J. Biosci., Vol. 7, Number 1, March 1985, pp. 49–59. © Printed in India.

Complimentary DNA sequence data analysis of prokaryotic systems

A. S. KOLASKAR and B. V. B. REDDY
Centre for Cellular and Molecular Biology, Hyderabad 500007, India

MS received 13 September 1984; revised 17 November 1984

Abstract. Complimentary DNA sequence data of $\phi \times 174$, fd, f1, G4, M13, MS2, λ and T7 phages of *Escherichia coli* are analysed at mono-, di-, tri- and tetranucleotide levels. Our analysis shows that, (i) mononucleotides have certain preferences to occur at specific positions X_1, X_2, X_3 of codon. (ii) These nucleotides interact nonlinearly to form dinucleotide and this dinucleotide also interacts nonlinearely with a third nucleotide to form codon. (iii) However, nonlinear interactions are negligible at tetranucleotide level suggesting that, coding regions of complimentary DNA are Markov chains of order two. Trinucleotide potential values in three frames have suggested that, at least thirteen different trinucleotides can be used as a marker to locate coding regions in DNA of prokaryotes. (iv) Parallel paired codons are expressed in such a way that one of the codons in the pair expresses with high frequency while the other with low frequency. On the other hand the complimentary codon pairs express with small frequency difference. (v) In the synonymous codon groups, codon ending with T are found to express with more frequency.

Keywords. DNA sequence data; mono-, di-, tri- and tetranucleotide level.

Introduction

In the last few years, due to availability of fast DNA sequencing methods, large amount of DNA sequence data has been accumulated. Analysis of this sequence data is carried out to understand and gain insight in certain problems in molecular biology. The usefulness of this type of approach and the results of these studies can be seen from some of the recent publications of Nussinov (1980, 1981a, b, 1984), Granthem (1978), Modiano *et al.* (1981), Shepherd (1981), Gouy and Gouteir (1982), Konigsberg and Godson (1983), Lipman *et al.* (1983) and Staden (1984). These studies have pointed out that in a given set of synonymous codons the usage does not seem to be random. Further, Nussinov (1980, 1981a) has also pointed out the presence of asymmetry at dinucleotide level. Thus there is a need to understand the pattern, if present, in codon usage and the choice of synonymous codons. With this in mind we have undertaken the studies of DNA sequence data analysis with the hope to correlate the usage of codons with three dimensional structure of DNA as well as with the rate of polypeptide chain elongation.

In such studies the choice of data base is very important. We have chosen, therefore, those systems for which complete genome sequence is known. This, we hope, will avoid any undue weightage to codons which have been used more predominantly in specific type of genes. Fortunately, all such systems also fall under one category, namely *Escherichia coli* phages. Thus, there is an internal consistency in the data base, but

generalisation of the results is difficult. However, we feel this choice reduces considerably the noise level, and, thus helps in retrieving useful information. This data base has also helped us to concentrate our efforts to understand the specific questions, *viz*, the distinction between the coding frame from the noncoding frame and, to distinguish translational initiator from internal ATG/GTG. In this communication we will discuss only the logic used to develop a method to predict coding sequence in prokaryotic systems using the results of our analysis at mono-, di-, tri- and tetranucleotide level. Our simple analysis has pointed out that the complimentary DNA (cDNA), molecules are in general Markov chain of order two and codons are the units of information transfer, which has already been intuitively accepted by the molecular biologist. The method used for analysis along with results are briefly discussed below. To our knowledge, no analysis has been carried on such a large data base at levels discussed below, though there are reports on analysis of prokaryotic phage systems. This large data base helped us to distinguish clearly statistical fluctuations from those which might have occurred due to their specific roles in biology.

Materials and methods

Complete genome sequences are known for MS_2, M13, $\phi \times 174$, fd, fl, λ and T7 phages of *E. coli*. The data are obtained from EMBL and the original papers (Sanger *et al.*, 1982; Dunn 1983). The number of proteins coded by each of these genomes as reported in literature and the total number of codons are given in table 1. Using cDNA sequence

Table 1. Total number of genes and codons in each of the *E. coli* phage systems considered.

Name of the systems	Number of codons	Number of genes
Phage MS2 (RNA)	1,071	3
Phage M13	2,066	10
Phage ϕx 174	1,996	10
Phage fd	2,066	10
Phage G4	2,038	11
Phage λ	13,737	64
Phage f1	1,885	10
Phage T7	13,552	58
Total	38,431	174

data of these 174 genes we have determined the nucleotide composition. In addition to this, frequencies of nucleotides occurring at 5′-end (X_1), middle (X_2) and at 3′-end (X_3) of codon are calculated and are given in table 2. Further, occurrence of dinucleotides were calculated in three different frames. If $X_1X_2X_3$ represents a codon, (where each digit represents the position of the nucleotide in the codon) and, X_1'' represents the 5′ nucleotide of the succeeding codon, then frequencies of various doublets X_1X_2, X_2X_3 and X_3X_1'' are computed. Using observed mononucleotide frequencies $f(X_i)$ and dinucleotide frequencies $f(X_iX_j)$ potential values are calculated for each type of

Table 2. Observed mononucleotide frequencies in various positions in a codon, potential values [$f(X_1)/f(X)$] (in square brackets) and the corresponding estimated standard errors.

Nucleotide	Nucleotide frequency in total data base f(X)	Nucleotide frequency $f(X_1)$ and potential	Nucleotide frequency $f(X_2)$ and potential	Nucleotide frequency $f(X_3)$ and potential
T	0·2555	0·178(5)	0·270(6)	0·319(9)
		[0·70(1)]	[1·06(2)]	[1·25(2)]
C	0·2275	0·203(4)	0·238(4)	0·242(5)
		[0·89(2)]	[1·04(2)]	[1·06(2)]
A	0·2627	0·276(4)	0·317(6)	0·195(5)
		[1·05(1)]	[1·21(2)]	[0·74(1)]
G	0·2542	0·344(6)	0·176(5)	0·244(6)
		[1·35(2)]	[0·69(2)]	[0·96(2)]

dinucleotides in various frames by the simple relation:

$$P(X_iX_j) = \frac{f(X_iX_j)}{f(X_i)\cdot f(X_j)} \quad (1)$$

where i and j vary from 1 to 3. For example, the potential of a pair of nucleotides AA in codons of type AAX_3 is the frequency of AA pair occurring in position X_1X_2 (0·080) divided by the product of frequency of A occurring at X_1 position (0·276) and the frequency of A occurring at X_2 position (0·317) of the codon. Calculated potential values for 16 types of dinucleotides in various frames along with the estimated standard errors are given in the table 3.

At triplet level, codon frequencies are obtained from cDNA sequence data along with the amino acid frequencies (figure 1). Codon potentials are calculated using observed dinucleotide and mononucleotide frequencies and these values are plotted in figure 3.

In order to study codon nucleotide interactions, two types of tetranucleotides $X'_3X_1X_2X_3$ and $X_1X_2X_3X''_1$, where X'_3 is the wobble nucleotide preceding the codon $X_1X_2X_3$ under consideration and X''_1 is the first nucleotide of the codon succeeding $X_1X_2X_3$, were considered. Frequencies of occurrence of these two types of tetranucleotides are obtained, and, potential values are calculated using observed codon frequencies and positional mononucleotide frequencies. These values are analysed and discussed below.

Results and discussions

In the data base we have used the complimentary DNA sequences corresponding to each mRNA. Our data include even the overlapping genes because of which certain oligonucleotides have been considered more than once during analysis as they have occurred in two different genes.

Table 3. Observed dinucleotide potential values in various frames.

Dinucleotide (XX)	Ps(XX)	$P(X_1X_2)$	$P(X_2X_3)$	$P(X_3X_1'')$
TT	1·04(2)	1·33(4)	1·03(3)	0·92(4)
TC	0·99(2)	1·25(4)	0·94(3)	0·91(3)
TA	0·78(2)	0·60(3)	0·79(3)	0·94(2)
TG	0·91(2)	0·88(4)	1·19(3)	1·15(2)
CT	1·10(2)	1·09(3)	1·11(3)	1·12(3)
CC	0·90(2)	0·79(4)	0·93(4)	0·97(4)
CA	0·99(2)	0·91(4)	1·13(4)	1·01(3)
CG	0·99(2)	1·30(5)	0·82(3)	0·95(3)
AT	0·97(2)	1·07(3)	0·78(3)	1·06(3)
AC	0·99(2)	0·91(2)	1·00(3)	1·04(4)
AA	1·16(2)	1·21(4)	1·25(4)	1·04(4)
AG	0·87(2)	0·62(3)	1·08(3)	0·91(3)
GT	0·92(2)	0·72(2)	1·19(4)	1·05(4)
GC	1·11(2)	1·06(3)	1·19(4)	1·12(4)
GA	1·04(2)	1·09(3)	0·70(4)	0·97(2)
GG	0·94(2)	1·18(4)	0·80(3)	0·92(2)

$$Ps = \frac{f(XX)}{f(X)\cdot f(X)};\ P(X_1X_2) = \frac{f(X_1X_2)}{f(X_1)\cdot f(X_2)};\ P(X_2X_3) = \frac{f(X_2X_3)}{f(X_2)\cdot f(X_3)}\ P(X_3X_1'') = \frac{f(X_3X_1'')}{f(X_3)\cdot f(X_1)}$$

The estimated standard errors are given in parenthesis.

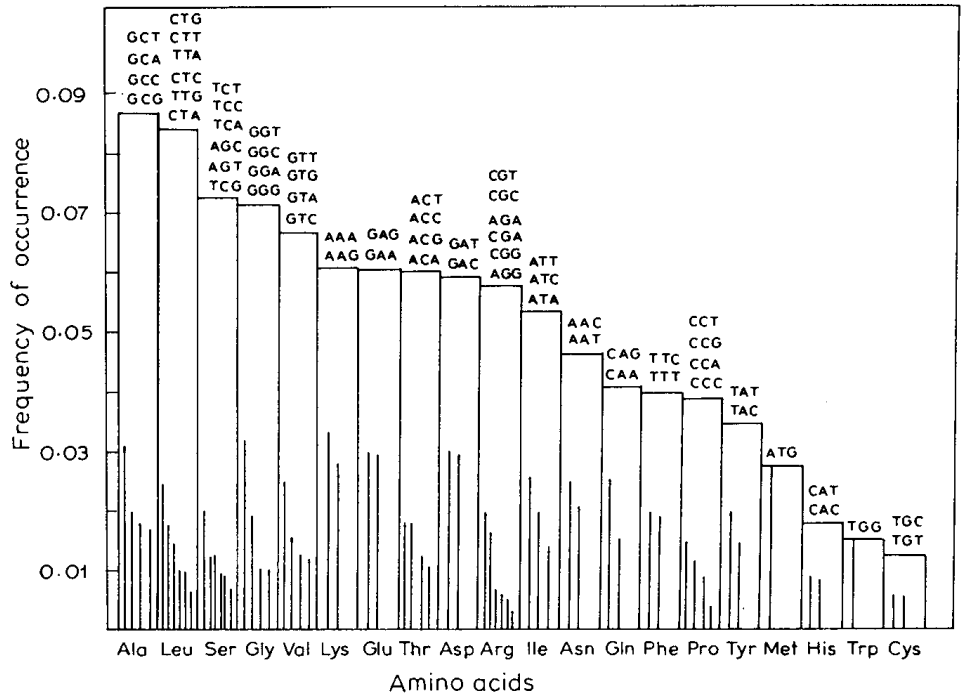

Figure 1. Histogram of amino acids along with the bar frequency diagram of codons observed in data base. The amino acid histogram shown here is very similar to the one obtained by Dolittle (1981) from a large number of eukaryotic and prokaryotic data. Codons ending with T are expressed with high frequency in each set of synonymous codons.

Analysis of mononucleotide data

Four nucleotides A, T, G and C occur with almost the same frequency in the data base under consideration, as is evident from the frequency values given in table 2. However, the study of distribution of these nucleotides in various positions X_1, X_2 and X_3 of the codon shows certain bias. Calculated potential values given in table 2 suggest that nucleotide G prefers X_1 position and occurs less frequently in X_2 position. On the other hand, T has least preference for X_1 position and maximum preference for X_3 position. X_2 position is preferred by A. Nucleotide C does not seem to have any positional preference. Estimated standard error associated with positional frequencies and potential values are also given in table 2. Mononucleotide frequencies which take into consideration the occurrence of nucleotides in specific positions of the codon can be used to calculate the expected codon frequencies. Comparison of the calculated and observed codon frequency values indicate that positional effect at mononucleotide level is important but not enough to give observed codon frequencies. This analysis suggests the need to analyse data at dinucleotide level.

Analysis of dinucleotide data

As mentioned in the method, we made use of observed frequencies of the single nucleotides to study the interactions at dinucleotide level. The potential values in table 3, point out that interactions between the nucleotides not only vary with the type of nucleotide that constitutes the pair but also vary with the position which they occupy at codon level. Therefore, simple analysis is carried out as has been done earlier for pair of amino acids in polypeptides (Kolaskar and Ramabrahmam, 1983) to study the presence of nonlinear interactions at dimer level. Linear interactions between nucleotides of the pair give rise to a simple relation of the type:

$$\frac{f(T_1T_2)}{f(C_1T_2)} = \frac{f(T_1C_2)}{f(C_1C_2)} = \frac{f(T_1A_2)}{f(C_1A_2)} = \frac{f(T_1G_2)}{f(C_1G_2)}$$

$$\frac{f(T_1T_2)}{f(T_1C_2)} = \frac{f(C_1T_2)}{f(C_1C_2)} = \frac{f(A_1T_2)}{f(A_1C_2)} = \frac{f(G_1T_2)}{f(G_1C_2)}$$

$$\frac{f(T_2G_3)}{f(T_2T_3)} = \frac{f(C_2G_3)}{f(C_2T_3)} = \frac{f(A_2G_3)}{f(A_2T_3)} = \frac{f(G_2G_3)}{f(G_2T_3)}$$

$$\vdots$$

etc.,

However, the equality is generally not obeyed by dinucleotides in X_1X_2 and X_2X_3 frames. This can be seen also from potential values given in table 3 for dinucleotides in various frames. The potential values in X_3X_1'' being close to unity for almost all types of dinucleotides indicate negligible nonlinear interactions among nucleotides in this frame at dinucleotide level. Similarly the Ps values in table 3, calculated without taking into consideration the specific position of any nucleotide, are pointers to suggest that the asymmetry reported by Nussinov (1980, 1981a) may be due to limited database in his analysis. The asymmetry present mainly in the coding frame, may be due to selection of nucleotides for transforming useful information and nonlinear interactions among them. This is quite clear from our potential values of dinucleotides in the frame X_1X_2 and X_2X_3. This analysis therefore, points out that probably nonlinear interactions in

polynucleotides are prominent up to trinucleotide level particularly in coding sequences and thus cDNA can be considered as Markov chain of order two. In order to get a better idea about this and the patterns which might be present at trinucleotide level in the coding frame, the data were analysed at tri- and tetranucleotide level.

Analysis of trinucleotides

The histogram of amino acid residues shown in figure 1, is quite similar to the one obtained by Dolittle (1981), where he has used amino acid compositions of large number of proteins from eukaryotic and prokaryotic systems. This suggests that codon frequency pattern reported in our analysis is also general in nature. The bar diagram of frequency of codons (figure 1) indicates that, in all the synonymous codon groups, those codons which end with T are expressed with maximum frequency, which was also pointed out earlier by Staden (1982). Non random usage of codons is also clear from figure 1 which was earlier observed by Shepherd (1981).

Further, analysis is carried out by arranging codons in ascending order of their frequency of occurrence. These frequency values are plotted in figure 2. Our data base being sufficiently large and the phages which we have used being quite different, the observed codon frequency indicate the thermodynamic stability of these codons. Variation from these frequency values can be due to specific environment and specific biological function. Therefore, if thermodynamics is the only criterion for usage of codons, then the frequency of codons in any genome should follow the pattern which we have observed for total data base. The variation from this value should therefore indicate mutability of the codon under consideration. Frequency of occurrence of codons in individual phage systems such as MS2, G4 and λ phage are also shown in figure 2 for comparison. As seen in figure 2, and observed by earlier workers, the codons

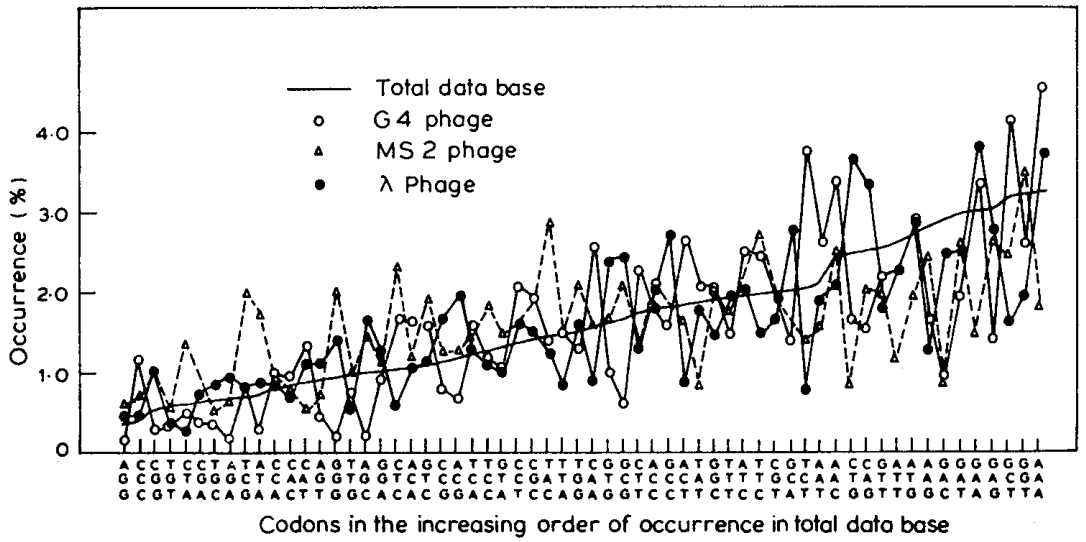

Figure 2. Frequency of occurrence of codons in total data base. The codons are arranged in their ascending order of frequency of occurrence in total data base. The codon frequencies are also shown as an example for individual phage systems G4, MS_2 and λ phages for comparision.

which occur maximally, irrespective of the amino acids which they code, do not have a common property and the only rational explanation comes from studies of Ikemura (1981a, b). However, when one studies the pairs of parallel and antiparallel codons we have observed following weak correlation. The pair of codon and complimentary codon occur either with high or low frequency indicating that in cDNA molecules the structure is stabilised due to pairing of such codons. The structure can be stabilised through the hydrogen bonds by pairing the codons in parallel direction with complimentary bases. For example, ATT can pair in parallel direction with TAA, while in antiparallel direction with AAT. Thus, if one studies the occurrence of parallel paired triplets, one observes again a weak correlation, namely, if one triplet occures with high frequency its parallel paired codon occurs with low frequency, indicating that, such pairing even if it were to exist, is present rarely. We have also observed that, even at individual genome level the correlation is valid. This we are mentioning mainly because recently Root-Bernstein (1982), has suggested that amino acids, which are coded by pair of parallel, paired codon triplets interact maximally. Further, we could observe this correlations mainly due to the selection of proper data base consisting of complete genome sequences.

Potential values are calculated for trinucleotides $X_1X_2X_3$, $X_2X_3X_1''$ and $X_3X_1''X_2''$ using positional dinucleotide and mononucleotide frequencies. Potential values in three frames are plotted in figure 3. In the coding frame $X_1X_2X_3$ there are more than 50% triplets which have potential values outside the range $0{\cdot}8 \geqslant P_1 \geqslant 1{\cdot}20$. These values point out that it is not sufficient to take into consideration the corresponding dinucleotide and mononucleotide frequencies to get the frequency of codons. Our analysis further points out that interactions between the dinucleotide and mononu-

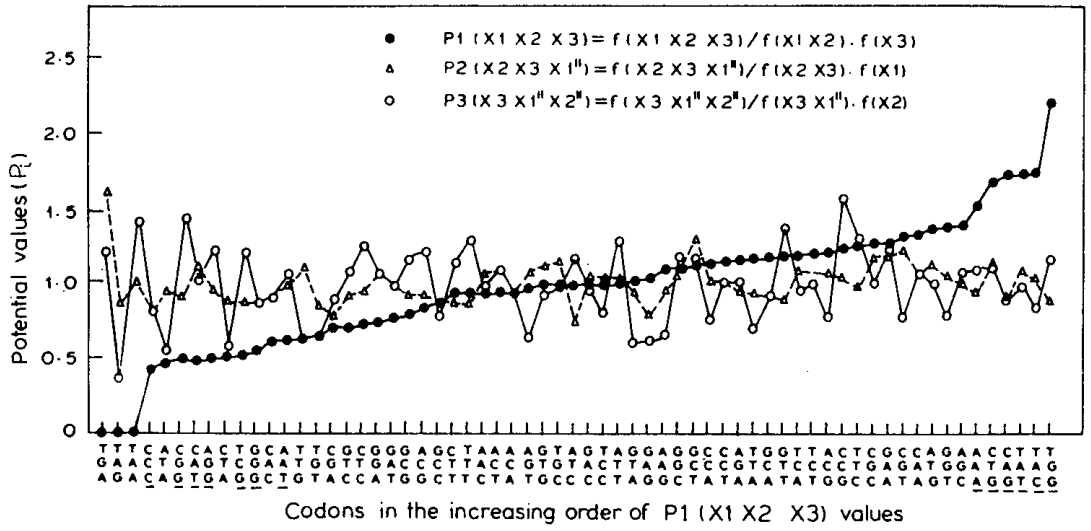

Figure 3. Plot of potential values against codons. Trinucleotides are arranged in the ascending order of potential values in coding frame $[P_1(X_1X_2X_3)]$ along the X-axis and the potential values along the Y-axis. Potential values in the other two frames are also plotted. Trinucleotides having potential values $P_1 \geqslant 1{\cdot}5$ and $P_1 \leqslant 0{\cdot}64$ in coding frame, but have P_2 and $P_3 \approx 1$ (in noncoding frames) are underlined.

cleotide forming a triplet are nonlinear in nature. However, the nonlinearity is least in the frame $X_2X_3X_1''$.

From the figure 3 it can also be seen that there are at least 13 triplets which have either low ($\leqslant 0{\cdot}64$) or high ($\geqslant 1{\cdot}5$) potential values in coding frame and in other frames the potential values of these triplets are either near unity or quite different from the corresponding values in coding frame. This pattern we have used to distinguish protein coding sequences from noncoding sequences. The algorithm developed and used to predict coding sequences in prokaryotic systems is discussed elsewhere.

Analysis of tetranucleotide data

The effect of neighbouring nucleotide on the codon can be seen from the data given in table 4, where the number of times the tetranucleotides $X_1X_2X_3 \cdot X$ and $X \cdot X_1X_2X_3$ occur is given. These data suggest that when X = A or G the tetranucleotide $X_1X_2X_3 \cdot X$ occur more number of times than $X \cdot X_1X_2X_3$. On the other hand when X = T or C the tetranucleotide $X \cdot X_1X_2X_3$ occur more number of times than the corresponding tetranucleotides of type $X_1X_2X_3 \cdot X$. It may be mentioned here that T and C mononucleotides have higher preference to occur in position X_3 while A and G have potentials greater than unity for X_1 position in codon. This suggests that in general tetranucleotide occurrence can be related to single nucleotide occurrence.

Table 4. Occurrence of tetranucleotides in the data base for the frames considered.

XXX	A·XXX	XXX·A	T·XXX	XXX·T	G·XXX	XXX·G	C·XXX	XXX·C
AAA	247	317	394	225	304	425	280	258
AAT	179	214	312	128	173	321	153	154
AAG	171	283	400	188	167	372	246	241
AAC	165	246	312	190	245	298	214	202
ATA	73	44	68	44	87	89	52	56
ATT	159	153	344	124	261	392	233	226
ATG	171	302	219	144	273	328	210	258
ATC	216	249	244	114	48	230	233	148
AGA	63	90	47	27	57	83	93	60
AGT	71	78	76	66	101	160	111	55
AGG	38	39	19	27	40	38	44	37
AGC	84	110	67	57	80	142	162	84
ACA	121	111	114	86	101	152	93	80
ACT	138	166	237	122	154	272	174	143
ACG	98	95	109	97	144	145	115	129
ACC	149	207	193	130	170	230	175	120
TAA	30	0	19	1	25	0	11	1
TAT	136	147	248	142	175	280	169	159
TAG	4	0	2	1	6	0	4	0
TAC	88	158	171	98	137	201	160	99
TTA	137	155	179	111	126	160	116	134
TTT	122	206	227	114	187	309	203	110
TTG	103	109	110	67	87	95	80	109
TTC	138	193	182	149	262	221	168	188
TGA	29	0	5	1	14	0	31	0
TGT	37	61	60	44	53	73	68	40

Table 4. (*Continued*)

XXX	A·XXX	XXX·A	T·XXX	XXX·T	G·XXX	XXX·G	C·XXX	XXX·C
TGG	129	190	102	113	170	162	178	114
TGC	56	67	55	60	50	72	94	56
TCA	116	138	133	97	109	151	114	86
TCT	126	185	258	114	161	335	240	151
TCG	58	51	74	79	69	71	64	64
TCC	97	185	161	84	105	206	136	78
GAA	204	337	419	177	298	387	223	243
GAT	172	345	435	234	282	481	241	70
GAG	241	277	474	250	224	384	180	245
GAC	195	273	479	220	208	344	225	270
GTA	91	135	154	109	143	175	125	94
GTT	138	241	346	164	217	345	276	220
GTG	92	185	193	108	182	197	138	131
GTC	83	152	159	71	104	132	97	88
GGA	87	134	119	83	72	98	123	86
GGT	221	351	493	176	181	472	329	225
GGG	72	112	97	17	86	100	118	91
GGC	128	206	297	191	115	253	194	84
GCA	169	237	210	152	203	220	187	160
GCT	168	309	416	160	315	515	299	221
GCG	124	155	231	136	197	200	92	153
GCC	147	216	264	129	130	243	158	111
CAA	113	142	200	123	156	178	120	146
CAT	73	85	95	69	92	115	86	77
CAG	220	286	306	179	281	312	154	184
CAC	63	96	100	72	89	95	85	72
CTA	58	72	56	52	67	48	56	65
CTT	113	189	210	105	186	266	172	121
CTG	209	280	238	139	323	323	184	212
CTC	88	122	123	81	98	128	100	78
CGA	62	79	46	50	75	71	67	50
CGT	123	187	213	121	191	285	231	165
CGG	46	45	41	40	52	75	73	52
CGC	109	129	188	151	115	150	144	128
CCA	75	105	80	68	97	95	75	59
CCT	116	138	157	103	142	185	141	113
CCG	77	96	151	98	126	155	90	95
CCC	33	33	46	39	25	60	47	19

Potential values are calculated for two types of tetranucleotides $X'_3 \cdot X_1X_2X_3$ and $X_1X_2X_3 \cdot X''_1$ using observed mononucleotide and codon frequencies, which constitute a particular tetranucleotide. It was observed that the potential values, Pq″ and Pq′ for most of the tetranucleotides are near unity. Only a small number (less than 25%) of tetranucleotides have potential values differing from unity as seen from the table 5a, b. The values indicate that the interaction between nucleotide and a codon is linear in nature, in the direction of 5′- to 3′-end. In other words this analysis suggests that mRNA or cDNA are Markov chains of order two and non-linear interactions vanish beyond trinucleotides. The analysis at tetranucleotide level also points out that the unit of

Table 5a. Tetranucleotides in the frame $X_1X_2X_3 \cdot X_1''$ whose potential values (P_q'') are different from unity.

Quarterates with $P_q'' \leqslant 0{\cdot}80$				Quarterates with $P_q'' \geqslant 1{\cdot}20$			
ATT·T	ATG·T	AGA·T	AGT·A	ATC·A	AGA·A	AGT·G	AGG·C
AGT·C	AGG·G	TTT·C	TTT·G	ACG·C	TTT·G	TTG·C	TTC·C
TCG·A	TCG·G	TCC·C	GAT·C	TGC·T	TCT·G	TCG·T	GAT·G
GGA·G	GGG·G	GGC·C	GCT·T	GAG·T	GTC·A	GGA·A	GGC·T
CTA·G	CGG·A	CGC·G	CCG·A	GCT·G	CAA·C	CTA·T	CTA·C
CCC·A	CCC·C			CGG·C	CCG·T	CCC·T	

Table 5b. Tetranucleotides in the frame $X_3' \cdot X_1X_2X_3$ whose potential values (P_q') differ from unity.

Quarterates with $P_q' \leqslant 0{\cdot}80$				Quarterates with $P_q' \geqslant 1{\cdot}20$			
A·AAG	A·GAT	A·GTT	A·GTG	A·ATA	A·ATC	A·AGA	A·AGG
A·GCT	T·ATA	T·ATG	T·AGA	A·ACA	A·TTA	A·TTG	A·TCA
T·AGT	T·AGG	T·AGC	T·ACG	A·CTA	A·CGA	T·GAT	T·GAG
T·TTC	T·TGG	T·TGC	T·CTA	T·GAC	T·GGT	T·GGC	G·ATA
T·CTG	T·CGA	T·CGG	T·CGA	G·ACG	G·TTC	G·TGG	G·GTG
G·AAC	G·ATC	G·GAG	G·GAC	G·GCG	G·CAG	G·CTG	G·CGA
G·GGA	G·GGT	G·GGC	G·GCC	G·CCA	C·ATC	C·AGA	C·AGT
G·CCC	C·AAT	C·ATA	C·GAC	C·AGG	C·AGC	C·TGT	C·TGG
C·GCG	C·CAG	C·CTG		C·TGC	C·TCT	C·GGA	C·GGG
				C·CGT	C·CGG	C·CCC	

information storage and transfer in the mRNA (or cDNA) seems to be trinucleotide. Further there are non-negligible interactions between a codon and mononucleotide on either side, thus suggesting a role for the wobble nucleotide in the selection of the neighbouring codon. To get more insight into the role of wobble nucleotide and neighbouring codons the analysis is being carried out at the level of pair of codons. The results will be discussed in the succeeding communication of this series.

Conclusions

Thus, from the analysis of data on cDNA from bacteriophage system one can see that the selection of mononucleotide in codon formation is not enough to provide dinucleotide pattern or codon patterns. The interactions between di- and mononucleotides being nonlinear in nature, the information content and structural preferences will be difficult to determine using mono- or dinucleotide as models. The occurrence of certain triplets in coding frame with very high or low potential values indicate that nonlinear interactions should be determined as they might give information regarding three-dimensional structural features of DNA which are likely to be recognised at transcriptional and translational level and thus distinguish the fortutious open reading frame from coding sequences.

Acknowledgements

One of the authors (B.V.B.R.) acknowledges the Council of Scientific and Industrial Research, New Delhi for providing financial assistance. We also acknowledge EMBL for providing us the data on magnetic tape.

References

Dolittle, R. F. (1981) *Science*, **214**, 149.
Dunn, J. J., Studier, F. W. (1983) *J. Mol. Biol.*, **166**, 477.
Grantham, R. (1978) *FEBS. lett.*, **95**, 1.
Gouy, M. and Goutier, C. (1982) *Nucl. Acids Res.*, **10**, 7055.
Ikemura, T. (1981a) *J. Mol. Biol.*, **146**, 1.
Ikemura, T. (1981b) *J. Mol. Biol.*, **151**, 389.
Kolaskar, A. S. and Rambrahmam, V. (1983) *Int. J. Pept. Pro. Res.*, **22**, 83.
Konigsberg, W. and Godson, G. N. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 687.
Lipman, D. J. and Wilbur, W. J. (1983) *J. Mol. Biol.*, **163**, 363.
Modiano, G., Battistuzzi, G. and Motulsky, A. G. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 1110.
Nussinov, R. (1980) *Nucl. Acids Res.* **8**, 4545.
Nussinov, R. (1981a) *J. Mol. Biol.*, **17**, 237.
Nussinov, R. (1981b) *J. Biol. Chem.*, **256**, 8458.
Nussinov, R. (1984) *Nucl. Acids Res.*, **12**, 1749.
Root-Bernstein, R. S. (1982) *J. Theor. Biol.*, **94**, 885.
Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. and Petersen, G. B. (1982) *J. Mol. Biol.*, **162**, 729.
Shepherd, J. C. W. (1981) *Proc. Natl. Acad. Sci. USA*, **78**, 1596.
Staden, R. (1982) *Nucl. Acids Res.*, **10**, 141.
Staden, R. (1984) *Nucl. Acids Res.*, **12**, 551.