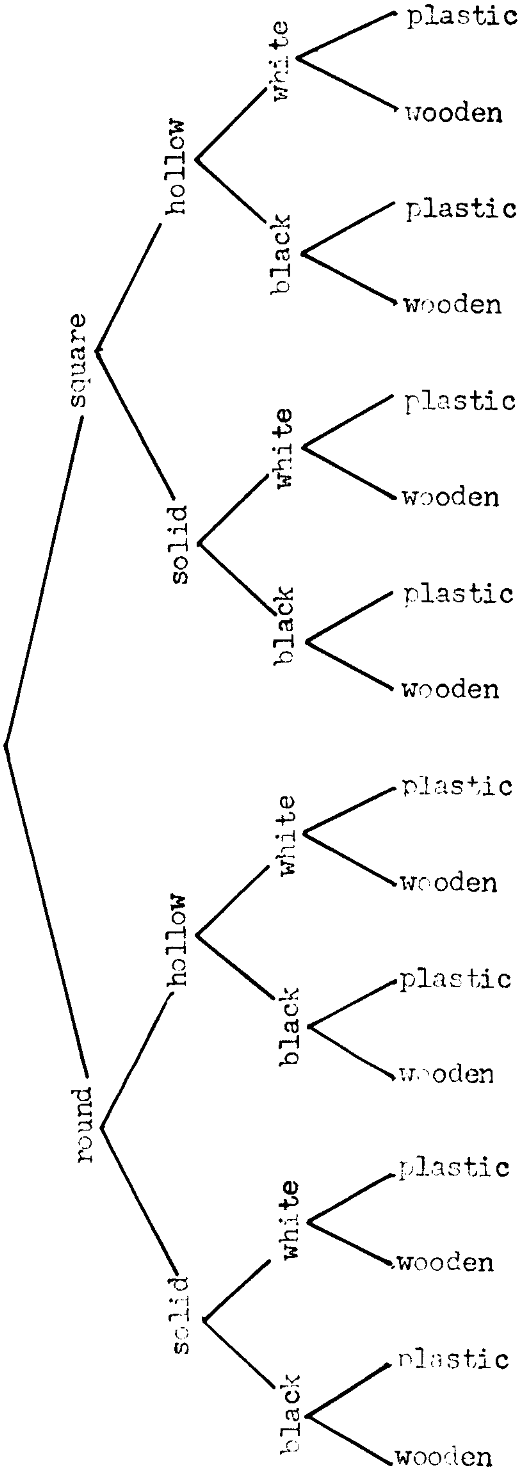
At each trial the player uses some criterion to divide a class into daughter classes. If the size of the daughter classes deviates greatly from equality, the information the player can receive when the referee compares them will be too small in quantity to permit him to win the game. He must therefore be capable of choosing a criterion of division which will result in an adequate approximation to equal sized daughter classes.

With some sets of criteria of division, this requirement can be met only if the criteria are used in a fixed order. Consider, for example, a dichotomous classification by size of the integers from 1 to 1024. If the criterion "larger or not larger than 4" is used at any level before the eighth, it serves to divide a class into two very unequal parts. Similarly, if the criterion "dog or non-dog" were used too early in classifying the set of living organisms, it would serve to divide a class into two very unequal parts, and to reduce the quantity of information obtainable from the referee's comparison of the two daughter classes. A hierarchy in which the quantity of information obtainable per trial is reduced by using criteria of division in any but one order may be called a non-permutable hierarchy.

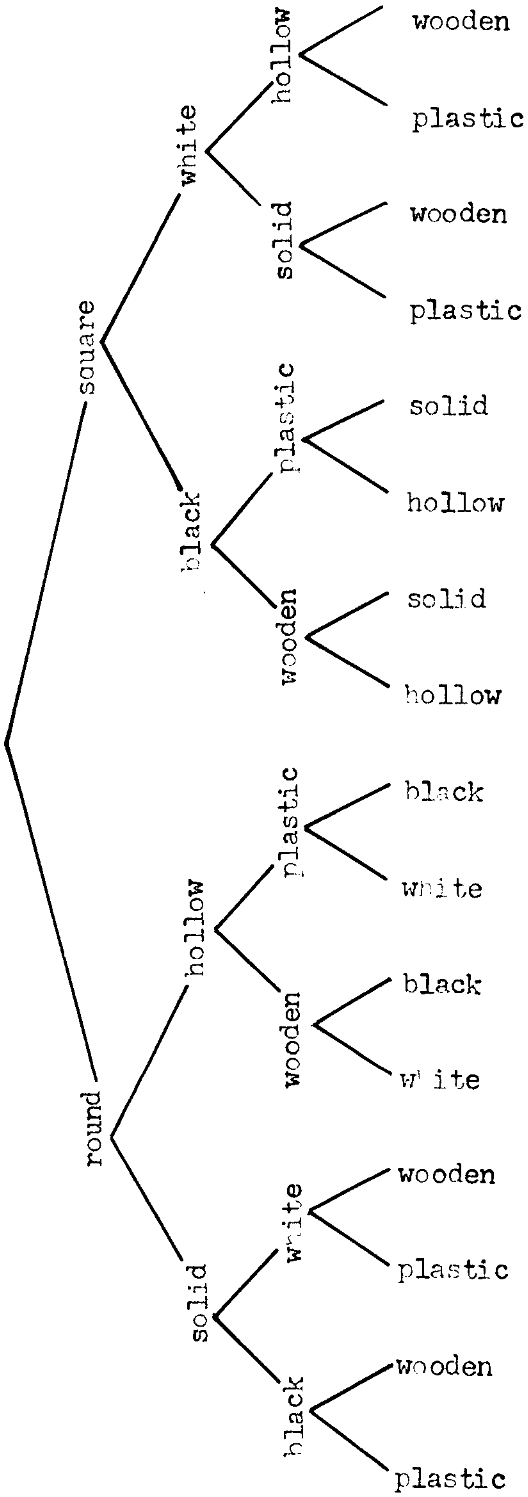
On the other hand, consider the set of sixteen objects, each of which is either round or square, solid or hollow, black or white, and wooden or plastic. Any of the criteria of division (colour, shape, etc.) will divide this set into two equal classes, and any of the other criteria will divide either class into two equal daughter classes. Either of the classifications shown in Figure 4, and many others as well, would be equally useful as a classification of the set in a guessing game. A hierarchy in which the quantity of information obtainable per trial is not affected to any important extent by using the criteria of division in different orders may be called a permutable hierarchy.

In a Darwinian game, the player can keep a record only of the best past guess. This is always a member of the class which should be divided next, but it is also a member of some class at every level of the hierarchy. As he has no t-record, therefore, the player cannot identify the class to be divided, nor infer what level in the hierarchy it is on. He must therefore be free to use criteria of division which could be usefully used on any level

Figure 4 Two of the many possible permutations of a permutable hierarchy.



square
hollow
white
plastic
wooden
black
plastic
wooden
solid
white
plastic
wooden
black
plastic
wooden
round
hollow
white
plastic
wooden
black
plastic
wooden
solid
white
plastic
wooden
black
plastic
wooden



square
white
hollow
wooden
plastic
solid
wooden
plastic
black
plastic
solid
hollow
wooden
solid
hollow
round
hollow
plastic
black
white
wooden
black
white
solid
white
wooden
plastic
black
wooden
plastic

in the hierarchy; as he has no record of rejected past guesses, however, he does not know what criteria of division have already been used, hence, he must be free to use any criterion of division on any level in the hierarchy. Therefore, a player can win a Darwinian game only if he can classify S, the set of possibilities, in a permutable hierarchy.

In a permutable hierarchy, if one follows a line to any member of the set one is confronted by the same series of criteria of division at the forks, although they do not necessarily occur in the same order on any two lines (see Figure 4). Therefore, each criterion occurs once in each line, and there are as many criteria as there are levels in the hierarchy. The criteria of division may be listed in some fixed, though arbitrary, order and an ordinal number 1, 2, ...i,...L, assigned to each. Each criterion gives rise to a small number, not exceeding the order r of the selector, of alternatives. Let the i'th criterion of division give rise to k_i alternatives. Then any member of the set of possibilities, S, can be represented by a sequence of L places in which the i'th place can be filled by any symbol from an alphabet of k_i symbols, and the alphabet available to fill the j'th place is not altered by the previous filling of any other place. For example, any of the set of sixteen objects described above, each of which must be round (R) or square (r), solid (S) or hollow (s), black (B) or white (b), and wooden (W) or plastic (w), can be represented by a sequence of four symbols, each of which is taken from an alphabet of two symbols. In the sequence RSBW, any symbols can be replaced by their alternatives and the sequence still represents a member of the set. If there were defects in the permutability of the classification -- if any criterion used at one level divided a class in a ratio different from that in which it would divide a class at another level -- this would not be true. Deviations from complete permutability would result in the absence of some members from the set of possible sequences. Given a sequence of symbols representing a member of a permutably classifiable set, one may freely change any symbol without necessarily changing any other, and the altered sequence continues to represent a member of the set.

In any Darwinian game, therefore, S can be represented by a set of $\prod^{L} k_i$ sequences of symbols, each L symbols long, with the i'th place filled by a symbol from an alphabet of k_i symbols.

The set of all known organisms, classified by the phenotypic criteria used by systematists, certainly does not fall into a permutable hierarchy. No one knows whether the set of all possible organisms could be permutably classified, and it is hard to imagine how the question could be decided. On the other hand the set of all possible genomes, each considered as a linear sequence of nucleotide pairs, would obviously be classifiable in a permutable hierarchy if all possible sequences were the same length (and if RNA viruses were ignored), as any nucleotide pair can presumably be changed without a necessary change in any other. Differences in the length of the nucleotide sequence from species to species is a complication which cannot be ignored indefinitely, but it can be ignored until simpler cases have been examined.

Notice that the model lends itself equally well to analysis of evolution using the "classical" concept of the genome as a series of independent loci, in which the i'th locus can be filled by any of k_i different alleles.

A fundamental characteristic of a Darwinian game is that the set of possibilities, S can be represented by the set of all possible sequences of L places in which the i'th place can be filled by any symbol from an alphabet of k_i symbols, and the alphabet available to fill each place is not altered by the previous filling of any other place. The assumption that the genetic material could be represented by such a linear sequence of symbols was made by the writer in several earlier models of natural selection, and justified by analogy with empirical genetics. Here, however, it has been shown to be a necessary condition for evolution by selection.

IV An Analysis of Simple Darwinian Games

The ranking system adopted by the selector, and the size of the alphabet available to fill each place in the sequences which represent the set of possibilities, have fundamentally important effects on the course of a Darwinian game. Nevertheless, many of the conclusions to be drawn from an analogy between organic evolution and a Darwinian game can be illustrated by games in which these two factors are kept as simple as possible. This section will therefore be devoted to discussion of variations on a game in which $\underline{S}$ consists of all possible sequences $\underline{L}$ places long with an alphabet of two symbols available to fill each place, in which there is only one right answer, and in which the referee nominates the symbol filling the $\underline{i}$'th place in the right answer as the correct $\underline{i}$'th symbol, and ranks the possible sequences by the number of correct symbols each contains.

In organic evolution, there are four symbols available to fill each place in the genetic material, if one uses the DNA molecule as a model, or a variable number of "alleles" if one treats the genetic material as a sequence of classical loci. There are many "right answers", and the environment ranks genotypes in a much more complex way. One will therefore expect the present model to lead to incorrect conclusions from time to time, but each such error should be explicitly attributable to one or more of the oversimplified assumptions stated above, and should be capable of correction by substituting more realistic though more complex assumptions.

At each trial the player must submit his best past guess (BPG) and an altered copy of it to the selector. There are many ways in which an altered copy could be prepared, but (unless otherwise stated) let us here assume the one most amenable to mathematical analysis. Let the player delete $\underline{m}$ symbols from randomly chosen places in the BPG, and replace them by $\underline{m}$ symbols chosen at random from the appropriate alphabets. Let us say he deletes $\underline{m}$ symbols forming the deleted set s_d, and in their places inserts $\underline{m}$ symbols forming the inserted set s_i. Sometimes, of course, the same symbol will occur in both s_d and s_i, so that in general the next guess will differ from the BPG in fewer than $\underline{m}$ places.

At a given stage in the game let us say that the BPG contains $W = wL$ incorrect symbols, and $R = (1 - w)L$ correct symbols.

The Simplest Possible Darwinian Game.

The simplest possible case is that in which a single player submits his BPG and one altered copy of it to a perfect selector at each trial.

The probability that a symbol in s_d is correct is $(1 - w)$. The probability that a symbol in s_i is correct is 1/2. A gain will occur if and only if s_i contains more correct symbols than s_d. The size of the gain, measured by the increase in $\underline{R}$, is the difference between the number of correct symbols in s_i and the number in s_d. For any value of $\underline{m}$, therefore, G_m, the expected gain per trial can be calculated from a table like Table 2, in which:

1. The probability $p(x)$ that s_i contains x correct symbols is given at the head of each column;

2. The probability $p(y)$ that s_d contains y correct symbols is given at the left of each row;

3. The difference $(x - y)$ is given in the body of the table, when it is positive;

4. $G_m = \sum p(x).\ p(y).(x - y)$ for all positive values of $(x - y)$.

Table 2 Method of calculating the exact value of the expected gain per trial, G_m, for a given value of m (number of symbols deleted and inserted) and w (frequency of incorrect symbols).

See p. 49.

Number of Correct Symbols, (y), in s_d		Number of Correct symbols, (x), in s_i							
		m	m-1	m-2	...	x	...	1	0
	Probability of Occurrence	$(1/2)^m$	$m(1/2)^m$	${}^mC_2(1/2)^m$	...	${}^mC_x(1/2)^m$	...	$m(1/2)^m$	$(1/2)^m$
m	$(1-w)^m$	0							
m-1	$mw(1-w)^{m-1}$	1	0						
m-2	${}^mC_2\left[w^2(1-w)^{m-2}\right]$	2	1	0					
⋮	⋮	⋮	⋮	⋮					
y	${}^mC_y\left[w^{m-y}(1-w)^y\right]$	m-y	m-y-1	m-y-2	...	x-y	...		
⋮	⋮	⋮	⋮	⋮		⋮			
1	$m\left[w^{m-1}(1-w)\right]$	m-1	m-2	m-3	...	x-1	...	0	
0	$m(w)^m$	m	m-1	m-2	...	x	...	1	0

More formally,

$$G_m = \sum_{x=1}^{m} \sum_{y<x} {}^{m}C_x \left(\tfrac{1}{2}\right)^m \cdot {}^{m}C_y \left[w^y (1-w)^{m-y} \right] \cdot (x-y) .$$

For example, from Tables 3, 4, 5 and 6, it can be computed that:

$G_1 = w/2;$

$G_2 = (w^2 + w)/2;$

$G_3 = (3w + 9w^2)/8;$

$G_4 = (w + 6w^2 + 2w^3 - w^4)/4;$

and so on.

If m is large, this form of computation becomes impractical. However, if m is large both x and y will be approximately normally distributed, and so will $(x - y)$. In a normal distribution with unit variance, let

$$P(u) = \frac{1}{\sqrt{2\pi}} \int_u^{\infty} e^{\frac{-v^2}{2}} . dv$$

i.e. $P(u)$ is the one-tailed normal probability integral, as given for example in Table III of Fisher and Yates (1957).

Let $P^{-1}(p)$ represent that value of u for which $P(u) = p$; i.e. $P^{-1}(p)$ is that normal deviate which is exceeded by a fraction p of the population. $P^{-1}(p)$ is closely related to the probit of p, but because of the way probits are conventionally tabulated $P^{-1}(p)$ can be more conveniently evaluated by finding p in the body of a table of the normal probability integral and reading the corresponding value of $P^{-1}(p)$ from the margins of the table. E.g. Fisher and Yates' Table III shows that $P^{-1}(0.30854)$

Tables 3, 4, 5, 6 Data for calculating G_1, G_2, G_3, G_4.

TABLE 3

			x	
			1	0
		probability	1/2	1/2
y	1	$(1-w)$	0	
	0	w	1	0

$$\boxed{G_1 = w/2}$$

TABLE 4

			x		
			2	1	0
		probability	1/4	1/2	1/4
	2	$(1-w)^2$	0		
y	1	$2w(1-w)$	1	0	
	0	w^2	2	1	0

$$\boxed{G_2 = (w+w^2)/2}$$

TABLE 5

		x			
		3	2	1	0
	probability	1/8	3/8	3/8	1/8
y 3	$(1-w)^3$	0			
2	$3(1-w)^2w$	1	0		
1	$3(1-w)w^2$	2	1	0	
0	w^3	3	2	1	0

$$\boxed{G_3 = (3w+9w^2)/8}$$

TABLE 6

		x				
		4	3	2	1	0
	probability	1/16	4/16	6/16	4/16	1/16
y 4	$(1-w)^4$	0				
3	$4(1-w)^3w$	1	0			
2	$6(1-w)^2w^2$	2	1	0		
1	$4(1-w)w^3$	3	2	1	0	
0	w^4	4	3	2	1	0

$$\boxed{G_4 = (w+6w^2+2w^3-w^4)/4}$$

= 0.5 and, conversely, P(0.5) = 0.30854.

If m is large, the number of correct symbols, x, in the inserted set has a mean m/2 and variance m/4; the number of correct symbols in the deleted set has a mean (1 - w)m and variance mw(1 - w); and their difference, (x - y) has a mean m(w - 1/2) and a standard deviation

$$\sigma = \pm\sqrt{m\left[\frac{1}{4} + w(1-w)\right]}$$

A gain occurs only when $(x - y) > o$, i.e. when (x - y) deviates from its mean by $u = \dfrac{m\left(\frac{1}{2} - w\right)}{\sqrt{m\left[\frac{1}{4} + w(1-w)\right]}}$ standard deviations. Therefore the frequency of a gain is given by

$$f = P(u) = P\left[\frac{m\left(\frac{1}{2} - w\right)}{\sqrt{m\left[\frac{1}{4} + w(1-w)\right]}}\right]$$

Those members of a population which exceed the mean of the population by more than u standard deviations, exceed it on the average by

$$P^{-1}\left[\frac{1}{2}\;P(u)\right] \qquad \text{standard deviations.}$$

Therefore, when m(x - y) exceeds zero, its mean size is

$$g = m\left(\frac{1}{2} - w\right) + \sigma\;P^{-1}\left[\frac{1}{2}\,P(u)\right] \qquad \text{where}$$

$$u = \frac{m\left(\frac{1}{2} - w\right)}{\sqrt{m\left[\frac{1}{4} + w(1-w)\right]}} \qquad \text{and} \qquad = \sqrt{m\left[\frac{1}{4} + w(1-w)\right]}$$

Hence $G_m = fg = P(u)\left\{m\left(\frac{1}{2}-w\right) + \sigma\, P^{-1}\left[\frac{1}{2}\, P(u)\right]\right\}$

and since $u = \dfrac{m\left(\frac{1}{2}-w\right)}{\sigma}$

$$G_m = P(u)\sigma\left\{P^{-1}\left[\frac{1}{2}\, P(u)\right] - u\right\}$$

Though cumbersome in appearance, this expression can fairly readily be evaluated by reference to a table of the one-tailed normal probability integral.

G_m is displayed graphically as a function of (1 - w) for various values of m in Figure 5.

Figure 5 can serve as a basis for several biologically suggestive conclusions.

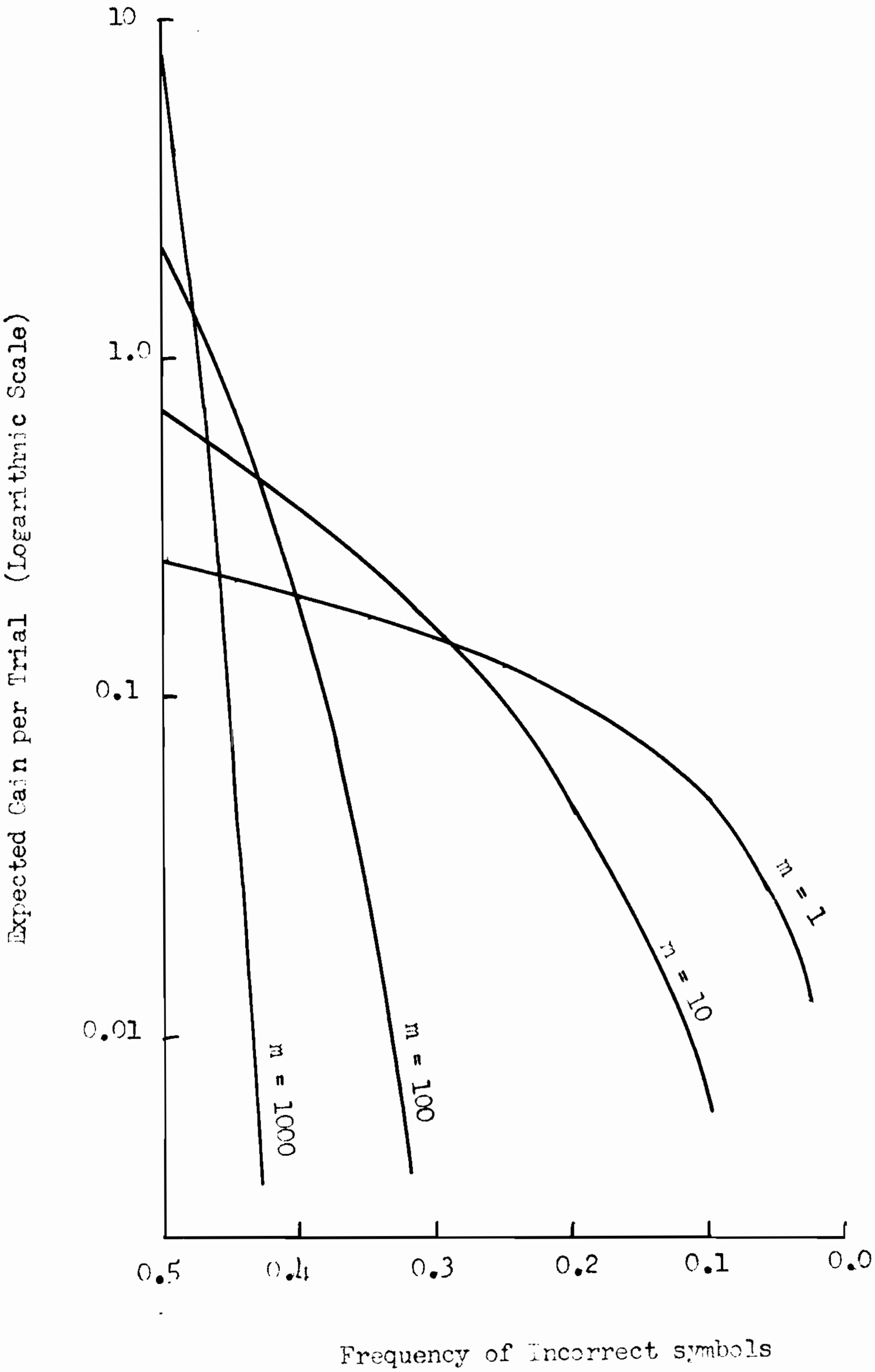
First, whether m remains constant throughout the course of the game or changes with time in such a way that G_m always has its maximum possible value, the rate of increase in R, the number of correct symbols, decreases as R increases. If m is constantly 1, for example,

$$\frac{dR}{dt} = G_1 = \frac{w}{2} = \frac{L - R}{2L} \quad .$$

Similarly, beginning with the relation $\frac{dW}{dt} = \frac{-W}{2L}$, if t is the number of past trials, one obtains by cross-multiplying, integrating both sides, taking antilogs, and settling the constant of integration = 1/2, since W/L = 1/2 when t - 0,

$$W = \frac{1}{2}\, e^{\frac{-t}{2L}} \quad .$$

Figure 5 The effect of changing m, the number of deleted and inserted symbols, on the relation between w, the frequency of incorrect symbols, and G_m, the expected gain per trial, in a simple Darwinian game.


Expected Gain per Trial (Logarithmic Scale)
10
1.0
0.1
0.01
m = 1000
m = 100
m = 10
m = 1
0.5
0.4
0.3
0.2
0.1
0.0
Frequency of Incorrect symbols

The number of incorrect symbols, W, therefore "decays" exponentially like atoms in a radioactive substance.

Therefore, although this particular guessing game is a greatly oversimplified model of organic evolution, it serves to suggest that evolution must have been a great deal faster, in some sense, in its earliest stages than it is now. Of course this is what one would expect, on the common sense basis that as organisms become more highly adapted, it becomes increasingly difficult for them to become still more highly adapted. Among the most primitive organisms, shortly after the first self-reproducing structures had somehow been spontaneously generated, even a large random change had a fairly high probability of being a biological improvement. Among horses, butterflies, and jellyfish, however, most random changes must render their victims less adapted, and large changes are almost always deleterious.

Simpson (1944) has attempted to explain the many gaps between large taxonomic groups, such as orders, in the fossil record by postulating that "...the basic differentiation of each order took a much shorter time than its later adjustment, spread, and diversification. In the magnitude of structural change involved, this basic part is comparable, on the average, to the subsequent changes... The change, for instance, from a carnivore or insectivore to an early cetacean is much more profound than the recorded change from early to late cetaceans... It follows that the basic differentiations must have proceeded, on the average, more rapidly than the later recorded evolution, almost surely twice as fast and probably more, quite possibily ten or fifteen times as rapidly in some cases." Concerning the wing of bats, Simpson states "...its recorded rate of evolution is effectively zero. The bat's wing has not essentially progressed since the Middle

Eocene... Extrapolation of this rate in an endeavour to estimate the time of origin from a normal mammalian manus might set that date before the origin of the earth."

Simpson attempts to explain this more rapid evolution in the early stages of newly arising orders by various assumptions about the structure of the populations involved. Possibly, however, at least part of the phenomenon can be explained as a result of the fact that almost any change in the wing of a modern bat is sure to be detrimental, but the imperfect wings of those Eocene insectivores which were just beginning to flutter from branch to branch must have been much more easily and rapidly improvable by random alterations.

A second conclusion to be drawn from Figure 5 is that the optimum strategy in a Darwinian game requires m to decrease with time. The player, however, has no record of the number of trials he has made, and therefore cannot adjust m to its optimum value (that which maximizes G_m) at each stage in the game. The only strategy available to a single player is to choose a value for m and retain it as a constant throughout the game. If the game is difficult, m must be small.

Now imagine that the game is being played competitively. There are several players, and from time to time one or more of them, those with the BPG's of lowest rank, are expelled from the game. If the players have chosen different values of m, and if eliminations take place early and often, the players who survive the eliminations will be ones who have chosen large values of m, and therefore have the highest expected rates of gain early in the game. These players, however, will be unable to go on and find the right answer within the required number of trials, as a large value of m will permit only very slow progress later in the game.

Organic evolution can be analogized to a competitive Darwinian game, in which the players are reproductively isolated populations of organisms in competition with each other. Very early in evolution, if one such population adopted the strategy of changing several places at once in the sequence of nucleotides making up its genome, it would improve more rapidly than competing populations and eliminate them if competition was severe. Later, however, the surviving population would be able to improve only very slowly, and could not give rise to the highly adapted organisms we see about us. In other words, the strategy leading to most rapid early improvement would lead to extremely slow later improvement, but one would expect the fast starters to eliminate the slow starters, so that the survivors of the earliest period of evolution would be ill-suited for rapid later evolution.

One can suggest three hypotheses as an escape from this problem:

a). Only one kind of primitive organism existed, which just by chance was a "slow starting" type capable of fairly rapid evolutionary progress continuing for a long time.

b). Several populations of early organisms existed, but they were isolated geographically or ecologically, or for other reasons did not come into competition with each other until the "slow starters" had had time to catch up and surpass the fast starters.

c). <u>m</u>, which is a sort of index of mutation size and mutation frequency combined, is to some extent a genetic variable itself, so that a population with large <u>m</u> which survived early competition would in turn give rise to several populations with different values of <u>m</u>, of which the one best suited for that stage of the evolutionary process would in turn survive, to give rise to further populations differing in the value of <u>m</u>.

The third of these hypotheses seems the most attractive.

Given a t-record, an intelligent player could compute the value of m to use for most rapid progress. The imaginary players in a competitive Darwinian game have no t-records, and cannot do this. The only source of information available for assistance in finding the best mutation size is therefore the selector. Selection must therefore go on not only between alternative guesses made by the same player -- i.e. between individuals within a population -- but between players as well, i.e. between whole populations.

A good strategy in a Darwinian game can therefore be discovered by a sort of second order Darwinian game, in which each guess is a player in the ordinary game. One cannot continue to expand indefinitely in this way, however. Each guess in the evolutionary game is an organism; billions of organisms exist simultaneously, each with an average duration of perhaps only a few hours, so that an enormous number of guesses can have been made in the few billion years the game has gone on. Each guess in the second order game is a population, of which perhaps only a few hundred million can exist simultaneously, and with an average duration of perhaps many years. The possible number of guesses which might have been made in a third order game, in which each trial involves a population of populations, is again much smaller. One must keep in mind that by choosing between two second or third order guesses, the selector does not supply any sort of super-information; at most it provides one bit of information. The discovery of a good value for m may be the result of selection between populations; the discovery of a good method of controlling the variation of m might, just possibly, be the result of selection between populations of populations; but the discovery of a genetic material whose own structure serves to control the parameters of its variability must have been mostly a matter of blind chance.

A third conclusion suggested by Figure 5 is that mutation size, in addition to mutation frequency, is an important parameter in assessing the effect of a mutagenic agent. In highly evolved organisms (those in which $(1 - w)$ has a high value) the probability that a mutation is deleterious increases rapidly with the size of the mutation. "Size" in this context of course means the size of the change in the genetic material itself; the number of nucleotide pairs disarranged, for example. The fact that mutations with large phenotypic effects are more likely to be deleterious than those with small phenotypic effects has long been accepted, but the classical concept of the gene does not allow one to distinguish mutations of different genotypic sizes, except in the case of chromosomal aberrations.

An important quantity to be calculated in a Darwinian game is the actual number of trials required to win the game. Whether or not each guess leads to an improvement is a matter of chance. One may therefore suspect that if several players operate at once, independently of each other (neither competing nor exchanging information) one of them may win in considerably fewer trials than the expected number required by an average player. To test this suspicion requires determination of the variance of $\underline{W}$ in a population of independent players.

Assuming that each player retains a constant value $\underline{m} = 1$, and that each player begins the game with a random sequence of $\underline{L}$ symbols, for each player:

Probability that the i'th symbol is initially incorrect = 1/2.

Probability that the i'th symbol, if initially incorrect, is not corrected at the first trial $= \frac{L - 1/2}{L}$.

Probability that the i'th symbol, if initially incorrect, remains incorrect after the t'th trial = $\left(\frac{L - 1/2}{L}\right)^{t}$.

The probability that the i'th symbol is incorrect after t trials is therefore $\frac{1}{2}\left(\frac{L - 1/2}{L}\right)^{t}$.

If W is the number of incorrect symbols in the BPG, the mean ($\overline{W}$) and standard deviation (σ_W) of W after t trials are

$$\overline{W} = \frac{L}{2}\left(\frac{L-\frac{1}{2}}{L}\right)^{t} \pm \sqrt{\frac{L}{2}\left(\frac{L-\frac{1}{2}}{L}\right)^{t} - \frac{1}{2}\left(\frac{L-\frac{1}{2}}{L}\right)^{2t}}$$

When t is large and W therefore small, the number of incorrect symbols approaches a Poisson distribution with mean and variance $\overline{W} = \sigma_W^2 = \frac{L}{2}\left(\frac{L-\frac{1}{2}}{L}\right)^{t}$.

Hence, if n independent players are operating, there exists an even chance that one will win when $t = t^{*}$ such that

$$e^{-\overline{W}} = \frac{1}{n} \quad ;$$

$\overline{W} = \log n$ (where "log" without subscript means natural logarithm).

$$\frac{L}{2}\left(\frac{L-\frac{1}{2}}{L}\right)^{t^{*}} = \log n$$

$$t^{*} = \frac{\log L - \log 2 - \log\log n}{\log L - \log\left(L - \frac{1}{2}\right)} .$$

Since $\frac{d \log L}{dL} = \frac{1}{L}$, if L is large $\log L - \log\left(L - \frac{1}{2}\right) \simeq \frac{1}{2L}$;

hence $t^{*} = 2L(\log L - \log 2 - \log\log n)$.

Only if log n is of the order of L (where n is the number of independent players) is t^{*} appreciably reduced below the number of trials required by a single player to win. But if log n is of the order of L, n is of the same order as e^L. The total number of members of $\underline{S}$, the set of possibilities one of which is the right answer, is 2^L. Hence, many players acting simultaneously but independently will discover the right answer in an appreciably shorter time than a single player only if the number of players is an appreciable fraction of the total number of possibilities among which the right answer lies. Biologically, this implies that a large population of organisms which are neither interbreeding, cooperating, nor competing with each other will not yield highly adapted forms much more rapidly than a single line of organisms.

An interesting expression arises if one repeats the above computation for the case of a player who always changes a single symbol, at every trial; i.e. who does not take the symbol to be inserted at random from the appropriate alphabet, thus failing to make the new guess differ from the BPG in half the trials, but specifically inserts the alternative to the deleted symbol. The mean and variance of $\underline{W}$ can be calculated in a manner parallel to the case above, yielding

$\overline{W} = \sigma_w^2 = \frac{L}{2}\left(\frac{L-1}{L}\right)^t$, if W is large; and

$$\frac{L}{2}\left(\frac{L-1}{L}\right)^{t^*} = \log n.$$

From the identity

$$\log\left(\frac{x+1}{x}\right) = 2\left[\frac{1}{2x+1} + \frac{1}{3(2x+1)^3} + \frac{1}{5(2x+1)^5} + \cdots\right],$$

it follows that $\log\left(\frac{L}{L-1}\right)$ is a trifle larger than $\frac{1}{L-\frac{1}{2}}$, if L is very large. Hence $t^* \simeq (L-\frac{1}{2})(\log L - \log 2 - \log\log n)$.

Notice that $\underline{L} = \log_2 N = H$, where $\underline{N}$ is the number of possibilities in the original set of possibilities, $\underline{S}$, and $\underline{H}$ is the entropy of $\underline{S}$ in bits. Therefore, very roughly,

$t^* = H \log H.$

It can easily be shown that if the player could retain complete records of all past trials, he could win this game with certainty in $\underline{H}$ trials. The deficiencies in his data storage facilities increase the time required to win by a factor of about $\log H$, and do not permit the player to be certain of winning in any number of trials.

Nevertheless, to achieve a 50% probability of winning this game by blind chance would require 2^{H-1} guesses. If $\underline{H}$ is large, the difference between $H \log H$ and 2^{H-1} is the difference between plausibility and silliness. In a mathematical model of a physically real Darwinian system, it is plausible to imagine that the "player" may make 230,000,000,000 trials, but ridiculous to imagine that he may make $2^{9,999,999,999}$.

A Game Permitting Many Simultaneous Trials.

$\underline{T}$, the total number of trials the player is permitted to make in a Darwinian game, may consist of two components, T_s, the number of <u>simultaneous</u> trials he may make, and T_c, the number of <u>consecutive</u> trials he may make. If the game is to be non-trivial, the product T_sT_c must of course be very much smaller than $\underline{N}$, the total number of members in the set of possibilities.

Simultaneous trials are in general less useful to the player than the same number of consecutive trials. If there are 10^{12} possibilities, for example, the player may easily be able to win in 10^6 consecutive trials, but cannot win if he must make 10^6 guesses all at once. In real Darwinian systems, however, T_c is limited by the amount of time available to the player, and T_s by the amount of available space, and the two are independent; reducing T_c does not compensatorily increase T_s, A species which fails to expand to fill its available habitat does not receive an increase in the number of generations available for its evolution as a reward. The player in a Darwinian game should therefore use a strategy which utilizes the available number of simultaneous trials as effectively as possible.

If the selector is of order greater than 2, it is clearly an advantage to the player to be able to make as many guesses per comparison as the order of the selector. In general, however, the order of the selector will be considerably smaller than T_s. It is therefore important to discuss the possibility of a procedure in which it is useful to make a number of guesses which exceeds the order of the selector at each comparison.

It has already been shown above that running T_s independent games simultaneously (each with a binary selector) only very slightly reduces the expected number of consecutive trials required to find the right answer. It may be stated without proof that running T_s/r independent games simultaneously, each with a selector of order r, has a similarly unimportant effect.

Let us therefore consider a game with a selector of order r. A perfect selector of order r will reject all but the best member of a group of r sequences simultaneously presented to it, or all but the best n/r members of a group of n sequences. The player must therefore submit r sequences to the selector at each comparison, and a comparison will be considered to comprise (r - 1) trials. m retains its former meaning as the number of symbols deleted and replaced in each altered copy of the BPG; f is the probability that an altered copy is higher in rank than the BPG; and g is the average number of correct symbols by which an improved sequence exceeds R, the number in the BPG.

Without attempting a detailed proof, which seems unnecessary and would be complicated by the need for exactitude where approximations will be given below, the writer may state the following conclusions:

If $f < \frac{1}{r-1}$ when m has that value which maximizes G_m for a binary selector, the optimum strategy is to retain the same value of m for the higher value of r. The expected rate of gain per comparison is then very nearly (r - 1) times as large as with a binary selector, and the expected rate of gain per trial is practically unchanged.

If $f > \frac{1}{r-1}$ when m has the value which maximizes G_m for a binary selector, the optimum strategy is to increase m until $f = \frac{1}{r-1}$. The expected rate of gain per comparison is then larger than with a binary selector, though by a factor of less than (r - 1), and the expected rate of gain per trial is decreased. The reason is simply that an increase in m decreases f by a larger factor than it increases g.

These conclusions rest on the fact that it is never an advantage to obtain more than one improved sequence per comparison, as only one can be retained. If f is so large that more than one improved sequence occurs at a comparison, the optimum strategy requires an increase in m, decreasing f to the value at which only one improvement occurs per comparison, but increasing the average magnitude of the improvements which occur. The complications required for exactitude arise from the fact that if f is in the vicinity of $\frac{1}{r-1}$, for a given value of m, more than one improvement will inevitably occur at some comparisons and none at others. To optimize the value of m requires finding that value of f which minimizes the joint effect of these two deviations from expectation. To do so is mathematically difficult, and would not alter the approximations given above enough to affect their usefulness.

An important problem arises if the player may submit $n > r$ sequences to the selector at each comparison. The selector, of order r, passes the best n/r sequences from the group. The player then has two possible courses:

a) Choose one of the passed sequences to be stored as the BPG, and discard the others. Since this choice must be made without guidance from the selector, it is effectively at random, and the result is exactly the same as if the player submitted only r sequences of which the selector passed only one.

b) Store all n/r passed guesses as BPGs. If n is already as large as the rules of the game permit (i.e. $n = T_s$), the player can then make only $(r - 1)$ copies of each BPG for submission to the selector.

The latter is obviously the only one of the two alternatives which can conceivably give the player an advantage from using more than r sequences per comparison.

As a simple case, assume that R, the number of correct symbols per BPG, is already so large that when m has its optimum value the probability of a gain larger than one correct symbol is negligible. Assume also that L is so large that a small increase in R does not appreciably change f. Assume too that the player changes m symbols in each altered copy of a BPG, i.e. does not merely delete m symbols and insert m chosen at random, but inserts m alternative symbols, so that the copy actually differs from the BPG at m places.

At a given comparison, the number of altered copies of BPGs submitted to the selector is $\frac{n(r-1)}{r}$. Among these, $\frac{fn(r-1)}{r}$ will be improvements, and will be stored as BPGs of rank R+1. A further improvement, i.e. a sequence of rank R+2, can occur only among copies of BPGs of rank R+1. At the next comparison, therefore, $\frac{fn(r-1)^2}{r}$ opportunities for improvement will occur, and $\frac{f^2n(r-1)^2}{r}$ BPGs of rank R+2 will be stored. By a similar argument, at the third comparison $\frac{f^3n(r-1)^3}{r}$ BPGs of rank R+3 will be stored.

The outcome is clear. If $f > \frac{1}{r-1}$, a situation will soon be reached in which an improvement is made and stored at every comparison, no matter what value n may have (as long as $n > r$). If $f < \frac{1}{r-1}$, a situation will soon be reached in which at every comparison only one BPG exists among copies of which opportunities for improvement can arise. This situation too will arise no matter what value n may have; a large value of n may delay it, but cannot prevent it. The remaining $\frac{n}{r} - 1$ BPGs are then useless, contributing nothing to the speed at which the game can be won. It would appear that the expected rate of gain per comparison can be only transiently increased by an increase in n/r, the number of BPGs stored (or, in biological terms, the size of the breeding population).

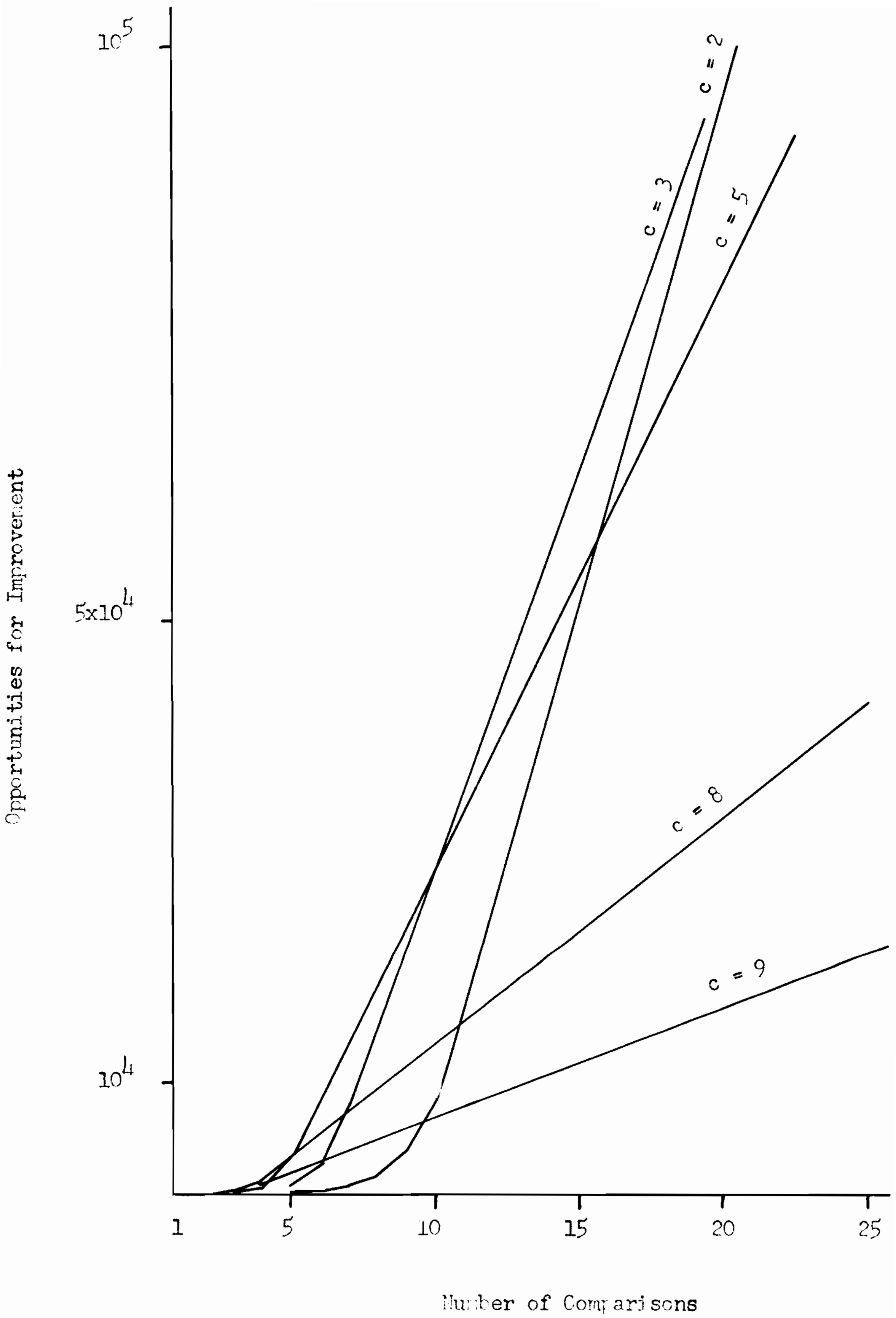
If $f < \frac{1}{r-1}$, however, there is a procedure by which the player can make good use of the possibility of submitting simultaneous guesses. It consists of submitting some true copies of each BPG to the selector at each comparison, along with a number of altered copies; let us say, c true copies and r-c altered copies. If one of the BPGs is superior to the others -- let us say it has a rank R + 1 -- the result of this stratagem is that after one comparison, there will exist c BPGs of rank R+1, c^2 after two comparisons, and so on, until all of the n/r BPGs have a rank of R+1, and are true copies of the first to achieve that rank, or until a sequence of rank R+2 arises among the altered copies of the BPGs of rank $R^{+}1$. In other words, an improved sequence will proliferate geometrically, at the expense of those which have not improved, until true copies of it constitute the entire "breeding population" or until a further improvement arises among its altered "descendants".

In the first comparison after a BPG of rank R+1 has been stored, (r-c) opportunities for improvement will occur; c(r-c) opportunities will occur at the next comparison, $c^2(r-c)$ at the next, and so on. This geometric increase in the number of opportunities for improvement can continue, however, only until the number of PBGs of rank $R^{+}1$ reaches n/r, after which opportunities for improvement occur at a constant rate cn per comparison. The faster the number of opportunities grows during the geometric phase, the more slowly it grows during the linear phase.

Figure 6 shows the way opportunities for improvement accumulate in the case in which $n = 10{,}000$, and $r = 10$, for various values of c. If f is rather large, so that an improvement will be expected to occur about once in 1,000 opportunities, the optimum value of c appears to be about 8; i.e. the "mutation rate", (r-c)/r, should be low. If f has certain intermediate values, so that an improvement will be expected once in about 10,000 opportunities, it appears that c should have an intermediate value of about 5. If f is very small, so that 100,000 or more opportunities must accumulate before an improvement is expected, c must be small, i.e. the mutation rate should be high. At intermediate values of f (in the neighbourhood of 1/10,000) choice of the wrong value of c may increase the number of comparisons required to find an improvement by a factor of about 2, but the same sort of deviation from the optimal strategy may increase the time requirements by a much larger factor if f is much smaller than 1/10,000.

The exact value of c, as a function of n, r, and f, which minimizes

Figure 6 The accumulation of opportunities for improvement in a Darwinian game in which 1,000 BPGs are stored at each comparison, and c true copies and (10 - c) altered copies of each are submitted to the selector at each comparison.


10^5
5x10^4
10^4
Opportunities for Improvement
c = 2
c = 3
c = 5
c = 8
c = 9
1
5
10
15
20
25
Number of Comparisons

the expected number of comparisons required to find an improvement, is difficult to compute. Nevertheless it is clear that if $f < \frac{1}{r-1}$, an improvement cannot be expected to occur until about $\frac{-\log f}{\log r}$ comparisons have taken place. The time required for an improvement to occur will be a minimum when _n_ is of the order of 1/f; longer times will be required if _n_ is smaller than this, but shorter times will not be expected if _n_ is greater than 1/f.

Let us now try to place a biological interpretation on these results. It must be emphasized that this game is a model of a hypothetical population of completely asexual organisms, in which all deaths and sterility are selective; no such thing as random death or accidental failure to reproduce occurs.

n/r, in this model, is the size of the reproductive population, each member of which produces _r_ offspring. Since the population is fixed at a maximum size of _n_, r-1 offspring per parent must die before the next act of reproduction. As all death is selective, therefore, r-1 serves as a measure of the severity of selection. With a perfect selector, the severity of selection increases directly with the reproductive rate of the population. $\frac{r-c}{r}$ is the mutation rate, the fraction of offspring differing from their parents.

Early in evolution, while the population was still so imperfectly adapted to its environment that a high probability existed of an improvement occurring among the offspring of a single parent, such a population would evolve fastest if its mutation rate were high: $\frac{r-1}{r}$. During this period, evolutionary progress would go on just about as rapidly with a breeding population of only one as it would in a much larger breeding population. This suggests the speculation that early evolutionary progress may have taken place as readily in splash pools by the seashore, in damp places under rocks, and so on, as in the great bulk of the ocean. Improvement may have been as fast in each of innumerable small populations as in a single large population, and, in the former, would be accompanied by tremendous diversification as well.

As the population became better adapted, the probability that a mutation would be an improvement would decrease until an advantageous mutation could no longer be regularly expected to occur among the offspring of a single parent. The rate of evolutionary improvement would then inevitably fall, but its decrease could be retarded and reduced by a reduction in the mutation rate, which would permit those advantageous mutations which did occur to proliferate geometrically. During this period it would become an advantage for the breeding population to increase, until the total number of offspring per generation was approximately of the order of the reciprocal of the probability that a single mutation would be advantageous. A larger breeding population would confer no additional advantage.

Eventually, as the adaptedness of the population continued to increase, the population would attain its maximum possible size, and the probability that a mutation would be advantageous would fall below the reciprocal of the population size. At this stage, it would be advantageous for the mutation rate to begin to rise again, although it should never return to its earliest high value of $\frac{r-1}{r}$.

Throughout the process, an increase in the number of offspring per parent can lead to a proportional increase in the rate of improvement, as long as the number of offspring per parent does not exceed the reciprocal of the probability of an advantageous mutation. At no stage in the process, however, does an increase in the size of the breeding population result in a proportional increase in the rate of increase of adaptedness.

It is hard to say how far these conclusions could be applied to a more realistic population model. They do make it clear, however, that reproductive rate, mutation rate, and mutation size must be independently adjustable if organisms are to achieve anything like the optimal evolutionary strategy made possible by large population sizes.

A Game with an Imperfect Selector.

To be a more realistic model of natural selection, a Darwinian game must incorporate an imperfect selector. An imperfect selector is one which does not always repeat the same response when called upon to repeat the same comparison, i.e. which indicates that sequence A is better than sequence B at one trial, but indicates that B is better than A at another. Clearly, if the selector is to supply any information at all to the player it must give the correct or true response more than half the time.

The simplest case is a game with a binary selector, called upon to compare the BPG with a single variant at each trial. Let the probability that the selector will correctly pass the better of the two sequences at each trial be p, and the probability that it will erroneously pass the worse of the two sequences (which will then be stored as the BPG) be $(1-p) = q$. Let the player make his guesses differ from the BPG at a single place. Then the probability is $w/2$ that a variant is better than the BPG, and $(1-w)/2$ that it is not as good. Therefore:

The probability that a variant is not as good as the BPG and is passed by the selector $= pw/2$;

The probability that a variant is not as good as the BPG, but is nevertheless passed by the selector $= q(1-w)/2$.

The BPG will be expected to improve until w reaches that value at which guesses worse than the BPG are being retained as frequently as guesses better than the BPG, i.e. until

$$pw = (1-p)(1-w),$$

$$\frac{p}{1-p} = \frac{1-w}{w},$$

i.e., until $R = pL$. The rank of the BPG will rise to a value of pL, and will then fluctuate about that value, but cannot rise consistently above it.

The behaviour of an imperfect selector of higher than binary order, or in a game in which several BPGs are stored per comparison, is more complicated. It seems to be most easily approached through a modification of Fisher's (1930) "fundamental theorem of natural selection": the rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time.

The <u>fitness</u>, <u>F</u>, of a type of organism is the ratio of its frequency in the next generation to its frequency in the present generation. In a Darwinian game, it may be defined as the average number of sequences of a given kind stored as BPGs out of <u>r</u> identical copies submitted to the selector. The fitness of a sequence may therefore range from 0 to <u>r</u>, and the mean fitness of all sequences submitted to the selector at a given comparison is $\bar{F} = 1$.

One may demonstrate Fisher's theorem thus (following Li 1955):

Consider a set of guesses of fitness $F_1, F_2, \ldots F_i \ldots$, with frequencies $f_1, f_2, \ldots f_i \ldots$, where

$$\sum f_i F_i = \bar{F} = 1.$$

The variance in fitness $= V = \sum f_i(F_i - 1)^2 = \sum f_i F_i^2 - 1.$

After selection, f'_i, the new frequency of BPGs of fitness F_i, is given by

$$f'_i = F_i f_i.$$

Hence the new mean fitness = $\sum f_i' F_i = \sum f_i F_i^2$,

and the increase in mean fitness ; $\Delta \overline{F} = \sum f_i F_i^2 - 1 = V$.
Hence, the increase in mean fitness is equal to the variance in fitness.

For the purposes of this thesis, let us extend Fisher's theorem by defining the superiority, Z_i, of the i'th kind of sequence, as

$$Z_i = F_i - 1.$$

The superiority, Z_i, of the i'th kind of sequence, is therefore the number of additional BPGs of the i'th kind expected to be stored, over and above the one previously stored, for each r sequences of the i'th kind presented to the selector. The mean superiority, $\overline{Z}$, of all guesses presented to the selector is of course $\overline{Z} = 0$, and Z_i cannot be less than -1 nor greater than r-1.

If guesses of superiority Z_i have frequency f_i, BPGs of superiority Z_i will be stored with frequency $f_i(1 + Z_i)$. It follows that the changes due to selection are:

$$\overline{Z} = \sum f_i(1 + Z_i)Z_i = \sum f_i Z_i + \sum f_i Z_i^2$$

$$\Delta \overline{Z} = \sum f_i Z_i^2$$

$$\Delta \sum f_i Z_i^2 = \sum f_i Z_i^3$$

$$\Delta \sum f_i Z_i^x = \sum f_i Z_i^{x+1}$$

If the frequency distribution of guesses with various superiorities presented to the selector is given, these relations allow the frequency distribution of the BPGs which will be stored to be computed in detail. Notice that the selector always increases the average superiority, but may increase the mean squared superiority, decrease it, or leave it unchanged, depending on whether or not the frequency distribution is skewed. If a series of comparisons is made without mutation, so that every guess presented to the selector is a true copy of a BPG, the improvement, $\Delta\overline{Z}$, made at each comparison may be larger than, smaller than, or the same as the improvement made at the previous comparison.

If $\overline{Z} = 0$, $\overline{Z^2}$, $\overline{Z^3}$, etc., are the second, third, etc. moments of the frequency distribution. If $\overline{Z} \neq 0$, this simple relation is no longer true. Nevertheless, $\overline{Z^2}$, $\overline{Z^3}$, etc. are still simple properties of the distribution, and in fact must be calculated before the moments can be found.

In a Darwinian game, the superiority of a sequence must be an increasing function of its rank. Let us say that if a sequence of rank R has a superiority Z_i, a sequence of rank R+1 has a superiority of $Z_i + a_i$, and a sequence of rank R-1 has a superiority $Z_i - b_i$. Notice that if $a_i + b_i = r$, the selector is perfect. An imperfect selector therefore is one in which a_i and b_i have small values.

Consider a set of BPGs in which $\overline{Z} = z$, where z is a small positive quantity to be defined below. Take those BPGs, forming a fraction f_i of the entire set, in which $Z = Z_i$. Among the copies of these submitted to the selector:

a fraction $\frac{r-c}{r}$ w are of superiority $Z_i + a_i$;

a fraction $\frac{r-c}{r}$ (1-w) are of superiority $Z_i - b_i$;

a fraction $\frac{c}{r}$ are of superiority Z_i, where $$\sum f_i \left\{ Z_i + \frac{r-c}{r} \left[wa_i - (1-w)\, b_i \right] \right\} = 0,$$

i.e. $$\frac{r-c}{r} \sum f_i \left[w(a_i + b_i) - b_i \right] = -z.$$

$\underline{z}$ is therefore the quantity by which mutation reduces the mean superiority of the guesses below the mean superiority of the BPGs from which they were derived. Notice that as $\underline{w}$ becomes smaller, i.e. as the mean rank of the BPGs increases, $\underline{z}$ assumes progressively larger values, if other things remain equal. As a population becomes more highly adapted, mutation occurring at a constant rate exerts an increasingly severe effect on its mean superiority. Among the BPGs, $\overline{Z^2} = \sum f_i Z_i^2$.

Among the guesses submitted to the selector,

$$\overline{Z^2} = \frac{c}{r} \sum f_i Z_i^2 + \frac{r-c}{r} \sum f_i \left[w(Z_i + a_i)^2 + (1-w)\,(Z_i - b_i)^2 \right]$$

$$= \frac{c}{r} \sum f_i Z_i^2 + \frac{r-c}{r} \sum f_i \left\{ Z_i^2 + 2Z_i \left[w(a_i + b_i) - b_i \right] + w(a_i^2 - b_i^2) + b_i^2 \right\}$$

$$= \sum f_i Z_i^2 + \frac{r-c}{r} \sum f_i \left\{ 2Z_i \left[w(a_i + b_i) - b_i \right] + w(a_i^2 - b_i^2) + b_i^2 \right\}.$$

Hence mutation changes $\overline{Z^2}$ by

$$\frac{r-c}{r} \sum f_i \left\{ 2Z_i \left[w(a_i + b_i) - b_i \right] + w(a_i^2 - b_i^2) + b_i^2 \right\}.$$

Now, a_i must in general nearly equal b_i, since

$$a_i = b_i + 1$$

Also, since $\sum f_i Z_i = z$, the mean value of $2Z_i$ must tend to increase as $\underline{w}$ decreases.

Hence, by inspection, the quantity by which mutation changes $\overline{Z^2}$ will tend to assume decreasing negative values as <u>w</u> becomes smaller.

Among the BPGs, $\overline{Z^3} = \sum f_i Z_i^3$,

and by a similar procedure, it can be shown that mutation tends to change $\overline{Z^3}$ by the quantity

$$\frac{r-c}{r} \sum f_i \left\{ 3Z_i^2 \left[w(a_i + b_i) - b_i \right] + 3Z_i \left[w(a_i^2 - b_i^2) + b_i^2 \right] + w(a_i^3 + b_i^3) - b_i^3 \right\}$$

This quantity, too, will tend to attain progressively more extreme negative values as <u>w</u> becomes smaller.

As <u>w</u> decreases in size, therefore, mutation has the following effects on the distribution of superiority among the guesses submitted to selection:

a) It tends to produce progressively larger decrements in mean superiority, thus increasing the increment in superiority which must be produced by selection if selection is to recoup the losses due to mutation.

b) it tends to reduce $\overline{Z^2}$, the quantity by which selection can increase the mean superiority. Unless $\overline{Z^2}$ exceeds <u>z</u> (the decrement in mean fitness produced by mutation), the selector cannot pass a set of sequences in which the mean superiority exceeds that of the set of BPGs from which the guesses were derived.

c) It tends to skew the distribution toward the left. If the distribution becomes skewed to the left (i.e. if it initially had no skew to the right), then selection as well as mutation will tend to reduce $\overline{Z^2}$, still further reducing the ability of the selector to restore the mean superiority lost by mutation.

The analysis of these relations has not yet been completed. Nevertheless, enough has been stated here to justify the tentative hypothesis that as <u>w</u> becomes small, the distribution of superiority among the BPGs may reach a condition in which $\overline{Z^2}$ is small, $\overline{Z^3}$ is negative, etc., and the selector is unable to add more to the mean superiority than mutation subtracts from it at each comparison. Improvement, the increase in the mean rank of the BPGs, could not continue once such a situation was established.

The Value of Recombination.

The difficulties facing the players in the games described above all stem from the fact that in preparing an altered copy of a BPG, the deleted symbols are drawn at random from a population in which the frequency of correct symbols is (1-w), but the inserted symbols which replace them are drawn from a population in which the frequency of correct symbols is only 1/2. Consequently, any change in m which increases the sizes of those gains which occur, simultaneously and usually disproportionately reduces the probability of any gain at all. If the selector is imperfect, any system of mutation which increases the dispersion of superiority among the guesses necessarily also causes a decrease in mean superiority. As a result there are severe restrictions on the rate at which improvement can go on, and if the selector is imperfect there are limits beyond which improvement cannot be carried.

There is one, and only one, escape from these restrictions. Somehow the player must be permitted to draw the inserted symbols from a population of symbols in which the frequency of correct symbols is (1-w). Only one such population of symbols exists; the other BPGS, in a game in which the number of guesses submitted to the selector per comparison exceeds the order of the selector.

Any system of guessing in which each guess submitted to the selector consists of a combination of parts from two or more BPGs may be called recombination.

Consider two BPGs L symbols long containing R_1 and R_2 correct symbols, respectively, and W_1 and W_2 incorrect symbols, where $(R_1 + R_2)/2 = \bar{R}$, and $(W_1 + W_2)/2 = \bar{W}$. Recombination between the two is carried out by combining L/2 symbols from each to form the recombinant. The expected number of correct symbols in the recombinant is $\bar{R}$ and the expected number of incorrect symbols is $\bar{W}$. Among a number of recombinants formed in this way between the same two parent BPGs, the ranks of the recombinants will be symmetrically distributed around their mean $\bar{R}$ in such a way that, if D is the deviation of a recombinant from the mean,

$$\frac{D^2}{\bar{R}} + \frac{D^2}{\bar{W}} = \chi_1^2 ,$$

from which it follows by simple substitutions that

$$D^2 = \bar{w}\,(1 - \bar{w})L\ \chi_1^2 \quad .$$

Since the mean value of χ_1^2 is one, it follows that the ranks of the recombinants are symmetrically distributed around the mean rank of the two parent BPGs with variance $\bar{w}(1 - \bar{w})L$.

If the BPGs have a mean rank $\hat{R}$, and a variance in rank V, and recombine in random pairs, the ranks of the recombinants will be distributed around values of $\bar{R}$ having a mean $\hat{R}$ and variance 2V/4, or V/2. If V is very large relative to $\bar{w}(1 - \bar{w})L$, recombination alone, without mutation or selection, will tend to reduce V by about half at each repetition, until V approaches $\bar{w}(1 - \bar{w})L$. On the other hand if V is smaller than $\bar{w}(1 - \bar{w})L$, recombination alone will raise its value to $\bar{w}(1 - \bar{w})L$ in a single step. Recombination therefore will tend to produce a set of guesses with a variance in rank $\bar{w}(1 - \bar{w})L$, with the same mean rank as the BPGs from which they were derived, and with ranks distributed symmetrically about the mean, at every comparison. The distribution of fitness or superiority among the guesses will depend on the details of the relation between rank and superiority, but one would generally expect the mean superiority of the guesses to be the same as the mean of the BPGs.

The advantage of recombination over mutation as a mechanism for deriving guesses from BPGs is therefore that it produces guesses with a rather large variance in rank but with the same mean rank as the BPGs. A fairly large gain can therefore be made at every comparison. Mutation, on the other hand, either yields guesses with a very small variance in rank, or guesses with a mean rank far below the mean rank of the BPGs. In either case a gain can be expected infrequently, if <u>w</u> is small, and large gains are extremely rare.

The disadvantages of recombination as a mechanism for deriving guesses from BPGs are serious, however. Obviously recombination cannot produce new symbols. If none of the set of sequences with which the player started the game contained the correct i'th symbol, no recombinant can ever contain it and the game cannot be won.

Secondly, imagine that each sequence consists of several segments, which are interchanged between BPGs in recombination. Then if

sequence 1: $A_1B_1C_1D_1\ldots$

and sequence 2: $A_2B_2C_2D_2\ldots$

recombine to produce four recombinants, of which the selector passes two, half the kinds of segments present in the BPGs must be absent from the passed guesses which form the next set of BPGs, if the selector is perfect. Half the remaining kinds must be lost at the next comparison, and so on. An imperfect selector will tend to produce the same result, though perhaps more slowly.

$\overline{Z^2}$, and hence the rate of improvement, increases with the variance in rank of the guesses presented to the selector. The variance $\bar{w}(1 - \bar{w})L$ calculated above rests on the assumption that the recombining BPGs are unrelated, i.e. that the probability of one parental BPG containing the correct i'th symbol depends solely on w, and is independent of whether the other parental BPG contains the correct i'th symbol. Two BPGs are related, however, if they have one or more "ancestral" BPGs in common, and there will therefore exist a "correlation between recombining BPGs" exactly analogous to Wright's (1922) "correlation between uniting gametes". Such correlations must come to exist unless the number of BPGs with which the player started the game was of the order of N, the total number of possibilities, and we have excluded this case by assuming that the game is non-trivial. To discuss this matter further here would merely be a repetition in unfamiliar terminology of the theory of inbreeding.

If the selector is imperfect, the superiority of a sequence, Z_i, is not in fact the absolute increase in frequency per comparison of the i'th kind of sequence but the statistical expectation of its increase. In a finite population of BPGs, actual quantities will vary around their expectations. If guesses are derived from BPGs by recombination an inevitable result is that some segments of sequences will be lost in spite of being better than the segments which are retained. The frequency and magnitude of such contradictions of expectation will increase as the number of BPG decreases. This is the phenomenon of random fixation or "drift" (wright 1931).

In order to procure those correct symbols which were absent from his initial set of guesses, and to replace those which are lost by inbreeding or drift, the player must utilize both mutation and recombination to derive guesses from his BPGs.

The Relation of Darwinian Games to Population Genetics.

This analysis of Darwinian games appears to have led into population genetics through the back door. Population genetics has of course already demonstrated that the major features of Mendelian genetic systems are sufficient conditions for evolution by natural selection. This examination of Darwinian games indicates that many of them are also necessary conditions. If living organisms exist on Mars, for example, (an idea which is made very plausible by Salisbury, 1962) one may be confident that the Martian organisms will have genetic systems analogous to those of earthly organisms in many ways, even though they may have a radically different biochemical basis.

The following conditions must be met if a difficult Darwinian game is to be won:

First, the members of the set of possibilities must be capable of being represented by linear sequences of symbols, in which the alphabet available to fill the i'th place is not changed by the previous filling of other places. Except for mechanical reasons, however, it does not matter whether the possibilities are sequences of nucleotides in nucleic acids, amino acids in polypeptides, magnetic impressions on spools of tape, or letters on paper.

Second, these sequences must duplicate themselves with mutations. At least in the later stages of the process, each mutation must be small, changing only one or a few symbols. It does not matter whether the production of such changes is left to random external causes, or whether they are "deliberately" produced during the copying process; in either case, the lack of adequate records forbids a systematic series of changes, and the mutations will appear to be random and undirected.

Third, if the selector is imperfect, the game cannot proceed very far unless a fairly large population of such sequences exists, and each produces true copies of itself as well as mutated copies. As a result, advantageous mutations will tend to proliferate exponentially at the expense of the parental sequences.

Finally, progress with an imperfect selector also appears to require a process of recombination between sequences, so that each sequence has two (or more) parents. This inevitably lead to analogues of such familiar genetic phenomena as inbreeding, drift, and so on.

The study of guessing games adds a dimension to population genetics, however, by indicating the ways in which the evolutionary process will change over long periods of time. The optimum population size for progressive evolution, the severity of the deleterious effects of mutation, the frequency of advantageous mutations, and even the necessity of recombination are all functions of the length of the evolutionary history of a population, and this is a quantity of which population genetics takes no account.

V Some Additional Complexities in Darwinian Games

Discussion of the Darwinian games chosen as examples in the previous section sheds no light on some problems of evolution, and actually suggests false conclusions about others. For example, it suggests that recombination should by preference take place between the most distantly related parents available, although we know that in organic evolution selection has tended to erect barriers preventing recombination beyond the limits of the species. This deficiency of the model is obviously due to its explicit assumption that there is only one "right answer" toward which the evolutionary player is striving. It and other defects could be corrected by introducing a number of further complications into the model. Most of these would fall under one of the following categories:

1. Larger alphabets and unequal alphabets.
2. The existence of more than one right answer.
3. Misleading ranking systems, involving
    a. Interaction between symbols,
    b. Unequal weighting of different places in the sequence.
4. Indirect selection.
5. Secular changes in the selector.

Larger Alphabets and Unequal Alphabets.

Darwinian games in which more than two symbols are available to fill each place in the sequence should be studied. Of special interest would be a game with four symbols available for each place, corresponding to the four nucleotide pairs in the Watson-Crick model of DNA, and a game with different numbers of symbols available for each place, corresponding to the different numbers of alleles available to fill different loci in classical genetics.

The subject is a large one, however, and will be omitted from this thesis.

A ten-symbol alphabet allows a particularly simple model of a Darwinian game to be construdted, which will be used in some examples below. The initial guess is a ten-digit sequence from Fisher and Yates' (1957) table of random numbers, for example 0 3 4 7 4 3 7 3 8 6. "Mutations" can also be read from the table, 9, 7 being interpreted as "insert 7 in the ninth place" and 0, 7 as "insert 7 in the tenth place". A ranking system must be set up to decide whether the "parent" or "offspring" sequence is to be retained as the basis for the next trial.

For example, the following game simply ranks sequences in order of size. Underlined symbols are "advantageous mutations".

$\underline{t}$	
0	0 3 4 7 4 3 7 3 8 6
3	0 $\underline{4}$ 4 7 4 3 7 3 8 6
4	0 4 4 7 4 $\underline{7}$ 7 3 8 6
6	$\underline{6}$ 4 4 7 4 7 7 3 8 6
	...
179	9 9 9 9 9 8 9 $\underline{8}$ 9 8
215	9 9 9 9 9 $\underline{9}$ 9 8 9 8
300	9 9 9 9 9 9 9 8 9 $\underline{9}$
572	9 9 9 9 9 9 9 $\underline{9}$ 9 9

This example strikingly illustrates how the frequency of favourable mutations decreases in the later stages of the game; the last two improvements in this sequence required more time than all previous improvements combined. Nevertheless, the right answer was found in 572 trials; one would expect to need 5 x 10^9 trials to find it by blind chance.

Closely related to unequal alphabets is the complexity introduced by unequal probability of mutation at different places in the sequence, and unequal probabilities of substitutions, i.e cases in which one symbol is more likely than another to be inserted in a given place. Models in the style of the game recorded above could be made to include such complications, but a more direct analytically mathematical approach should not be difficult, and would be more informative.

More than One Right Answer.

The process of selection becomes rather more complex if more than one right answer exists. Imagine that the right answers in the game above, for example, were

0 0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1 1
8 8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9 9 .

The Best Past Guesses would soon become divisible into a "high" group and a "low" group, between which recombination would be undesirable. Recombination between the "8" group and the "9" group would be useful for a longer time, until sequences like

9 8 8 9 9 9 7 8 9 6

began to appear.

A more interesting situation arises when the selector does not nominate a definite set of right answers, but establishes certain relations among sequences which, if met, make all of them right answers. Let the selector call two ten-digit sequences right answers if the sum of the figures filling the same place in each is ten. There are 8,888,888,889 possible pairs meeting the criterion, but as there are 10^{20} possible pairs altogether the chance of finding one by pure luck is remote. A record of such a game is given below:

t										
0	0	3	4	7	4	3^{*}	7^{*}	3	8	6
	3	6	9	6	4	7	3	6	6	1
1	0	3	4	7	4	3^{*}	7^{*}	3	$\underline{7}$	6
	3	6	9	6	4	7	3	6	6	1
3	0	$\underline{4}^{*}$	4	7	4	3^{*}	7^{*}	3	7	6
	3	6	9	6	4	7	3	6	6	1
6	0	4^{*}	4	7	4	3^{*}	7^{*}	3	7	6
	$\underline{6}$	6	9	6	4	7	3	6	6	1
11	$\underline{2}$	4^{*}	4	7	4	3^{*}	7^{*}	3	7	6
	6	6	9	6	4	7	3	6	6	1
12	2	4^{*}	4	7	4^{*}	3^{*}	7^{*}	3	7	6
	6	6	9	6	$\underline{6}$	7	3	6	6	1
24	2	4^{*}	4	7	4^{*}	3^{*}	7^{*}	3	7	6
	6	6	9	6	6	7	3	6	$\underline{4}$	1
27	2	4^{*}	4	$\underline{2}$	4^{*}	3^{*}	7^{*}	3	7	6
	6	6	9	6	6	7	3	6	4	1
28	2	4^{*}	4	2	4^{*}	3^{*}	7^{*}	3	7	6
	$\underline{7}$	6	9	6	6	7	3	6	4	1
45	2	4^{*}	4	$\underline{4}^{*}$	4^{*}	3^{*}	7^{*}	3	7	6
	7	6	9	6	6	7	3	6	4	1
46	2^{*}	4^{*}	4	4^{*}	4^{*}	3^{*}	7^{*}	3	7	6
	$\underline{8}$	6	9	6	6	7	3	6	4	1
62	2^{*}	4^{*}	4^{*}	4^{*}	4^{*}	3^{*}	7^{*}	3	7	6
	8	6	$\underline{6}$	6	6	7	3	6	4	1

etc.

The asterisk (*) indicates pairs of symbols with sums equal to 10.

Apparently it would take no longer to find two "mutually adapted" sequences than to find a single right answer. In the former case, however, nobody -- not even the selector, if the selector is a person -- can predict the nature of the products which will be produced by selection.

Organic evolution involves a combination of these two kinds of "right answers". Certain requirements must be met by any organism which is to live on land, and can be met in rather few ways: these constitute a sort of "fixed" part of the criteria of a right answer. However, living on land in the presence of grass and wolves is a different matter from living on land in the presence of leopards and banana trees. All organisms must be adapted to their inorganic environments, and organisms living together must be mutually adapted to each other as well. The development of photosynthesis, the colonization of fresh water and land, and adaptation to deserts and cold climates have been major developments in evolution, but so have the rise of grass-eating cursorial herbivores, the mutual adaptations of hosts and parasites, and the co-ordinated evolution of insects and flowering plants. The model suggests that the latter kind of evolution would go on just as rapidly as the former, but that its course is much less predictable.

Each "right answer" is an organism adapted to a particular ecological niche. Elucidation of the number of right answers which may exist, and the relationships between them, therefore lies within the field of modern mathematical ecology, which is concerned with the diversity of niches, the trophic structure of communities, and so on. Given a satisfactory analysis of these problems, the study of Darwinian games with multiple right answers might be of considerable aid in understanding the phenomena of adaptive radiation and speciation.

Misleading Ranking Systems.

By keeping the context of the i'th place (i.e. the symbols filling all other places) constant, and systematically trying each of the k_i symbols available in the i'th place, one may order the i'th alphabet, ranking its symbols from worst to best. The ranking system of the selector is then non-misleading if and only if the order of each alphabet is unchanged by a change in context. For example, if 9 9 9 is the right answer, 9 is clearly better than 7 in the second place in the context 9...9. If the sequence 7 7 7 is better than the sequence 7 9 7, the ranking system is misleading since in a changed context the order of the symbols available for the second place is changed. A misleading ranking system always tends to retard progress in a Darwinian game, since the player may discard the correct i'th symbol midway through the game, and then have to rediscover it later. With a non-misleading ranking system (and a perfect selector, and $\underline{m} = 1$) the player will always keep the correct i'th symbol as soon as he happens to try it.

A misleading ranking system is always the result of interaction between places in the sequence, i.e. of the fact that the ordering by rank of the symbols available to fill the i'th place depends on the symbols already filling other places. Notice that the variety of symbols available to fill the i'th place is not affected by the previous filling of other places, only the order in which the k_i symbols will be ranked by the selector.

The weight of the i'th place in a sequence is the maximum quantity by which the rank of a sequence may be changed by changing the symbol filling the i'th place. If ten-digit numbers are ranked by size, for example, the first place has a weight of 9×10^9, while the last place has a weight of only 9. This is an example of unequal weighting of places. The first game recorded in this section of the thesis utilizes such unequal weighting, which however did not cause interaction between places and therefore did not result in a misleading ranking system. Under certain conditions, unequal weighting can cause interaction and produce a misleading ranking system, although not all instances of interaction are caused by unequal weighting.

Interaction with Equal Weighting.

Consider a game using ten-digit sequences, in which the rank *R* of each sequence is the sum of the absolute differences between adjacent digits. The two sequences of highest rank, R = 81, are then

0 9 0 9 0 9 0 9 0 9
and 9 0 9 0 9 0 9 0 9 0.

The actual course of such a game, using Fisher and Yates' Table XXXIII (III) is:

t		*R*
0	2 2 1 7 6 8 6 5 8 4	20
1	$\underline{9}$ 2 1 7 6 8 6 5 8 4	27
9	9 2 1 7 6 8 6 5 8 $\underline{2}$	29
14	9 2 1 7 6 8 $\underline{1}$ 5 8 2	34
23	9 2 1 7 $\underline{3}$ 8 1 5 8 2	43
25	9 2 $\underline{9}$ 7 3 8 1 5 8 2	45
35	9 2 9 7 3 8 1 5 8 $\underline{0}$	47
39	9 2 9 7 3 8 $\underline{0}$ 5 8 0	49
40	9 2 9 7 3 8 0 5 $\underline{9}$ 0	51
52	9 2 9 7 $\underline{0}$ 8 0 5 9 0	57
128	9 $\underline{0}$ 9 7 0 8 0 5 9 0	61
148	9 0 9 7 0 $\underline{9}$ 0 5 9 0	63.

The sequence obtained on the 148'th trial cannot possibly be improved by a mutation affecting only one place. The fourth and eighth places in the sequence have both beomce irrelevant to its rank, and may be filled by any symbol without increasing or decreasing the rank of the sequence. The sequence has in effect worked its way into an evolutionary blind alley, from which it could escape only by a large mutation changing places 5, 6 and 7 all at once, or by recombination with another different sequence.

Interaction due to Unequal Weighting.

A simple but important rule about the weighting of places in the sequence follows from a consideration of the way ten-digit sequences can be ranked.

In general, let 9 represent the best symbol at each place, 8 the next best, and so on, and 0 the worst. If the set is completely ordered (in 10^{10} ranks) it follows that if the first place (from the left) is the place of greatest weight, and the second place is the place of next greatest weight, and so on, and if each sequence is read as a number, the ranking system must be misleading if any sequence has a rank exceeding its number. Hence, the weight of the place with greatest weight cannot exceed 9×10^9, the next highest weight cannot exceed 9×10^8, and the n'th highest weight cannot exceed $9 \times 10^{10-n}$.

The result can be generalized. A set of sequences of $\underline{L}$ places, each of which may be occupied by any of $\underline{k}$ symbols, and which are completely ordered by primary rank, is misleadingly ranked if the weight of the place with the n'th largest weight exceeds $(k - 1)k^{L-n}$.

The same proposition is true of a set of sequences which are incompletely ordered by primary rank; the ranking system is misleading if the n'th largest weight exceeds $(k - 1)k^{L-n}$; but some even more strict constraint must apply, and the writer's analysis has not yet gone far enough to allow him to state it. Consider the following game, for example, in which the right answer is 6 3 5 9 4 0 4 0 3 2, and sequences are considered as ten-digit numbers ranked in inverse order of their absolute deviation from the right answer. Ranking is incomplete (7 3 5 ...has the same rank as 5 3 5...). The weights of places meet the restriction stated above, but the ranking system is still obviously misleading.

Right answer: 6 3 5 9 4 0 4 0 3 2

t	
0	0 3 4 7 4 3 7 3 8 6
3	0 $\underline{4}$ 4 7 4 3 7 3 8 6
4	0 4 4 7 4 $\underline{7}$ 7 3 8 6
6	$\underline{6}$ 4 4 7 4 7 7 3 8 6

Previously, every increase in a digit after the first was an improvement. A change in the first digit has reversed this, and now a decrease is an improvement in every digit after the second. A non-misleading ranking system cannot lead to such reversals of the direction of selection at certain places.

7	6 4 4 7 4 7 $\underline{6}$ 3 8 6
8	6 4 4 7 4 $\underline{2}$ 6 3 8 6
	• • •
55	6 4 1 2 3 0 $\underline{5}$ 3 2 1
56	6 $\underline{3}$ 1 2 3 0 5 3 2 1

A change in the second digit has again reversed this tendency, and it now becomes an improvement for digits after the second to increase.

59	6 3 1 2 3 $\underline{4}$ 5 3 2 1
62	6 3 $\underline{6}$ 2 3 4 5 3 2 1

Once more the number has become too large, and further improvements will come from reducing the digits to the right of the third place. It is also clear that no single change which inserts the correct symbol in either the third or fourth place is an improvement; both would have to occur together if the change was to be retained. The sequence is approaching an evolutionary cul-de-sac.

71 6 3 6 2 3 4 $\underline{0}$ 3 2 1

...

214 6 3 6 0 0 0 0 0 $\underline{1}$ 0

237 6 3 6 0 0 0 0 0 0 0

No change in a single digit can improve this sequence; the player cannot find the right answer. He has been misled by the ranking system.

Another game, using the same initial sequence but using section II of the table of random numbers as a source of "mutations" led to the result

6 4 0 0 0 0 0 0 0 0

which also cannot be improved by any single change.

The selector in organic evolution has obviously used a misleading ranking system. Gene interaction, much of it associated with unequal selective weighting of loci, is a familiar phenomenon. The complexities which arise from dominance and recessiveness, special kinds of gene interaction which occur only in diploids, have not even been mentioned in this thesis, although they form much of the subject matter of population genetics.

Games like the ones recorded above could be made immensely complex, in an attempt to incorporate all the complexities of real genetic systems. Such games could of course only be played on an electric computer. Various workers have programmed electronic computers to mimic complex situations in population genetics, involving the effects of linkage and epistasis on selection. The writer shares with Martin and Cockerham (1960) the view that such models, whether simple or complex, are useful for discovering where present theory breaks down, and for obtaining suggestions about the effects of various parameters, but that a thorough understanding of the system requires a formulated analytical theory. In the present stage of development of Guessing Theory, the simple pencil and paper models recorded above are probably more instructive than more complex and therefore more confusing models would be.

Many groups of organisms appear to have evolved into evolutionary cul-de-sacs. The gigantic Titanotheres, the Irish elk, and the several independently arisen series of Gryphaea-type oysters, are examples in which species or larger groups have become extinct after a period of "orthogenetic" evolution during which they seemed to flourish. Some such instances may be the result of secular changes in the selector (see below), but others are undoubtedly due to the misleading nature of the ranking system imposed by the environment. Analysis of the ways in which ranking systems can be misleading may therefore be informative in the study of extinction. It may be valuable, too, in clarifying the roles of hybridization, non-adaptive diversification resulting from drift in fragmented populations, etc., as possible means of escape from the traps set by a misleading selector.

Indirect Selection.

Lima-de-Faria (1962) has pointed out that natural selection must have some effect at the level of the molecular structure of the genetic material. "Every structural feature of the chromosome and every chromosomal property which has been carefully investigated, reveals a distribution along the chromosome according to a defined pattern of organization." Hence ... "it is logical to conclude that a structural or molecular rearrangement can only survive if it conforms to the organized system of interactions already present in the chromosome body."

This kind of situation, in which the selector acts directly on the sequence of symbols forming each guess, may be called direct selection. In organic evolution, however, the selector is more commonly considered to act on the phenotype, which is a transformation of the genotype; i.e., the selector acts indirectly, comparing guesses only after each has been subjected to a process of transformation. There must be certain constraints on the kinds of transformations which are permissible if a difficult Darwinian game is to be won.

The transformation of the genotype into a phenotype may be thought of as a process of decoding the genetic material. From this point of view the subject has received a vast amount of attention in recent years (reviewed by Crick 1963; Jukes 1963).

The sequence of symbols forming the genetic material may be called the code message, and the resultant phenotype called the clear message. The relationship between code and clear messages can be discussed in a general way in terms of two properties of the code, its redundancy and its complexity.

The code is redundant if there are more possible code messages than clear messages, and ambiguous if there are more possible clear messages than code messages. The DNA-protein code appears to be redundant, since each of 20 different amino acids is coded for by more than one triplet of nucleotide pairs. The effect of a redundant code in a Darwinian game with an indirect selector is the same as the effect of a smaller alphabet. For example, if the code sequence 6 3 5 9 4 0 4 0 3 2 is decoded into E O O O E E E E O E

(Even, Odd, etc.) before selection, the course of the game will be exactly the same as if the alphabets available to fill the places in the code sequence consisted only of two symbols, instead of ten.

The complexity of a code is a measure of the size of the dictionary required to translate code messages into clear messages. The code is of minimum complexity when the number of entries in such a dictionary is equal to the number of symbols in the code alphabet, and is of maximum complexity when the number of entries in the dictionary must equal the number of possible code messages. The substitution ciphers used by schoolboys, in which, for example M = A, N = B, O = C, etc., are codes of minimum complexity. The DNA-protein code is of greater, though small, complexity; a dictionary of 64 entries is required although the code alphabet has only four symbols. Written languages are codes of great complexity; a French-English dictionary requires thousands of entries. Commercial telegraphic codes are codes of maximum complexity, and every possible message must be listed in the dictionary.

The complexity of the code is important in a Darwinian game with an indirect selector. As the code becomes more complex, the relation between the size of a change in the code message and the size of a change in the clear message becomes more complex and may approach apparent randomness. In a schoolboy's cipher, changing a single symbol in the code message always changes a single symbol in the clear message. In the DNA-protein code, changing a single nucleotide pair may or may not change more than one amino acid. Changing a single letter in a French word may change every letter in its English translation (pain = bread, paix = peace). In the later stages of a Darwinian game the player must control the size of the changes in the guesses on which the selector acts, but he can exert this control over the clear messages only by controlling the sizes of the changes he makes in the code messages. For a

Darwinian game to be won, therefore, there must be a rather low maximum limit on the complexity of the code.

A complex code always leads to symbol interaction, and therefore to a misleading ranking system.

Discussions of developmental genetics in terms of a process of decoding have been confined to the synthesis of the protein text, because of the ease with which the structure of either DNA or a polypeptide can be represented symbolically as a sequence of symbols. This unfortunately leads to the notion that its protein text is the phenotype of an organism (Lederberg 1960). From an evolutionist's viewpoint this notion is reasonable enough if selection is exerted only by chemical means, as by antibiotics or absent nutrients in the medium in which micro-organisms are grown, and the biochemistry of the organism is the only aspect of its phenotype which matters. For higher organisms, however, physiology, anatomy, and behaviour are also selectively relevant. The notion of development as a decoding of the genetic material is difficult to apply to these aspects of the phenotype, which cannot be abstractly represented by sequences of symbols. Nevertheless,one can state that the processes of development must be so organized that in general, and on the average, a small genetic change causes only a small change in any selectively important phenotypic variable.

Secular Changes in the Selector.

The attempt to construct a complete analogy between a Darwinian game and the process of organic evolution is further complicated by the fact that in organic evolution, the selector changes its ranking system as time goes on. Mountain building and erosion, and climatic changes, including glacial periods, are obvious examples of such secular changes, which are not caused by evolving organisms but which affect their adaptedness to the environments in which they live. The more indirect effects of organisms on each other might be treated in the same way, mathematically; the axe, the plough, and the spraying machine have evolutionary effects which might better be analogized to the effects of earthquakes and glaciers than to the slow interaction over millions of years between predators and their prey or hosts and their parasites. The photosynthetic transformation of a reducing atmosphere into an oxidizing atmosphere is also a phenomenon of such magnitude, and involving so many kinds of organisms, that it might more easily be treated as a secular change in the environment than as the interaction of organisms with each other.

The prediction or description of secular changes in the environment does not lie within the field of mathematical evolution theory. The contribution of a theory of Darwinian games to their study would consist chiefly of an investigation of the rates of change which must not be exceeded if the overall adaptedness of the fauna and flora is not to decrease because of the change.

VI The Source of Genetic Information

The question of the source of the large quantities of information carried by the genetic material of living organisms is one of profound importance to biology and philosophy. It has been answered in several ways since the concept of information became well enough defined to allow the question to be asked in that form.

The central assumption throughout this thesis was pointed out previously in a brief communication (Warburton 1958): a selector is a source of information, and the process of selection is a process of transferring information from the selector to a receiver. In a Darwinian game, information is transmitted from the referee, who originally stores it, to the player, who stores it in the form of his best past guess. In organic evolution, the selector is the environment, and natural selection is a process for transferring information from the environment to the genetic material.

A simple illustration may clarify the notion of a selector as an information source. Imagine that a teletype machine has been modified so that it generates a random sequence of letters and spaces, and that a man with a switch determines which of these will be allowed to pass out along the wire and which will not. By operating his switch, the man can send a message (such as "The card is the Queen of Hearts") along the wire. What is the source of the information in the message? Obviously, the man with the switch, in whose brain the information must have been stored. If the man operated the teletype machine directly, he could choose the message to be sent from a hypothetical set of possibilities, existing only in his mind. Operating solely as a selector,however, he can choose only from the set of real possibilities generated by the machine; the teletype machine is not a source of information, but a necessary source of variety from which the selector can choose. Similarly, in organic evolution, mutation and recombination are not sources of information, but sources of variety. The environment is the source of information, but since it acts solely as a selector, it cannot choose among some hypothetical set of possibilities but only among the real entities produced by the variety source.

Other writers have explained the source of genetic information in other ways.

One possible answer to the question of the origin of an entity is to deny that the entity exists. Elsasser (1958) devoted several chapters of his book to an attempt to demonstrate that living organisms possess no structures capable of storing large quantities of information, and must develop and function without the benefit of stored information, controlled by "biotonic laws" which are not the laws of physics, though they are necessarily compatible with the laws of physics. Such a hypothesis escapes from the problem of the source of genetic information by denying that genetic information exists. Elsasser's views have been thoroughly rebutted by Raven (1962). Raven's arguments are elegant and detailed. Here it is enough to mention that he shows that chromosomal DNA and the cortex of the fertilized egg are both capable of carrying tremendous quantities of information.

Raven (1962) devotes only a few lines to the origin of genetic information. "So, when we say that all order in living organisms is a consequence of previous order, ... it follows that ... the idea of living organisms not possessing this intrinsic order is inconceivable. Order is a necessary requisite of all life."

"We may understand how the intrinsic order of living organisms, once generated, was perpetuated and transmitted throughout the generations. If we admit, in accordance with the neo-Darwinian thesis, that occasional errors in the process of the conversion of order could acquire survival value, and in this way be perpetuated, we may account for the differentiation of the numerous forms of life in the course of evolution. But our equation gives no clue as regards the first appearance of organismic order. The only thing we can say is: if we want to adhere to a scientific explanation of the origin of life, we must presume that this order was derived, somehow or other, from the primordial negentropy of the universe. At present we can only conjecture as to the way in which this has occurred."

In short, Raven seeks to push the source of genetic information back in time to the mysterious origin of life. His statements seem to imply that the entire process of evolution has added no information (or if one wishes, "order") to living organisms, but has merely variegated them. This is a conclusion with which almost all evolutionists will heartily disagree. In blunt language, almost any DNA molecule presumably contains enough "order" to duplicate itself, and hence be called "living" or nearly so, in the lukewarm soup of the primordial ocean, but only a very special DNA molecule can be the genome of as highly adapted an organism as a bird, a tapeworm, or an oak. This "specialness" -- the narrower and narrower restriction of the set of acceptable possibilities -- has been brought about, literally bit by bit, by billions of generations of selection.

Kimura (1962) concluded that information is created by the process of natural selection, and accumulated by the genetic material. In other words, genetic information has no source, in the sense of a previous carrier from which it has been transferred to the genetic material, but has been created out of nothing. The meaning of this notion of "creating" information is not clear.

Kimura remarks, "What is the source of such genetic information? If the Lamarckian concept of the inheritance of acquire characters were accepted, one might be justified in saying that it was acquired from the environment. However, since both experimental evidence and logical deductions have entirely failed to corroborate such a concept, we must look for its source somehwere else." This is clearly a <u>non sequitur</u>. Lamarckianism implies some sort of direct transfer of information from the environment to the genetic material of an organism. There is no evidence that any such direct transfer occurs. However, the environment acts as a selector, and a selector is a source of information; the process of selection is not a process of creating information, but of transferring it from the environment to the genetic material. The dismissal of Lamarckianism does not imply that the environment is not the source of genetic information, but only that this information has been transferred by a Darwinian mechanism, not a Lamarckian mechanism.

Ashby (1952) argued that noise serves as a source of information in evolution; natural selection is a process for transforming noise into information. The random variations to which genomes are subject may be considered noise, and selection reveals the changes which are improvements and retains them. The conclusion, that information is here being manufactured out of noise, was concurred with by J.B.S. Haldane (1953). Hick (1953) quickly contradicted this hypothesis, pointing out that the noise source in Ashby's model was simply a device for ensuring that, in time, all possible states of the system will occur. A noise source, in other words, is merely a variety generator, like the randomized teletype in the author's analogy given above. A variety source is a necessary part of a system in which a selector is to supply information, since the selector can choose only among possibilities which really exist, not among potentialities. A noise source cannot supply information, however. It is conceivable that a randomized teletype machine would sometimes produce true statements, such as "The card is the Queen of Hearts", but these would fail to carry information simply because they would be indistinguishably mixed with statements like "The card is the Ace of Spades", "The Heart is the Card of Queens", "Tomorrow is Tuesday", and "All the previous statements are untrue".

Ashby (1952) also expressed the idea that the environment imposes a very simple rule, "The dead shall not breed", and the process of selection then consists of working out the consequences of this simple axiom. Ashby's opinion was that the body of inferences drawn from the axiom contains far more information than the axiom itself. The fact is, of course, that this "rule" is imposed as much by the structure of the English language as by any real physical system. The environment imposes an immensely more complex set of rules of the form "Organisms of the following genotypes are forbidden to breed:...".

A closely analogous idea was expressed by Waddington (1961) who wrote "The type of phenomena which we are confronted with in embryonic development should be compared with the elaboration of a whole corpus of theorems in Euclidean geometry or in arithmetic from certain axioms. In this sense the whole of formal mathematics is a set of tautologies. Are we to conclude that their information content is no more than that of the axioms from which they are deduced?"

The answer to this query is "Yes; and no". Euclid's contribution to mathematics did not consist of deducing all the consequences of his axioms, but of carefully selecting a peculiarly elegant and interesting set of consequences asworthy of deduction. It is rather painful to imagine an unfortunate mathematician trying to deduce <u>all</u> the consequences of Eculid's axioms. First, perhaps, he would prove that if two straight lines intersect at an angle of 1", opposite angles are equal; he might then go on to angles of 2", 3" and so on. If ever he completed this series (which would still leave an infinite number of propositions of the same form unproved) he might go on to three intersecting straight lines. He would certainly never reach Pythagoras' theorem, or any other conclusion of interest. Formal mathematics contains more information that its axioms because it consists of a relatively small set of deductions carefully selected from the enormous and decidedly uninteresting set of all possible deductions.

It is therefore hard to see exactly what Waddington may have meant by his remark. Either developmental mechanisms are mechanisms for working out all the consequences of some set of axioms, or they are that plus a mechanism for sorting the useful consequences from the others. In either case, either the axioms or the axioms plus the rules defining usefulness jointly contain all the information embodied in the end result. Either the rule "The dead shall not breed", or that and a set of rules determining which organisms shall be dead and which shall not, contain all the information required to "design" a viable and fertile organism.

Having examined all of these alternatives and found them wanting, the writer sees no reason to change his opinion of 1958. Natural selection is a process by which information is transferred from the environment to the genetic material of organisms.

A consequence of this deduction is that the genetic material of organisms can be considered as coded descriptions of the environments in which they evolved. (One must be extremely wary of the "nothing but..." fallacy (Julian Huxley, cited by Simpson 1949) in this context, however.) A detailed description of the anatomy and chemistry of the organisms inhabiting some region of space allows one to infer much about the geology, climate and topography of the region, the texture and composition of its soils, the chemistry of its waters, and the purity of its air. If we knew how to decode them, a list of the genotypes of its inhabitants would presumbably permit the same inferences. The genetic material of organisms is in fact a coded description of their environments.

Inferences cannot be made with equal precision in the other direction; a detailed description of a newly discovered continent would not allow one to make anything but rather general predictions about the probable nature of its fauna and flora. This is partly because the process of evolution is incomplete, and its exact direction has been determined by chance at many points, so that the genetic material of organisms is a sort of randomly abridged description of their environments. Now an abridgment of a book, in which, say, half the words have been omitted at random, will allow one to describe the unabridged original in considerable detail, but the original does not allow one to make anything but probabilistic statements about the abridgment. The difficulty of inference from environment to organisms is also a result of the fact that faunas and floras consist of mutually adapted organisms. The environment does not determine the nature of each organism in detail, but only the nature of the relations between them.

The expected objection to this interpretation of evolution is that the environment, the inorganic world in general, is "disorderly", "unorganized", "simple", and so on, and cannot possibly contain or ever have contained as much information as the genetic material of the highly organized, orderly, and complex living organisms which now inhabit it.

In spite of innumerable attempts to define them, these words retain rather vague meanings. They may even be more emotive than descriptive in some contexts. While this is true it is useless to use them in an essentially quantitative argument. It can at least be pointed out, however, that geologists, geographers, geomorphologists, physical oceanographers, climatologists, meteorologists, and pedologists might object to having the systems they study dismissed as "simple" and "disorderly". Perhaps one can truthfully say that a plant or an animal contains more "order" or "complexity", if one must use such words, than the same weight of rock or sea water, but this does not disagree with the writer's hypothesis that natural selection is a mechanism for trans-coding information from a large, highly redundant carrier (the inorganic environment) to a carrier of molecular dimensions in which it is much more compactly coded.

The strongest argument for the idea that genetic information was previously stored in the structure of the inorganic world rests on a consideration of the requirements which must be met by the selector in a Darwinian game which can be won. The environment has sometimes been compared to a sieve in the action it exerts on living organisms, separating the fit from the unfit. Now a sieve, or some other very simple mechanical device, may often serve as a secondary referee in a guessing game, but a selector must be a far more complex mechanism than a secondary referee. A secondary referee applies the same criterion repeatedly; if the same guess is repeated, the same response will always occur. A selector applies different criteria at different trials, changing its criteria by a sort of feedback mechanism so that it becomes progressively "tougher" as the candidates it has already passed become progressively better. Archaeapteryx was passed when it had only reptiles with which to compete, but it would no doubt be rapidly eliminated by competition with modern birds. A selector cannot simply divide the set of possibilities into two classes,-but must order them in a large number of ranks.

Now one may order a set of things in several ranks in many different ways. The selector in organic evolution, however, must not merely have ranked the possibilities, but must have ranked them in a way which allowed the player to win. This imposes severe constraints on the kind of ranking system the selector may have used. To allow organic evolution to occur, the environment must use a "special" kind of ranking system. The environment must therefore be a "special" kind of environment.

Two objects can be ordered in 2 ranks in 2 ways, and in 1 rank 1 way, for a total of 3 ways of ranking 2 objects. Three objects can be ordered in 3 ranks 6 ways, in 2 ranks 6 ways, and in 1 rank 1 way, for a total of 13 possible ways of ranking 3 objects. Similarly, 4 objects can be ranked in 81 ways, 5 objects in 621 ways, and 6 objects in 10, 823 ways. For comparison, notice that 6! is only 720. The number of different ways in which N objects can be ranked cannot easily be computed, but it is certainly far greater than N!. If the player is to win a Darwinian game, however, the selector must rank the members of the set of possibilities in such a way that almost any permutation of the hierarchy in which the player classifies the possibilities is a relevant classification and is at worst only slightly misleading. In short, the selector is free to rank the members of the set in only a minute fraction of the number of ways they could be ranked. If one assumes that there is a possible selector for every possible ranking system, it follows that only a small fraction of all possible selectors will permit the player to win a Darwinian game. To define (or design and construct) a selector which will permit a Darwinian game to be won therefore requires a large amount of information.

It will require a much more detailed analysis to place the foregoing argument on a quantitative basis, and to estimate the quantity of information which must be carried by the selector in a Darwinian game. The writer suggests as a tentative hypothesis that the quantity equals or exceeds the quantity which can be accumulated by the player, i.e. the quantity required to specify the right answer.

Scientific Value of the Theory

This thesis is not offered as a completed mathematical theory of evolution. Many aspects of the mathematical model described in it have been only mentioned, as suggestions for future investigation. Nevertheless enough has been said to indicate that the concept of a Darwinian game, which formalizes the way information is transferred and stored in systems evolving by selection, gives rise to a mathematical model which permits quantitative discussion of the conditions which must be met by any system in which large quantities of information are to be transferred in a limited time by selection.

A study of Darwinian games with imperfect selectors shows that the set of possibilities must be able to be represented by a set of linear sequences of symbols, and that a winning player's strategy must include the use of large populations, guesses which reproduce themselves with small "mutational" changes, a rather low and controllable mutation rate, and recombination. This result -- the inference that some of the major features of genetic systems are necessary conditions of evolution by selection, not merely sufficient conditions -- is a useful confirmation of the validity of the model.

Darwinian games therefore lead to a sort of generalized population genetics, independent of the physical or chemical nature of the evolving "guesses". Practically, this result is of limited value, at least until living organisms or something like them are discovered on other worlds. Population genetics is a well developed branch of mathematics, and is capable of solving its own problems without translating them into Guessing Theory.

Theoretically or philosophically the result is of greater importance. It shows that the major generalizations of population genetics are independent of the physical basis on which population genetics has been built. The major features of genetic systems are not by-products of the chemistry and physics of nucleic acids and proteins, but are necessary properties of all systems evolving by natural selection. To exemplify the reductionist approach to biology, Simpson (1962) quotes this statement (which is paraphrased and may therefore distort the intent of the original): "Genetics, the central discipline of biology, has as its ultimate subject a group of chemical compounds. Genetics requires knowledge of the physical sciences but no other knowledge in the field of biology itself". On the contrary, much can be inferred from the central fact that biology deals with entities which have evolved by natural selection, and these inferences can be made with no knowledge of the chemical compounds of which organisms are composed.

These remarks are not meant to belittle the contributions which molecular biology will make to evolution theory. After all, questions of alphabet size, symbol interaction, the redundancy and complexity of the genetic code, etc., must all be answered by molecular genetics. They are meant to belittle the attitude that evolution theory has nothing fundamental to contribute, and its practitioners are merely wasting time until someone is able to "reduce" their discipline to biochemistry. To reduce evolution theory to the chemistry of DNA would be an interesting philosophical exercise, but not much more, because the basic principles and mathematical regularities of evolution theory would remain the same whether it was reduced to DNA, magnetic tape, or ink and paper.

In addition to permitting this philosophical conclusion about population genetics, the analysis of Darwinian games supplements population genetics as a mathematical approach to the study of evolution. Its value rests largely on the fact that rank, one of the fundamental quantities entering into every equation, is a quantity representing the <u>adaptedness</u> of the evolving organisms. The basic feature of organic evolution is an increase in the adaptedness of organisms, but no quantity representing adaptedness enters into the formulation of classical mathematical studies of evolution.

Adaptedness must be clearly distinguished from fitness. The fitness of a kind of organism is a measure of its proliferative ability compared to its contemporary competitors. The mean fitness of all living populations of organisms is approximately one, and the mean fitness of all earlier populations was also approximately one. The adaptedness of an organism may be thought of as a measure of what its fitness would be if it were somehow placed in competition with all its own ancestors. It may also be thought of as the inverse of the probability that an organism spontaneously generated in some chance way would be as fit as or fitter than the organism in question. Adaptedness might be measured, in principle and possibly in practice, by the probability that a random mutation is selectively disadvantageous.

J.B.S. Haldane (1954) has pointed out that most of mathematical evolution theory deals with forces which are very nearly in equilibrium. Indeed, most of the text of books like Li (1955) and Falconer (1960) is devoted to calculating the conditions under which the genetic structure of a population will achieve equilibrium or fixity. Fisher's (1930) "fundamental theorem of natural selection" yields a measure of the rate of change of fitness in a population, but in order to avoid the conclusion that every organism is undergoing a population explosion, the application of this theorem always involves the assumption that the average fitness of a population is "re-set" at about unity in each generation by density-dependent controlling factors of population size. Adaptedness, therefore, is the only important quantity entering into mathematical evolution theory which increases throughout the entire evolutionary process, never achieving equilibrium or fixity except as a result of defects in the genetic mechanism of the evolving organisms.

The inclusion of a parameter representing adaptedness in the mathematical expressions describing the course of a Darwinian game allows inferences about the way the process of natural selection would change with time. As adaptedness increases, mutations must become smaller, and if the selector is imperfect breeding populations must become larger, the mutation rate must be small and adapted to population size, and recombination becomes necessary. The rate of increase in adaptedness decreases as adaptedness increases.

Furthermore, the parameters desdribing mutation size and rate, population size, etc., must lie within increasingly narrow limits if evolution is to continue as adaptedness increases. The genetic mechanism itself must therefore have evolved in order to meet these increasingly severe demands. This probably occurred by competition between populations. However, those genetic systems best suited for rapid evolution while adaptedness was still small would be badly suited for continuing progress later in the process. It therefore appears that the parameters of the genetic system must themselves be genetically controlled to some extent, and must be continually adjusted by selection.

The analysis of Darwinian games relates evolutionary genetics more closely to modern ecology than population genetics has been able to do. The study of niche diversity, competitive exclusion, the trophic structure of ecological communities, and so on, is essentially an attempt to find the number of "right answers" existing in a Darwinian game and the relationships between them. If the results of such investigations can be mathematically formulated in a suitable way and incorporated in a Darwinian game, they will lead to a potentially valuable mathematical model of the processes of adaptive radiation and speciation.

A model of natural selection based on Guessing Theory also seems to be more directly related to molecular genetics than population genetics does. The complexities due to symbol interaction and unequal selective weighting of mutational sites can be investigated either from the viewpoint of classical genetics or that of molecular genetics, but the questions of code redundancy and complexity, and symbol interaction due to the complexity of the code, must be investigated by molecular genetics. The nature of the genetic code is a subject which does not seem very relevant to classical population genetics, but it will be basic to any attempt to work out a detailed model of organic evolution taking account of indirect selection.

The writer therefore contends that the study of evolution has a most useful role to play as a framework for unifying the many fields of biological investigation into a single comprehensive science, and for relating them to the inorganic sciences.

In the writer's opinion, however, the most interesting conclusion of this thesis is the answer it provides to the question of the origin of genetic information. As Raven (1962) said, it "was derived, somehow or other, from the primordial negentropy of the universe". However, we need not merely conjecture as to the way this has occurred; we have known the basic principles since 1859. The conclusion that natural selection is a process which transfers information from the environment to the genetic material of organisms therefore serves to relate organic evolution to the history of the universe as a whole.

Summary

Organic evolution can be analogized to a guessing game in which a player draws guesses from an initial set of possibilities known to include the right answer, and submits them to a referee who passes the best of the submitted guesses and rejects the others. As a record of the previous course of the game the player can retain only the guesses last passed by the referee; he cannot maintain a record of the number of trials which have occurred. The total number of guesses the player may submit is much smaller than the number of possibilities.

If the player is to find the right answer, it must be possible to represent each possibility by a linear sequence of symbols in which the alphabet from which the i'th symbol must be chosen is unchanged by the previous filling of other places in the sequence. The player must prepare new guesses by changing a number of symbols in the past guesses; this number must be small, or must decrease as the game progresses. If the referee sometimes makes errors, the player must submit large numbers of guesses at once, must reduce the "mutation rate" so that a large fraction of new guesses are identical to the past guesses, and must utilize a kind of "recombination" among past guesses to prepare new ones. Influences analogous to inbreeding, genetic drift, etc., then act to complicate the player's strategy. The model therefore leads to a sort of generalized population genetics, which is independent of the physical nature of the sequences representing guesses.

A basic parameter in the model is the rank assigned to each guess by the referee. This represents a measure of the adaptedness of an organism, a quantity which does not occur in classical population genetics. Roughly,

it is a relative measure of the fitness an organism would possess in competition with its ancestors. As rank increases the requirements of a winning strategy become increasingly stringent. The rate of increase in rank becomes smaller as rank increases.

Complications are added if the ranks of guesses are influenced by interaction between symbols, and if the "selective value" of different places in the sequences representing guesses are different. The model relates the genetic theory of natural selection closely to modern ecology, by lending itself to discussion of cases in which there are several right answers which are not uniquely defined, but are defined by the relationships between them, corresponding to the mutual adaptations which must exist among the members of an ecological system. It also closely relates natural selection and molecular genetics, by permitting discussion of indirect selection. Indirect selection exists when the referee does not directly judge the sequences representing guesses (the genotypes), but judges transformations of them (phenotypes). The course of the game is then strongly influenced by the redundancy and complexity of the genetic code.

The model is basically a formalization of the way information is transmitted from the referee, who originally stored it, to the player, who stores it as his best past guess, in a particular kind of guessing game. Thus it represents the way natural selection serves to transmit information from the environment to the genetic material of evolving organisms. The source of genetic information is therefore the environment. Natural selection has not created information, but has transferred it from the environment, transformed it, and concentrated it in a smaller material carrier, the genetic material of organisms.

BIBLIOGRAPHY

Ashby, W. Ross, 1952. Can a mechanical chess-player outplay its designer? Brit. J. Philos. Sci., 3: 44-57.

Ashby, W. Ross, 1957. An Introduction to Cybernetics. Chapman and Hall, London.

Bennett, Alfred W., 1870. The theory of natural selection from a mathematical point of view. Nature, 3: 30-33.

Blum, Harold F., 1955. Time's Arrow and Evolution. Princeton University Press, Princeton, N.J.

Brillouin, Leon, 1956. Science and Information Theory. Academic Press, New York.

Burnet, Sir Macfarlane, 1958. The Clonal Selection Theory of Acquired Immunity. Vanderbilt Press, New York.

Colbert, Edwin H., 1955. Evolution of the Vertebrates. John Wiley and Sons, New York.

Crick, F.H.C., 1963. On the genetic code. Science, 139: 461-464.

Darwin, Charles, 1859, etc. The Origin of Species. Variorum text, ed. Morse Peckham, 1959. University of Pennsylvania Press, Philadelphia.

Darwin, Francis, 1887 ed. The Life and Letters of Charles Darwin. 3 vol. John Murray, London.

Elsasser, Walter M., 1958. The Physical Foundation of Biology. Pergamon Press, London.

Falconer, D.S., 1960. Introduction to Quantitative Genetics. Ronald Press, New York.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection. Clarendon Press, Oxford.

Fisher, Sir Ronald A. and Frank Yates, 1957. Statistical Tables for Biological, Agricultural and Medical Research. 4th ed. Oliver and Boyd, Edinburgh.

Florkin, M., 1960. (ed.) Aspects of the Origin of Life. Pergamon Press, London.

Friedman, George J., 1959. Digital simulation of an evolutionary process. General Systems, 4: 171-184.

Gardner, Martin, 1962. Mathematical games. Scientific American, 206: 138-144.

Gause, G.F., 1934. The Struggle for Existence. Williams and Wilkins, Baltimore.

Goldman, Stanford, 1953. Information Theory. Prentice-Hall, New York.

Haldane, J.B.S., 1932. The Causes of Evolution. Harper and Bros., New York and London.

Haldane, J.B.S., 1953. The mechanical chess-player. Brit. J. Philos. Sci., 3: 189-191.

Hardy, G.H., 1908. Mendelian proportions in a mixed population. Science, 28: 49-50.

Hick, W.E., 1953. The mechanical chess-player. Brit. J. Philos. Sci., 3: 353-355.

Hutchinson, G. Evelyn, 1957. Concluding remarks. Cold Spring Harbor Symp. Quant. Biol., 22: 415-427.

Hutchinson, G. Evelyn, 1959. Homage to Santa Rosalia, or Why are there so many kinds of animals? Amer. Nat., 93: 145-159.

Huxley, Julian, 1942. Evolution: the Modern Synthesis. Allen and Unwin, London.

Jenkin, Fleeming, 1867. Darwin and the origin of species. Reprinted as pp. 215-263 in Papers, Scientific and C., by the late Fleeming Jenkin. Sidney Colvin and J.A. Ewing (eds.), with a memoir by R.L. Stevenson, 1887. Longmans, Green and Co., London.

Jukes, Thomas H. 1963. The genetic code. American Scientist, 51: 227-245.

Kalmus, Hans, 1950. A cybernetical aspect of genetics. J. Heredity, 41: 19-22.

Kimura, Motoo, 1961. Natural selection as the process of accumulating genetic information in adaptive evolution. Genet. Res. Camb., 2: 127-140.

Langer, Susanne K., 1953. An Introduction to Symbolic Logic. 2nd ed. Dover Publications, New York.

Lederberg, Joshua, 1960. A view of genetics. Science, 131: 269-276.

Lewontin, R. C., 1961. Evolution and the theory of games. J. Theoret. Biol., 1: 382-403.

Li, C. C., 1955. Population Genetics. University of Chicago Press, Chicago.

Lima-de-Faria, A., 1962. Selection at the molecular level. J. Theoret. Biol., 2: 7-15.

Lotka, Alfred J., 1922. Contribution to the energetics of evolution. Proc. Nat. Acad. Sci., U.S.A., 8: 147-151.

MacArthur, Robert, 1957. On the relative abundance of bird species. Proc. Nat. Acad. Sci., U.S.A., 43: 293-295.

Marchant, James, 1916. Alfred Russel Wallace; Letters and Reminiscences. 2 vol. Cassell & Co., London.

Martin, Frank G., and C. Clark Cockerham, 1960. High speed selection studies. pp. 35-45 in Biometrical Genetics, Oscar Kempthorne, ed. Pergamon Press, New York.

Michie, Donald, 1961. Trial and error. Penguin Science Survey, 1961, 2: 129-145. Penguin Books, Harmondsworth.

Moore, Edward F., 1956. Artificial living plants. Scientific American, 195: 118-126.

Muller, H. J., 1929. The method of evolution. Scientific Monthly, 29: 481-505.

Muller, H.J., 1958. Evolution by mutation. Bull. Amer. Math. Soc., 64: 137-160.

Neumann, John von, and Oskar Morgenstern, 1944. Theory of Games and Economic Behaviour. Princeton University Press, Princeton, N.J.

Oparin, A.I., 1931. The Origin of Life. MacMillan, London. Translated by S. Morgulis.

Pringle, J.W.S., 1956. On the parallel between learning and evolution. General Systems, 1: 90-110. (Reprinted from Behaviour, 3: 174-215).

Quastler, Henry, 1958. A primer on information theory. In, Yockey, Platzman and Quastler, 1958, pp. 3-49.

Rapoport, Anatol, 1961. The perfect learner. Bull. Math. Biophys., 23: 321-335.

Raven, Chr. P., 1961. Oogenesis: the Storage of Developmental Information. Pergamon Press, London.

Romer, Alfred Sherwood, 1945. Vertebrate Paleontology. University of Chicago Press, Chicago.

Salisbury, Frank B.,1962. Martian biology. Science, 136: 17-26.

Shannon, Claude, and Warren Weaver, 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana, Ill.

Simpson, George Gaylord. 1944. Tempo and Mode in Evolution. Columbia University Press, New York.

Simpson, George Gaylord, 1949. The Meaning of Evolution. Yale University Press, New Haven.

Slobodkin, Lawrence B., 1961. Growth and Regulation of Animal Populations. Holt, Rinehart and Winston, New York.

Stern, Curt, 1943. The Hardy-Weinberg Law. Science, 97: 137-138.

Waddington, C.H., 1961. Architecture and information in cellular differentiation. pp. 117-126 in The Cell and the Organism: Essays presented to Sir James Gray. J.A. Ramsay, V.B. Wigglesworth eds. University Press, Cambridge.

Wallace, Alfred Russel, 1870. Reply to Mr. Bennett. Nature, 3: 49-50.

Warburton, Frederick E., 1958. Natural selection as the accumulation of information from the environment. Proc. X Congr. Genet., 2: 308.

Wiener, Norbert, 1948. Cybernetics. John Wiley and Sons, New York.

Wright, Sewall, 1922. Coefficients of inbreeding and relationship. Amer. Nat., 56: 330-338.

Wright, Sewall, 1931. Evolution in Mendelian populations. Genetics, 16: 97-159.

Yčas, Martynas, 1958. The protein text. In Yockey, Platzman and Quastler, 1958, pp. 70-102.

Yockey, Hubert P., Robert L. Platzman, and Henry Quastler, 1958. Symposium on Information Theory in Biology. Pergamon Press, New York.