J.Mol.Evol.7,185-195 (1976)

Journal of Molecular Evolution

Counter-Examples to a Neutralist Hypothesis

LILA L. GATLIN*
Space Sciences Laboratory, University of California, Berkeley

Received September 16 and December 8, 1975

Summary. Specific counter-examples are derived theoretically to the hypothesis that a random amino acid composition signifies a random evolutionary process.

Key words: Evolution - Randomicity - Counter-Example

INTRODUCTION

In 1969, King & Jukes chose a sample of 53 vertebrate proteins which they regarded as representative and estimated the base composition of the DNA coding for this protein under the assumptions that degenerate codons are used equally and that the base composition at the third position of the codon was the same as the first two. Codon frequencies, p(ijk), were calculated from the relation

$$(1) \qquad p(ijk) = p_i \, p_j \, p_k$$

where the p_i are the estimated DNA base composition values. The codon frequencies were then summed under the code to give the expected amino acid frequencies in protein. These expected frequencies were plotted versus those observed as shown in Fig.1 of King & Jukes (1969) who described the agreement as "fairly good" except for arginine.

Kimura & Ohta (1971) presented a variation of this calculation in which the expected frequencies were replaced by the eigenvalues of the transition probability matrix of Dayhoff (1973) which were called "equilibrium" frequencies. Even arginine was found on the line. These authors then concluded:

* Present address: Department of Genetics, University of California, Davis, Cal. 95616, USA

Through these analyses we have been led to the view that the amino acid composition of proteins is determined largely by the existing genetic code and the random nature of base changes in evolution.
A specific counter-example to this argument is given in this study.

Although it is obvious that if the amino acid distribution, or any part thereof, differs significantly from code expectations, this implies a definite selective process; the logical point of this paper is that *even if* the amino acid distribution of a given protein or group of proteins turns out to be perfectly random, it still could have been produced from a highly non-random DNA sequence by a highly non-random selective process because of the highly biased nature of the genetic code.

Specifically this work demonstrates by means of information theoretic arguments the mathematical existence of DNA sequences with highly non-random pair correlations which nevertheless could produce not only a distribution with the same randomness as the King-Jukes sample but even perfectly random amino acid distributions.

THEORY

Perspective

For a more rigorous and complete presentation of the theory the reader is referred to Gatlin (1972, 1974). Only the concepts relevant to this work will be summarized below.

In studying the randomicity of biological sequences the problem of how to deal with relatively short sequences arises. Classical statistics have not solved this problem in any complete sense, and it is quite possible that it can only be approached with Monte Carlo methods. If we take an idealized parameter P measuring in some way departure from randomicity in the sequence, generate a large number of random sequences on a computer and plot the average value of P as a function of the length of the sequence generated, functions of the type shown in Fig.1 are often obtained.

As the sequence becomes longer, the ideal randomness upon which classical statistical concepts rest is approached asymptotically, and any parameter measuring departure from randomicity approaches zero. However, for very short sequences, $\bar{P}$ is finite, not because the generator or source is biased, but simply because the length of the sequence is so short that in this region classical statistical concepts of randomicity fail.

Not only can this generalized failure of classical statistical concepts be demonstrated for very short sequences, there is also empirical evidence that for very long DNA sequences,

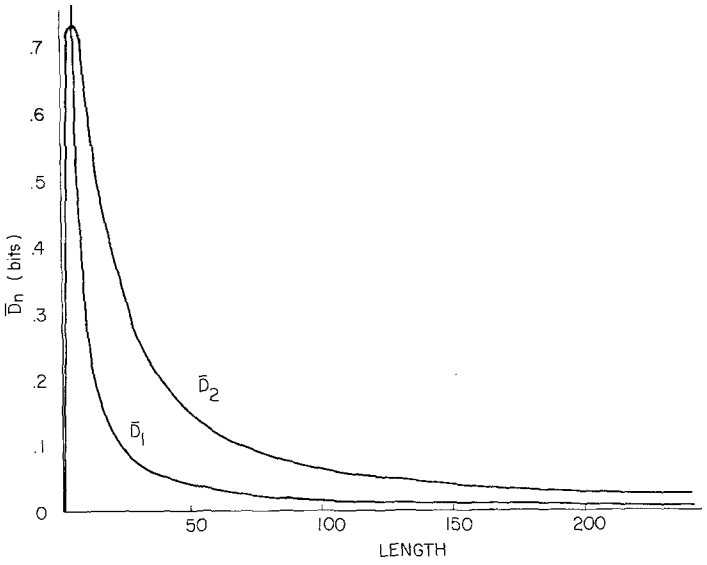

Fig.1
$\bar{D}_1$ and $\bar{D}_2$ for 100 iterations versus length for a random DNA source from which the STOP codons have been excluded

statistical parameters are not as sensitive and meaningful in demonstrating evolutionary relationships among living organisms (see p. 85 of Gatlin, 1972). For these reasons this work builds on Monte Carlo methods and information theoretic parameters as the basic tools in studying the randomicity of biological sequences. This is not to say that statistical tools are not useful, but rather that a blend of both statistics and information theory is a more powerful tool than either of them separately.

Information Parameters

A source is any apparatus or process that emits a sequence of symbols from a specified alphabet according to specified probabilities of emission of the single letters, doublets, triplets, etc. A stationary source is characterized by the time invariance of its n-tuple emission probability distribution; and for an ergodic source, which is a special kind of stationary source, this distribution is unique.

Since a random source is totally unconstrained, the emission probabilities of all n-tuples are equal for a given n. For example, all single letters are emitted with equal a priori probability as are all doublets, triplets, etc. We may describe the departure from this completely random state in the following hierarchical manner.

The divergence from equiprobability of the single letters in the sequence is given by

(2) $$D_1 = \log a - H_1$$

where a is the number of letters in the alphabet and H_1, the zero memory entropy of the source, is given by

$$(3) \qquad H_1 = - \sum_{i=1}^{a} p_i \log p_i$$

where p_i is the probability of occurrence of letter i in the sequence.

The divergence from equiprobability of the doublet sequences beyond that fixed by D_1 is given by

$$(4) \qquad D_2 = H_1 - H_M$$

where H_M is the entropy of a Markov source of memory one. In general, the divergence from perfect randomicity is hierarchically structured and the general increment of divergence is given by

$$(5) \qquad D_{m+1} = H_M^{(m-1)} - H_M^{(m)}$$

where m is the memory of the Markov source and D_{m+1} is the departure from randomicity at the level of an n-tuple of length m+1.

The total departure from ideal randomicity is related to Shannon's (1949) redundancy, R, by

$$(6) \qquad R \log a = D_1 + D_2 + D_3 + \ldots D_{m+1}$$

A random sequence of a given length will be defined as one where any D_n value lies outside $\pm 2\sigma$ of the average value $\bar{D}_n$ for that length from a random source. This mean value and its standard deviation can be determined by the Monte Carlo methods described previously (Gatlin, 1974, 1975).

THE UNCONSTRAINED DNA-TO-PROTEIN CHANNEL

The process of protein synthesis may be regarded as an information processing channel with the base sequence of DNA at the input, the amino acid sequence of protein at the output, and, connecting the two, the transfer function of the genetic code. If a completely unconstrained or random model of this process can be constructed, then various evolutionary constraints upon the channel can be studied separately and quantitatively.

Table 1. Means and σ for $\bar{D}_1$ $\bar{D}_2$ and $\bar{R}$ for a random[a] DNA source

	$\bar{D}_1$ (bits)	$\bar{D}_2$ (bits)	$\bar{R}$
L = 30	0.06846	0.25035	0.15940
σ	0.05307	0.10217	0.05783
L = 60	0.03586	0.11771	0.07678
σ	0.03041	0.05222	0.03101
L = 90	0.02849	0.07268	0.05059
σ	0.02509	0.03478	0.02122
L = 120	0.01921	0.05751	0.03836
σ	0.01469	0.02336	0.01379
L = 150	0.01564	0.04493	0.03029
σ	0.01234	0.02447	0.01366
L = 180	0.01122	0.03547	0.02334
σ	0.00830	0.01519	0.00856
L = 210	0.01120	0.03509	0.02314
σ	0.00992	0.01636	0.00933
L = 240	0.01079	0.02876	0.01977
σ	0.00883	0.01377	0.00801
L = 270	0.00755	0.02557	0.01656
σ	0.00627	0.01065	0.00615
L = 300	0.00868	0.02345	0.01606
σ	0.00587	0.01081	0.00596

[a]The STOP codons have been excluded.

Fig.1 is a plot of $\bar{D}_1$ and $\bar{D}_2$ as a function of length for a random DNA source from which the STOP codons have been excluded. Such DNA sequences could code for real proteins. Table 1 lists the average values of the parameters and their standard deviations as a function of length. These sequences can then be mapped to protein via the code and the same parameters calculated for the protein sequences. The protein functions are reported elsewhere (Gatlin, 1974). This procedure constitutes a mathematical model of a completely unconstrained DNA-to-Protein channel under no other regulation except the genetic code. If real DNA or protein sequences in the given length ranges have observed parameter values outside ± 2 σ of the random channel, these sequences are non-random with respect to the information parameters, which constitutes quantitative evidence of evolutionary constraints on the protein synthesis process.

In real systems these constraints must be extremely complex. Our objective here is to find a reasonably realistic set of constraints which will allow the location of a unique DNA informational state at the channel input which is significantly non-random but which, under our biased genetic code, can give rise to a random amino acid distribution at the output.

In particular the second level of departure from randomicity as measured by D_2(DNA) is of special interest since it is obvious that there is a spectrum of departure from randomicity at the D_1(DNA) level in naturally occurring DNA, particularly in lower organisms where the base composition may range anywhere from approximately 0.20 - 0.80% (C+G). D_1(DNA) measures only this compositional non-randomness but D_2(DNA) is the first measure of *sequential* non-randomness.

At the output of the channel the randomicity of the amino acid distribution can be monitored directly by H_1(P), the zero memory entropy of protein. Therefore we seek a channel state where D_2(DNA) is outside the limits of a random channel but H_1(P) is not.

CONSTRAINED CHANNELS

Smith's Channel

Temple F. Smith (1969) was the first to suggest an algorithm which performs this feat of monitoring informational states simultaneously at both the input and output of the DNA-to-Protein channel.

Smith assumed a Watson-Crick constraint, i.e., C=G, A=T, leaving only one free DNA variable, which is customarily expressed as % (C+G) and is varied systematically over its natural range of approximately 0.20 - 0.80% (C+G). Smith next assumed D_n(DNA) = 0 for $n \geqslant 2$ and calculated the codon frequencies from Eq.1. The codon frequencies for a given amino acid were then summed under the code to give the corresponding amino acid frequencies in protein from which H_1(P) was calculated.

A plot of H_1(P) versus % (C+G) displays an absolute maximum at approximately 42% (C+G) which coincides with the natural base composition range of all vertebrates (Sueoka, 1965). Since H_1(P) is a direct measure of amino acid variety, approximately 42% (C+G) is an informationally optimal base composition under the code because it permits maximal freedom of choice, in an evolutionary sense, of a wide variety of amino acids in protein.

The general significance of Smith's calculation for our purposes is that the device of monitoring H_1(P) as a function

of % (C+G) can be freed of many of the restrictive constraints Smith used and developed as a more general informational tool.

A Doublet-Variable Constrained Channel

Let us place at the input of the DNA-to-Protein channel 16 free variables representing the 16 DNA doublet frequencies. The normalization condition leaves 15 analytically independent variables without further constraint. Beginning with any arbitrary initial distribution the codon frequencies are calculated from the relation

$$(7) \qquad p(ijk) = \frac{p(ij)\ p(jk)}{p_j}$$

where $p(ij)$ is the doublet probability and p_j is the singlet probability. This relation essentially assumes $D_n(\text{DNA}) = 0$ for $n \geq 3$ but leaves D_1 and D_2 free to vary. The codon frequencies are summed as in Smith's algorithm and $H_1(P)$ calculated. A Watson-Crick constraint is imposed and this procedure is iterated while constraining the p_j to a given base composition value until the maximum value of $H_1(P)$ is reached. The global optimization program of Bremerman (1970) was adapted for this algorithm.

In channel calculations of this type the STOP codons can be included or omitted. In this study they have been omitted and all other codon frequencies renormalized since we wish to model real genes coding for protein.

The p_j must be calculated from the $p(ij)$ according to the standard summation

$$(8) \qquad p_j = \sum_i p(ij)$$

However, in any arbitrary matrix of doublet frequencies the row-column sums for a given base do not always agree, i.e., the matrix is non-ergodic and

$$(9) \qquad \sum_i p(ij) \neq \sum_i p(ji)$$

In such a case the doublet frequencies can be adjusted until these contradictory sums agree. Let us call this adjustment the E-constraint which is then applied at each iteration of the algorithm.

Fig.2 is a plot of $H_1(P)^{\text{Max}}$ versus % (C+G) for the above algorithm. The constraints are: (1) Watson-Crick symmetry (2) E-adjustment (3) $D_n(\text{DNA}) = 0$ for $n \geq 3$ and (4) STOP codons omitted.

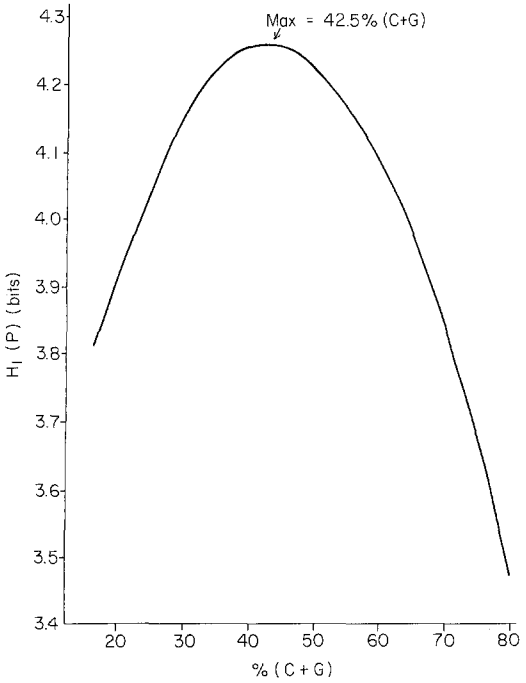

Fig.2
$H_1(P)$ versus % (C+G) for the con strained doublet-variable channe described in the text

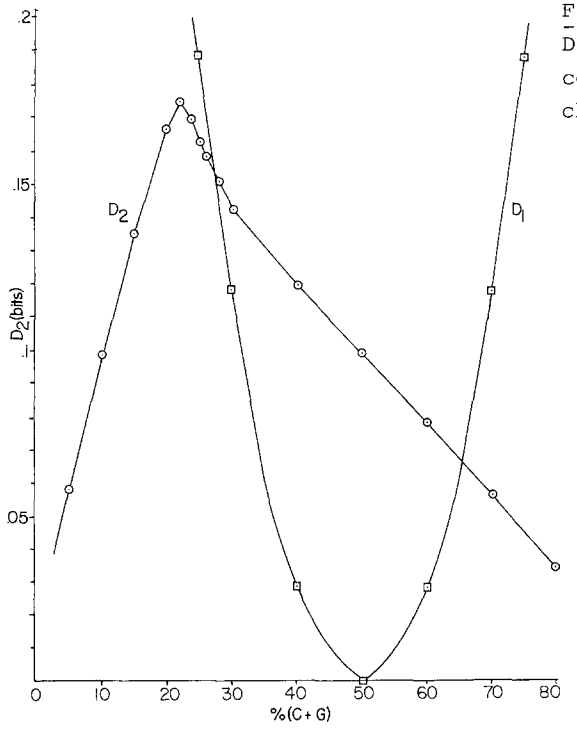

Fig.3
$\bar{D}_1$ and $\bar{D}_2$ versus % (C+G) for the constrained doublet-variable channel described in the text

The result of particular interest is that regardless of initial conditions or random perturbation of the final solutio the program always returns rapidly to an apparently unique sol ution set of DNA doublet frequencies from which D_1(DNA) and D_2(DNA) can be calculated. Fig.3 is a plot of D_1(DNA) and

Table 2. Information parameters at both the input and output of the constrained doublet-variable DNA-to-Protein channel

% (C+G)	D_1(DNA) (bits)	D_2(DNA) (bits)	R(DNA)	H_1(P) (bits)
10	0.531	0.099	0.315	3.480
20	0.278	0.167	0.223	3.923
30	0.119	0.143	0.131	4.159
40	0.029	0.120	0.075	4.251
42.5	0.016	0.115	0.066	4.254 [a]
50	0.000	0.100	0.050	4.223
60	0.029	0.079	0.054	4.088
70	0.119	0.057	0.088	3.844
80	0.278	0.034	0.156	3.478
90	0.531	0.015	0.273	2.943

[a] Absolute maximum in H_1(P).

D_2(DNA) versus % (C+G) corresponding to the maximum in H_1(P) at the given base composition. Table 2 lists the information parameters at both the input and output. This calculation achieves our objective of locating a spectrum of unique informational states of DNA where D_2(DNA) $\neq$ 0 and their corresponding amino acid distributions. Let us now examine the randomicity of these informational states.

THE COUNTER-EXAMPLES

The value of H_1(P) for the King-Jukes (1969) sample is 4.200 bits. By interpolation from Table 2 this corresponds to a value of D_2(DNA) = 0.133 bit at the channel input. The departure from randomicity of this value may be evaluated with the question: "At what length of DNA sequence does this value of D_2(DNA) become statistically significant?" Since the critical values of D_2 decrease with increasing length, at any sequence length greater than this minimum significant length the specified D_2 value would only become more and more non-random.

From Table 1 we find that the value of D_2(DNA) corresponding to the King-Jukes sample becomes significant at lengths of only about 100 DNA bases. This means that for sequences as long as the genomes of the shortest viruses these values represent highly non-random DNA sequences. Let us check this result more carefully.

Let us take an ideally random amino acid distribution where the frequency of each amino acid is given by $n_i/61$ where n_i is the number of codons coding for that particular amino acid.

Table 3. The counter-examples

	$H_1(P)$ (bits)	D_2(DNA) (bits)	Minimum significant length
King-Jukes sample	4.200	0.133	≈100
Ideal randomness	4.139	0.147	≈ 90
Unconstrained $H_1(P)^{Max}$	4.300	0.201	≈ 80

The second line in Table 3 lists the values for this distribution. They become significant at lengths of only about 90. This result shows vividly how, under the code, the same random amino acid distribution can be obtained from either a perfectly random channel or from a constrained or non-random channel.

This result is not highly dependent on the nature of the constraints chosen. For example, let us relax the Watson-Crick constraint and the E-constraint. The p_i may be calculated as the average of the row-column sums when they disagree. We may still calculate a free maximum in $H_1(P)$ which occurs at $H_1(P)$ = 4.300 bits, % (C+G) = 43.27. The corresponding value of D_2(DNA) 0.201 bit for which the minimum significant length is approximately 80.

The order of magnitude of the minimum significant lengths in Table 3 is surprisingly low. They show how drastically non-random DNA sequences at the input can become and yet, under our biased code, still give rise to completely random or nearly random amino acid distributions at the output. The values in Table 3 constitute quantitative counter-examples to the original concept of King & Jukes (1969) and Kimura & Ohta (1971) that a random amino acid distribution constitutes evidence for the random nature of base changes in DNA.

Acknowledgement. I thank the Department of Applied Science, University of California, Davis, for use of the CDC 3400.

REFERENCES

Bremermann, H.J. (1970). Math.Biosci. 9, 1

Dayhoff, M.O. (1973). Atlas of protein sequence and structure. Silver Spring, Md.: Natl.Biomed.Res.Found.

Gatlin, L.L. (1972). Information theory and the living system. New York: Columbia Univ.Press

Gatlin, L.L. (1974). J.Mol.Evol. 3, 189
Gatlin, L.L. (1975). J.Mol.Evol. 6, 147
Kimura, M., Ohta, T. (1971). In: Proc.6th Berkeley Symp.on Math.Stat.and Prob., L.M. Le Cam, J. Neyman, E.L. Scott, eds. Berkeley: U.C. Press
King, J.L., Jukes, T.H. (1969). Science 164, 788
Shannon, C.E. (1949). The mathematical theory of communication. Urbana: Univ.of Ill.Press
Smith, T.F. (1969). Math.Biosci. 4, 179
Sueoka, N. (1965). In: Evolving genes and proteins, V. Bryson, H.J.Vogel, eds. New York: Academic Press