A Study of the Purine/Pyrimidine Codon Occurrence with a Reduced Centered Variable and an Evaluation Compared to the Frequency Statistic

CHRISTIAN J. MICHEL
Friedrich Miescher Institut, Bioinformatic Group, CH-4002 Basel, Switzerland

Received 26 September 1988; revised 21 June 1989

ABSTRACT

With the three-letter alphabet {R,Y,N} (R = purine, Y = pyrimidine, N = R or Y), there are 26 codons (NNN being excluded): RNN,...,NNY (six codons at two unspecified bases N), RRN,...,NYY (12 codons at one unspecified base N), RRR,...,YYY (eight specified codons). A statistical methodology that uses the codon frequency and a reduced centered variable leads to similar results for a codon occurrence study, regardless of gene function and regardless of a particular protein coding gene taxonomic population. Therefore, this variable can be considered a new codon usage index, whose use removes certain nonsignificant results found with the frequency statistic. This methodology identifies the common and rare codons (i.e., the codons having the highest and lowest occurrence) and leads to a model of codon evolution at three successive states: RNN, then RNY, and finally RYY. Some biological relations between this model and the $YRY(N)_6YRY$ preferential occurrence are also presented.

INTRODUCTION

With the three-letter alphabet {R,Y,N} (R = purine, Y = pyrimidine, N = R or Y), there are 26 codons (NNN being excluded): RNN,...,NNY (six codons at two unspecified bases N), RRN,...,NYY (12 codons at one unspecified base N), RRR,...,YYY (eight specified codons). A statistical methodology is developed in order to study the codon occurrence, regardless of gene function and regardless of one or a few particular protein coding gene populations (see the 33 available taxonomies in Table 1). Two approaches seem to be possible to analyze this genetic information regardless of molecular evolution. A first approach consists of using a unique data sample constituted of several gene populations of approximately the same size. Unfortunately, in release 15 of the EMBL Nucleotide Sequence Data Library, the primates, the rodentia, and the gram-negative prokaryotes are strongly overrepresented compared to the 30 other gene populations (see

655 Avenue of the Americas, New York, NY 10010 0025-5564/89/$03.50

TABLE 1

Convention and Size (Number of Open Reading Frames ORF and of Codons c) in the 33 Gene Populations P of the Library L

AMPEUKO: amphibia eukaryote ORF, 188 ORF, 15884 c.
ARIEUKO: artiodactyla eukaryote ORF, 387 ORF, 66158 c.
ARTEUKO: arthropoda eukaryote ORF, 431 ORF, 73297 c.
ASCEUKO: aschelminthe eukaryote ORF, 31 ORF, 5192 c.
AVEEUKO: ave eukaryote ORF, 413 ORF, 44226 c.
ECHEUKO: echinodermata eukaryote ORF, 42 ORF, 5603 c.
LAGEUKO: lagomorpha eukaryote ORF, 164 ORF, 27209 c.
MYCEUKO: mycophyta eukaryote ORF, 541 ORF, 174664 c.
PISEUKO: pisce eukaryote ORF, 111 ORF, 12618 c.
PRIEUKO: primate eukaryote ORF, 2668 ORF, 431387 c.
RODEUKO: rodentia eukaryote ORF, 2526 ORF, 307883 c.
SPEEUKO: spermatophyta eukaryote ORF, 564 ORF, 80694 c.
CILEUKO: ciliata protozoa eukaryote ORF, 16 ORF, 5021 c.
FLAEUKO: flagellata protozoa eukaryote ORF, 31 ORF, 10520 c.
SPOEUKO: sporozoa protozoa eukaryote ORF, 59 ORF, 16872 c.
ACTPROO: actinomycete prokaryote ORF, 54 ORF, 16052 c.
CYAPROO: cyanobacteria prokaryote ORF, 60 ORF, 14355 c.
ENDPROO: endospore-forming rods and cocci prokaryote ORF, 213 ORF, 66412c
GRNPROO: gram-negative prokaryote ORF, 1219 ORF, 351633 c.
GRPPROO: gram-positive prokaryote ORF, 101 ORF, 25262 c.
METPROO: methane producing prokaryote ORF, 22 ORF, 6791 c.
PHOPROO: phototrophic prokaryote ORF, 26 ORF, 4999 c.
DDEVIRO: double strand-DNA enveloped virus ORF, 308 ORF, 121392 c.
DDNVIRO: double strand-DNA nonenveloped virus ORF, 481 ORF, 105814 c.
DRNVIRO: double strand-RNA nonenveloped virus ORF, 46 ORF, 18296 c.
NCLVIRO: nonclassified virus ORF, 97 ORF, 29677 c.
SDNVIRO: single strand-DNA nonenveloped virus ORF, 88 ORF, 18809 c.
SREVIRO: single strand-RNA enveloped virus ORF, 715 ORF, 265184 c.
SRNVIRO: single strand-RNA nonenveloped virus ORF, 322 ORF, 134924 c.
PHYCHLO: phycophyta chloroplast ORF, 49 ORF, 7830 c.
SPECHLO: spermatophyta chloroplast ORF, 156 ORF, 36433 c.
METMITO: metazoa mitochondria ORF, 72 ORF, 20222 c.
PLAMITO: planta mitochondria ORF, 106 ORF, 25426 c.

Table 1). This situation is due to the past priorities of biologists working on DNA sequencing. To avoid this biased study, a second approach has been taken:

(1) For each gene population, the occurrence of the 26 codons is observed with the codon frequency.

(2) At the library level, the comparison of the occurrence of two codons is tested by analyzing the signs and the absolute values of their frequency differences obtained with the 33 gene populations.

A significant occurrence comparison for a codon pair is then independent of gene function and of a particular gene population (see Method).

When applied to all the codon pairs, this methodology using the codon frequency cannot identify the common codons (i.e., the codons having the highest occurrence regardless of gene function and regardless of a particular gene population). This problem can be removed if the same methodology uses, instead of the codon frequency, a reduced centered variable based on the average fraction of R and Y in the three codon sites. Compared to the frequency, this variable will not only lead to similar results for the codon occurrence study but will also allow its extension and the removal of certain nonsignificant results (see Method and Results sections).

This statistical methodology is also developed to study the primordial codons, which are assumed to be related to the alphabet {R, Y, N} [4, 5]. The primordial codon formation is mainly associated with the RNY codon (a codon at one unspecified base N) because its chemical and genetic properties may explain the interaction between the codon and the anticodon as well as the structure of the primordial proteins [4, 5]. On the other hand, this RNY model was in agreement with several statistical observations at the DNA sequence level [12] and at the DNA population level (e.g., [1–3]), which showed a higher frequency of the RNY codon compared to the codons RNR, YNR, and YNY. The methodology developed herein will show that the RNY codon has a greater occurrence compared to all other codons at one unspecified base N. Furthermore, the statistical results obtained will lead to a model of codon evolution at three successive states: RNN, then RNY, and finally RYY.

On the other hand, in some previous studies, a large-scale statistical analysis of a perturbation [1] in the coding (modulo 3) periodicity [8, 10] led to the following result [2]:

The mean occurrence probability of the i-motif $YRY(N)_iYRY$ *is not uniform with i in the range* [1, 99] *but presents a maximum at* $i = 6$ *with the following gene populations: protein coding genes of eukaryotes, of prokaryotes, of chloroplasts, and of mitochondria, also with viral introns and ribosomal and transfer* RNA *genes.*

The exception that has been found with the eukaryotic introns has been solved [3] by showing that the $YRY(N)_6YRY$ preferential occurrence is hidden by an alternating purine/pyrimidine (modulo 2) periodicity. It was suggested that the 6-motif $YRY(N)_6YRY$ may be a code of the DNA double helix pitch [1, 2]. This study also focuses on the problem of coexistence of the 6-motif $YRY(N)_6YRY$ with the primordial codons. In particular, a weak occurrence of the YRY codon would be expected in order to avoid ambiguities between a primordial transcription–translation code and a DNA double helix pitch code.

METHOD

STATISTICAL DATA

The EMBL Nucleotide Sequence Data Library L (release 15) is divided into m ($m = 33$ in Table 1) different gene taxonomic populations P having at least 5000 codons, all the populations being protein coding genes. This partition represents almost all the available genetic information (in total: 12,307 open reading frames and 2,546,739 codons) taken from a great number of species. This codon study incorporates all the open reading frames (ORF) that belong to a given P of L. Only genes with unspecified bases or with inappropriate information about the ORF have been excluded from this survey. Several tests also verify this information (e.g., the length of a given ORF must be congruent to 0 modulo 3, etc.). By construction, this partition is independent of gene function.

OBSERVATION OF THE CODON OCCURRENCE WITH THE CODON FREQUENCY AND A REDUCED CENTERED VARIABLE

Let a codon c be three successive bases $b_1b_2b_3$ (coding an amino acid) on the three-letter alphabet {R,Y,N} (see above). To give some meaning to the codon frequency comparisons and also for biological reasons (see Discussion), a codon is associated with a state according to the number of base specifications in its codon sites: A codon at two unspecified bases N belongs to state 1; a codon at one unspecified base N, to state 2; and a specified codon, to state 3. Let n be the total number of codons in a given P of L. Let $p(c; P)$ be the observed and $p_t(c; P)$ the theoretical (below defined) occurrence probability of c in P of L. Then the reduced centered variable $q(c; P)$ is given by

$$q(c; P) = \sqrt{n}\,\frac{p(c; P) - p_t(c; P)}{\{p_t(c; P)[1 - p_t(c; P)]\}^{1/2}}$$

The variable $q(c; P)$ measures (in standard deviation units), for a given codon c in P, the deviation of its observed probability from its theoretical probability; a positive (negative) value means that the codon c occurs with a probability greater (less) than the expected one. This variable depends on the number n of codons in P, leading thereby to a better differentiation between the codons with the large P. Therefore, this variable gives a stronger weight for a statistical result obtained with the large populations.

The average fraction of R and Y in the three codon sites is used to define $p_t(c; P)$. For a given P of L, let $p(b_i; P)$ be the observed occurrence probability of the base b_i in c. If we suppose that the probabilities $p(b_i; P)$,

with i varying between 1 and 3, are independent variables; then

$$p_t(c; P) = \Pi_{i=1,3} p(b_i; P)$$

Note $q(\mathrm{RNN}; P) = -q(\mathrm{YNN}; P)$, $q(\mathrm{NRN}; P) = -q(\mathrm{NYN}; P)$, and $q(\mathrm{NNR}; P) = -q(\mathrm{NNY}; P)$.

STATISTICAL METHODOLOGY

Let $r(c; P)$ be an observation [$p(c; P)$ or $q(c; P)$] of the occurrence of a codon c in a gene population P of a library L. In order to compare the occurrence of two codons c and c' ($c' \neq c$) in L, the null hypothesis H_0: $r(c; L) = r(c'; L)$ is tested by arranging in order the m nonzero observed $|r(c; P) - r(c'; P)|$ values from the smallest absolute value to the largest. Note: With our data, no zero absolute value has been observed, and the equal absolute values have been randomly arranged (very few cases have been observed). Each absolute value is assigned to a rank that corresponds to its position in this ordering. The statistical law $T(c, c'; L)$ used is the positive rank sum. Since $m > 20$, the Wilcoxon signed-rank test follows a normal law $Z(c, c'; L)$ if

$$Z(c, c'; L) = \frac{T(c, c'; L) - m(m+1)/4}{[m(m+1)(2m+1)/24]^{1/2}}$$

Therefore, the occurrence of the codon c in L is greater than the occurrence of the codon c' at the 1% statistical level if $Z(c, c'; L) \geqslant 2.58$ and less if $Z(c, c'; L) \leqslant -2.58$. A significant occurrence comparison for a codon pair (c, c') is then independent not only of gene function (by construction of the partition), but also of a particular gene population (by construction of the test).

Note: $Z(c, c'; L) = -Z(c', c; L)$ and $Z(c, c'; L) \in [-k, k]$ with

$$k = \frac{m(m+1)/4}{[m(m+1)(2m+1)/24]^{1/2}}$$

The statistic $Z(c, c'; L)$ so that $r(c; P) = p(c; P)$ and $r(c'; P) = p(c'; P)$ is called the frequency statistic (FS), while the statistic $Z(c, c'; L)$ so that $r(c; P) = q(c; P)$ and $r(c'; P) = q(c'; P)$ is called the reduced centered variable statistic (RCVS). Therefore, the occurrence comparison for a codon pair is evaluated by FS and RCVS.

The two statistics FS and RCVS are applied to all the codon pairs in order to identify the codons that have the highest and lowest occurrence in L. More precisely, for a given state and for a given statistic, a codon c is called first common in L if $Z(c, c'; L) \geqslant 2.58$ and first rare in L if

$Z(c, c'; L) \leqslant -2.58$ whatever the codon c' of the same state ($c' \neq c$). A less strong condition is also used: for a given state and for a given statistic, a codon c is called second common in L if $Z(c, c'; L) \geqslant 2.58$ and second rare in L if $Z(c, c'; L) \leqslant -2.58$ whatever the codon c' of the same state or different from the first codon. If this methodology cannot identify a codon for one of the above-mentioned conditions, the set of the best codon candidates is given. By convention, these conditions C are described as follows ("first" and "second" are denoted by "1" and "2," respectively):

$$C(\{1,2\};\{\text{common,rare}\};\text{state }\{1,2,3\};\{\text{FS,RCVS}\};L):\text{codon}$$

For example, a codon c of state 2 that is first common in L with the RCVS statistic is denoted

$$C(1;\text{common};\text{state }2;\text{RCVS};L):c$$

The RCVS statistic permits extension of the codon study with the occurrence comparison of the codon pairs constituted of two codons of different states because it achieves an analysis independent of the choice of unity. This property is restricted to the most interesting biological case with the codon pairs constituted of only first common codons.

Table 2 gives the observed occurrence probability $p(c; P)$ (in percentage and denoted Prob.) and the reduced centered variable $q(c; P)$ (denoted RCV) for the 26 codons in the 33 gene populations P (see Table 1) of the library L. Table 3 shows the occurrence comparison results with the FS statistic (top value) and with the RCVS statistic (bottom value) for the 109 codon pairs constituted of two codons of the same state. Only the results of the top triangular table are given because $c' \neq c$ and $Z(c, c'; L) = -Z(c', c; L)$ (see above). A result that is nonsignificant at the 1% statistical level is mentioned as NS.

RESULTS

THE RCVS RESULTS ARE NOT IN CONTRADICTION WITH THE FS RESULTS

For the 109 codon pairs, no (c, c') codon pair has a $Z(c, c'; L)$ value greater than 2.58 with one statistic and a $Z(c, c'; L)$ value less than -2.58 with the other statistic (see Table 3). The RCVS statistic eliminates some nonsignificant results found by using the FS statistic for the following codon pairs: (NRN,NYN), (RRN,RNY), (RYN,RNR), (RYN,NYY), (YYN,NRR), (YYN,NYR), (RNR,NYY), (YNY,NRR), (NRR,NRY), (RRR,RYY), (RRY,RYY), and (RYR,YYY) (see Table 3). Otherwise, the RCVS statistic introduces nonsignificant results for the following codon pairs: (RRN,NYY), (YYN,NRY), (RNR,NRY), and (RYR,YYR) (see Table 3). For three codon pairs, the results are nonsignificant with both statistics: (YNY,NYR), (NRR,NYR), and (RRR,RRY) (see Table 3).

TABLE 2

Observed Occurrence Probability $p(c; P)$ (Prob; in Percent) and Reduced Centered Variable $q(c; P)$ (RCV) for the 26 Codons in the 33 Gene Populations P of the Library L (See Method). Percentages of R and Y are given in columns 1 and 2.

	R	Y	RNN		YNN		NRN		NYN		NNR		NNY	
			Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV
AMPEUKO	54.73	45.27	63.54	22.3	36.46	-22.3	54.70	-0.1	45.30	0.1	45.96	-22.2	54.04	22.2
ARIEUKO	51.27	48.73	59.08	40.2	40.92	-40.2	49.94	-6.8	50.06	6.8	44.79	-33.4	55.21	33.4
ARTEUKO	52.25	47.75	61.11	48.0	38.89	-48.0	51.12	-6.1	48.88	6.1	44.52	-41.9	55.48	41.9
ASCEUKO	52.07	47.93	60.77	12.5	39.23	-12.5	51.85	-0.3	48.15	0.3	43.61	-12.2	56.39	12.2
AVEEUKO	53.63	46.37	63.48	41.5	36.52	-41.5	50.56	-12.9	49.44	12.9	46.85	-28.6	53.15	28.6
ECHEUKO	52.82	47.18	65.23	18.6	34.77	-18.6	51.90	-1.4	48.10	1.4	41.32	-17.2	58.68	17.2
LAGEUKO	51.59	48.41	59.44	25.9	40.56	-25.9	49.41	-7.2	50.59	7.2	45.91	-18.7	54.09	18.7
MYCEUKO	51.81	48.19	62.86	92.4	37.14	-92.4	48.67	-26.3	51.33	26.3	43.91	-66.1	56.09	66.1
PISEUKO	51.58	48.42	59.61	18.0	40.39	-18.0	49.90	-3.8	50.10	3.8	45.23	-14.3	54.77	14.3
PRIEUKO	51.74	48.26	58.40	87.5	41.60	-87.5	50.28	-19.1	49.72	19.1	46.53	-68.4	53.47	68.4
RODEUKO	51.81	48.19	59.60	86.6	40.40	-86.6	49.70	-23.4	50.30	23.4	46.12	-63.2	53.88	63.2
SPEEUKO	50.39	49.61	54.76	24.9	45.24	-24.9	48.75	-9.3	51.25	9.3	47.64	-15.6	52.36	15.6
CILEUKO	52.21	47.79	67.02	21.0	32.98	-21.0	49.39	-4.0	50.61	4.0	40.21	-17.0	59.79	17.0
FLAEUKO	54.51	45.49	64.13	19.8	35.87	-19.8	49.84	-9.6	50.16	9.6	49.55	-10.2	50.45	10.2
SPOEUKO	59.78	40.22	67.82	21.3	32.18	-21.3	57.75	-5.4	42.25	5.4	53.76	-15.9	46.24	15.9
ACTPROO	50.01	49.99	61.93	30.2	38.07	-30.2	47.00	-7.6	53.00	7.6	41.11	-22.6	58.89	22.6
CYAPROO	49.07	50.93	59.78	25.7	40.22	-25.7	46.80	-5.5	53.20	5.5	40.64	-20.2	59.36	20.2
ENDPROO	54.86	45.14	64.70	51.0	35.30	-51.0	50.87	-20.6	49.13	20.6	49.01	-30.3	50.99	30.3
GRNPROO	51.85	48.15	61.75	117.5	38.25	-117.5	48.36	-41.4	51.64	41.4	45.44	-76.1	54.56	76.1
GRPPROO	55.81	44.19	66.36	33.8	33.64	-33.8	52.42	-10.8	47.58	10.8	48.64	-22.9	51.36	22.9
METPROO	56.50	43.50	70.62	23.5	29.38	-23.5	49.08	-12.3	50.92	12.3	49.79	-11.2	50.21	11.2
PHOPROO	48.71	51.29	60.61	16.8	39.39	-16.8	44.59	-5.8	55.41	5.8	40.93	-11.0	59.07	11.0
DDEVIRO	49.07	50.93	55.23	43.0	44.77	-43.0	45.77	-23.0	54.23	23.0	46.20	-20.0	53.80	20.0
DDNVIRO	53.26	46.74	62.30	58.9	37.70	-58.9	51.01	-14.7	48.99	14.7	46.47	-44.2	53.53	44.2
DRNVIRO	54.15	45.85	62.08	21.5	37.92	-21.5	46.59	-21.1	53.61	21.1	53.98	-0.5	46.02	0.5
NCLVIRO	53.33	46.67	62.21	30.7	37.79	-30.7	49.02	-14.9	50.98	14.9	48.75	-15.8	51.25	15.8
SDNVIRO	48.86	51.14	59.34	28.8	40.66	-28.8	48.26	-1.6	51.74	1.6	38.98	-27.1	61.02	27.1
SREVIRO	53.88	46.12	60.54	68.7	39.46	-68.7	49.33	-47.0	50.67	47.0	51.79	-21.7	48.21	21.7
SRNVIRO	51.84	48.16	61.87	73.7	38.13	-73.7	47.42	-32.5	52.58	32.5	46.23	-41.2	53.77	41.2
PHYCHLO	49.74	50.26	59.76	17.7	40.24	-17.7	45.94	-6.7	54.06	6.7	43.51	-11.0	56.49	11.0
SPECHLO	50.57	49.43	59.76	35.1	40.24	-35.1	45.91	-17.8	54.09	17.8	46.03	-17.3	53.97	17.3
METMITO	43.92	56.08	52.50	24.6	47.50	-24.6	32.82	-31.8	67.18	31.8	46.45	7.2	53.55	-7.2
PLAMITO	48.93	51.07	58.75	31.3	41.25	-31.3	42.27	-21.3	57.73	21.3	45.77	-10.1	54.23	10.1

TABLE 2 *Continued.*

	RRN		RYN		YRN		YYN		RNR		RNY		YNR		YNY	
	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV
AMPEUKO	38.89	24.6	24.65	-0.4	15.81	-26.2	20.64	0.5	31.53	4.3	32.01	21.1	14.42	-30.2	22.03	4.8
ARIEUKO	33.03	39.4	26.05	6.4	16.91	-48.0	24.01	1.6	27.60	7.7	31.49	38.6	17.19	-46.3	23.73	-0.1
ARTEUKO	33.91	40.2	27.20	14.1	17.21	-48.4	21.68	-7.2	27.04	-1.5	34.06	57.0	17.47	-46.8	21.42	-8.9
ASCEUKO	35.11	13.0	25.65	1.2	16.74	-13.7	22.50	-0.8	28.31	1.9	32.45	12.5	15.29	-16.1	23.94	1.7
AVEEUKO	36.02	33.7	27.46	12.6	14.54	-50.2	21.98	2.5	31.31	11.8	32.17	35.5	15.55	-45.4	20.98	-2.7
ECHEUKO	35.14	12.1	30.09	8.9	16.76	-14.1	18.01	-7.7	29.64	2.9	35.59	18.5	11.67	-22.9	23.09	1.5
LAGEUKO	32.40	21.6	27.04	7.9	17.01	-30.4	23.55	0.4	27.84	4.6	31.60	25.2	18.07	-26.3	22.49	-3.7
MYCEUKO	35.46	81.2	27.40	23.5	13.21	-113.5	23.92	7.0	26.63	-2.0	36.23	108.7	17.27	-74.3	19.86	-33.2
PISEUKO	33.02	16.3	26.59	4.2	16.89	-21.0	23.51	0.2	28.71	5.4	30.89	15.4	16.52	-22.0	23.88	1.1
PRIEUKO	32.91	91.1	25.49	7.9	17.38	-115.2	24.23	14.5	28.32	23.0	30.08	77.5	18.21	-102.5	23.39	1.5
RODEUKO	33.15	79.0	26.45	19.1	16.55	-107.9	23.84	8.1	28.83	25.0	30.77	74.4	17.28	-98.5	23.11	-1.5
SPEEUKO	29.00	23.6	25.77	5.0	19.75	-34.4	25.48	5.7	24.82	-3.7	29.94	32.4	22.82	-14.3	22.42	-14.5
CILEUKO	33.52	10.0	33.50	14.0	15.87	-14.9	17.11	-9.7	28.70	2.3	38.32	21.9	11.51	-22.0	21.47	-2.3
FLAEUKO	34.83	11.5	29.31	10.7	15.01	-23.2	20.86	0.4	34.43	10.6	29.71	11.7	15.12	-23.0	20.74	0.1
SPOEUKO	44.72	24.4	23.09	-2.9	13.02	-33.5	19.16	10.5	37.51	4.8	30.30	19.0	16.25	-23.7	15.93	-0.9
ACTPROO	30.71	16.7	31.22	18.2	16.29	-25.5	21.78	-9.4	23.69	-3.9	38.24	38.8	17.42	-22.2	20.65	-12.7
CYAPROO	28.62	12.7	31.17	17.1	18.18	-18.8	22.03	-10.7	22.19	-5.3	37.59	34.9	18.45	-18.1	21.77	-11.4
ENDPROO	36.15	34.0	28.56	22.7	14.73	-59.9	20.57	1.2	32.47	13.3	32.23	44.6	16.53	-49.1	18.76	-10.3
GRNPROO	31.21	57.8	30.54	76.4	17.15	-107.0	21.10	-29.3	26.86	-0.4	34.89	136.0	18.58	-87.5	19.67	-49.4
GRPPROO	38.44	25.0	27.93	12.0	13.98	-39.4	19.65	0.5	32.63	5.1	33.73	33.4	16.00	-31.9	17.63	-7.6
METPROO	38.32	11.3	32.31	14.8	10.76	-26.4	18.61	-0.7	37.80	10.4	32.82	15.8	11.99	-24.1	17.39	-3.2
PHOPROO	27.81	6.8	32.81	12.8	16.78	-13.4	22.60	-5.9	22.76	-1.6	37.85	21.0	18.16	-11.1	21.22	-8.2
DDEVIRO	27.31	26.3	27.93	23.6	18.47	-52.5	26.30	2.9	26.54	20.1	28.69	29.7	19.66	-42.9	25.11	-6.6
DDNVIRO	34.40	43.5	27.90	22.6	16.61	-62.3	21.09	-5.9	30.12	12.6	32.18	54.8	16.36	-64.2	21.35	-3.9
DRNVIRO	31.37	6.1	30.72	18.4	15.02	-30.7	22.90	6.2	33.20	11.5	28.88	12.7	20.78	-12.7	17.14	-12.9
NCLVIRO	33.12	17.9	29.09	16.7	15.90	-35.8	21.89	0.4	31.72	12.5	30.49	22.3	17.03	-31.3	20.76	-4.3
SDNVIRO	31.70	25.2	27.64	8.4	16.56	-26.7	24.10	-6.4	22.82	-3.4	36.53	36.6	16.16	-28.0	24.50	-5.2
SREVIRO	33.32	48.6	27.22	28.2	16.01	-105.3	23.45	27.5	32.87	43.5	27.67	33.6	18.92	-70.6	20.54	-9.1
SRNVIRO	32.48	46.4	29.39	37.6	14.94	-85.1	23.18	0.0	29.93	25.3	31.94	59.2	16.30	-73.5	21.82	-11.9
PHYCHLO	29.63	10.0	30.13	10.5	16.31	-17.8	23.93	-2.7	26.53	3.7	33.23	16.8	16.99	-16.4	23.26	-4.1
SPECHLO	30.42	21.2	29.34	19.2	15.49	-41.9	24.75	1.4	28.41	12.4	31.36	28.0	17.62	-32.5	22.61	-8.1
METMITO	18.88	-1.5	33.62	29.7	13.94	-35.3	33.56	6.5	23.26	14.3	29.24	15.2	23.19	-4.7	24.31	-21.9
PLAMITO	29.04	19.1	29.71	17.4	13.23	-43.3	28.02	7.0	26.11	8.1	32.65	28.2	19.66	-19.6	21.58	-16.3

TABLE 2 *Continued.*

	NRR		NRY		NYR		NYY	
	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV
AMPEUKO	27.46	-6.9	27.25	7.2	18.50	-18.3	26.79	19.7
ARIEUKO	25.07	-7.1	24.87	-0.7	19.72	-31.3	30.34	39.9
ARTEUKO	22.88	-26.8	28.24	20.6	21.63	-20.7	27.25	28.7
ASCEUKO	30.57	5.6	21.28	-6.1	13.04	-19.8	35.11	20.8
AVEEUKO	25.98	-12.9	24.58	-1.4	20.87	-19.4	28.56	36.2
ECHEUKO	27.29	-1.0	24.61	-0.5	14.03	-18.8	34.07	21.2
LAGEUKO	24.25	-8.8	25.15	0.7	21.66	-12.6	28.94	21.4
MYCEUKO	22.72	-38.9	25.96	9.6	21.19	-36.5	30.14	68.5
PISEUKO	24.43	-5.5	25.48	1.3	20.80	-10.8	29.29	15.5
PRIEUKO	25.51	-18.6	24.77	-3.0	21.02	-59.9	28.70	83.9
RODEUKO	25.23	-20.2	24.48	-6.3	20.89	-52.3	29.41	81.2
SPEEUKO	26.08	4.5	22.67	-15.3	21.56	-22.6	29.69	33.5
CILEUKO	22.55	-7.5	26.85	3.1	17.67	-11.9	32.94	17.0
FLAEUKO	24.17	-12.4	25.67	2.1	25.38	1.4	24.78	10.3
SPOEUKO	28.45	-19.7	29.30	16.0	25.31	3.9	16.94	2.7
ACTPROO	18.89	-17.9	28.11	9.1	22.22	-8.1	30.77	16.9
CYAPROO	18.88	-14.6	27.92	8.1	21.76	-8.9	31.44	15.0
ENDPROO	24.26	-32.8	26.61	11.0	24.74	-0.1	24.38	25.6
GRNPROO	19.87	-93.8	28.49	48.3	25.57	8.2	26.07	40.6
GRPPROO	25.07	-20.9	27.35	9.9	23.57	-4.0	24.01	18.0
METPROO	25.98	-10.5	23.10	-2.8	23.81	-1.5	27.11	17.2
PHOPROO	17.18	-10.9	27.41	4.0	23.74	-2.0	31.67	8.6
DDEVIRO	21.11	-24.2	24.67	-2.6	25.09	0.8	29.13	25.4
DDNVIRO	24.22	-29.9	26.78	14.2	22.25	-19.9	26.74	38.5
DRNVIRO	22.74	-19.6	23.65	-3.7	31.24	20.1	22.37	4.5
NCLVIRO	23.73	-18.0	25.30	1.6	25.02	0.5	25.95	17.4
SDNVIRO	20.01	-12.4	28.25	10.3	18.97	-19.1	32.77	20.7
SREVIRO	27.04	-22.7	22.29	-30.5	24.75	-1.2	25.92	58.6
SRNVIRO	22.40	-37.1	25.02	0.5	23.83	-9.6	28.75	48.4
PHYCHLO	19.14	-11.5	26.79	3.7	24.37	-1.3	29.69	9.0
SPECHLO	22.82	-12.0	23.09	-8.4	23.20	-7.9	30.89	28.7
METMITO	15.20	-14.8	17.62	-23.1	31.25	21.9	35.93	13.7
PLAMITO	17.80	-23.0	24.47	-1.9	27.97	11.0	29.76	13.4

TABLE 2 *Continued.*

	RRR		RRY		RYR		RYY		YRR		YRY		YYR		YYY	
	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV	Prob.	RCV
AMPEUKO	21.19	16.3	17.70	15.2	10.34	-11.8	14.31	12.4	6.26	-26.9	9.55	-6.6	8.16	-12.2	12.48	13.9
ARIEUKO	17.64	31.4	15.39	19.9	9.96	-22.0	16.10	30.9	7.43	-41.4	9.48	-21.2	9.76	-19.0	14.24	21.5
ARTEUKO	16.27	15.6	17.63	37.0	10.77	-18.2	16.43	37.7	6.61	-51.7	10.60	-11.0	10.86	-8.8	10.82	-0.6
ASCEUKO	23.38	19.2	11.73	-2.7	4.93	-17.3	20.72	19.5	7.18	-12.5	9.55	-5.3	8.11	-8.6	14.39	7.8
AVEEUKO	19.92	26.2	16.10	17.1	11.39	-12.1	16.07	29.9	6.06	-45.0	8.48	-20.1	9.48	-13.5	12.50	17.7
ECHEUKO	21.43	14.2	13.71	1.2	8.21	-11.0	21.88	23.5	5.85	-16.2	10.90	-2.0	5.82	-13.8	12.19	4.1
LAGEUKO	16.49	13.2	15.91	14.9	11.36	-7.5	15.69	18.2	7.77	-25.2	9.24	-14.4	10.30	-9.1	13.25	9.9
MYCEUKO	17.49	43.2	17.97	62.7	9.14	-47.3	18.26	80.0	5.23	-96.0	7.99	-52.0	12.05	0.2	11.88	9.1
PISEUKO	17.74	13.1	15.28	8.0	10.98	-6.4	15.61	12.1	6.69	-20.8	10.20	-6.5	9.83	-7.8	13.68	8.2
PRIEUKO	17.68	73.0	15.22	45.1	10.63	-44.8	14.86	56.6	7.82	-99.7	9.55	-50.4	10.39	-33.5	13.84	54.0
RODEUKO	17.81	62.6	15.34	39.8	11.03	-31.6	15.43	57.9	7.42	-91.2	9.13	-49.4	9.86	-37.0	13.98	49.0
SPEEUKO	14.73	16.5	14.27	14.3	10.09	-21.4	15.67	28.2	11.35	-10.7	8.40	-34.5	11.47	-8.1	14.01	15.6
CILEUKO	17.96	7.6	15.55	5.3	10.73	-4.8	22.76	23.7	4.58	-17.8	11.29	-1.4	6.93	-10.9	10.18	-1.7
FLAEUKO	18.45	6.3	16.38	8.6	15.98	7.4	13.33	6.6	5.72	-23.4	9.29	-6.5	9.40	-6.1	11.45	7.2
SPOEUKO	24.01	8.4	20.71	23.5	13.50	-3.2	9.59	-0.4	4.44	-36.8	8.58	-4.8	11.81	9.4	7.35	4.4
ACTPROO	11.87	-2.4	18.84	24.3	11.81	-2.7	19.41	26.5	7.01	-21.0	9.28	-12.3	10.41	-8.0	11.37	-4.3
CYAPROO	11.17	-2.4	17.45	18.9	11.03	-4.5	20.14	26.6	7.71	-16.6	10.47	-8.1	10.73	-7.2	11.30	-6.8
ENDPROO	17.89	9.6	18.25	35.1	14.58	7.5	13.98	22.9	6.37	-54.3	8.36	-23.0	10.17	-8.3	10.40	10.7
GRNPROO	13.26	-11.5	17.94	88.3	13.59	11.4	16.95	89.8	6.61	-112.0	10.55	-26.9	11.98	-0.8	9.13	-38.4
GRPPROO	19.54	9.1	18.89	23.7	13.09	-3.1	14.84	20.1	5.53	-38.0	8.46	-12.5	10.48	-2.1	9.18	3.1
METPROO	22.22	9.0	16.09	5.3	15.58	4.0	16.73	16.1	3.75	-24.1	7.01	-9.8	8.23	-6.6	10.38	6.4
PHOPROO	11.06	-1.1	16.74	9.9	11.70	-1.0	21.10	17.5	6.12	-13.1	10.66	-4.6	12.04	-1.6	10.56	-6.1
DDEVIRO	13.52	18.4	13.79	16.2	13.03	8.1	14.90	22.7	7.59	-49.6	10.88	-19.3	12.07	-6.9	14.23	10.5
DDNVIRO	17.51	21.9	16.88	34.8	12.60	-6.3	15.30	37.1	6.71	-62.8	9.90	-17.6	9.65	-20.2	11.45	13.3
DRNVIRO	15.89	0.0	15.48	8.1	17.32	15.3	13.40	8.6	6.85	-26.2	8.17	-13.7	13.93	10.8	8.97	-3.1
NCLVIRO	17.44	10.9	15.68	12.2	14.27	5.1	14.81	17.2	6.28	-35.5	9.62	-10.7	10.75	-4.7	11.14	5.6
SDNVIRO	13.19	6.5	18.51	26.4	9.63	-10.8	18.01	21.5	6.82	-22.6	9.74	-12.5	9.34	-14.1	14.76	5.6
SREVIRO	19.44	53.7	13.88	7.4	13.43	0.6	13.79	37.7	7.60	-87.5	8.41	-49.3	11.32	-2.3	12.13	40.2
SRNVIRO	16.73	29.7	15.75	30.7	13.20	2.8	16.19	47.1	5.67	-79.6	9.27	-31.1	10.63	-15.7	12.55	16.1
PHYCHLO	13.69	3.7	15.94	9.4	12.84	1.1	17.29	12.6	5.45	-18.7	10.86	-4.6	11.53	-2.8	12.40	-0.8
SPECHLO	15.93	17.1	14.50	10.7	12.48	-0.9	16.86	26.1	6.89	-33.0	8.59	-21.8	10.73	-9.5	14.02	11.4
METMITO	8.72	1.3	10.16	-3.0	14.54	17.0	19.08	21.7	6.48	-19.9	7.46	-26.2	16.71	12.0	16.85	-2.9
PLAMITO	13.12	7.0	15.92	18.0	12.98	3.7	16.73	19.0	4.67	-36.8	8.55	-20.1	14.99	10.6	13.03	-1.4

TABLE 3

Occurrence Comparison Results with the FS Statistic (Top Value) and with the RCVS Statistic (Bottom Value) for the 109 Codon Pairs Constituted of Two Codons of the Same State.

c' / c	RNN	YNN	NRN	NYN	NNR	NNY
RNN		5.01 5.01	5.01 5.01	4.55 4.98	5.01 5.01	4.89 5.01
YNN			-4.55 -4.98	-5.01 -5.01	-4.89 -5.01	-5.01 -5.01
NRN				NS -5.01	3.35 3.24	-4.23 -5.01
NYN					4.23 5.01	-3.35 -3.24
NNR						-4.39 -4.98
NNY						

Only the results of the top triangular table are given because $c' \neq c$ and $Z(c, c'; L) = -Z(c', c; L)$.
NS = a result that is nonsignificant at the 1% statistical level (see Method).

TABLE 3 *Continued.*

c' / c	RRN	RYN	YRN	YYN	RNR	RNY	YNR	YNY	NRR	NRY	NYR	NYY
RRN		3.74 2.97	5.01 5.01	4.55 4.96	4.60 4.65	NS -3.39	4.99 5.01	4.98 5.01	5.01 5.01	5.01 5.01	4.62 4.71	3.23 NS
RYN			5.01 5.01	5.01 4.53	NS 3.17	-4.60 -4.65	5.01 5.01	5.01 4.89	4.15 4.98	4.17 4.35	4.94 4.94	NS -3.08
YRN				-5.01 -5.01	-5.01 -5.01	-5.01 -5.01	-2.97 -3.65	-5.01 -5.01	-5.01 -5.01	-5.01 -5.01	-4.76 -4.92	-5.01 -5.01
YYN					-4.30 -4.12	-4.96 -5.01	5.01 5.01	2.97 3.67	NS 4.71	-3.37 NS	NS 2.94	-4.94 -4.92
RNR						-3.17 -4.89	5.01 5.01	4.89 4.98	4.96 4.94	3.46 NS	4.46 4.53	NS -4.74
RNY							5.01 5.01	5.01 5.01	5.01 5.01	5.01 5.01	4.96 4.96	4.32 2.99
YNR								-4.73 -4.82	-4.46 -4.55	-4.87 -4.89	-4.96 -4.94	-5.01 -5.01
YNY									NS 4.17	-4.32 -3.24	NS NS	-5.01 -5.01
NRR										NS -4.23	NS NS	-4.23 -5.01
NRY											3.35 3.24	-3.67 -4.73
NYR												-4.06 -4.82
NYY												

TABLE 3 *Continued.*

c' / c	RRR	RRY	RYR	RYY	YRR	YRY	YYR	YYY
RRR		NS NS	4.35 4.06	NS -3.55	5.01 5.01	5.01 5.01	4.46 4.49	4.26 4.90
RRY			4.60 4.65	NS -3.42	5.01 5.01	5.01 5.01	4.67 4.82	4.32 2.87
RYR				-4.39 -4.92	4.96 4.94	4.19 3.89	3.06 NS	NS -3.08
RYY					5.01 5.01	5.01 5.01	4.94 4.92	5.01 4.80
YRR						-4.71 -4.65	-4.99 -5.01	-5.01 -5.01
YRY							-2.97 -3.67	-4.60 -4.76
YYR								-2.94 -3.62
YYY								

IDENTIFICATION OF THE COMMON AND RARE CODONS IN THE LIBRARY L WITH THE TWO STATISTICS FS AND RCVS

According to the results in Table 3, the common and rare codons identified are:

C(1;common;state 1;FS;L): RNN
C(2;common;state 1;FS;L): NNY
C(1;common;state 2;FS;L): {RRN,RNY}
C(2;common;state 2;FS;L): {RRN,RNY}
C(1;common;state 3;FS;L): {RRR,RRY,RYY}
C(2;common;state 3;FS;L): {RRR,RRY,RYY}
C(1;rare;state 1;FS;L): YNN
C(2;rare;state 1;FS;L): NNR
C(1;rare;state 2;FS;L): YRN
C(2;rare;state 2;FS;L): YNR
C(1;rare;state 3;FS;L): YRR
C(2;rare;state 3;FS;L): YRY
C(1;common;state 1;RCVS;L): RNN
C(2;common;state 1;RCVS;L): NNY
C(1;common;state 2;RCVS;L): RNY
C(2;common;state 2;RCVS;L): {RRN,NYY}
C(1;common;state 3;RCVS;L): RYY
C(2;common;state 3;RCVS;L): {RRR,RRY}
C(1;rare;state 1;RCVS;L): YNN
C(2;rare;state 1;RCVS;L): NNR
C(1;rare;state 2;RCVS;L): YRN
C(2;rare;state 2;RCVS;L): YNR
C(1;rare;state 3;RCVS;L): YRR
C(2;rare;state 3;RCVS;L): YRY

TABLE 4

Occurrence Comparison Results with the RCVS Statistic for the Codon Pairs Constituted of two Codons among RNN, RNY, and RYY (See Method and Table 3).

c / c'	RNN	RNY	RYY
RNN		... NS1.7	... 4.51
RNY			... 4.14
RYY			

The RCVS statistic applied to the codon pairs constituted of two codons among the first common codons of state 1 (RNN), of state 2 (RNY), and of state 3 (RYY), leads to the following first common codon order: *RNN, then RNY, and finally RYY* (see Table 4). However, $Z(\text{RNN},\text{RNY}; L) = 1.7$ is significant only at the 9% statistical level; it should also be observed that the four codons of state 2 with an R in the first codon site (i.e., RRN, RYN, RNR, and RNY) have observed occurrence probabilities greater than 0.25 and positive reduced centered variables with most of the gene populations (see Table 2).

Finally, another partition of the library L in 26 gene populations having at least 10,000 codons leads, with both statistics FS and RCVS, to identical results for the common and rare codons, for the codon pairs having non-significant results, and for the first common codon order (data not shown).

DISCUSSION

STATISTICAL REMARKS

This statistical study was carried out without any a priori choice of the data concerning both the gene populations and the open reading frames that belong to a given gene population. These data represent almost all the available genetic information taken from a great number of gene species (see Method).

The statistical methodology that uses the frequency and a reduced centered variable based on the average fraction of R and Y in the three codon sites leads to similar results for a codon occurrence study regardless of molecular evolution. Therefore, this variable can be considered a new index of codon usage. Furthermore, compared to the frequency statistic, this variable permits extension of the codon occurrence study and removal of certain nonsignificant results (see Results).

That two different statistics lead to a similar codon distribution is no light assertion. For example, another RCVS statistic has been compared to the frequency statistic by making use of the same methodology. The fractions of R and Y in each codon site have been used to define the theoretical occurrence probability $p_i(c; P)$ of this new reduced centered variable; more precisely, $p(b_i; P)$ is the observed occurrence probability of the base b_i in the ith site of the codon c in the gene population P. This new RCVS statistic contains several contradictions of the FS statistic; for example, with the specified codons, there are six codon pairs that have a $Z(c, c'; L)$ value greater than 2.58 with one statistic and a $Z(c, c'; L)$ value less than -2.58 with the other statistic (data not shown). The transformation of biological information into statistical data leads to a loss of information that can be minimized by choosing different statistical "parameters": frequency, observed/expected ratios, reduced centered variables, autocorrelation functions, etc. Nevertheless, to obtain a coherent biological interpretation of the observations, these parameters should be classified, and the frequency could be used as the reference parameter. It should be noted that the methodology developed herein allows the comparison of quantitative observations and could be applied to such a classification.

A MODEL OF CODON EVOLUTION AT THREE SUCCESSIVE STATES: RNN, THEN RNY, AND FINALLY RYY

Biological evolution can be defined as a transformation of one state into a more specific state. In particular, this definition can be applied to the primordial codon evolution. A preliminary condition for this codon evolution is the existence of an alphabet with at least two letters (see, for example, some analogies of information theory). According to the chemical properties of adenine, cytosine, guanine, and thymine, the purine/pyrimidine alphabet is the best-adapted type for this evolution. It verifies the previous condition in a minimal way (only two letters), and it is chemically unambiguous. Indeed, the large purine base is well differentiated from the smaller pyrimidine base, their respective molecular weights being different, and so on. These chemical properties are important for the spatial organization of the DNA molecule. Based on these remarks, a model of codon evolution can be proposed in terms of three successive states, each state being a base specification of a codon site. Such a model leads to the following prediction:

> A codon is older in ancestry than another if it is less specified and consequently its presence must be stronger in the present-day genes.

Note: This model is in opposition to the phylogenetic model according to which the primordial features are lost in the present-day genes.

Obviously, statistical methods are one approach to testing the proposed model by identifying the common and rare codons.

The first common codon order—RNN, then RNY, and finally RYY (see results)—agrees well with the three successive states of the proposed model (see above).

Parallel to the formation of a "codon world," a "rare-codon world" (comprising the rare codons) is also selected by biological evolution. The YRR codon is rare (see Results) because it encodes the three stop codons. On the other hand, the rarity of the YRY codon (see Results) may be explained by the $YRY(N)_6YRY$ preferential occurrence (see Introduction and below).

The methodology developed here generalizes the preferential occurrence of the RNY codon to all the codons at one unspecified base N (RNY is the first common codon at state 2).

Two results have to be added to the primordial codon evolution theories.

(1) The RNN codon is more common than the RNY (see Results). This result may have the following interpretations. An RNN series can contain an alternating purine/pyrimidine stretch every two triplets, while an RNY series always conflicts. It should be remembered that the eukaryotic introns are characterized by alternating purine/pyrimidine stretches [3]. Therefore, the RNN structure (or one of its combinations) could well be a common ancestor of protein coding genes and introns. This observation also suggests that a single base (in our case, the first codon site) may interact with an amino acid, in agreement with several recent biological experiments: A specific amino acid binding site is detectable on the intron of the Tetrahymena self-splicing ribosomal precursor RNA [14], a single base pair can direct an amino acid to a specific transfer RNA [9], and so on.

(2) The RYY codon is the first common codon at state 3. The importance of RYY is explained below by stressing that the RRY codon was preferred to the RYY for protein synthesis [5].

Over the past few years, the main approach has been to analyze genetic information in terms of protein synthesis (according to, for example, the type and number of residues), while aspects concerning DNA spatial organization were neglected. The DNA spatial constraints have preceded those of protein synthesis because a small amount of DNA codes proteins. In agreement with this hypothesis, the 6-motif $YRY(N)_6YRY$ is older (according to the above model) than the RNY codon, because (1) only six out of 12, or one of every two bases are specified in the 6-motif $YRY(N)_6YRY$ while two out of three bases are specified in the RNY codon and (2) the 6-motif $YRY(N)_6YRY$ is found in a larger gene diversity. Indeed, the $YRY(N)_6YRY$ preferential occurrence is verified in the eukaryotic introns [3] and in the ribosomal and transfer RNA genes [2], while the RNY coding periodicity is

not verified in these three populations [2, 3]; however, an RNY message was identified with a phylogenetic analysis (this concept is different; see the note earlier in this section) with some 5S RNA sequences [7] and some tRNA sequences [6]. The first common codons RNN, RNY, and RYY, but not the RRY codon, permit the coexistence of a DNA double helix pitch code with a primordial transcription–translation code as well as their unambiguities, the 6-motif $YRY(N)_6YRY$ being shifted one base in the 5′ direction from the first common codons.

Otherwise, the 6-motif $YRY(N)_6YRY$ may still have a function. Indeed, experiments [13] showed that the RNA polymerase process unwinds in an open complex, a 12-base segment that starts three bases after the first base position of the TATA box of the lac UV5 promoter, that is, after a motif YRY. These experiments [13] also proved that the start position of the mRNA synthesis is preceded by a motif YRR (one base mutation from YRY) that occurs exactly six bases after the previous 5′ motif YRY. Furthermore, other promoter sequences have also, exactly at the mRNA upstream start site, a 6-motif of the general form $YRY(N)_6YRY$: $YRY(N)_6YRY$ for λc17, $RRY(N)_6YRY$ for λp_L and λP_R, $YRR(N)_6YRY$ for λPRM, $YRY(N)_6RRY$ for λcin and P_{ant}, $YRY(N)_6YYY$ for araC, $galP_1$, fdII, and φX174B, $YRY(N)_6YRR$ for lacP115 [11, 15], etc. It remains to be verified if these features are fortuitous or not.

Finally, these statistical results may lead to an interesting application to determine an open reading frame. Indeed, since RNN is the most common codon (see results), the open reading frame may be assigned to the modulo 3 frame having the highest number of purine bases in the first position of the triplets (Michel, in preparation).

We thank Professors Didier Arquès, Max Burger, and Jacques Streith, Dr. Christoph Nager, Alan Horsfield, Thomas Nyffenegger, Nouchine Soltanifar and the referees for their advice, and members of the Bioinformatic Group for their assistance. This work was supported by grants from the Friedrich Miescher Institute to C.J.M.

REFERENCES

1 D. G. Arquès and C. J. Michel, Study of a perturbation in the coding periodicity, *Math. Biosci. 86*:1–14 (1987).

2 D. G. Arquès and C. J. Michel, A purine-pyrimidine motif verifying an identical presence in almost all gene taxonomic groups, *J. Theor. Biol. 128*:457–461 (1987).

3 D. G. Arquès and C. J. Michel, Periodicities in introns, *Nucleic Acids Res. 15*:7581–7592 (1987).

4 F. H. C. Crick, J. S. Griffith, and L. E. Orgel, Codes without commas, *Proc. Natl. Acad. Sci. U.S.A. 43*:416–421 (1957).

5 M. Eigen and P. Schuster, The hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle, *Naturwissenschaften 65*:341–369 (1978).

6 M. Eigen and R. Winkler-Oswatitsch, Transfer-RNA: The early adaptor, *Naturwissenschaften 68*:217–228 (1981).
7 M. Eigen, B. Lindemann, R. Winkler-Oswatitsch, and C. H. Clarke, Pattern analysis of 5S rRNA, *Proc. Natl. Acad. Sci. U.S.A. 82*:2437–2441 (1985).
8 J. W. Fickett, Recognition of protein coding regions in DNA sequences, *Nucleic Acids Res. 10*:5303–5318 (1982).
9 Ya-M. Hou and P. Schimmel, A simple structural feature is a major determinant of the identity of a transfer RNA, *Nature 333*:140–145 (1988).
10 C. J. Michel, New statistical approach to discriminate between protein coding and non-coding regions in DNA sequences and its evaluation, *J. Theor. Biol. 120*:223–236 (1986).
11 M. Rosenberg and D. Court, Regulatory sequences involved in the promotion and termination of RNA transcription, *Ann. Rev. Genet. 13*:319–353 (1979).
12 J. C. W. Shepherd, Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification, *Proc. Natl. Acad. Sci. U.S.A. 78*:1596–1600 (1981).
13 U. Siebenlist, R. B. Simpson, and W. Gilbert, *E. coli* RNA polymerase interacts homologously with two different promoters, *Cell 20*:269–281 (1980).
14 M. Yarus, A specific amino acid binding site composed of RNA, *Science 240*:1751–1758 (1988).
15 P. Youderian, S. Bouvier, and M. M. Susskind, Sequence determinants of promoter activity, *Cell 30*:843–853 (1982).