EVOLUTIONARY INDICES

LILA L. GATLIN
UNIVERSITY OF CALIFORNIA, BERKELEY

1. Introduction

As *Homo sapiens* we have always believed that we are higher organisms. After all, we are more complex, more differentiated, more highly ordered than lower organisms. As thermodynamicists we recognize these words and realize that the concept of entropy must somehow enter into the explanation of this.

We have always had the vague notion that as higher organisms have evolved, their entropy has in some way declined because of this higher degree of organization. For example, Schröedinger made his famous comment that the living organism "feeds on negative entropy." We reason that this decreasing entropy of evolving life, if it exists, does not in any way, violate the second law of thermodynamics which states that the entropy of an isolated system never decreases. The living system is not isolated and the reduction in entropy has been compensated for by a correspondingly greater increase in the entropy of the surroundings. It does not violate the letter of the second law, and yet something about it seems to make us uneasy. Why should the evolution of the living system constantly drive in the direction of increasing organization while all about us we observe the operation of the entropy maximum principle, which is a disorganizing principle? I know of no other system except the living system which does this.

First of all, can we establish that the entropy has, in fact, declined in higher organisms? No one has ever proved this quantitatively. In fact, one can argue that it is impossible to establish this thesis by classical means because of the uncertainty principle in its broadest sense. In particular, if we were to make the precise and extensive measurements necessary to determine accurately the entropy difference between a higher and a lower organism, these measurements would disturb the living systems so much that they would kill them. So, it is impossible by classical means to even establish this proposition in which almost all of us seem to believe.

When concepts break down like this they are of little use to us. I think that our classical notions of entropy are totally inadequate in dealing with the living system. This does not mean that there is anything mysterious, supernatural, or vitalistic about the living system. It simply means that our classical notions of entropy are inadequate, just as the laws of Newtonian mechanics were inadequate in dealing with the interior of the atom.

I shall extend the entropy concept primarily through the apparatus of information theory, but I shall extend this also. Shannon [10] gave the most general definition of entropy to date and I shall extend the concept of Shannon. Specifically, I shall show that the entropy function which Shannon called the redundancy is composed of two parts which I call D_1 and D_2. We must characterize the redundancy of a sequence of symbols by two independent numbers, one describing the amount and the other the kind of redundancy of the sequence. I can state this in terms of entropy. I shall show that phrases like, increasing entropy or decreasing entropy, are not completely definitive. We must ask, in what way the entropy has increased or decreased or what kind of entropy is it? We do not encounter such questions in either classical thermodynamics or information theory. I shall develop a theory which can answer these questions.

In classical thermodynamics we dealt with the ordering of three dimensional aggregates of matter, but in information theory we begin to grapple with the concept of ordering of one dimensional sequences of symbols. This is very significant because we now know that the DNA molecule is a linear sequence of symbols which stores the primary hereditary information from which the entire living organism is derived just as a set of axioms and postulates stores the primary information from which a mathematical system is deduced. Therefore, if we wish to investigate the organization of living systems, we must investigate the ordering of the sequences of symbols which specify them.

DNA stores the hereditary information in a sequence of symbols from an alphabet of four letters, the four DNA bases, A, T, C and G. DNA stores its information in the particular sequential arrangement of these four letters just as any language. Therefore we are dealing with language in general although we will apply it to DNA in particular.

2. Theory

We must first define the alphabet. We let

$$S_1 = \{X_i : i = 1, a\}, \tag{1}$$

where the X_i are the letters of the alphabet and a is the number of letters. For DNA, $a = 4$.

With each X_i there is associated a probability $0 \leqq P_i \leqq 1$;

$$\sum_i P_i = 1. \tag{2}$$

Thus, S_1 is a finite probability space. The entropy of S_1 according to Shannon [10] is

$$H_1 = -K \sum_i P_i \log P_i. \tag{3}$$

When $K = 1$ and the logarithm base is 2, the units of H_1 are bits. It can be shown under very reasonable postulates (Khinchin [6]) that (3) is unique and takes on its maximum value, $\log a$, if and only if all the P_i are equal. Thus, $H_1^{\max} = \log a$.

The maximum entropy state for a sequence of symbols is characterized by equiprobable, independent single letter elementary events. This statement is not difficult to justify. Almost any game situation illustrates that the most "random" state is characterized by equiprobable, independent events.

We all know that in any language the single letter frequencies diverge from equiprobability. For example, in the English language the letter e occurs more frequently than any of the others. The divergence from the maximum entropy state due only to this divergence from equiprobability is given by

$$D_1 \equiv \log a - H_1, \tag{4}$$

where the P_i in H_1 are the experimentally observed values for a given language. Biologists call the distribution of the P_i on S_1 the "base composition" of DNA.

We are interested in the sequential arrangement of the letters in a sequence. Therefore, we define a space of n tuples:

$$S_n = \{X_iX_j \cdots X_n : i, j \cdots n = 1, a\}. \tag{5}$$

There are a^n n tuples in S_n.

If the letters in the sequence are independent of each other,

$$H_n^{\text{Ind}} = -\sum_i \sum_j \cdots \sum_n P_iP_j \cdots P_n \log P_iP_j \cdots P_n \tag{6}$$

or

$$H_n^{\text{Ind}} = nH_1. \tag{7}$$

Let m be the memory of a Markov source. If $m = 1$, the probability of occurrence of a given letter depends only on the letter immediately preceding it in the sequence. Then the entropy of S_n is given by

$$H_n^{\text{Dep}} = -\sum_i \sum_j \cdots \sum_n P_iP_{ij} \cdots P_{(n-1)n} \log P_iP_{ij} \cdots P_{(n-1)n}, \tag{8}$$

where P_{ij} is the one step Markov transition probability from letter i to letter j. Utilizing the summations

$$\sum_j P_{ij} = 1 \tag{9}$$

and

$$\sum_i P_iP_{ij} = P_j, \tag{10}$$

equation (8) reduces to

$$H_n^{\text{Dep}} = H_1 + (n-1)H_M^1, \tag{11}$$

where

$$H_M^1 = -\sum_i \sum_j P_iP_{ij} \log P_{ij}. \tag{12}$$

This is just the well-known form for the entropy of a first order Markov source. If $m = 2$,

$$H_n^{\text{Dep}} = -\sum_i \sum_j \cdots \sum_n P_iP_{ij}P_{ijk} \cdots P_{(n-2)(n-1)n} \log P_iP_{ij}P_{ijk} \cdots P_{(n-2)(n-1)n}, \tag{13}$$

$$H_n^{\text{Dep}} = H_1 + H_M^1 + (n-2)H_M^2, \tag{14}$$

where

$$H_M^2 = -\sum_i \sum_j \sum_k P_i P_{ij} P_{ijk} \log P_{ijk}. \tag{15}$$

Following this same pattern, we generalize for an mth order Markov source:

$$H_n^{\text{Dep}} = H_1 + H_M^1 + H_M^2 + \cdots H_M^{(m-1)} + (n - m)H_M^m. \tag{16}$$

If the sequence of symbols diverges from the maximum entropy state due only to a divergence from independence of the symbols, this divergence must be a function of the difference between H_n^{Ind} and H_n^{Dep}.

Since $H_M^{m+1} \leqq H_M^m$ (this is a generalized form of Shannon's fundamental inequality; its proof is in Khinchin [6]),

$$H_n^{\text{Ind}} \geqq H_n^{\text{Dep}}, \tag{17}$$

and since both H_n^{Ind} and H_n^{Dep} are monotonically increasing functions of n, I define the divergence from independence

$$D_2 \equiv \lim_{n \to \infty} \frac{1}{n} (H_n^{\text{Ind}} - H_n^{\text{Dep}}). \tag{18}$$

From (7),

$$\lim_{n \to \infty} \frac{1}{n} H_n^{\text{Ind}} = H_1, \tag{19}$$

and from (16) if we impose the condition that $m \ll n$,

$$\lim_{n \to \infty} \frac{1}{n} H_n^{\text{Dep}} = H_M^m. \tag{20}$$

Therefore,

$$D_2 = H_1 - H_M, \tag{21}$$

where the order of H_M is understood.

Our condition that $m \ll n$ holds for DNA. DNA molecules are very long. For human DNA, $n \cong 4 \times 10^9$ and for even a small bacteria such as *Eschericia coli,*

$$n \cong 4 \times 10^6. \tag{22}$$

At the present tıme there is no conclusive evidence that the m is any greater than one for DNA. There are good theoretical arguments for expecting future evidence that it is greater than this, but it is highly unlikely that m will be of any greater magnitude than a small integer. Therefore, $m \ll n$ will almost certainly hold for DNA. If one knows m for any given language, one can always impose the condition $m \ll n$ simply by considering sequences of sufficient length.

The total divergence from the maximum entropy state is $D_1 + D_2$, which we shall call the information density of the sequence I_d,

$$I_d = D_1 + D_2. \tag{23}$$

I previously called this quantity the "information content" of DNA (Gatlin [2]), but these are poorly chosen words for a number of reasons.

Now let us show the relationship between this quantity I_d and the redundancy of Shannon. According to Shannon's [10] definition, the redundancy of a sequence of symbols is given by

$$R \equiv 1 - \frac{H_M}{\log a}. \tag{24}$$

From (21) and (4),

$$R = \frac{D_1 + D_2}{\log a} \tag{25}$$

or

$$R = \frac{I_d}{\log a}. \tag{26}$$

The redundancy of Shannon is just the information density expressed as a fraction of its maximum value, $\log a$. This entire picture is illustrated in Figure 1 which is an entropy scale.

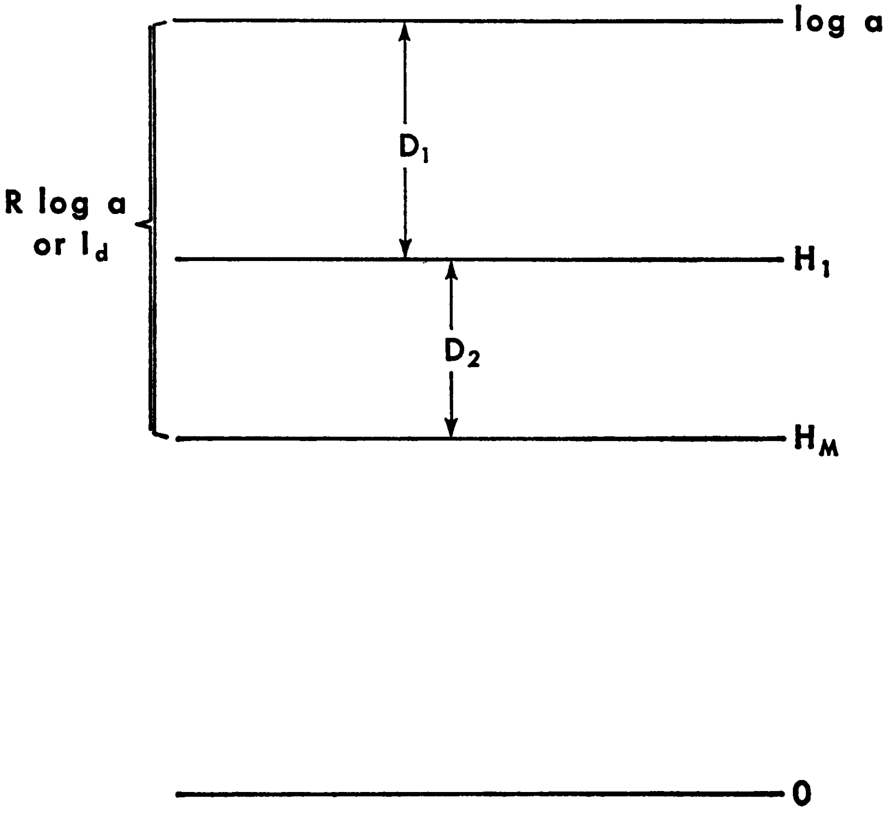

FIGURE 1

The entropy scale.

It is clear from the entropy scale that a given value of I_d may be achieved with different relative contributions from D_1 and D_2. Hence, I define the divergence indices or simply the D indices to characterize this relative contribution:

$$RD1 \equiv \frac{D_1}{D_1 + D_2} = \frac{D_1}{I_d} = \frac{D_1}{R \log a} \tag{27}$$

or

$$RD2 \equiv \frac{D_2}{D_1 + D_2} = \frac{D_2}{I_d} = \frac{D_2}{R \log a}. \tag{28}$$

There is, of course, only one independent index being defined and

$$RD1 + RD2 = 1. \tag{29}$$

Thus, we have in R and $RD1$ or $RD2$ two independent parameters, both dimensionless fractions with a range of zero to one. The R tells us how much the system has diverged from the maximum entropy state and the D index tells us in what way this divergence has taken place, that is, primarily through D_1 or D_2.

Let us now observe how these two parameters describe living systems and what they tell us about their evolution.

3. Results

Figure 2 is a plot of R versus $RD2$. There are 34 organisms represented. The basic data is obtained from the nearest neighbor experiment of Kornberg's group (Josse, Kaiser, and Kornberg [4] and Swartz, Trautner, and Kornberg [13]) which measures the basic P_{ij} for a given DNA. We are considering therefore only a first order Markov dependence, that is, $m = 1$. Also we should note that these data do not include any values for satellite DNA's but must be assumed to represent primarily the main DNA of the organism which carries the hereditary information.

The circles are bacteriophage, viruses which invade bacteria. They follow an empirical functional dependence. The squares are bacteria and follow a similar empirical curve. The vertebrates, however, do not exhibit a similar functional behavior between R and $RD2$ but fall into a rather restricted domain. R lies between about 0.02 to 0.04 and $RD2$ lies between about 0.6 to 0.8. There are some lower organisms with R values as high as or even higher than vertebrates, but whenever this occurs the $RD2$ value invariably drops quite low. This means that whenever lower organisms achieve R values in the vertebrate range they do so primarily by increasing D_1, the divergence from equiprobability of the DNA symbols (or bases). This confirms a well established experimental fact that the base composition of lower organisms, particularly bacteria, has a wide variational range from almost 20 to 80 per cent cytosine plus guanine while the base composition of vertebrates lies within the restricted range of about 40 ± 4 per cent $(C + G)$ (Arrighi, Mandel, Bergendahl, and Hsu [1]). Therefore, vertebrates have achieved their higher R values by holding D_1 relatively constant and in-

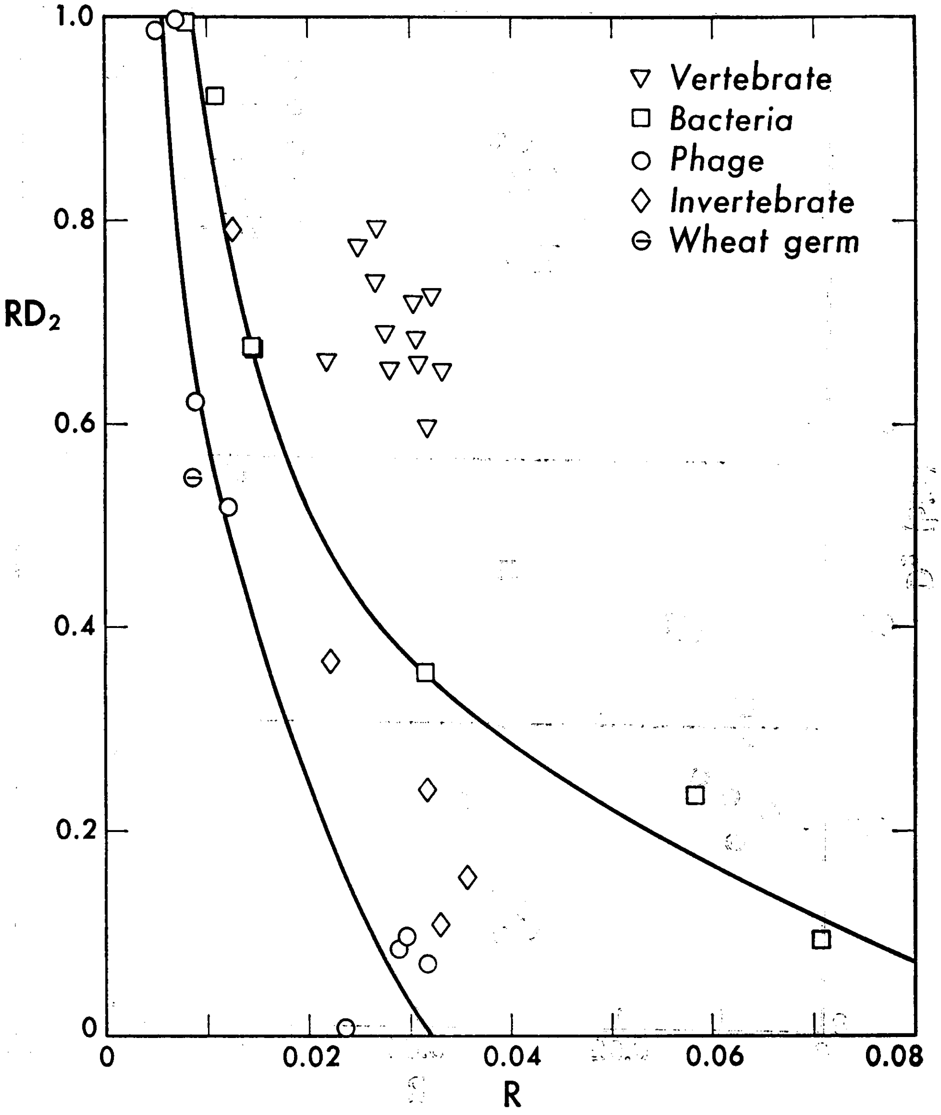

FIGURE 2

R versus RD_2

creasing D_2 whereas lower organisms use D_1 as the primary variable. The mechanism is fundamentally different.

If this is the case, we should expect to find the vertebrate D_2 values higher in general than for lower organisms. This is what we observe. Figure 3 is a plot of R versus D_2.

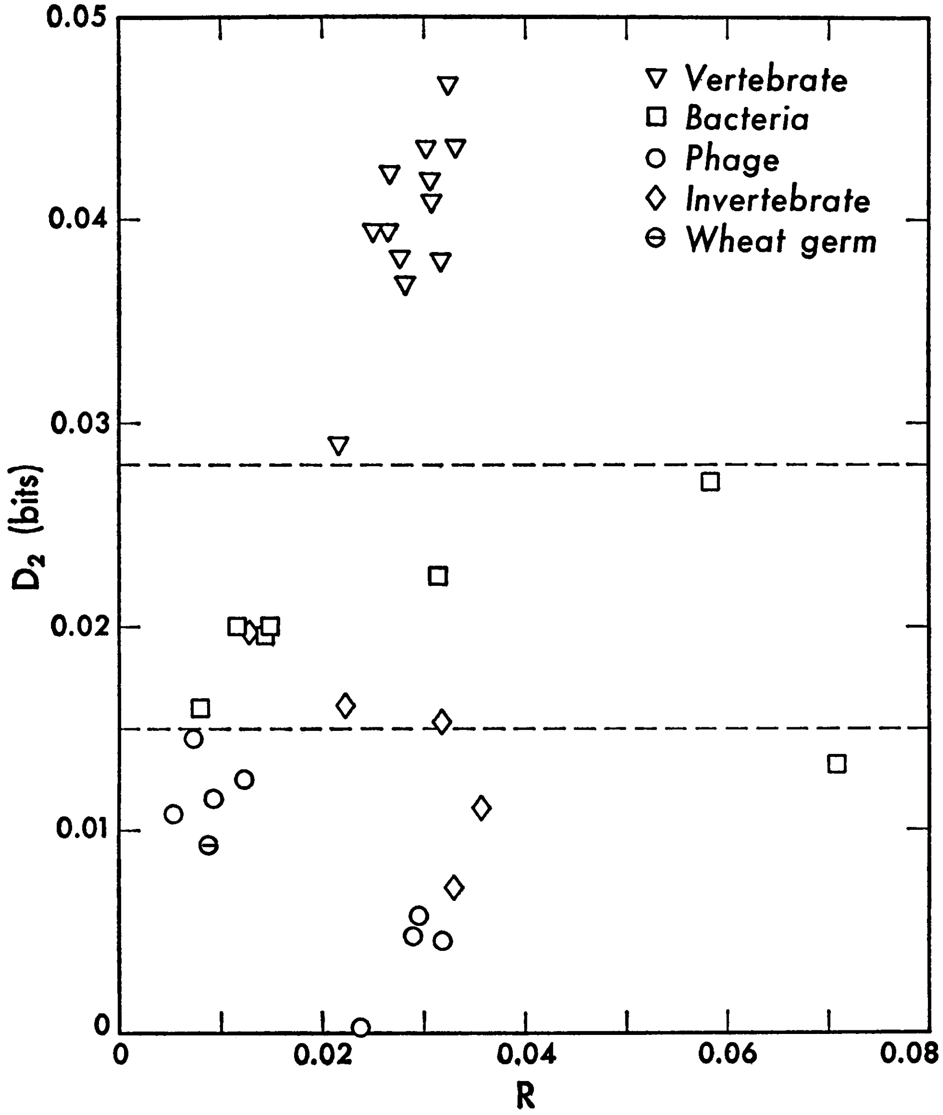

FIGURE 3

R versus D_2.

The vertebrates are characterized by the highest absolute magnitudes of D_2, the divergence from independence of the bases. We might say that D_2 is an evolutionary index which separates the vertebrates from all other "lower" organisms. In terms of entropy, vertebrate DNA does not necessarily have the lowest values of H_M, but they have the lowest values of H_M relative to H_1, that is, they have the highest values of D_2 which is a more important measure of the ordering of a sequence of symbols than any *single* entropy value.

This situation can be described in terms of game theoretic limits. Table I lists the maximum and minimum values of D_1 and D_2 for the groups of bacteriophage, bacteria, invertebrates and vertebrates. The vertebrates display a max-min of D_2 and a min-max of D_1. From the relations we have derived and from an inspection of the entropy scale this can be stated in terms of entropy. The vertebrates display a max-min of H_1 and a min-max of H_M. These game theoretic limits are indicative of an optimization between the opposing elements of variety versus reliability which must occur in any sophisticated language (Gatlin [3]).

TABLE I

GAME THEORETIC LIMITS OF D_1 AND D_2

The † indicates min-max, the asterisk indicates max-min.

		D_1	D_2
Phage	max	.059	.015
	min	.000	.000
Bacteria	max	.129	.027
	min	.000	.013
Vertebrates	max	.026†	.047
	min	.011	.029*
Invertebrates	max	.211	.020
	min	.005	.007

4. Other evolutionary indices

A statistician might be interested in whether or not the result that vertebrates have the highest values of D_2 could be duplicated by classical statistical procedures. After all, D_2 is a measure of the "deviation from random" of the base sequence in DNA. Usually when we speak of a "random" sequence, we mean one where there has been no divergence from independence of the symbols as separate and distinct from the divergence from equiprobability.

Figure 4 is a plot of R versus σ, where

$$\sigma = \{\sum_i \sum_j \tfrac{1}{16} (P_iP_j - P_iP_{ij})^2\}^{1/2}. \tag{30}$$

Therefore, σ is the standard deviation from the random of the base sequence in DNA taking into consideration only a first order Markov dependence. This should be a classical counterpart of our D_2 measure. However, σ cannot begin to duplicate the results of the D_2 index. There is significant overlap of the bacterial and vertebrate domains.

It is possible to define arbitrarily other classical evolutionary indices using the standard root mean square form with a slightly different base. T. F. Smith [11]

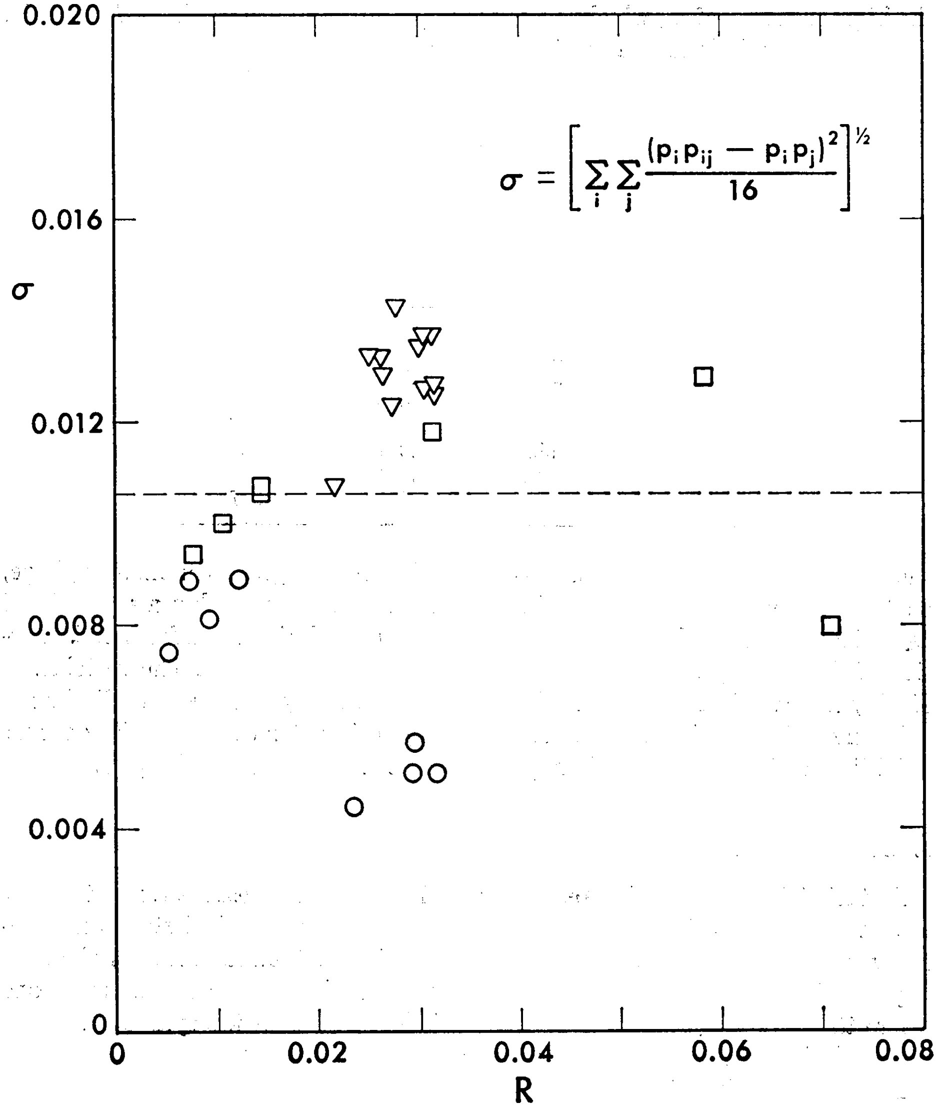

FIGURE 4

R versus σ.

has defined the following evolutionary index. Figure 5 is a plot of R versus e, where

$$e = \{\sum_i \sum_j (P_{ij} - P_j)^2\}^{1/2}. \tag{31}$$

Here the base of the index is the transition matrix element P_{ij} minus the base composition value P_j. The results are much better. There is separation of the

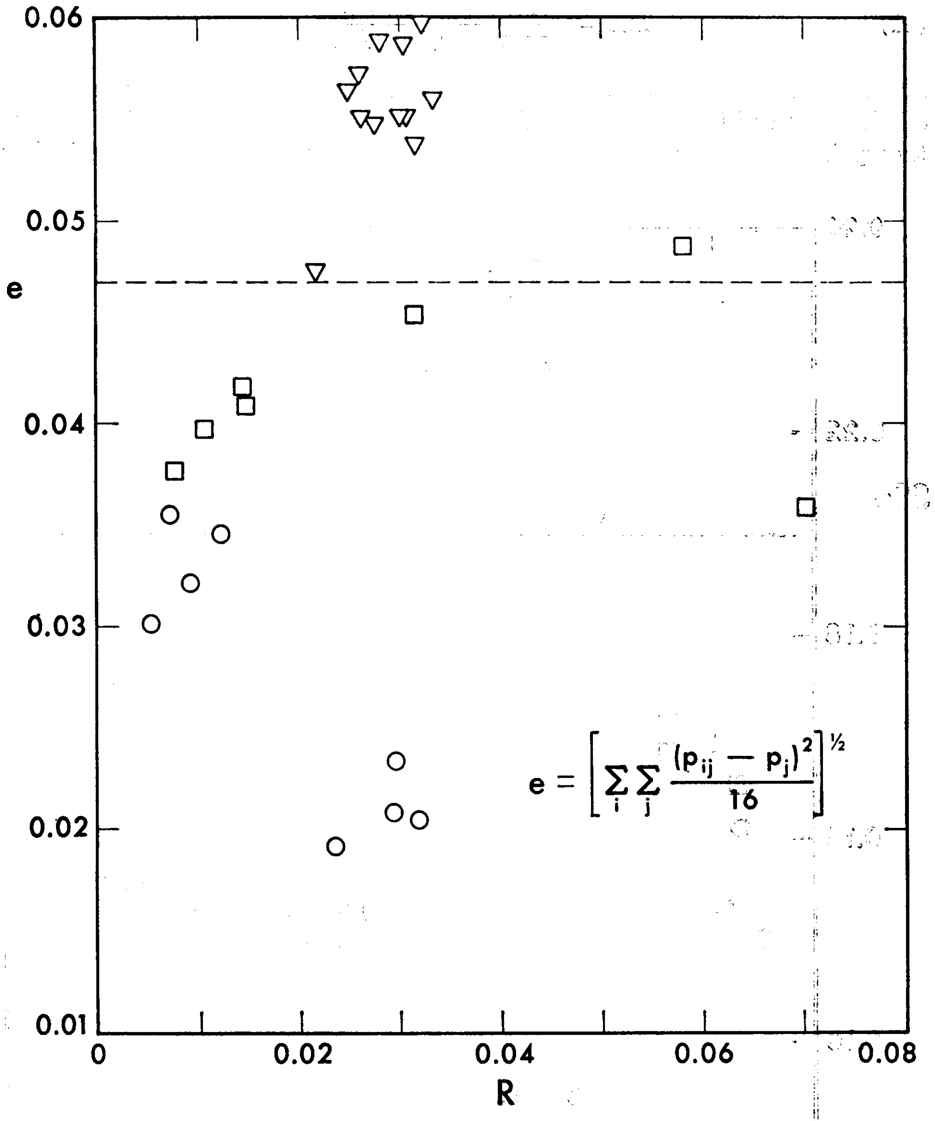

FIGURE 5

R versus e.

vertebrate and lower organism domains with only a very slight overlap at the boundary.

The experimentalist, Subak-Sharpe [12], who has worked extensively with the nearest neighbor data, has an intuitive, algorithmic procedure by which he analyzes the data. I have summarized his algorithm and defined the following evolutionary index:

$$SSe = \left\{\sum_i \sum_j \frac{1}{16}\left(1 - \frac{P_{ij}}{P_j}\right)^2\right\}^{1/2}. \tag{32}$$

Figure 6 is a plot of R versus SSe. The SSe index duplicates the result of Smith's e index. Both of these classical indices come close to mimicing the information theory index D_2. However, they are by no means mathematically equivalent

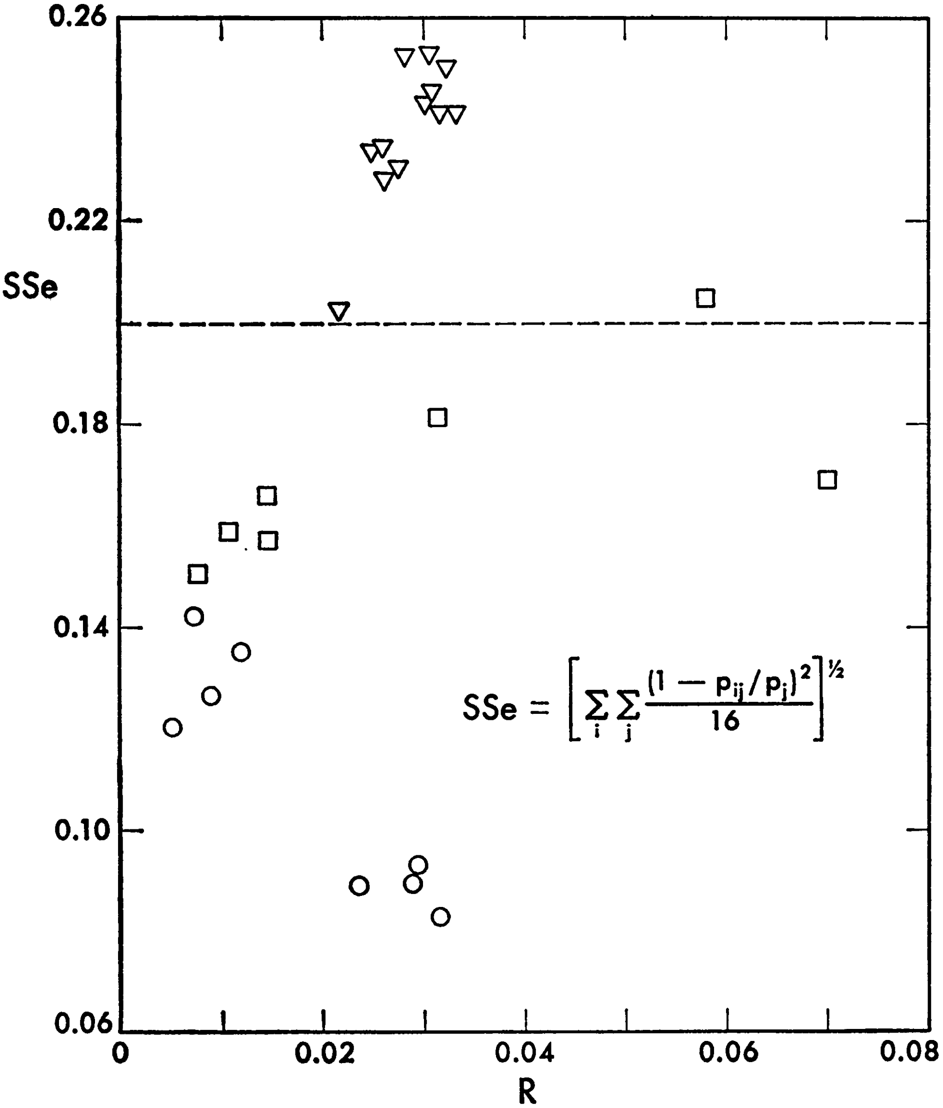

FIGURE 6

R versus SSe.

because nowhere either in the definition of e or SSe does the concept of entropy enter in, along with its inevitable logarithmic functional form.

The index D_2 is slightly better quantitatively than either of the classical indices but this is not the primary reason why D_2 is vastly superior as an evolutionary index. It is an entropy function and because of the structure with which it endows the entropy concept we are left not with just an isolated arbitrary result as we would have been with the classically defined indices, but with an explanation of our result and a workable theory which allows us to explore a vast and new conceptual area.

5. Shannon's second theorem

Let us speculate on the evolutionary implications of our observations. Let us assume that the first DNA molecules assembled in the primordial soup were random sequences, that is, D_2 was zero, and possibly also D_1. One of the primary requisites of a living system is that it reproduce itself accurately. If this reproduction is highly inaccurate, the system has not survived. Therefore, any device for increasing the fidelity of information processing would be extremely valuable in the emergence of living forms, particularly higher forms.

Redundancy, or information density, is a measure of all the constraints placed upon a sequence which make possible error detection and correction. Therefore, redundancy is in this sense a measure of the fidelity of a message. Lower organisms first attempted to increase the fidelity of the genetic message by increasing R primarily by increasing D_1, the divergence from equiprobability of the symbols. This is a very unsuccessful and naive technique because as D_1 increases, the potential message variety, the number of different words that can be formed per unit message length, declines. This is not difficult to show (Gatlin [3]) and in the limit at the maximum divergence from equiprobability, we would have the distribution where one of the p_i is one and all the rest are zero. This is a monotone, a sequence of only one letter which has no message variety at all. Hence, the lower organisms which have achieved R values in the vertebrate range or above have purchased them at the expense of a reduction in potential message variety. This is why they have remained "lower" organisms.

A much more sophisticated technique for increasing the accuracy of the genetic message without paying such a high price for it was first achieved by vertebrates. First they fixed D_1. This is a fundamental prerequisite to the formulation of any language, particularly more complex languages. We observe it in human languages. The particular distribution of the single letter frequencies in human language is so stable and characteristic of a given language that this is a fundamental tool used by cryptographers in decoding messages. When a cryptographer is faced with an unknown message, he first begins to count the single letter frequencies. If the message is in English, the letter e will always be the most frequently occurring providing the text is of sufficient length. The distribution of the P_i on S_1 is stable and characteristic of a given language. The vertebrates were the first living organisms to achieve the stabilization of D_1, thus laying the foundation for the formulation of a genetic language. Then they increased D_2 at relatively constant D_1. Hence, they increased the reliability of the genetic message without loss of potential message variety. They achieved a reduction in error probability without paying too great a price for it, and an information theorist would recognize this as the utilization of Shannon's second theorem, the coding theorem of information theory.

This kind of sophisticated reduction in the error of a message was first set forth in the second theorem of Shannon [10] which states that under certain conditions it is possible to reduce the error of a message to an arbitrarily small value even in a noisy channel and without reduction in transmission rate provided that the message has been properly encoded at the source. This statement still reflects the jargon of the communications engineer, but the second theorem principle is a broad, fundamental principle which can be stated in many ways. Let us state it in the language of the biologist.

It is possible within limits to increase the fidelity of the genetic message without loss of potential message variety provided that the entropy variables change in just the right way, namely, by increasing D_2 at relatively constant D_1. This is what the vertebrates have done. They have utilized Shannon's second theorem. This is why we are "higher" organisms.

6. Language in general

In review, the theory upon which the definition of D_1 and D_2 is based is perfectly general and could be applied to language in general. Then we observed in the genetic language the increase of D_2 at constant D_1 as a fundamental mechanism for increasing the fidelity of the genetic message. Now I ask the question: Is this mechanism a general mechanism for increasing the fidelity of any message? Is it used anywhere in human language? It is.

The human mind is an information processing channel, the most complex in the universe, and like any channel possesses a certain capacity, an upper limit to the rate at which it can receive and process information. If information is transmitted at a rate which overloads this capacity, the result is *not* that an amount of information up to the channel capacity is received and processed and the rest "spills over." The result of overloading the channel is utter confusion and chaos. Any good teacher knows this, and very carefully and with deliberation lays a firm foundation of fundamentals before increasing the rate of transmission of information to the student. It is extremely important in the initial stages of this process that error is held to an absolute minimum. Therefore, any device for increasing the fidelity of a message is extremely useful.

One of the most important learning processes which the human mind undergoes is when a little child learns to read the written language. He has spoken it for several years before he learns to read it and this is a major advancement. It is obvious that any safeguards against error in the early critical stages of this learning process would be invaluable.

I shall now show that the writers of children's textbooks intuitively utilize the basic device of increasing D_2 at constant D_1 to increase the fidelity of the message. I selected a series of well-known children's readers beginning with the primer and continuing through the sixth grade. The series selected is the Ginn Basic Reader series (Ginn and Company, New York). I calculated the redundancy of each book by taking texts of increasing length until the R value stabilized taking into consideration only the first order Markov effect. Figure 7 is a plot of R versus the grade of the reader. The R value is quite high in the primer and follows a very smoothly declining curve as the grade of the reader increases. In Figure 8, I have taken the R value apart into D_1 and D_2. It is very apparent that the high R value in the early readers has been achieved by increasing D_2 at constant D_1, just like the vertebrates. Therefore, this appears to be a fundamental mechanism, a general mechanism, for increasing the fidelity of a message.

7. Second theorem selection

We must now inquire into the detailed evolutionary mechanisms whereby the vertebrates have achieved a DNA message with higher R of the high $RD2$ type. The fundamental underlying mechanism is natural selection; but it is a different type of selection than we have considered previously. To define this type of selection we must be more explicit about the jargon of the communications engineer. This is diagrammed in Figure 9.

Here we have an information processing channel. The source or transmitter is just any mechanism for generating a sequence of symbols. The encoding of a message in a particular language occurs at the source. The channel is simply any medium over which the message is transmitted and finally received at the output of the channel. Conceptually, it is just anything one regards as intermediate between the transmitter and receiver, and hence may be sometimes somewhat a matter of definition. I define the base sequence of DNA as the encoded message at the source of the living channel and the amino acid sequence of proteins as the message which is finally received at the output. This is, of course, in a different language. The channel consists of the entire mechanics of protein synthesis which we know a great deal about today due to the massive experimental efforts expended in this area.

All evolutionary thought to date has focused its attention primarily upon the output of this channel, the protein. Natural selection acts because of the sequence of amino acids in proteins. Even the so-called "non-Darwinian" theories of evolution which have arisen recently still focus their attention on the output of the channel and it is here that they search for the reason why a mutation is selectively neutral, the ultimate reason being that the amino acid in the protein is not critical to the function of the protein.

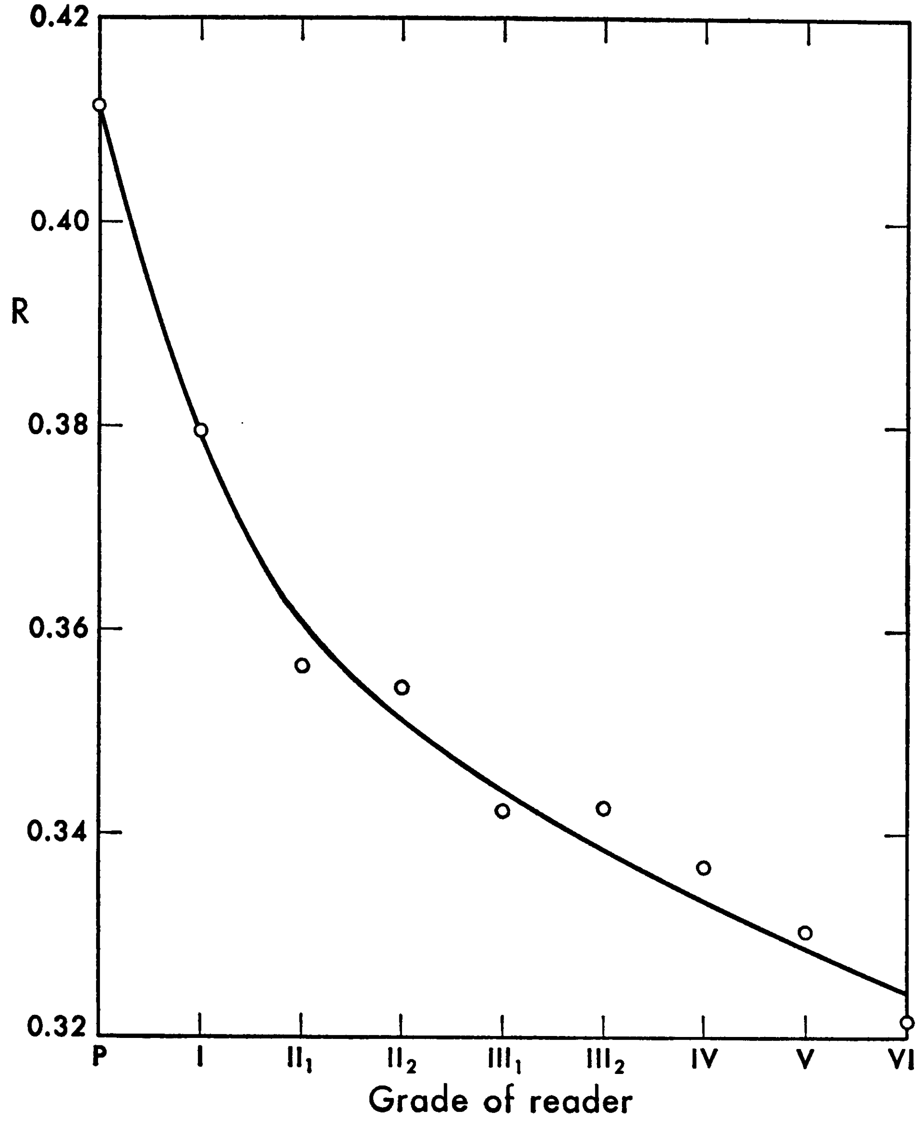

FIGURE 7

R versus grade of reader.

One can pick up any paper in the evolutionary literature, particularly the more recent ones, and confirm this preoccupation with the output of the channel. For example, I quote from Ohta and Kimura [9]: "From the point of view of survival probability, the amino acid substitution between a particular pair has a certain average probability of being accepted by natural selection." Even survival

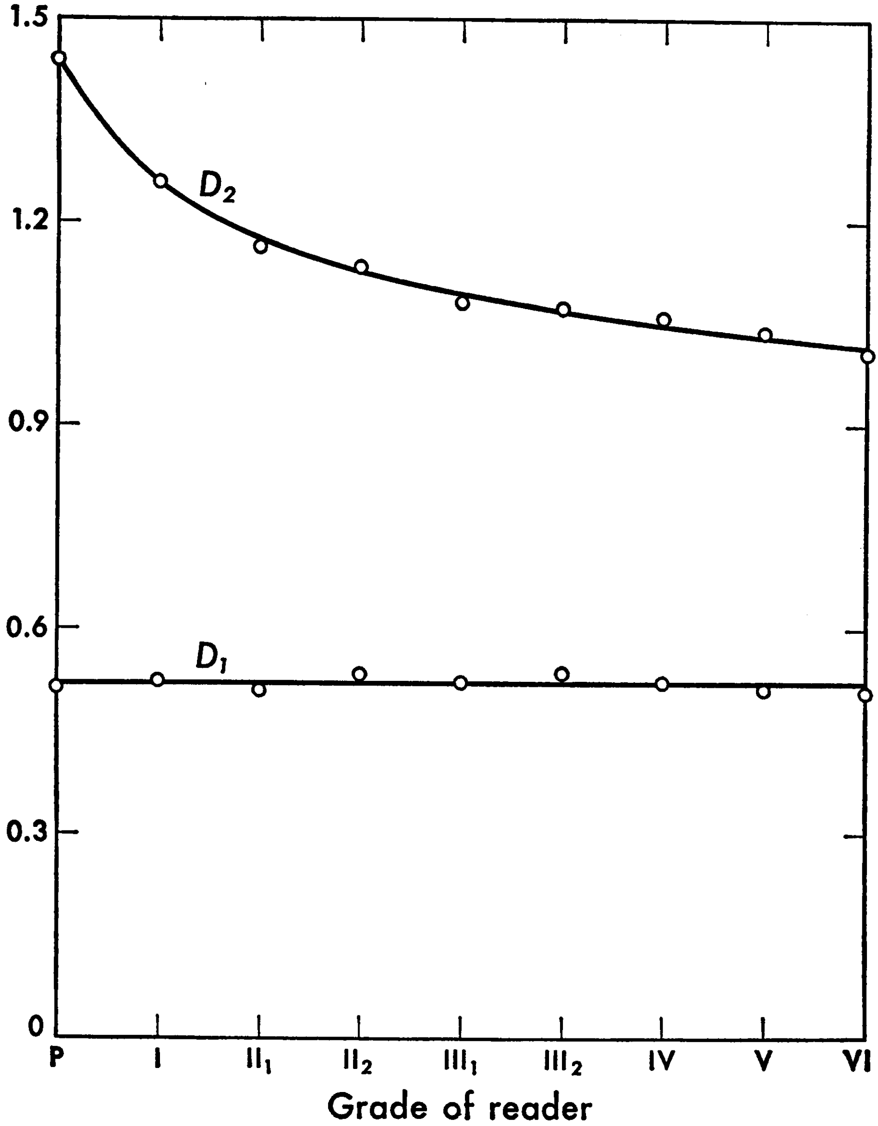

FIGURE 8

D_1 and D_2 versus grade of reader.

probabilities are conceived of in terms of amino acid substitutions in the protein at the output of the channel.

I wish to consider a new type of selection which I shall call second theorem selection because this is the basic principle under which it acts. Second theorem selection directs our attention for the first time to the input of the channel. I

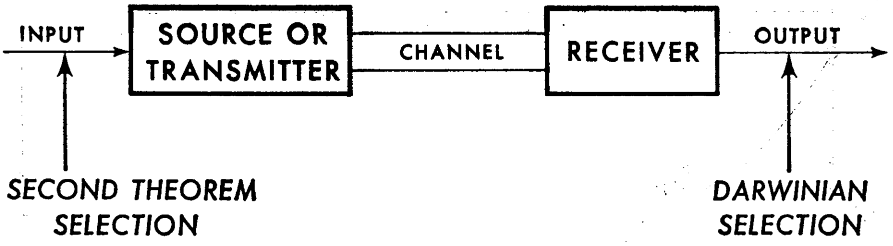

FIGURE 9

The engineer's diagram.

define second theorem selection as natural selection which acts not because of the sequence at the output but because of the informational efficiency with which this sequence has been encoded at the source under the second theorem principle.

As vertebrates have evolved, they have selected for DNA sequences at the input of the channel with a higher information density of the high $RD2$ type because these sequences have a lower probability of error in the information processing channel, and they achieve this higher measure of fidelity without paying an excessive price for it. This type of selection is made possible because of the extensive degeneracy of the genetic code.

We know that several codons can code for the same amino acid. This means that for a given protein message at the output of the channel there are a large number of possible DNA sequences at the input of the channel all of which could code for it. Under current concepts these sequences are all selectively neutral. I quote from King and Jukes [8]: "Because of the degeneracy of the genetic code, some DNA base-pair changes in structural genes are without effect on protein structure. . . . As far as is known, synonymous mutations are truly neutral with respect to natural selection." This is not the case with respect to second theorem selection. The different DNA sequences coding for the same amino acid sequence could have significantly different R and D values, and hence different probabilities of error in the channel.

I have calculated the R and D values for a set of DNA base sequences, all of which could code for the same amino acid sequence in protein (Gatlin [2]). I chose arbitrarily a sequence of equiprobable, independent amino acids in protein and constructed from the genetic code, a dozen arbitrary types of DNA base sequences, all of which could code for this same amino acid sequence. For this small sample of DNA sequences the R value ranged from 0.021 to 0.224, a variation of 20.3 per cent of the entire theoretical range of R. This is a very significant variation. The $RD2$ values ranged from 0.58 to 0.97 which is very close to the vertebrate range of $RD2$ values. Thus, there is adequate variation for second theorem selection to act upon. Therefore, second theorem selection can distinguish between different DNA base sequences all of which give rise to the same amino acid sequence in protein. This is a new concept in evolutionary thought.

Let us go on a step further. It is possible that second theorem selection can distinguish between different DNA sequences which code for slightly different amino acid sequences which are selectively neutral in the Darwinian sense.

We now believe that selectively neutral mutations are fixed by random drift (Kimura [7]). However, there are certain discrepancies between this random model and the experimental data (Jukes [5]). If we impose the concept of second theorem selection as a constraint upon this random model, perhaps this will improve the agreement. This possibility is totally unexplored.

In conclusion, we see that D_2 is not *just* another evolutionary index which can distinguish between vertebrates and lower organisms. It is an entropy function. It extends the entropy concept and endows it with structure. It defines a fundamental mechanism for increasing the fidelity of a message which we observed in the genetic language and in human language. We are led into the consideration of a new evolutionary principle which is the confluence of Darwin's principle of natural selection and Shannon's second theorem. This is an organizing principle in contrast to the disorganizing principle of thermodynamics. And finally, we are left with the rather satisfying explanation that it is the second theorem of information theory rather than the second law of thermodynamics which has given the evolution of life its unique direction.

REFERENCES

[1] F. E. Arrighi, M. Mandel, J. Bergendahl, and T. C. Hsu, "Buoyant densities of DNA of mammals," *Biochem. Genet.*, Vol. 4 (1970), pp. 367–376.

[2] L. L. Gatlin, "The information content of DNA. II," *J. Theor. Biol.*, Vol. 18 (1968), pp. 181–194.

[3] ———, *Information Theory and the Living System*, New York, Columbia University Press, 1972, in press.

[4] J. Josse, A. D. Kaiser, and A. Kornberg, "Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid," *J. Biol. Chem.*, Vol. 236 (1961), pp. 864–875.

[5] T. H. Jukes, "Comparison of polypeptide sequences," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1972, Vol. 5, pp. 101–127.

[6] A. I. Khinchin, *Mathematical Foundations of Information Theory*, New York, Dover, 1967.

[7] M. Kimura, "Evolutionary rate at the molecular level," *Nature*, Vol. 217 (1968), pp. 624–626.

[8] J. L. King and T. H. Jukes, "Non-Darwinian evolution," *Science*, Vol. 164 (1969), pp. 788–798.

[9] T. Ohta and M. Kimura, *Nature*, 1971, in press.

[10] C. Shannon and W. Weaver, *The Mathematical Theory of Communication*, Urbana, University of Illinois Press, 1949.

[11] T. F. Smith, "The genetic code: information density and evolution," *Math. Biosci.*, Vol. 4 (1969), p. 179.

[12] H. Subak-Sharpe, R. R. Burk, L. V. Crawford, J. M. Morrison, J. Hay, and H. M. Keir, "An approach to evolutionary relationships of mammalian DNA viruses through

analysis of the pattern of nearest neighbor base sequences," *Symp. Quant. Biol.*, Vol. 31 (1966), p. 737.

[13] M. N. SWARTZ, T. A. TRAUTNER, and A. KORNBERG, "Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids," *J. Biol. Chem.*, Vol. 237 (1962), pp. 1961–1967.