THE AMOUNT OF INFORMATION STORED IN PROTEINS AND OTHER SHORT BIOLOGICAL CODE SEQUENCES

THOMAS A. REICHERT
CARNEGIE-MELLON UNIVERSITY

1. Introduction

These remarks were made to the conference assembly in a context not unlike that of a surprise witness for the defense. This work was so hot off the press that there had been no time to communicate it before the conference itself. I am grateful to Dr. Lila Gatlin for the opportunity to make this presentation. All of the work to be discussed here has been done in collaboration with A. K. C. Wong, also of the Biotechnology Program at Carnegie-Mellon University.

In the last year, we have developed a measure of the amount of information required to perform genetic mutations, together with an algorithm utilizing these measures, for aligning amino acid and RNA code sequences [9], [11]. In the process of this development, we attempted to calculate the amount of information which was stored in a protein's amino acid sequence. We had, at the time, only the tools of the conventional communications form of information theory. Thus, we attempted the calculation using the two expressions:

$$H = -\sum_{i=1}^{a} p(i) \log p(i) \tag{1}$$

or

$$I_{\text{self}} = -\sum_{i=1}^{a} n_i \log p(i). \tag{2}$$

Equation (1) is the expression for the entropy of a discrete information source operating with an alphabet of a letters. This quantity is also interchangeably called the information content of such a source. Since information and entropy are more nearly opposites than synonyms, this equivalence has always been confusing. Indeed, the values of H obtained for a set of cytochrome c sequences, by allowing the amino acid frequencies in each sequence to determine the alphabet character probabilities used in equation (1), displayed the then embarrassing trend of higher information content with lower organism complexity. Since the method used to estimate these probabilities is not generally known, let me describe it here.

It was Laplace, I believe, who first noted that the frequency limit estimate of the probability of an event's occurrence was applicable only in the limit of infi-

nite frequency. In the anecdote reported by many biographers concerning the probability of the morrow's sunrise, he showed the importance of the real difference from this limit. Laplace's formula for the probability of the sun's not rising was $p = 1 - (x + 1)/(n + 2)$, where x is the number of times the sun has previously risen and n is the number of times a new day has presented itself (x apparently $= n$).

R. A. Christensen [2] has formally generalized this estimate. Probabilities so constructed are called relatively unbiased probabilities and are defined by the expression

$$p(i) = \frac{x_i + t}{n + t + f}, \tag{3}$$

where x_i is the number of occurrences of the event i (successes), n is the total number of occurrences of any event (trials), t is the number of possible different realizations of the event i, and f is the number of events in the event space which are not realizations of the event i.

In the absence of any experimental data, this expresssion reduces to the condition of maximum ignorance $p(i) = t/(t + f)$, and assumes the frequency limit form as the data acquisition proceeds to infinity.

This formula has proved especially useful in our case because occasionally one or more amino acids would be completely absent from a particular sequence, and the resulting zero frequency would otherwise have been difficult to handle in a fashion not arbitrary.

Equation (2) is the sum of the self information associated with each character in a sequence. We reasoned that since H is simply the average value of the self information of the particular alphabet in use, $I = -\log p(i)$; then, if we were to use the entire ensemble of homologous sequences to determine the character probabilities, the total self information for each sequence would give us the elements of the distribution having H as its first moment. The set of ensemble based probabilities characterize what we call the "super source" for the particular protein. The values of I_{self} so obtained were, if anything, more blatant in the inverse correlation of information content with intuitive notions of complexity. Faced with this apparently incomprehensible result, we took the only course open to conscientious scientists, and placed the results in an appendix to an overlong paper where it was referred to only obliquely.

2. Formulae

The matter hung thus in limbo until we discovered the work of L. L. Gatlin [6], [7], [8]. She has explained how, in her formulation, the information storage ability of an information source is measured by the deviations of its entropy from the theoretical maximum. For a source whose next emission depends only on the last character emitted, a first order Markov source, the information density I_d is given by $I_d = H_{max} - H_{Markov}$. This difference she has further decomposed into two components

$$D_1 = H_{\max} - H_1, \qquad D_2 = H_1 - H_{\text{Markov}}, \tag{4}$$

where D_1 is the deviation of the source entropy from equiprobability and D_2 is its deviation from independence.

DNA sequences are of enormous length, effectively infinite, so that the source information density and the average information per sequence character are essentially identical quantities. The short sequences of proteins are, however, another matter. You will remember that H_{Markov} is given by

$$\sum_i p(i) \sum_j p(j/i) \log p(j/i), \tag{5}$$

where $p(j/i)$ is the probability of occurrence of the $(i - j)$th pair of characters.

The probabilities, both marginal and conditional, are determined by the entire ensemble of homologous sequences so that H_1, H_{Markov}, and thereby I_d would, utilizing the two expressions given above, have the same value for every such sequence. The key to this dilemma lies in recalling that H_1 and H_{Markov} are both averages of the form $\langle f(i) \rangle = \sum_i p(i) f(i)$. Replacing this formulation by that for obtaining a simple mean of the correspondent self information measures, we obtain, for equations (1) and (5),

$$\bar{I}_{\text{self}_1} = -\sum_i \frac{n_i}{\sum_i n_i} \log p(i) = -\frac{1}{N} \sum_i n_i \log p(i),$$

$$\bar{I}_{\text{self}_M} = -\sum_i{}' \frac{n_i}{\sum_i{}' n_i} \sum_j \frac{n_{ij}}{\sum_j n_{ij}} \log p(j/i) \tag{6}$$

$$= -\frac{1}{N} \sum_i{}' \frac{n_i}{\sum_j n_{ij}} \sum_j n_{ij} \log p(j/i),$$

where n_{ij} is the number of $i \to j$ directed pairs in the sequence. The $\sum'$ indicates that only those residues which can form pairs are to be included in the average. The last residue in the sequence has no following residue. Thus, $\sum_i' n_i = \sum_i n_i - 1 = N'$, and

$$\bar{I}_{\text{self}_M} = -\frac{1}{N} \sum_i{}' \sum_j n_{ij} \log p(j/i), \tag{7}$$

where $\bar{I}_{\text{self}_1}$ and $\bar{I}_{\text{self}_M}$ are the short sequence analogs of Gatlin's H_1 and $H_{M=\text{Markov}}$. Thus, we may define the analogous deviations

$$D_1 \triangleq \log a - \bar{I}_{\text{self}_1}, \qquad D_2 \triangleq \bar{I}_{\text{self}_1} - \bar{I}_{\text{self}_M}. \tag{8}$$

3. Applications

3.1. *Cytochrome c.* When these measures are assembled for the set of available cytochrome c sequences (the ensemble presently contains 33 sequences), the values in Table I are obtained. Figure 1 is a plot of D_2, the deviation of the amino

TABLE I

INFORMATION MEASURES FOR CYTOCHROME *c* SEQUENCES

D_2 and R are calculated only for the 27 sequence ensemble.

Cytochrome *c*	Length	I_{self_1}	I_{self_M}	D_2	R	TD_2	TI
1. Human	104	3.998	3.017	1.008	0.3104	105.1	138.8
2. *Rhesus* monkey	104	3.993	2.985	1.042	0.3195	107.9	142.1
3. Pig and bovine	104	3.958	2.902	1.105	0.3431	112.7	150.6
4. Horse	104	3.956	2.994	1.025	0.3253	103.1	141.1
5. Donkey	104	3.969	2.94	1.093	0.3377	110.	146.7
6. Dog	104	3.956	2.861	1.138	0.3508	116.7	154.8
7. Rabbit	104	3.987	2.899	1.138	0.3433	116.	150.9
8. California gray whale	104	3.964	2.88	1.131	0.3472	115.6	152.9
9. Kangaroo	104	3.935	2.99	0.9671	0.3166	101.2	141.5
10. King penguin	104	4.007	2.962	1.116	0.3329	111.7	144.4
11. Chicken and turkey	104	4.016	3.001	1.077	0.3216	107.6	139.4
12. Pekin duck	104	4.015	2.948	1.137	0.3354	113.9	145.8
13. Pigeon	104	3.999	2.997	1.059	0.3217	107.1	140.7
14. Snapping turtle	104	3.955	3.003	0.983	0.3144	102.	140.1
15. Dogfish	104	4.05	3.14	0.9303	0.2773	97.8	126.1
16. Pacific lamprey	104	4.044	3.147	0.9318	0.2799	96.3	125.3
17. Silkworm moth	107	4.049	3.216	0.8433	0.256	92.4	121.6
18. Tobacco horn worm moth	107	4.071	3.2	0.8938	0.2626	96.3	123.2
19. Fruit fly	107	4.018	3.122	0.9254	0.2839	99.	131.5
20. Screw-worm fly	107	4.024	3.063	0.9926	0.2981	105.9	137.8
21. Wheat	112	4.110	3.44	0.4475	0.1454	78.5	102.2
22. *Neurospora crassa*	107	4.024	3.725	0.2615	0.1294	35.7	67.6
23. Baker's yeast	108	4.077	3.695	0.3914	0.1465	45.	71.4
24. *Candida krusei*	109	4.146	3.862	0.2952	0.1047	34.8	54.
25. Rattlesnake	103	4.071	3.219	0.8905	0.2631	91.	116.8
26. Tuna	104	3.996	3.153	0.8555	0.2746	90.8	124.7
27. Bullfrog	104	4.013	3.056	0.9951	0.3029	102.6	134.7
28. Mung bean	111	4.114	3.533			68.	91.1
29. Castor bean	111	4.087	3.499			68.8	94.9
30. Sesame	111	4.112	3.483			73.3	96.6
31. Sunflower	111	4.099	3.407			80.2	104.9
32. *Physarum*	108	4.132	3.531			68.4	88.9
33. *Euglena*	105	4.127	3.997			17.7	38.2

Hemoglobin α	Length	TD_2	TI
Human	141	123.3	185.5
Gorilla	141	117.6	180.9
Rhesus monkey	141	121.	184.3
Mouse C57 strain	141	107.7	168.1
Mouse NB strain	141	105.7	165.5
Rabbit	141	109.5	152.4
Horse (slow sequence)	141	118.2	183.6
Donkey	141	110.2	175.2
Sheep α strain	141	107.1	172.3
Sheep δ strain	141	105.8	170.9
Carp	142	48.3	77.1
Horse (fast sequence)	141	118.5	180.8
Bovine	141	106.1	176.6
Chicken	140	87.	125.7
Kangaroo	129	78.8	125.3

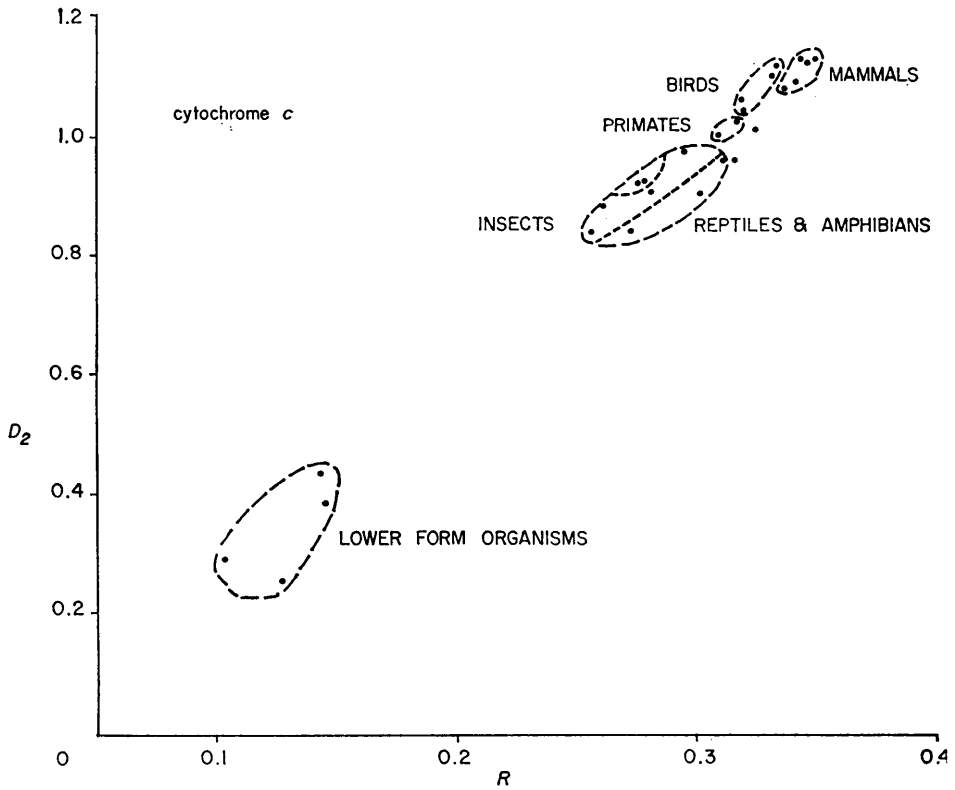

FIGURE 1

The average deviation of the amino acid sequence average self information from independence, D_2, *versus* the redundancy of that sequence for 27 species relative to the ensemble super source based on the same set of sequences.

acid code of each sequence from independence, *versus* R, the Shannon redundancy of the code based on an ensemble comprised of the first 27 cytochrome c sequences enumerated in Table 1 of [3]. The Shannon redundancy

$$R = \frac{D_1 + D_2}{\log a} \tag{9}$$

is proportional to the short sequence average information density. This plot is the short sequence analog of the figure just presented by Lila Gatlin for DNA sequences [8]. In her figure, you will remember that the DNA for the vertebrate was characterized by higher values of D_2 than that of bacteria which was, in turn, higher in D_2 than the DNA for phages. The same correlation with complexity is apparent in Figure 1. What are commonly called higher organisms exhibit higher values of D_2. There are, however, additional features present in Figure 1 which are absent in the DNA plot.

(1) The relationship appears to be linear with an intercept of zero. Unlike the DNA case, then, higher organisms are higher both in D_2 *and* in I_d.

(2) Data from many more higher organisms are available for the protein analysis so that a finer screening is possible. We note that it is possible to enclose nearly all of the members of established taxonomic groupings with convex boundaries. We hesitate, however, to ascribe taxonomic significance to these boundaries and feel that these values should be looked on more as the cytochrome *c* coordinate of the taxonomic vector of these organisms.

Since we have acknowledged that the length of our sequences is not infinite, it seems proper to ask if, in view of the fact that the average information per character is higher in higher organisms, the total amount of information stored is also different. The length of all of these sequences is tabulated in Table I. Note that lower form sequences are longer than vertebrate sequences by a significant amount. The expressions for the total information stored in such a sequence are

$$TI_1 = -\sum_{i=1}^{a} n_i \log p(i) = N\bar{I}_{\text{self}_1}, \tag{10}$$

$$TI_M = -\sum_{i=1}^{a}\sum_{j=1}^{a} n_{ij} \log p(j/i) = N'\bar{I}_{\text{self}_M}, \tag{11}$$

$$TD1 = N \log a - T1_1, \tag{12}$$

$$TD2 = TI_1 - T1_M, \tag{13}$$

and

$$TI = TD1 + TD2. \tag{14}$$

A plot of $TD2$ *versus* TI for all 33 sequences is presented in Figure 2. Note that TR is not suitable since it is length independent as seen in the analog expression $TR = TI/N \log a$.

The clusterings here become a bit more tortuous to construct. Particularly troublesome are the set of insect sequences. On the other hand, the data is even more highly linear, indicating that our earlier caution against the use of the information stored in a single protein as a monophyletic basis for taxonomy is probably well advised. The addition of the *Euglena* sequence [4] not only provides the lowest point thus far, but also makes the straight line a poor extrapolation to low values. One obvious interpretation of Figure 2 is that evolution appears to be a process of the acquisition of information. The curve best fitting the data would indicate that the lowest life forms have virtually all of their information stored as $TD1$. Thus, life would appear to have arisen merely from a properly asymmetrical distribution of amino acids with little or no interresidue correlation. This correlation, however, apparently developed quite rapidly after function first evolved. At the level of the slime mold, it would appear that the nature of the process changed, perhaps to refinement rather than development. In this mode both $TD1$ and $TD2$ are augmented at the same rate.

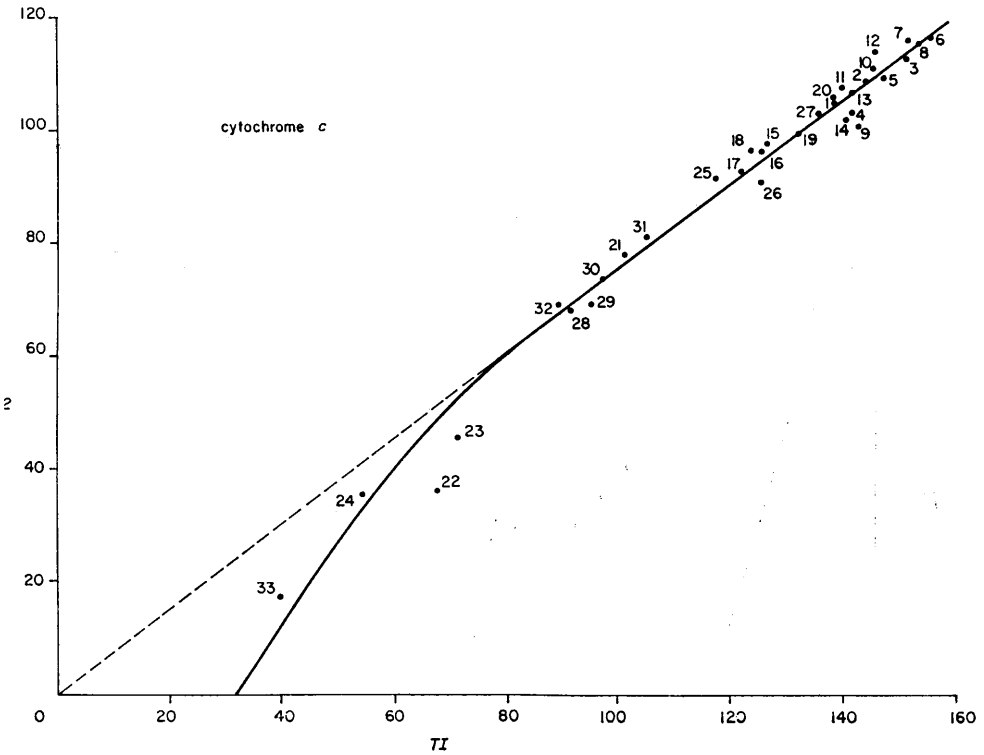

FIGURE 2

The total deviation from independence *versus* the total amount of information stored in the amino acid code of the 33 cytochrome *c* sequences of Table I. The numbers refer to the elements of the table. The slope of the limiting line is 0.75.

Figure 2 contains 33 sequence points based on the 33 sequence super source. Figure 1 contains 27 points based on the ensemble source of the first 27 sequences. This was intentional because it is necessary to examine the measures used for stability and bias.

Figure 3 is a plot of the ensemble entropies H_1 and H_M as a function of the number of sequences included in the ensemble. The sequences were added in two different orders. The approach to some stable value for H_M is clear only in the first case. From the figure, we note that H_1 stabilizes much more quickly than H_M. A curious fact relative to these measures is that those elements (both single residues and amino acid pairs) which are *most* common to the ensemble contribute most to the stored information. Thus, those sequences with the highest TI would utilize the elements most common to the ensemble super sources most often. Since the source entropies are, as noted above, the first moments of the distributions characterizing the super source, these high TI sequences will contribute the most to the final frequency distribution. Sequences, then, should be added to the

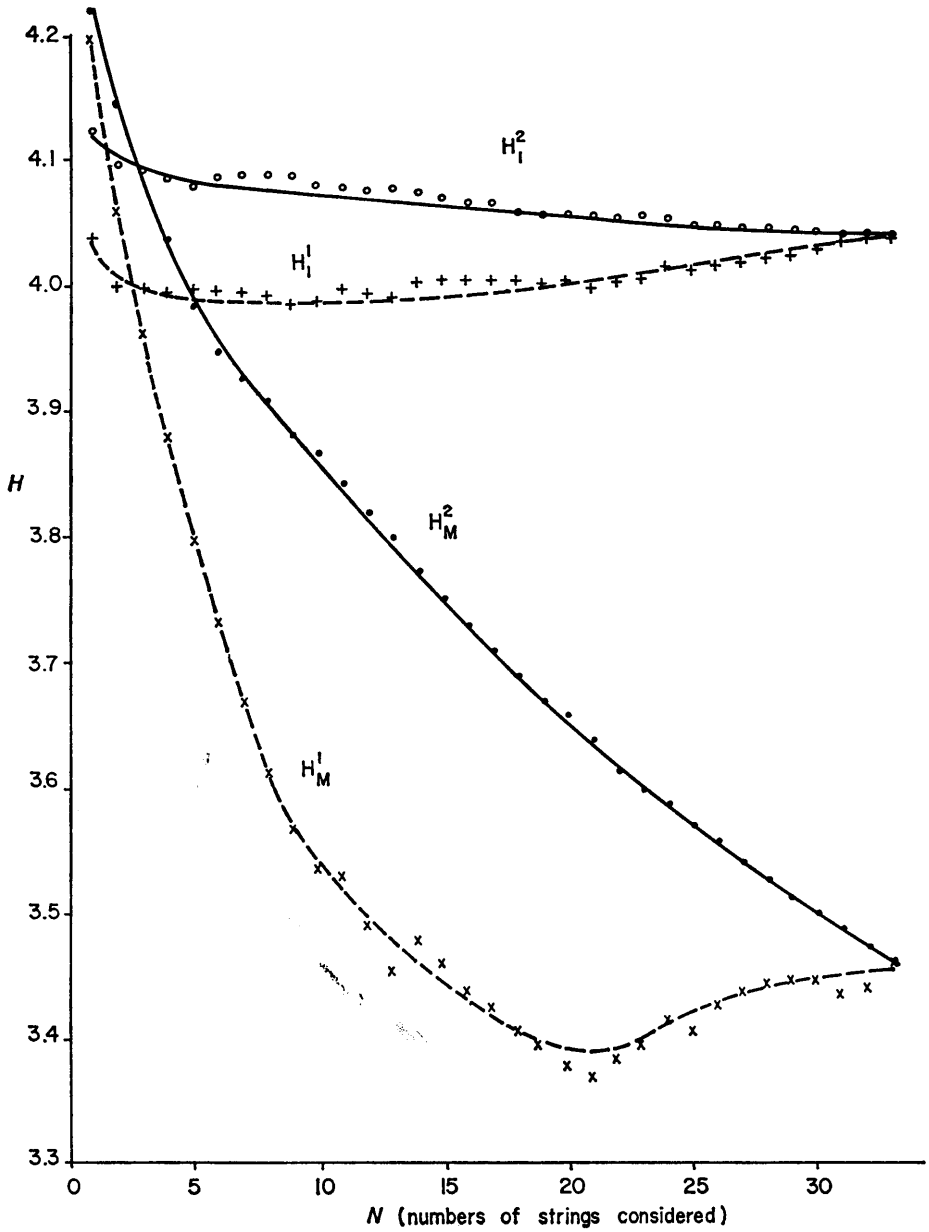

FIGURE 3

The development of the source entropies for cytochrome *c* with ensemble size showing the effect of altering the order of addition of sequences.

ensemble in order of decreasing *TI* to obtain the most rapid approach to the limiting values. This has been done for H_M^1 for the 33 sequence ensemble source appearing as the dashed line in Figure 3. The second ordering is the inverse of the first, that is, sequences were added in order of increasing *TI*. No limiting value is in evidence for this case; whereas, for the first ordering, the source entropy was within two to three per cent of its apparent limiting value after ten sequences. It would appear then, that high information sequences are the most important in establishing a protein super source.

The feature, more common-more information, is entirely in conflict with the established notion of the communications form of information theory in which information value and the amount of surprise the message element generates in the receiver are somehow correlated. Because of this feature, one might suspect that the high values accorded some of the taxonomic groupings might be due solely to their overrepresentation in the ensemble. Among the species represented by the first 27 sequences of Table I, there are nineteen vertebrates: 8 mammals, 1 marsupial, 4 birds, 6 reptiles, fish, or amphibians; and eight nonvertebrates: 4 insects, 1 higher plant, 3 lower organisms. Certainly the sample appears to be mammalian biased although the less well represented birds fare at least as well as the mammals. In the ensemble used in Figure 2, the added sequences are four higher plants [1], one slime mold and one protozoan [4]. The mammalian-vertebrate bias is mitigated and certainly the wheat value is substantially altered without changing materially the other orderings. (Typical changes are on the order of three per cent for *TD*2—the same order as the alteration in the source entropy H_M.)

3.2. *Hemoglobin.* The set of hemoglobin sequences which have been completely elucidated number 26 (15 α chains, 8 β chains, 2 δ chains and 1 γ chain). Only the entropies for the α chain source appear to be suitably stable, although it is possible to lump all of the hemoglobin sequences together in an all hemoglobin ensemble. We have already dealt with this possibility in another paper [10]. In Figure 4 we have presented the stability plot for the α chain source in the ordering of the entries in Table I. Figure 5 is a view of the same parameters presented earlier for cytochrome *c*. The limiting linearity is again in evidence. The slope of the line characterizing cytochrome *c* development in Figure 1 is nearly identical to that of a similar plot for hemoglobin. The limiting slopes of Figures 2 and 5 are, however, different. The values for β hemoglobins would fall at lower values along the same line in Figure 5.

4. Discussion

4.1. *Evolution.* The obvious interpretation of Figures 1, 2, and 5, in complete agreement with intuition, that evolution proceeds in the direction of increasing information, has a strong bearing on the construction of ancestor sequences. Fitch [5] has shown that a divergent evolution may be obtained using an algo-

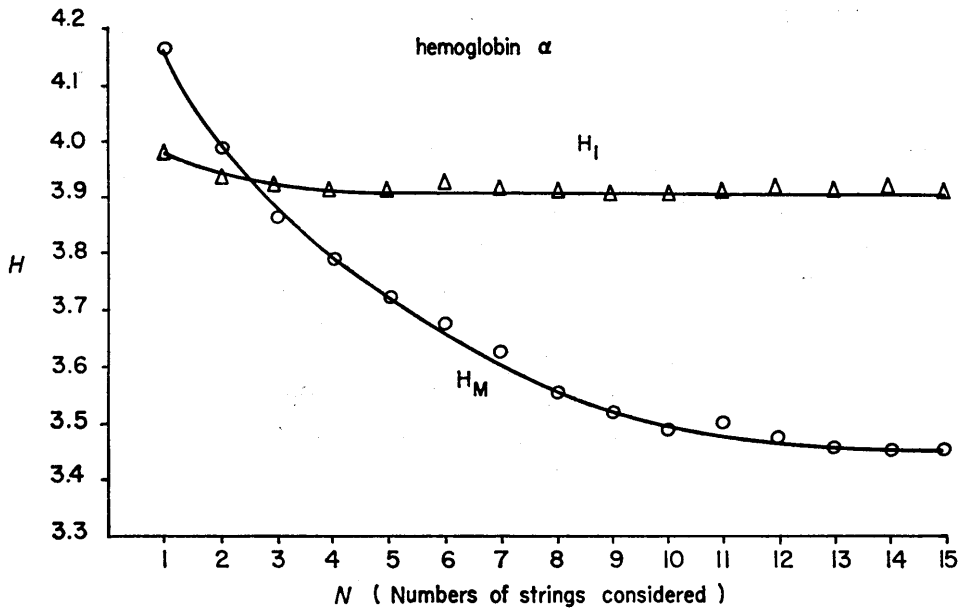

FIGURE 4

The development of the source entropies for hemoglobin α. The order of sequence addition is that of Table I.

rithm for constructing ancestor sequences which has as its basic element the following rule.

Given two homologous sequences which differ at a single site, the ancestral sequence will contain, at that site, that residue, of the two possible, which is most common in the remainder of the sequence.

From the discussion here, an ancestral sequence should possess less information than its progeny. Thus, the residue which would lower *TI* for the sequence should be incorporated in the ancestral sequence. This would be the *rarer* of the two residues and/or that which formed the rarer pairs—where the lowering of *TI* was the preeminent feature. Fitch's sequences are thus more nearly descendant than ancestral.

If *TI* for a cytochrome *c* sequence is calculated using the super source probabilities of another protein, very low, generally negative values of some of the parameters are obtained. The same is true of some myoglobin sequences relative to the hemoglobin α source. The effect is not unlike looking up a particular word, say an English word, in a dictionary for a language other than English. If the language is different enough from that of the word or set of words, no meaning will be assignable. It would seem, for example, that one cannot "say" cytochrome *c* in hemoglobin. This is, of course, very close to an operational definition of homology.

4.2. *The meaning of information.* That higher organisms store more information in homologous molecules, seems to be the message so far. Can we conclude that a plethora of this stuff is better? To begin to answer this question, we examined the set of variants of human hemoglobin α which differ from the normal α chain only by a single amino acid substitution. The informational parameters of this group relative to the ensemble of hemoglobin α sequences are presented in Figure 6. There are three subsets of variants of particular interest demonstrated in the figure.

The first contains the two variants which have *TI* greater than the normal sequence while *TD*2 is lower. Bearers of these deviant hemoglobins are apparently clinically normal (cn).

The second cluster is the group of variants with *TD*2 and *TI* closest to the normal sequence. Of these five are clinically normal variations and *one*, M Boston, produces severe complications.

The third cluster is the group of *four* variants which are the lowest in both *TD*2

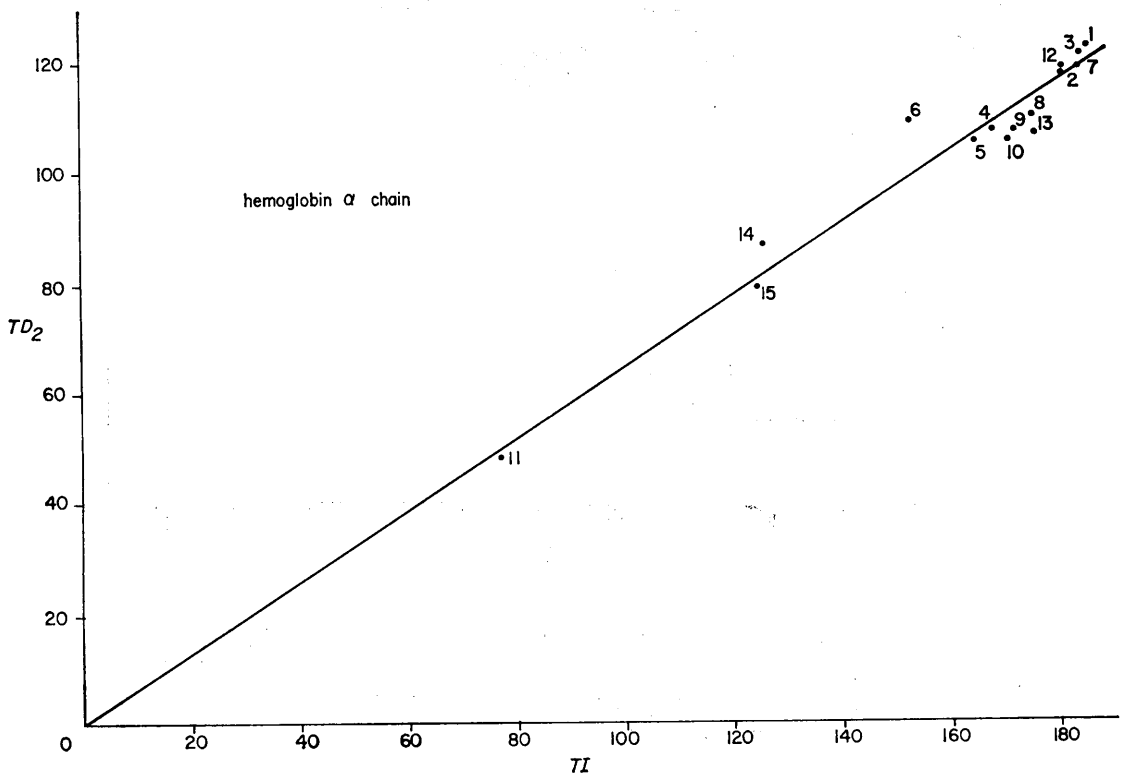

FIGURE 5

A plot of the informational parameters *TD*2 *versus TI* for the α chain of hemoglobin for an ensemble of 15 sequences. The slope of the limiting line is 0.65.

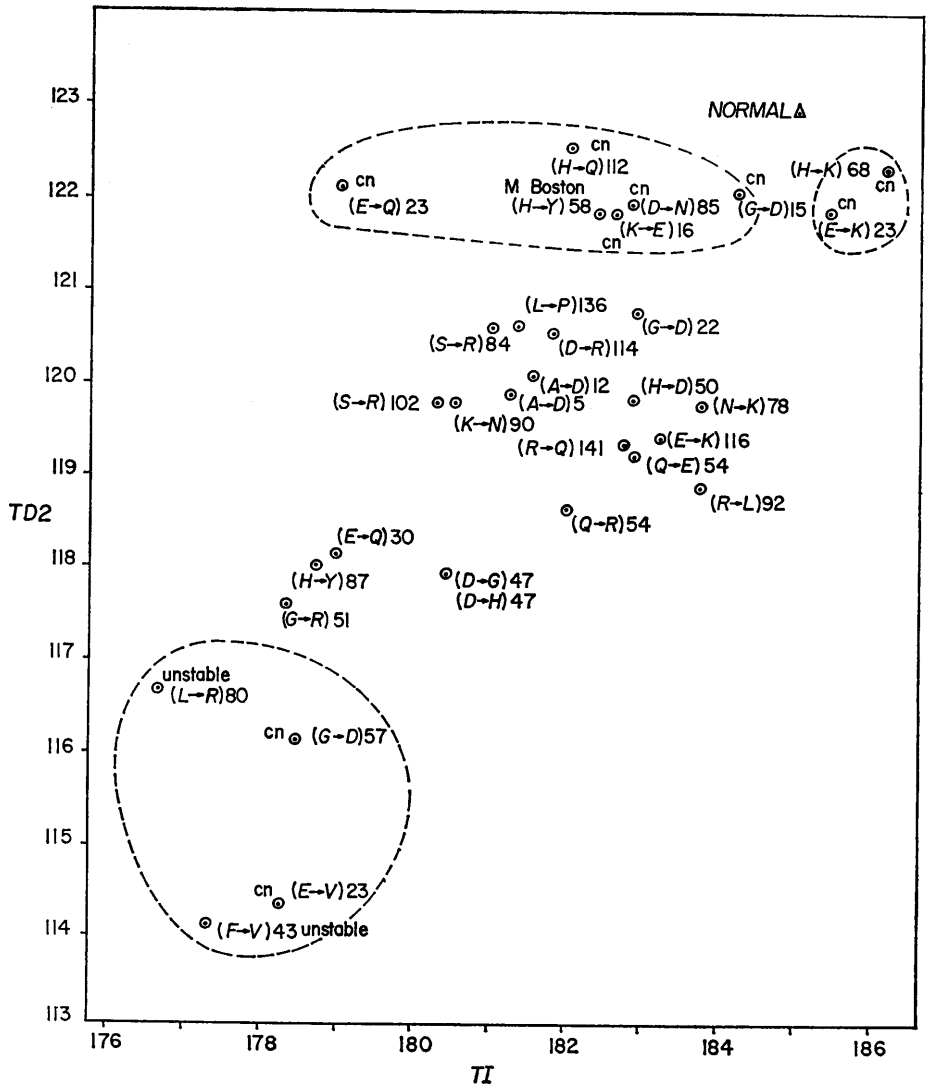

FIGURE 6

A plot of the information stored in the selection of amino acid pairs, *TD2*, *versus* the total amount of information stored in the sequence for all known single substitution variants in the α chain of human hemoglobin. The letters cn mean clinically normal in the heterozygous condition.

and *TI*. Of this group, *two* represent clinically normal variations while *two* are decidedly pathological. Thus, we have some evidence that indeed "more is better." A careful study of the function associated with each variant site should even further pigeonhole this "stuff," information.

We have recently completed an analysis [12] of the correlation of the informational parameters developed here with the known structural and functional significance of each amino acid in the normal hemoglobins (both α and β) of humans and horses. We find that the residues which are ascribed structural/functional significance are highly optimized, that is, any substitution at that site will produce a decrease in *TI*. The sites accorded no such importance, on the other hand, often have several possible information-improving substitutions.

By further refinements in these studies, we may hope to further localize this property heretofore spoken of only in its form as an average. The molecular biological format provides a testable basis for significance, and may even lead us, quite incidentally, to the meaning of meaning.

We have limited our discussion here to proteins. We have made some attempt to treat *t*-RNA premethylation sequences in a similar fashion. The results were not well defined, however, and it is clear that the modified bases must, in some way, be put into the formulation.

REFERENCES

[1] D. Boulter, E. W. Thompson, J. A. M. Ramshaw, and M. Richardson, "Higher plant cytochrome *c*," *Nature*, Vol. 228 (1970), pp. 552–554.

[2] R. A. Christensen, "A general approach to pattern discovery," University of California, Berkeley, Computer Center, Technical Report No. 20, 1967.

[3] M. O. Dayhoff, *Atlas of Protein Sequence and Structure*, Silver Spring, Md., National Biomedical Research Foundation, 1969.

[4] W. M. Fitch, Personal communication.

[5] ———, "Distinguishing homologous from analogous proteins," *Syst. Zool.*, Vol. 19 (1970), pp. 99–113.

[6] L. L. Gatlin, "The information content of DNA," *J. Theoret. Biol.*, Vol. 10 (1966), pp. 281–300.

[7] ———, "The information content of DNA II," *J. Theoret. Biol.*, Vol. 18 (1968), pp. 181–194.

[8] ———, "Evolutionary indices," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1972, Vol. 5, pp. 277–296.

[9] T. A. Reichert and A. K. C. Wong, "An application of information theory to genetic mutations and the matching of polypeptide sequences," submitted to *J. Theoret. Biol.*

[10] ———, "Toward a molecular taxonomy,'" *J. Molec. Evol.*, Vol. 1 (1971), pp. 99–111.

[11] A. K. C. Wong, T. A. Reichert, and B. Aygun, "A generalized method for aligning unambiguous code sequences," submitted to *J. Computers Biol. Med.*

[12] A. K. C. Wong and T. A. Reichert, "The structure and function of hemoglobin reflected in code sequence optimization," submitted to *J. Mol. Biol.*