Protein Constraints Induced by Multiframe Encoding

TEMPLE F. SMITH*
Physics Department, Northern Michigan University, Marquette, Michigan 49855

AND

MICHAEL S. WATERMAN
Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545

Received 8 February 1979; revised 6 July 1979

ABSTRACT

Recent sequencing of viral genomes supports the existence of multiframe codon reading. This study considers the restrictions imposed on proteins coded for in overlapping regions. Calculation of conditional probabilities shows that the restrictions imposed are severe, even at the amino acid–protein structural level. These conditional probabilities are used to calculate conditional information per codon, showing the nonequivalence of the various pairs of reading frames. In addition, the coevolution or primary-secondary evolution of two overlapping proteins should be distinguishable.

INTRODUCTION

The recent sequencing of the viral genomes SV40, ΦX174 [1–3] strongly supports the existence of multiframe codon reading. This overlapping of protein sequence information within the genome, while not the rule, was part of the original proposal by Gamow et al. in 1956 [4]. The fact that a minimum of three bases is required to code for anything over sixteen possible amino acids results in considerable degeneracy for the encoding of the known twenty amino acids by triplets.

The natural question which arises is what restrictions are imposed on pairs of proteins coded for within the same region of DNA. In order to approach this question it is conceptually convenient to view these restrictions as being imposed on one of the proteins given the other. In other words, to what degree does the specification of one amino acid sequence, even with the degeneracy of the code, constrain the possibilities available for the second sequence?

*Present address: Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520.

0025-5564/80/030017+10$02.25

The calculation of all amino acids in a particular reading frame compatible with a given amino acid in a different reading frame can be directly obtained from the genetic code. These frequencies allow the calculation of conditional probabilities. These probabilities in turn allow the calculation of the average information content per codon. This latter is the basis of measuring the degree of restriction imposed by overlapping coding.

These proposed methods of analysis have the advantage that they can be applied not only to the coding constraints imposed among various amino acids but to those imposed among various classes of amino acids. The importance of these classes has resulted from the protein structure–amino acid correlations studied by Anfinsen and Scheraga [5] and others [6,7] as well as from the molecular taxonomy studies [8], which show considerable amino acid substitution flexibility.

A number of important conclusions have been obtained. The first, and most intuitive, is the nonequivalence of the five different secondary reading frames. In addition, the constraints imposed even at the amino acid–protein structural class level are severe. Finally, it appears that one should be able to distinguish between fully coordinate evolution of two overlapping proteins and primary-secondary evolution of two such proteins.

A recent attempt by Figueroa et al. [12] was made to measure the information density of the ΦX174 genome directly from the nearest neighbor Markovian indexes defined by Gatlin [10]. While this study suggests an increased information density above randomness, the code's degeneracy and multi-reading-frame constraints were not investigated.

CALCULATION OF MULTIFRAME READING CONSTRAINTS

For any given codon there are six overlapping reading frames. These can be diagrammed for the Methionine codon, for example, as

Reading Frame					Description	Index
	A	U	G		Primary frame	0
		U	G	X	Same sense, one to the right	1
X	A	U			Same sense, one to the left	2
	C	A	U		Opposite sense, same frame	3
		X	C	A	Opposite sense, one to the right	4
A	U	X			Opposite sense, one to the left	5

Here X denotes any base.

A direct examination of the genetic code allows the listing of all compatible codons for any two amino acids read out of frame one to the left or right on either of the two DNA strands. To begin, consider the $4^3=64$ codons for the 20 amino acid and the terminator codons. These 64 codons are put into n classes—$C_1, C_2, \ldots, C_n$—where all codons for each amino acid and for the terminator occur in exactly one class C. Thus n cannot be larger than 21, the number of amino acids plus terminator. Initially, we let each class correspond to exactly one amino acid or terminator.

Under the *a priori* assumption that all 64 possible codons are equally probable, conditional probabilities can be calculated. The probability of class C_j is given by

$$P(C_j)=\frac{\#(C_j)}{64}, \tag{1}$$

where $\#(C_j)$ denotes the number of elements in the set C_j. Note that if C_j is the class of codons encoding for serine, $\#(C_j)$ is just the degeneracy for serine, or 6. The conditional probability of reading a codon frame shifted one to the right for any member of class C_i given any member of class C_j is just

$$P(C_i|C_j)=\frac{P(C_i\cap C_j)}{P(C_j)} \tag{2a}$$

$$=\frac{\#\{b_2b_3x\in C_i \quad \text{where} \quad b_1b_2b_3\in C_j \quad \text{and} \quad x \quad \text{is any base}\}}{\#\{b_2b_3x \quad \text{where} \quad b_1b_2b_3\in C_j \quad \text{and} \quad x \quad \text{is any base}\}}. \tag{2b}$$

Here b_1, b_2, and b_3 denote the particular bases encoding an amino acid in C_j, and x denotes any of the four possible bases.

For example; given the codons for tyrosine, of which there are two, (UAU and UAC), under the *a priori* assumption of codon equivalence, there are eight possible codons reading in the same sense one frame shift to the right. These eight codons are written AUX and ACX and encode for threonine (ACX), isoleusine (AUU, AUC, AUA), and methionine (AUG). Table 1 shows the probabilities calculated for the twenty amino acids and terminators in a similar manner using Eq. (2).

The conditional probabilities calculated in Table 1 assume the independence of successive codons in the primary reading frame. In order to calculate the reading frame constraints more realistically one should specify at least two successive codons in the primary reading frame. Identifying the

TABLE 1

A Priori Conditional Probabilities
of Reading One Amino Acid (Rows) Frame Shifted One to the Right
Relative to a Reading Frame Specifying Another Amino Acid (Columns)

	Phe	Leu	Ilu	Met	Val	Ser	Pro	Thr	Ala	Tyr	End	His	Gln	Asn	Lys	Asp	Glu	Cys	Trp	Arg	Gly
Phe	0.25	0.08	0.17	—	0.13	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Leu	0.25	0.08	0.17	—	0.13	0.17	0.25	0.25	0.25	—	—	—	—	—	—	—	—	—	—	—	—
Ilu	—	—	—	—	—	—	—	—	—	0.38	—	0.38	—	0.38	—	0.38	—	—	—	—	—
Met	—	—	—	—	—	—	—	—	—	0.13	—	0.13	—	0.13	—	0.13	—	—	—	—	—
Val	—	—	—	—	—	0.17	—	—	—	—	—	—	—	—	—	—	—	0.50	—	0.17	0.25
Ser	0.50	0.17	0.33	—	0.25	—	—	—	—	—	0.17	—	0.25	—	0.25	—	0.25	—	—	—	—
Pro	—	—	—	—	—	0.17	0.25	0.25	0.25	—	—	—	—	—	—	—	—	—	—	—	—
Thr	—	—	—	—	—	—	—	—	—	0.50	—	0.50	—	0.50	—	0.50	—	—	—	—	—
Ala	—	—	—	—	—	0.17	—	—	—	—	—	—	—	—	—	—	—	0.50	—	0.17	0.25
Tyr	—	0.17	0.17	—	0.13	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
End	—	0.25	0.17	0.25	0.19	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
His	—	—	—	—	—	0.08	0.13	0.13	0.13	—	—	—	—	—	—	—	—	—	—	—	—
Gln	—	—	—	—	—	0.08	0.13	0.13	0.13	—	—	—	—	—	—	—	—	—	—	—	—
Asn	—	—	—	—	—	—	—	—	—	—	0.17	—	0.25	—	0.25	—	0.25	—	—	—	—
Lys	—	—	—	—	—	—	—	—	—	—	0.17	—	0.25	—	0.25	—	0.25	—	—	—	—
Asp	—	—	—	—	—	—	—	—	—	—	0.17	—	—	—	—	—	—	—	—	0.17	0.13
Glu	—	—	—	—	—	—	—	—	—	—	0.17	—	—	—	—	—	—	—	—	0.17	0.13
Cys	—	0.17	—	0.50	0.13	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Trp	—	0.08	—	0.25	0.06	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Arg	—	—	—	—	—	0.17	0.25	0.25	0.25	—	0.17	—	0.25	—	0.25	—	0.25	—	—	—	—
Gly	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	1.00	0.33	0.25

reading frames, as above, for the methionine-proline sequence one has

Reading frame						Index
A	U	G	C	C	X	0
	U	G	C			1
		G	C	C		2
C	A	U				3
	G	C	A			4
		G	G	C		5

Consider the results of a reading frame shifted one to the right read on the successive occurrence of C_j and C_k, denoted by $C_jC_k = \{b_1b_2b_3b_4b_5b_6$ where $b_1b_2b_3 \in C_j$ and $b_4b_5b_6 \in C_k\}$. Then the conditional probability of reading a codon in C_i given C_jC_k is

$$P_1(C_i|C_jC_k) = \frac{\#\{b_2b_3b_4 \in C_i \quad \text{where} \quad b_1b_2b_3b_4b_5b_6 \in C_jC_k\}}{\#\{b_2b_3b_4 \quad \text{where} \quad b_1b_2b_3b_4b_5b_6 \in C_jC_k\}} \tag{3a}$$

$$= \frac{\#\{b_2b_3b_4 \in C_i \quad \text{where} \quad b_1b_2b_3b_4b_5b_6 \in C_jC_k\}}{\#(C_j)\#(C_k)}. \tag{3b}$$

These probabilities give a measure of the restrictions on a class of amino acids in one reading frame imposed by the two consecutive classes in the other. The *a priori* probability of the occurrence of the given consecutive classes C_j and C_k is just

$$P(C_jC_k) = \frac{\#(C_jC_k)}{(64)^2} = \frac{\#(C_j)\#(C_k)}{(64)^2}. \tag{4}$$

Next we consider the information per codon or triplet of bases in the DNA. The maximum information is given by

$$\bar{I}_{\text{codon}} = \log_2 64 = 6 \text{ bits}, \tag{5}$$

where $\log_2$ is logarithm to the base 2. If the genetic code, with its redundant coding for amino acids, is taken into account, the maximum average information carried per amino acid is 4.22 bits, obtained from

$$\bar{I}_{\text{AA}} = \max\left\{ -\sum_{\text{AA}} P(\text{AA})\log_2 P(\text{AA}) \right\}. \tag{6}$$

Here $P(\text{AA})$ is just the probability of amino acid AA given the GC to AT base ratio. The maximum occurs at a value of 42 percent GC in the genetic sequence. This calculation was first made by Smith [9], and the difference from the maximum of 6 bits is a measure of the informational restriction imposed by the structure of the genetic code itself.

Our interest is in the average conditional information per codon of a second protein given an initial protein. The definition of conditional information [10, 11] of a probability space **A** with elements A given knowledge of a probability space **B** with elements B is

$$\bar{I}(\mathbf{A}|\mathbf{B}) = -\sum_{A,B} P(B)P(A|B)\log_2 P(A|B). \tag{7}$$

Equation (7) can be applied to the conditional probabilities derived above. The symbol $\bar{I}_m$ will denote the average conditional information contained per codon for reading frame m. For all classes C_i from $\mathbf{C} = \{C_1, C_2, \ldots, C_n\}$ given reading frame m on each class C_j from $\mathbf{C}$, we have

$$\bar{I}_m(\mathbf{C}|\mathbf{C}) = \sum_{i,j=1}^{n} P(C_j)P_m(C_i|C_j)\log_2 P_m(C_i|C_j). \tag{8}$$

Note that the sums and thus the average are over all codon classes weighted by their probability of occurrence. The next case is for all classes C_j from $\mathbf{C} = \{C_1, C_2, \ldots, C_n\}$ given reading frame m on classes C_jC_k from $\mathbf{C}\times\mathbf{C}$. The formula for conditional information per codon then has the form

$$\bar{I}_m(\mathbf{C}|\mathbf{C}\times\mathbf{C}) = \sum_{i,j,k=1}^{n} P(C_jC_k)P_m(C_i|C_jC_k)\log_2 P_m(C_i|C_jC_k). \tag{9}$$

A general result from information theory [11, p. 124] implies $\bar{I}_m(\mathbf{C}|\mathbf{C}\times\mathbf{C}) < \bar{I}_m(\mathbf{C}|\mathbf{C})$. This can be observed in Table 2, which lists the values obtained

TABLE 2

The Average Conditional Information per Codon Obtained from Eqs. (8) and (9) When the Encoding of Each Amino Acid Defines a Codon Class.

Reading frame m	$I_m(\mathbf{C}|\mathbf{C})$	$I_m(\mathbf{C}|\mathbf{C}\times\mathbf{C})$
0[a]	4.218	—
1	2.144	1.709
2	2.144	1.729
3	1.532	—
4	3.424	1.832
5	0.821	0.644

[a] Given for reference only, as the maximum possible [9].

from Eqs. (8) and (9) when the encoding of each amino acid is used to define a codon class. The most evident conclusion is that the reading frames are not equivalent. Frame 3 (the opposite sense complement) is not the most restricted as one might suppose. However, the values obtained are compatible with our knowledge of the genetic code. That is, to first order, the middle base is most determining, the first base next, and the third least. Thus, in reading frame 4, the primary frame third base becomes the middle base. This results in a relatively high potential information per code, $\bar{I}_4(\mathrm{C}|\mathrm{C}) = 3.424$ bits. On the other hand, the low information obtainable in the combination of primary reading frame and frame 5—$[\bar{I}_0(\mathrm{C}|\mathrm{C}) + \bar{I}_5(\mathrm{C}|\mathrm{C})]/2 = 2.52$ bits per codon—suggests that this combination should be very rare in nature.

The protein structural classes listed in Table 3 were used to calculate the conditional probabilities defined by Eq. (2) and given in Table 4. These probabilities indicate that even when summing over these rather large codon classes there still exist various important structural constraints imposed on one protein by the other. It must be emphasized that throughout we have assumed equal likelihood for each codon. In nature this is, in general, not the case [9]. While the overall base composition varies widely [13], the soluble protein amino acid compositions are quite narrowly distributed [14], and this means that the restrictions apparent in Tables 2 and 4 are upper bounds. In addition, there are restrictions arising from the RNA polymerase initiation site requirements [15, 16]. Thus restrictions on functional proteins overlappingly coded for in the same regions will be even greater.

TABLE 3

Protein Structural Codon Classes[a]

[A]	[B]	[C]	[D]	[E]
α-helix	β-	α-helix	β-sheet	Terminator
incompatible	turners	compatible	compatible	codons
Asn	Asn	Ala	Ala	(UAA)
Asp	Asp	Arg	Asn	(UAG)
Cys	Cys	Gln	Asp	(UGA)
Gly	Gly	Glu	Gln	
Pro	Ilu	His	Gln	
Ser	Lys	Ilu	Gly	
	Pro	Lue	Pro	
	Ser	Lys	Ser	
	Thr	Met		
	Trp	Phe		
	Tyr	Trp		
		Val		

[a] From Anfinsen and Scheraga [5], Robson and Suzuki [6], and Chou and Fasman [7].

TABLE 4

Conditional Probabilities of Reading a Codon in Class C_j (Rows) in Frame m Relative to a Frame Specifying Any Codon in Class C_i (Columns)[a]

Reading Frame $m = 1$

i	*A*	*B*	*C*	*D*	*E*
j					
A	0.14	0.15	0.20	0.17	0.00
B	0.22	0.25	0.30	0.26	0.14
C	0.38	0.36	0.22	0.31	0.86
D	0.21	0.20	0.26	0.22	0.00
E	0.06	0.04	0.02	0.04	0.00

Reading Frame $m = 2$

i	*A*	*B*	*C*	*D*	*E*
j					
A	0.14	0.14	0.20	0.16	0.27
B	0.24	0.25	0.31	0.24	0.27
C	0.39	0.37	0.23	0.38	0.20
D	0.23	0.23	0.22	0.23	0.27
E	0.00	0.01	0.05	0.00	0.00

Reading Frame $m = 3$

i	*A*	*B*	*C*	*D*	*E*
j					
A	0.12	0.16	0.20	0.14	0.20
B	0.27	0.32	0.26	0.28	0.20
C	0.39	0.30	0.27	0.37	0.40
D	0.18	0.21	0.24	0.19	0.20
E	0.03	0.02	0.03	0.02	0.00

Reading Frame $m = 4$

i	*A*	*B*	*C*	*D*	*E*
j					
A	0.19	0.19	0.16	0.18	0.11
B	0.30	0.29	0.26	0.28	0.26
C	0.27	0.28	0.32	0.29	0.47
D	0.23	0.23	0.22	0.23	0.16
E	0.01	0.02	0.04	0.02	0.00

Reading Frame $m = 5$

i	A	B	C	D	E
j					
A	0.26	0.19	0.13	0.23	0.00
B	0.26	0.25	0.30	0.26	0.29
C	0.17	0.29	0.37	0.17	0.29
D	0.30	0.25	0.18	0.33	0.14
E	0.00	0.02	0.02	0.01	0.29

[a] Calculated using Eq. (2) and reading frames as defined in text. Note for example that in reading frame 3 two overlapping α-helical incompatible regions are very unlikely, while the overlapping of α-helical and non-α-helical regions is quite compatible. Similar calculations have been carried out using the contiguous codon conditional probabilities given in Eq. (3), and the constraints are still more restrictive.

Given that one region of DNA codes for two proteins, it is of interest to discover whether one protein, evolved under more rigorous constraints, preexisted the other or whether they have undergone linked coordinate evolution. The statistical restrictions we have derived can assist in a determination of this question. If the composition pattern of amino acids of one protein fits closely the conditional probabilities given here, then that protein must have come into existence after or be subsidiary to the other protein coded for in the same region.

REFERENCES AND NOTES

1 V. B. Reddy, B. Thimmappaya, R. Dhar, K. N. Subramanian, B. S. Zain, J. Pan, P. K. Ghosh, M. L. Clema, and S. M. Weissman, *Science* 200:494 (1978).

2 F. Sanger, G. M. Air, G. G. Barrell, N. L. Brown, A. E. Coulson, J. C. Fiddes, C. Hutchinson, P. M. Slocombe, and M. Smith, *Nature* (London) 265:681 (1977).

3 W. Fiers et al., *Nature* (London) 260:500 (1976).

4 G. Gamov, A. Rich, and M. Ycas, *Advances in Biol. and Med. Phys.* 4:23 (1956).

5 C. B. Anfinsen and H. A. Scheraga, *Advances in Protein Chemistry* 29:205 (1975).

6 B. Robson and E. Suzuki, *J. Mol. Biol.* 107:327 (1976).

7 P. Y. Chou and G. D. Fasman, *J. Mol. Biol.* 115:135 (1977).

8 E. Margoliash, W. M. Fitch and R. E. Dickerson, in *Structure, Function and Evolution of Proteins, Brookhaven Symp. Biol.* 21:259 (1969).

9 T. F. Smith, *Math. Biosci.* 4:179 (1969). It is of interest to note that the amino acid composition alone gives a value only slightly lower; see Ref. [14] below.

10 L. L. Gatlin, *Information Theory and the Living System*, Columbia U.P., New York, 1972, p. 54.

11 P. Billingsley, *Ergodic Theory and Information*, Wiley, New York, 1964, p. 77.

12 R. Figueroa, A. Sepulveda, M. A. Soto, and J. Toha, *Naturforsch. Sect. C. Biosci.* 32:850 (1977).

13 N. Sueoka, in *Evolving Genes and Proteins* (V. Bryson and H. J. Vogel, Eds.), Academic, New York, 1965, p. 480.

14 H. R. Branson, in *Information Theory in Biology* (H. Quastler, Ed.), 1953, pp. 84–104; for more data see M. O. Dayhoff, *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, 1969, 1973.

15 N. L. Brew and M. Smith, *Nature* 265:695 (1977).

16 G. N. Godson, B. G. Barrell, R. Staden, and J. C. Fiddes, *Nature* 276:236 (1978).