Bulletin of Mathematical Biology Vol. 51, No. 1, pp. 95–115, 1989.
Printed in Great Britain.

0092-8240/89$3.00+0.00
Pergamon Press plc

CODON PREFERENCE AND PRIMARY SEQUENCE STRUCTURE IN PROTEIN-CODING REGIONS

■ SIMON TAVARÉ and BRENDA SONG
Department of Mathematics,
University of Utah,
Salt Lake City, UT 84112, U.S.A.

The stochastic complexity of a data base of 365 protein-coding regions is analysed. When the primary sequence is modeled as a spatially homogeneous Markov source, the fit to observed codon preference is very poor. The situation improves substantially when a non-homogeneous model is used. Some implications for the estimation of species phylogeny and substitution rates are discussed.

1. Introduction. One of the central problems in evolutionary theory is the reconstruction of species trees (or gene trees), and the estimation of substitution rate and divergence times, using data from DNA or RNA sequences (Kimura, 1983; Nei, 1987). Statistical methods for doing this take as their starting point a model for the evolution of sequence structure through time, and then use the model to estimate phylogeny and substitution rates. Felsenstein (1983) reviews some of these methods.

When the data comprise homologous sequences from coding regions, it is usual to break the data into three constituent sequences, $\mathbf{X}^i$ ($i=1, 2, 3$) say, the bases in $\mathbf{X}^i$ comprising all the bases that occur in the i^{th} codon positions in the sequence. It is usual to assume that each $\mathbf{X}^i$ is a sequence of independent trials, and to analyse each of the $\mathbf{X}^i$ separately. As might be anticipated, the highest substitution rates occur in $\mathbf{X}^3$, the next highest in $\mathbf{X}^1$, and $\mathbf{X}^2$ is the least variable across homologous sequences. This method of analysis leads to a gene tree and a set of substitution rates for each of the $\mathbf{X}^i$. It is certainly not clear how to combine the separate results into a single tree.

What is really needed is a model for sequence evolution that describes with reasonable fidelity the key biological features of a coding region $\mathbf{X}$, without breaking it into its constituent subsequences $\mathbf{X}^1$, $\mathbf{X}^2$, $\mathbf{X}^3$. The features we considered important are: (a) amino acid usage; (b) codon preference bias.

We are currently developing a time-dependent model for the evolution of $\mathbf{X}$ through time. Of course, our choice of model has to be based on the (observable) present day sequence. With this in mind, we have analysed a collection of coding regions, with a view to assessing the type of dependencies that arise in the structure of $\mathbf{X}$ and the $\mathbf{X}^i$. In section 3, we describe some simple

codon-usage and amino acid usage statistics for our data. Section 4 discusses homogeneous Markov models for the primary sequence of **X**, and in section 5 we analyse a spatially inhomogeneous process.

2. The Data. We chose our coding regions from Release 10 of the EMBL (European Molecular Biology Laboratory) Nucleotide Sequence Data Library. A summary of the organisms and number of sequences from each is given in Table I.

TABLE I
Summary of Data Used in the Analysis

Organism (EMBL file)	Length (bases)	Number of coding regions used
Epstein-Barr virus (EBV)	172282	75
SV40 (SV40XX)	5243	6
Yellow fever virus (FLYF17DG)	10862	12
arv-2 (AIARV2)	9737	5
Bacteriophage λ (LAMBDA)	48502	64
$\phi X174$ (PHIX174)	5386	11
ms2 (LEMS2X)	3569	4
*Escherichia coli**	$\approx 4.704 \times 10^6$	188

Data taken from 142 files in EC

Altogether, there are 365 sequences, ranging in length from 45 codons (the rpmH gene coding for ribosomal protein L34 in *E. coli*) to 1381 codons (the BcLF1 reading frame of EBV). The sequences were selected only if they were complete genes (or unidentified/open reading frames), starting with an initiation codon, ending with one of the stop codons, and having length an integral multiple of 3. In the analyses that follow, the initiation codon and the stop codon were not used.

The sample of genes from *E. coli* contains sequences that code for structural proteins, for enzymes and for regulatory proteins. They may be viewed as representative of the whole genome.

3. Codon Usage. Since the pioneering work of Grantham *et al.* (1980a; 1980b; 1981), it has been realized that DNA coding sequences do not use synonymous codons with equal frequency. A substantial amount of codon preference data

has now been compiled (e.g. Maruyama *et al.*, 1986), from which several observations have been drawn.

Firstly, genes within a species usually adopt codon usage strategies that are closer to each other than to those adopted by genes from different taxonomic groups. For example, Ikemura (1985) showed that *E. coli* and *S. typhimurium* show similar codon usage patterns, but those in the unrelated *B. subtilis* are very different (Ogasawara, 1985). Secondly, there remain considerable differences in codon preference between genes in the same species.

Several explanations for these observations have been suggested. Gouy and Gautier (1982) demonstrate that among Bacteria, highly expressed genes exhibit more bias than do lowly expressed genes. Bernardi *et al.* (1985) showed that mammalian genomes seem to comprise very long stretches of rather homogeneous base composition, the regions being distinguished by different G+C content. Bernardi and Bernardi (1985) and Ikemura (1985) go on to correlate codon preference with the local G+C content. Ikemura (1981) and Ikemura and Ozeki (1982) correlate codon usage with abundance of *t*RHAs, and Grosjean and Fiers (1982) attribute it to the theoretical advantage of intermediate bond strength between *t*RNA and *m*RNA. Finally, Wilbur (1985) and Sharp and Li (1986) provide a more evolutionary perspective on the issue.

Codon usage statistics. To summarize different codon usage patterns in our data set, we have used several measures of entropy. For a given sequence, let p_i denote the observed fraction of amino acid i in the sequence. Since we are not counting the stop codons, there are 20 amino acids. The between amino acid or sequence entropy, S, is defined by:

$$S = -\sum_{i=1}^{20} p_i \ln p_i, \tag{1}$$

where ln denotes $\log_e$. S achieves its maximum value of $\ln(20) \approx 2.996$ when each amino acid used is with equal frequency. S has minimum value of 0, attained when the sequence is a polypeptide like poly-Ser, presumably having no significant biological role. Small values of S therefore correspond to biased amino acid usage.

The more interesting measure of entropy is one that assesses non-random synonymous codon usage. If we now let p_{ij} be the relative frequency of the j^{th} of k_i possible synonymous codons for the i^{th} amino acid, then:

$$\sum_{j=1}^{k_i} p_{ij} = p_i, \; i = 1, \ldots, 20.$$

The total entropy, T, of the sequence is defined by:

$$T = -\sum_{i=1}^{20} \sum_{j=1}^{k_i} p_{ij} \ln p_{ij}. \tag{2}$$

cf. Subba-Rao *et al.* (1982), Konopka (1985). If we define $p_{j|i} = p_{ij}/p_i$ to be the conditional relative frequency of synonymous codon j in the i^{th} amino acid, then:

$$\begin{aligned} T &= -\sum_{i=1}^{20} \sum_{j=1}^{k_i} p_i p_{j|i} (\ln p_{j|i} + \ln p_i) \\ &= -\sum_{i=1}^{20} p_i \ln p_i - \sum_{i=1}^{20} p_i \sum_{j=1}^{k_i} p_{j|i} \ln p_{j|i} \\ &= S + \sum_{i=1}^{20} p_i S_i, \\ &\equiv S + W, \end{aligned}$$

where:

$$S_i = -\sum_{j=1}^{k_i} p_{j|i} \ln p_{j|i},$$

is the entropy within the i^{th} amino acid, and:

$$W = \sum_{i=1}^{20} p_i S_i. \tag{3}$$

The term W is a natural measure of codon usage bias, or within-amino-acid entropy. We have also considered two other such measures. These are:

$$W_1 = \sum_{i=1}^{20} \frac{k_i}{61} S_i,$$

and:

$$W_2 = \sum_{i=1}^{20} p_i \frac{S_i}{\ln (k_i)},$$

W_1 is a measure that is independent of the amino acid frequencies in the sequence; the maximum value of W_1 is ≈ 1.242. W_2 is a weighted sum of standardized entropy measures; $0 \leqslant S_i/\ln(k_i) \leqslant 1$ for all i, and so $0 \leqslant W_2 \leqslant 1$.

Data analysis. We calculated the values of S, and W for the sequences in our data base. Codon usage bias is demonstrated in Fig. 1, a scatter plot of W versus sequence length. In the *E. coli* set, there is essentially no correlation between W and sequence length ($r=0.26$), whereas for MS2, $r=0.60$.

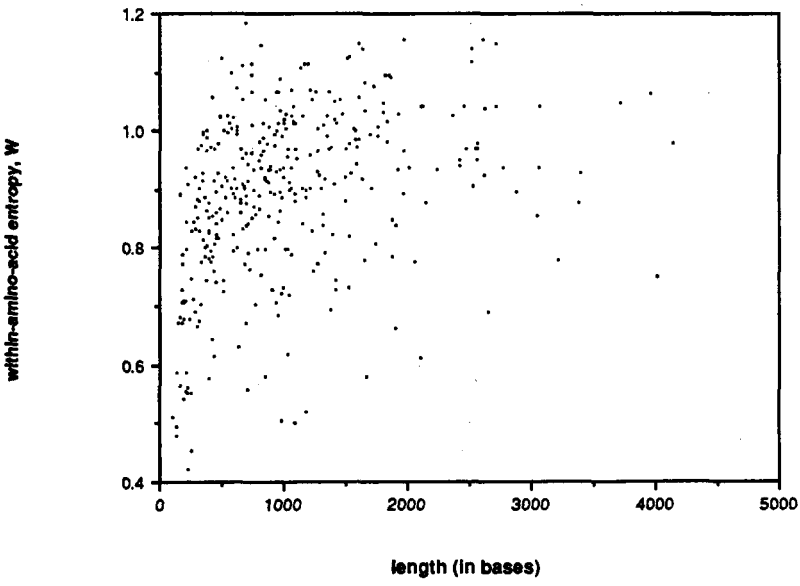

Figure 1. Within-amino-acid entropy W (from equation (3)) versus length of sequence.

MS2 exhibits the most unbiased codon usage, with an average value of $W=1.04$, whereas ϕX174 showed the most biased pattern, with W averaging 0.82. The overall average was 0.89. Using W_1 or W_2 as measures of usage bias gave qualitatively similar conclusions, so no detailed results are given here.

In contrast to within-amino-acid entropy, sequence entropy varies little over the genomes we considered (it is in any event a less interesting biological quantity!). The overall average value of S was 2.65, ranging from an average of 2.59 for ϕX174 to an average of 2.68 for rv2. Shorter sequences might be expected to have somewhat smaller values of S (just from the way it is calculated). Figure 2 gives a scatter plot of S versus sequence length to illustrate the relationship.

In Fig. 3 the relationship between W and S is shown. There is essentially no overall correlation between these variables ($r=0.19$), as is evident from the scatter plot.

4. Models for Primary Sequence Structure. The base-sequence structure of DNA or RNA sequences has often been modeled as the output of a Markov source. This provides a way to summarize the complexity of the sequence, to assess the local dependence of bases on their neighbors (e.g. Blaisdell, 1985;

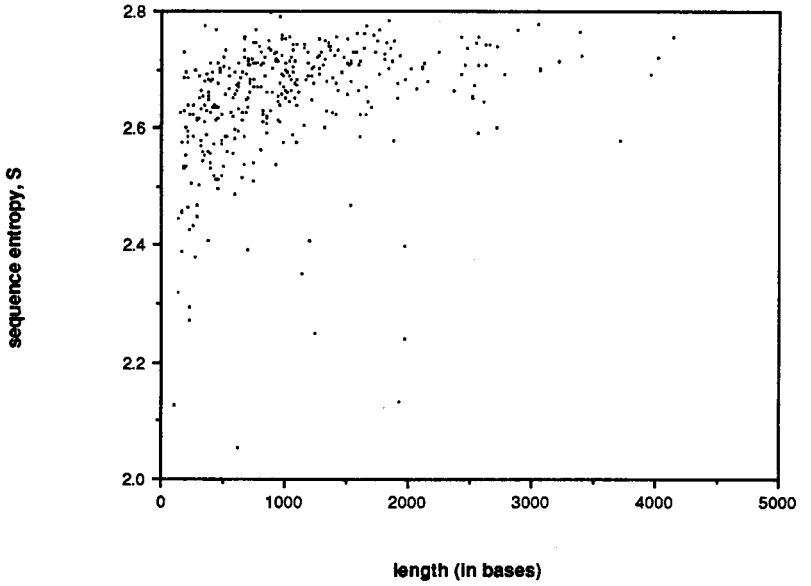

Figure 2. Sequence entropy S (from equation (1)) versus length of sequence.

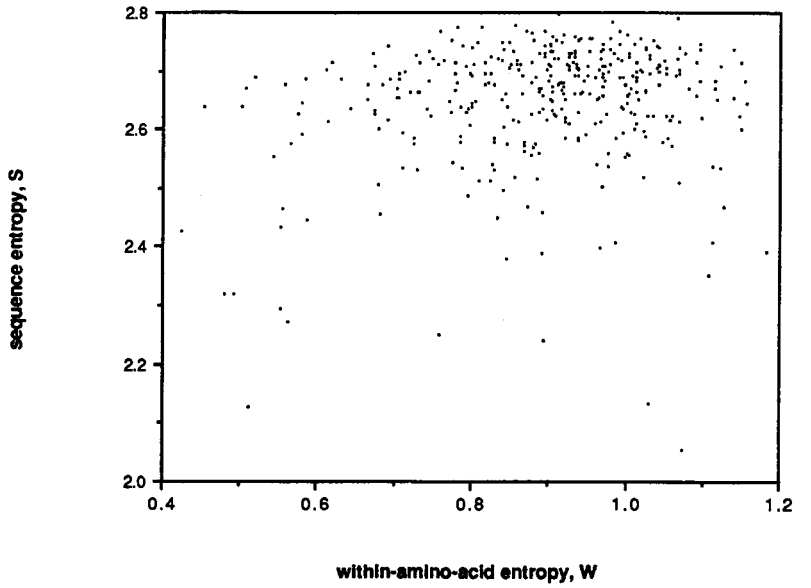

Figure 3. Sequence entropy S versus within-amino-acid entropy W.

Almagor, 1983; Phillips *et al.*, 1987a,b), to search for coding function (Shulman *et al.*, 1981), and to compare sequences from different groups (Erickson and Altman, 1979), perhaps with a view toward reconstructing gene trees (Blaisdell, 1986).

Estimating the order of dependence. For convenience, we number the bases in alphabetical order: $A(1)$, $C(2)$, $G(3)$ and $T(4)$. Let $\mathbf{X}=\{X_n, n=1, 2, \ldots\}$ denote the sequence of bases in the gene of interest. $\mathbf{X}$ is called a (time-homogeneous) Markov chain of order k if:

$$Pr\{X_{n+1}=i_{n+1}|X_n=i_n, \ldots, X_1=i_1\}$$
$$=Pr\{X_{n+1}=i_{n+1}|X_n=i_n, \ldots, X_{n-k+1}=i_{n-k+1}\},$$

for all $n \geqslant k$, and for all possible choices of states $i_1, \ldots, i_{n+1}$. A Markov chain of order 0 is a sequence of independent trials, and (by convention) a chain will be said to have order -1 if it corresponds to equally-likely independent trials.

Estimating the order of dependence, k, may be accomplished in several ways. One is a classical hypothesis-testing framework that tests whether the chain is of order l, within the hypothesis that it is of order $m>l$; Billingsley (1961), Chatfield (1973). Here we will compare two Bayesian methods: AIC, the Akaike Information Criterion (Tong, 1975; Garden, 1980); BIC, the Bayesian Information Criterion (Katz, 1981). For a sequence of N nucleotides, let $n(i_1, i_2, \ldots, i_r)$ denote the number of times the r-tuple $(i_1, \ldots, i_r)$ occurs in the sequence $\mathbf{X}$. For a k^{th} order chain, the natural estimator of the probability $P(i_1, \ldots, i_k; i_{k+1})$ of a transition from $(i_1, \ldots, i_k)$ to i_{k+1} is:

$$\hat{P}(i_1, \ldots, i_k; i_{k+1})=\frac{n(i_1, \ldots, i_k, i_{k+1})}{n(i_1, \ldots, i_k, +)},$$

where:

$$n(i_1, \ldots, i_k, +)=\sum_j n(i_1, \ldots, i_k, j).$$

The log-likelihood L_k is defined by:

$$L_k=\sum n(i_1, \ldots, i_{k+1}) \ln \hat{P}(i_1, \ldots, i_k; i_{k+1}), \tag{4}$$

the summation being over all $i_1, \ldots, i_{k+1}$ for which $n(i_1, \ldots, i_{k+1})>0$.

The AIC for order k is defined by:

$$AIC(k)=-2L_k+2p_k, \quad k=0, 1, \ldots \tag{5}$$

while BIC is defined by:

$$BIC(k)=-2L_k+p_k \ln n, \quad k=0, 1, \ldots \tag{6}$$

and $AIC(-1)=BIC(-1)=n \ln 4$. In equations (5) and (6), p_k is the number of independent parameters in the model ($p_k=3\times 4^k$ in the present case), and n is the number of (overlapping) subsequences used to compute the transition

counts. In all the examples analysed here, the transition counts start from the 7^{th} base in the sequence, so that $n=N-6$ for all values of k considered. The order estimated by AIC (or BIC) is that value k at which the minimum in equation (5) (respectively, equation (6)) occurs. Garden (1980) and Fuchs (1980) used AIC in their comparison of several genomes.

We estimated the order of dependence for the sequences in our database. Table II, a cross-classification of the results, shows the interplay between orders estimated by AIC and BIC. Notice that 100 of the 166 sequences classified by AIC as being of order 2 are classified as order -1 by BIC. BIC certainly favors simpler models.

TABLE II
Cross-Tabulation AIC and BIC

Estimated order AIC	BIC -1	0	1	2	Total
-1	19	0	0	0	19
0	20	8	0	0	28
1	85	39	23	0	147
2	100	25	40	1	166
3	0	0	4	1	5
Total	224	72	67	2	365

It has been observed that AIC estimates are highly correlated with sequence length (Fuchs, 1980). In Table III, we give the mean sequence length (in bases) for each estimated order for both AIC and BIC. These data confirm Fuchs' observation in both cases.

TABLE III
Dependence of Estimated Order on Length of Sequence

AIC	-1	0	1	2	3
Mean length	444.3	381.4	817.5	1400.6	2356.2
Number of sequences	19	28	147	166	5
BIC	-1	0	1	2	
Mean length	787.6	1050.8	1873.8	3000.0	
Number of sequences	224	72	67	2	

If we use the order estimated by AIC as a measure of sequence complexity, then the highest average complexity is the *E. coli* set (averaging 1.54), followed by EBV (1.28). The least complex sequence is ms2, averaging -0.75. There is no evidence, however, of systematic differences in codon usage (as measured by

W) among the different organisms. Using the order as estimated by BIC changes the picture somewhat. Now sv40 and rv 2 are the most complex, and there *is* some evidence that codon usage entropy increases with BIC order.

There seem to be no simple criteria for relating AIC and BIC results. For example, the 100 sequences that have order 2 (by AIC) and -1 (by BIC) appear in terms of entropies and lengths to be similar to the other observations. A more detailed analysis of some examples is necessary.

Residual analysis. To assess further the differences between the order estimated by AIC and that estimated by BIC, some form of residual analysis is appropriate (*cf.* Fuchs, 1980; Blaisdell, 1985; Phillips *et al.*, 1987a,b). Of course, the type of residuals to be analysed depend somewhat on the key biological features that the model is supposed to recover. For example, if the estimated order is $k=1$, then the second-order counts (or triplet frequencies) may be estimated by:

$$\hat{n}(i_1, i_2, i_3) \approx n(i_1, i_2)n(i_2, i_3)/n(i_2, +), \tag{7}$$

and these may be compared to the observed counts $n(i_1, i_2, i_3)$. (Formal hypothesis testing in this context is not straightforward, as the asymptotic distribution of the natural "χ^2" goodness-of-fit statistic is *not* χ^2 (see Chatfield, 1973 for example).

The features of particular biological importance here are firstly codon-usage and secondly amino acid usage. Under the stationarity assumption, the estimator $\hat{n}_c(i_1, i_2, i_3)$ of the frequency of the codon (i_1, i_2, i_3) is:

$$\hat{n}_c(i_1, i_2, i_3) = \hat{n}(i_1, i_2, i_3)/3, \tag{8}$$

while the estimated number of a particular amino acid may be found by summing equation (8) over synonymous codons.

Example. To illustrate, we chose the carB gene that encodes the carbomylphosphate synthetase large sub-unit in *E. coli* (Nyunoya and Lusty, 1983). This gene is 1072 codons in length (excluding initiation and termination codons). The order of dependence is estimated at 1 by BIC and 3 by AIC. We therefore used $k=1$ to estimate codon frequencies. The results are given in Table IV.

The column headed "Pearson residual" in Table IV gives the values of:

$$\text{Residual} = (\text{Observed} - \text{Expected})/\sqrt{\text{Expected}},$$

as a quick way to locate bad fits. Negative residuals correspond to overprediction, while positive residuals correspond to underprediction. It is clear from Table IV that the first order model is not a good predictor of codon

TABLE IV
Observed and Expected Codon-Usage for carB Gene of *E. coli* (First-Order Markov Model)

Amino acid	Codon	Observed frequency*	Expected frequency†	Pearson residual‡
Leu	UUA	3	6.27	−1.3
	UUG	5	16.71	−2.9
	CUU	2	11.01	−2.7
	CUC	5	14.74	−2.5
	CUA	0	8.79	−3.0
	CUG	68	23.45	9.2
Ser	UCU	10	11.27	−0.3
	UCC	18	12.57	1.5
	UCA	1	9.98	−2.8
	UCG	9	21.51	−2.7
	AGU	3	10.28	−3.2
	AGC	5	15.22	−2.6
Arg	CGU	37	23.38	2.8
	CGC	31	34.61	−0.6
	CGA	0	27.90	−5.3
	CGG	1	24.77	−4.8
	AGA	0	12.27	−3.5
	AGG	0	10.89	−3.3
Gly	GGU	28	15.07	3.3
	GGC	51	22.31	6.1
	GGA	2	17.98	−3.8
	GGG	2	15.97	−3.5
Thr	ACU	12	13.24	−0.3
	ACC	40	14.77	6.6
	ACA	1	11.72	−3.1
	ACG	7	25.26	−3.6
Ala	GCU	15	20.31	−1.2
	GCC	21	22.64	−0.3
	GCA	13	17.97	−1.2
	GCG	63	38.75	3.9
Lle	AUU	31	9.68	6.9
	AUC	39	12.96	7.2
	AUA	1	7.73	−2.4
Pro	CCU	1	13.18	−3.4
	CCC	0	14.69	−3.8
	CCA	7	11.66	−1.4
	CCG	35	25.14	1.7
Val	GUU	20	12.79	2.0
	GUC	13	17.12	−1.0
	GUA	14	10.21	1.2
	GUG	47	27.22	3.8

TABLE IV, continued

Amino acid	Codon	Observed frequency*	Expected frequency†	Pearson residual‡
Asn	AAU	4	17.27	−3.2
	AAC	32	22.01	2.1
Phe	UUU	8	7.85	0.1
	UUC	26	10.51	4.8
Tyr	UAU	9	6.76	0.9
	UAC	22	8.61	4.6
Gln	CAA	6	17.39	−2.7
	CAG	29	10.03	6.0
His	CAU	4	10.51	−2.0
	CAC	11	13.40	−0.7
Glu	GAA	66	27.21	7.4
	GAG	25	15.70	2.3
Lys	AAA	46	28.56	3.5
	AAG	9	16.48	−1.8
Asp	GAU	22	16.45	1.4
	GAC	42	20.97	4.6
Cys	UGU	6	18.59	−2.9
	UGC	8	27.52	−3.7
Trp	UGG	4	19.70	3.5
Met	AUG	30	20.62	2.1
Stop	UAA	0	11.18	−3.3
	UGA	0	22.18	−4.7
	UAG	0	6.45	−2.5

*Total is 1070 amino acids. First two are removed by estimation procedure.
†Calculated from equations (8) and (7).
‡(Obs-exp)/$\sqrt{\text{exp}}$.

usage in carB. Aside from obvious problems with stop codons, there are several amino acids poorly represented, for example Lle and Arg. We used a goodness-of-fit statistic:

$$F = \text{sum of squares of Pearson residuals}, \tag{9}$$

to compare model fits. For codon-usage, the data in Table IV gives a value $F = 834$.

If the model is not a good predictor of codon-usage, how well does it recover amino acid composition? The comparison of observed and expected frequencies are presented in Table V. Notice that the stop codon is not the most inaccurately represented "codon"; Glu and Lle are highly under-represented

by the model. The F statistic (equation (9)) for amino acid usage has value $F=310$.

While the first order model may not fit these important aspects of the sequence very well, the situation is not much improved by the third order model suggested by the AIC criterion. The predicted codon-usage frequencies will be (approximately) one third of the *observed* triplet frequencies. The resulting F-statistic for codon usage is $F=583$, while the F-statistic for amino acid usage is $F=282$.

Comments. AIC and BIC provide useful criteria with which to assess stochastic complexity. We noted, as have others, that the estimated order increases with sequence length. This is not surprising as there is a trade-off in these criteria between the number of parameters fitted and the length of the sequence.

Using BIC, a substantial fraction of the sequences *appeared* to be described by independent, identically distributed trials. In practice, BIC seems to be estimating orders too low. Perhaps a correction factor for use with small sample sizes should be used. This rather disconcerting observation prompted a more detailed analysis of the fits.

To assess whether the models describe key biological features of the regions faithfully, we used residual analysis to compare observed and expected codon-usage and amino acid frequencies. In the example cited here (and in many others not reported in detail here) it is clear that even high-order models do not do well, and fits to short sequences are not very good either. One of the reasons for this is to be found in the non-homogeneous structure of coding regions: there are local dependencies that vary in different positions in the codon. This lack of structural homogeneity mitigates against the use of homogeneous Markov chains for the analysis of patterns and structure within coding regions. In the next section, we investigate a class of models which make some allowance for the observed heterogeneity.

5. Non-Homogeneous Models. Several authors have noted that, in coding regions, transitions from the third base position of a codon to the first base of the following codon often appear to be random (*cf.* Shulman *et al.*, 1981; Erickson and Altman, 1979; Lipman and Wilbur, 1983; Lipman and Maizel, 1982; Smith *et al.*, 1983), while transitions from first position to second position, and second position to third position are often markedly non-random. Dependencies such as these induce dependencies of different types in, for example, the sequence of bases from successive first (or second or third) codon positions. In this section, we describe a stochastic model of sequence structure in which the transition matrix governing successive transitions depends on codon position.

TABLE V
Predicted Amino Acid Counts for CarB Gene of *E. coli* (First-Order Markov Model)

Amino acid	Observed frequency*	Expected frequency	Pearson residual†
Leu	83	80.97	0.2
Arg	69	133.82	−5.6
Ser	46	80.83	−3.9
Gly	83	71.33	1.4
Thr	60	64.99	−0.6
Pro	43	64.67	−2.7
Val	94	67.34	3.2
Ala	112	99.67	1.2
Lle	71	30.37	7.4
Gln	35	27.42	1.4
Asn	36	39.28	−0.5
Phe	34	18.36	3.7
His	15	23.91	−1.8
Glu	91	42.91	7.3
Tyr	31	15.37	−4.0
Lys	55	45.04	1.5
Asp	64	37.42	4.3
Cys	14	46.11	−4.7
Trp	4	19.70	−3.5
Met	30	20.62	2.1
Stop	0	39.81	−6.3

*Total is 1070 amino acids. First two are removed by estimation procedure.

†(Obs-exp)/$\sqrt{\text{exp.}}$

Structure of model. The process is a non-homogeneous Markov chain in which transitions from a first codon position to a second position are governed by transition matrix P_1, from a second codon position to a third codon position by P_2, and from a third codon position to the following first codon position by P_3. In this model, the sequence $\mathbf{X}^i$ of bases in the i^{th} position in the codon '($i=1, 2, 3$) form (first-order) Markov chains with transition matrices $P_1\ P_2\ P_3$ if $i=1$, $P_2\ P_3\ P_1$ if $i=2$ and $P_3\ P_1\ P_2$ if $i=3$. By contrast, for the homogeneous first-order model with transition matrix P each of the sequences $\mathbf{X}^i$ is a Markov chain with transition matrix P^3.

In the special case in which "transitions from position 3 to the following position 1 seem independent", we set:

$$P_3 = \mathbf{1}\alpha^T, \tag{10}$$

where $\mathbf{1}$ is a vector of 1's, and $\alpha^T = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ is a probability vector. P_3 is a matrix with identical rows. If equation (10) holds, then $P_1\ P_2\ P_3 = P_1\ P_2\ \mathbf{1}\alpha^T =$

$P_1\,\mathbf{1}\alpha^T = \mathbf{1}\alpha^T$. Similarly, $P_2\,P_3\,P_1 = \mathbf{1}(\alpha^T P_1)$ and $P_3\,P_1\,P_2 = \mathbf{1}(\alpha^T P_1\,P_2)$. It follows that (with the possible exception of the first observation) each of $\mathbf{X}^1, \mathbf{X}^2$ and $\mathbf{X}^3$ are independent trials processes. Of course, $\mathbf{X}^1, \mathbf{X}^2$ and $\mathbf{X}^3$ are not (in general) mutually independent.

Estimation of parameters. Denote the elements of P_r by $\|p_{r;i,j}\|$, and let $n_r(i,j)$ denote the number of times in the sequence $\mathbf{X}$ of bases that base i in the r^{th} codon position is followed by base j. The maximum likelihood estimators are given by:

$$\hat{p}_{r;i,j} = n_r(i,j)/n_r(i,+),\ r = 1, 2, 3, \tag{11}$$

where $n_r(i,+) = \sum_j n_r(i,j)$.

In Table VI the estimated transition matrices $\hat{P}_r$ for the carB gene of *E. coli* are given.

TABLE VI
Estimated Transition Matrices* $\hat{P}_r$ for CarB Gene of *E. coli*

		A	C	G	T
$\hat{P}_1$	A	0.35	0.23	0.03	0.39
	C	0.21	0.18	0.29	0.32
	G	0.35	0.25	0.19	0.21
	T	0.24	0.29	0.14	0.33
$\hat{P}_2$	A	0.36	0.33	0.19	0.12
	C	0.09	0.31	0.45	0.15
	G	0.01	0.53	0.04	0.42
	T	0.06	0.27	0.48	0.20
$\hat{P}_3$	A	0.27	0.18	0.47	0.08
	C	0.23	0.20	0.41	0.17
	G	0.25	0.28	0.37	0.10
	T	0.24	0.21	0.45	0.10

*Estimates based on transition counts starting from 6th base position in sequence to allow comparison with earlier models.

The marked heterogeneity in base position is clearly evident from these results. Notice also that the matrix $\hat{P}_3$ has approximately constant rows, as might be expected on biological grounds.

The results of Billingsley (1961) can be used to establish a statistical test for the hypothesis that $P_3 = \mathbf{1}\alpha^T$ for this non-homogeneous Markov chain. The test turns out to be equivalent to the usual test for homogeneity in a contingency table. For the example of the carB gene, the test statistic has a value of 22.1,

with 9 degrees of freedom, suggesting that the distribution of first-position bases in carB is not seriously influenced by the preceeding base.

Comparison with homogeneous models. Included with the "3-matrix model" are the homogeneous Markov models of order 1 (in which $P_1 = P_2 = P_3 = P$, say), the 0^{th} order case of independent trials ($P = \mathbf{1}\alpha^T$), and the order -1 model, in which $\alpha^T = \frac{1}{4}(1, 1, 1, 1)$. In particular, Billingsley's (1961) results may once more be used to test whether, within the "3-matrix model", any of these homogeneous sub-models is appropriate. For the carB gene of *E. coli*, the log-likelihood for the 3-matrix model is -4044.9, while the likelihood for the first-order Markov model is -4331.7. The log-likelihood-ratio statistic has a value of $-2(-4331.7+4044.9) = 573.6$, which should be compared to a χ^2 random variable with $24 (\equiv 36-12)$ degrees of freedom. In this example, the 3-matrix model provides a far superior description of the data, emphasizing once more the heterogeneity shown in Table VI.

In order to compare the fit of the 3-matrix model to the Markov models of section 4, we again use AIC and BIC. For each sequence in our database, we calculate the log-likelihood L^* for this model, and then compute $\text{AIC}^* = -2L^* + 72$ and $\text{BIC}^* = -2L^* + 36 \ln n$. The values of AIC* are then compared to the optimal AIC value in equation (5). The "best" model is the one that corresponds to the minimum of these two values. The same comparison is also made using the BIC criterion. The results are shown in Table VII.

TABLE VII
Comparison of "3-Matrix" Model and Homogeneous Models

		Number in which 3-matrix model is better	Number in which homogeneous model is better
AIC order	−1	8	11
	0	11	17
	1	122	25
	2	160	6
	3	3	2
BIC order	−1	24	200
	0	11	61
	1	35	32
	2	1	1

Using AIC, 304 of the 365 sequences are better modelled by the 3-matrix model, whereas by BIC it is only 71 sequences. (49 of these are from the *E. coli* set, and 15 from EBV.)

Residual analysis. We will examine how well the 3-matrix model describes codon-usage and amino acid usage. Let $N(i_1, i_2, i_3)$ be the (random) number of times the codon (i_1, i_2, i_3) appears in a sequence $\mathbf{X}$ of length $m = 3r$ bases. Then the expected value of $N(i_1, i_2, i_3)$ is:

$$EN(i_1, i_2, i_3) = \sum_{j=1}^{r} Pr\{X_{3j-2} = i_1, X_{3j-1} = i_2, X_{3j} = i_3\}$$

$$= \sum_{j=1}^{4} Pr\{X_{3j-2} = i_1\} p_{1;i_1,i_2} p_{2;i_2,i_3}. \tag{12}$$

Since $\mathbf{X}^1$ is itself a Markov chain with transition matrix $P_1\ P_2\ P_3$ it follows that:

$$\frac{1}{r} \sum_{j=1}^{r} Pr\{X_{3j-2} = i_1\} \to \pi_{i_1},$$

as $r \to \infty$, where $(\pi_1, \pi_2, \pi_3, \pi_4)$ is the stationary distribution for $P_1\ P_2\ P_3$. (We are assuming that $P_1\ P_2\ P_3$ is irreducible.) For large r, it follows that:

$$EN(i_1, i_2, i_3) \approx r\pi_{i_1} p_{1;i_1,i_2} p_{2;i_2,i_3}. \tag{13}$$

The right side in equation (13) may be estimated by:

$$\hat{n}_c(i_1, i_2, i_3) \approx n_1(i_1, i_2) n_2(i_2, i_3)/n_2(i_2, +). \tag{14}$$

Compare with equation (8). As in the homogeneous case, the amino acid counts may be estimated by grouping the appropriate codon frequencies from equation (14).

For the sequence carB from *E. coli*, the fitted codon frequencies are given in Table VIII. Notice that in contrast with the first order results in Table IV, the 3-matrix model provides a rather accurate description of codon-usage for Arg. This difference is due almost entirely to the very low frequency of $G \to A$ and $G \to G$ transitions from the second codon position. The homogeneous model predicts relative frequencies of about 25% for each of these transitions, compared to the observed frequencies of 1% and 4% respectively (Table VI). Notwithstanding this case, neither model can adequately describe the observed codon usage bias for Leu. The goodness-of-fit statistic equation (9) is, from Table VIII, $F = 360$, compared to $F = 831$ (first-order model) and $F = 583$ (third-order model).

TABLE VIII
Observed and Expected Codon-Usage for carB Gene of *E. coli* (3-Matrix Model)

Amino acid	Codon	Observed frequency*	Expected frequency†	Pearson residual‡
Leu	UUA	3	2.42	0.4
	UUG	5	20.19	−3.4
	CUU	2	14.66	−3.3
	CUC	5	19.95	−3.3
	CUA	0	4.33	−2.1
	CUG	68	36.06	−5.3
Ser	UCU	10	5.71	1.8
	UCC	18	11.87	1.8
	UCA	1	3.30	−1.3
	UCG	9	17.12	−2.0
	AGU	3	3.33	−0.2
	AGC	5	4.27	0.4
Arg	CGU	37	28.69	1.6
	CGC	31	36.83	−1.0
	CGA	0	0.78	−0.9
	CGG	1	2.71	−1.0
	AGA	0	0.09	−0.3
	AGG	0	0.32	−0.6
Gly	GGU	28	34.51	−1.1
	GGC	51	44.30	1.0
	GGA	2	0.93	1.1
	GGG	2	3.26	−0.7
Thr	ACU	12	9.01	1.0
	ACC	40	18.74	4.9
	ACA	1	5.22	−1.8
	ACG	7	27.04	−3.9
Ala	GCU	15	16.82	−0.4
	GCC	21	34.97	−2.4
	GCA	13	9.74	1.0
	GCG	63	50.47	1.8
Lle	AUU	31	19.75	2.5
	AUC	39	26.87	2.3
	AUA	1	5.83	−2.0
Pro	CCU	1	6.46	−2.1
	CCC	0	13.43	−3.7
	CCA	7	3.74	1.7
	CCG	35	19.38	3.5
Val	GUU	20	18.39	0.4
	GUC	13	25.01	−2.4
	GUA	14	5.42	3.7
	GUG	47	45.19	0.3

TABLE VIII, continued

Amino acid	Codon	Observed frequency*	Expected frequency†	Pearson residual‡
Asn	AAU	4	10.85	−2.1
	AAC	32	29.78	0.4
Phe	UUU	8	8.21	−0.1
	UUC	26	11.17	4.4
Tyr	UAU	9	3.70	2.8
	UAC	22	10.14	3.7
Gln	CAA	6	18.04	−2.8
	CAG	29	9.63	6.2
His	CAU	4	5.96	−0.8
	CAC	11	16.36	−1.3
Glu	GAA	66	55.93	1.3
	GAG	25	29.86	−0.9
Lys	AAA	46	32.84	2.3
	AAG	9	17.53	−2.0
Asp	GAU	22	18.49	0.8
	GAC	42	50.72	−1.2
Cys	UGU	6	7.48	−0.5
	UGC	8	9.61	−0.5
Trp	UGG	4	0.71	3.9
Met	AUG	30	48.56	−2.7
Stop	UAA	0	11.19	−3.3
	UGA	0	0.20	−0.4
	UAG	0	5.97	2.4

*Total is 1070 amino acids. First two are removed by estimation procedure.
†Calculated from equation (13).
‡(Obs-exp)/$\sqrt{\text{exp}}$.

The corresponding results for amino acid usage are shown in Table IX. Qualitatively, this provides a much more accurate representation than that shown in Table V. The F-statistic is $F=88$, compared to $F=310$ for the first-order model, and $F=282$ for the third order case.

Comments. The non-homogeneous process provides a better model for the generation of coding regions. However, it does not recover all the details of codon usage accurately; some averaging is taking place. For example, in constructing P_2 (the matrix governing the choice of degenerate base) for carB, the amino acid Threonine has codon usage ACA(1), ACC(40), ACG(7), and ACU(12). In contrast, the frequencies for Alanine are GCA(13), GCC(21),

TABLE IX
Predicted Amino Acid Counts for CarB Gene of *E. coli* (3-Matrix Model)

Amino acid	Observed frequency*	Expected frequency	Pearson residual†
Leu	83	97.61	−1.5
Arg	69	69.42	−0.1
Ser	46	45.60	0.1
Gly	83	83.00	0.0
Thr	60	60.01	−0.0
Pro	43	43.01	−0.0
Val	94	94.01	−0.0
Ala	112	112.00	0.0
Lle	71	52.45	2.6
Gln	35	27.67	1.4
Asn	36	40.63	−0.7
Phe	34	19.38	3.3
His	15	22.32	−1.5
Glu	91	85.79	0.6
Tyr	31	13.84	4.6
Lys	55	50.37	0.7
Asp	64	69.21	−0.6
Cys	14	17.09	−0.7
Trp	4	0.71	3.9
Met	30	48.56	−2.7
Stop	0	17.36	−4.2

*Total is 1070 amino acids. First two are removed by estimation procedure.

†(Obs-exp)/$\sqrt{\text{exp}}$.

GCG(63) and GCU(15). We are currently analysing a model that also makes allowance for this type of inhomogeneity. The 3-matrix model should work well for sequences whose degenerate base choices are similar.

6. Summary and Conclusions. Spatially homogeneous Markov chains are an attractive (and often used) model for the primary sequence structure of a coding region. We have seen, however, that such models often do not do an adequate job of describing the key biological features of the region: codon preference, and amino acid usage. The evident spatial heterogeneity in such sequences invalidates much of this Markov chain theory.

We presented a simple spatially heterogeneous Markov model that reflects more accurately the coding features of the regions. Several alternative models exist in the literature (Lipman and Wilbur, 1983). We are currently studying two other models, one described at the end of section 5, the other being a simple

nearest-neighbor interacting particle system. This latter model allows a particular base in the sequence to be determined by its "nearest" neighbors in both directions, rather than just its neighbors to the left. We are also analysing a much more extensive set of *E. coli* sequences, with a view to assessing which types of models apply best within classes of genes that code for regulatory proteins, for enzymes and for structural proteins.

While there is perhaps "too much biology" in these sequences for such simple models to reconstruct accurately, the search for better and more detailed models with which to estimate phylogeny seems well worthwhile.

Simon Tavaré was supported in part by National Science Foundation grants DMS 86-08857 and DMS 88-03284. We would like to thank Ron Lundstrom for helpful discussions on some statistical aspects of the model in Section 5.

LITERATURE

Almagor, H. 1983. "A Markov Analysis of DNA Sequences." *J. Theor. Biol.* **104**, 633–645.

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival and F. Rodier. 1985. "The Mosaic Genome of Warm-Blooded Vertebrates." *Science* **228**, 953–958.

Bernardi, G. and G. Bernardi. 1985. "Codon Usage and Genome Composition." *J. Molec. Evol.* **22**, 363–365.

Billingsley, P. 1961. *Statistical Inference for Markov Processes*. Chicago: University of Chicago Press.

Blaisdell, B. E. 1985. "Markov Chain Analysis Finds a Significant Influence of Neighboring Bases on the Occurrence of a Base in Eukaryotic Nuclear DNA Sequences Both Protein-Coding and Noncoding." *J. Molec. Evol.* **21**, 278–288.

———. 1986. "A Measure of the Similarity of Sets of Sequences Not Requiring Sequence Alignment." *Proc. Natn. Acad. Sci. U.S.A.* **83**, 5155–5159.

Chatfield, C. 1973. "Statistical Inference Regarding Markov Chain Models." *Appl. Statist.* **22**, 7–20.

Erickson, J. W. and G. G. Altman. 1979. "A Search for Patterns in the Nucleotide Sequence of the MS2 Genome." *J. Math. Biol.* **7**, 219–230.

Felsenstein, J. 1983. "Statistical Inference of Phylogenies." *J. R. Statist. Soc.* **146**, 246–272.

Fuchs, C. 1980. "On the Distribution of Nucleotides in Seven Completely Sequenced DNAs." *Gene* **10**, 371–373.

Garden, P. W. 1980. "Markov Analysis of Viral DNA/RNA Sequences." *J. Theor. Biol.* **82**, 679–684.

Gouy, M. and C. Gautier. 1982. "Codon Usage in Bacteria: Correlation with Gene Expressivity." *Nucleic Acids Res.* **10**, 7055–7074.

Grantham, R., C. Gautier and M. Gouy. 1980a. "Codon Frequencies in 119 Individual Genes Confirm Consistent Choices of Degenerate Bases according to Genome Type." *Nucleic Acids Res.* **9**, r43–r74.

———, ———, ———, R. Mercier and A. Pavé. 1980b. "Codon Catalog Usage and the Genome Hypothesis." *Nucleic Acids Res.* **8**, r49–r62.

———, ———, ———, M. Jacobzone and R. Mercier. 1981. "Codon Catalog Usage is a Genome Strategy Modulated for Gene Expressivity." *Nucleic Acids Res.* **9**, r43–r74.

Grosjean, H. and W. Fiers. 1982. "Preferential Codon Usage in Prokaryotic Genes—The Optimal Anticodon Interaction Energy and the Selective Codon Usage in Efficiently Expressed Genes." *Gene* **18**, 199–209.

Ikemura, T. 1981. "Correlation Between the Abundance of *Escherichia coli* Transfer RNAs and the Occurrence of the Respective Codons in its Protein Genes." *J. Molec. Biol.* **146**, 1–21.

———. 1985. "Codon Usage and the tRNA Content in Unicellular and Multicellular Organisms." *Molec. Biol. Evol.* **2**, 13–34.

——— and H. Ozeki. 1982. "Codon Usage and Transfer RNA Contents: Organism-Specific Codon-Choice Patterns in Reference to the Isoacceptor Contents." *Cold Spring Harbor Symp. Quant. Biol.* **49**, 1087–1097.

Katz, R. W. 1981. "On Some Criteria for Estimating the Order of a Markov Chain." *Technometrics* **23**, 243–249.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution.* New York: Cambridge University Press.

Konopka, A. 1984. "Is the Information Content of DNA Evolutionarily Significant?" *J. Theor. Biol.* **107**, 697–704.

Lipman, D. J. and J. Maizel. 1982. "Comparative Analysis of Nucleic Acid Sequences by their General Constraints." *Nucleic Acids Res.* **10**, 2733–2739.

——— and W. J. Wilbur. 1983. "Contextual Constraints on Synonymous Codon Choice." *J. Molec. Biol.* **163**, 363–376.

Maruyama, T., T. Gojobori, S. Aota and T. Ikemura. 1986. "Codon Usage Tabulated from the GenBank Genetic Sequence Data." *Nucleic Acids Res.* **14**, r151–r197.

Nei, M. 1987. *Molecular Evolutionary Genetics.* New York: Columbia University Press.

Nyunona, H. and C. J. Lusty. 1983. "The CarB Gene of *Escherichia coli*: A Duplicated Gene Coding for the Large Sub-unit of Carbamoyl-Phosphate Synthetase." *Proc. Natn. Acad. Sci. U.S.A.* **80**, 4529–4633.

Ogasawara, N. 1985. "Markedly Unbiased Codon Usage in *Bacillus subtilis*." *Gene* **40**, 145–150.

Phillips, G. J., J. Arnold and R. Ivarie. 1987a. "Mono-Through Hexanucleotide Composition of the *Escherichia Coli* Genome: A Markov Chain Analysis." *Nucleic Acids Res.* **15**, 2611–2626.

———, J. Arnold and R. Ivarie. 1987b. "The Effect of Codon Usage on the Oligonucleotide Composition of the *E. coli* Genome and Identification of Over- and Under-represented Sequences by Markov Chain Analysis." *Nucleic Acids Res.* **15**, 2627–2638.

Sharp, P. M. and W.-H. Li. 1986. "An Evolutionary Perspective on Synonymous Codon Usage in Unicellular Organisms." *J. Molec. Evol.* **24**, 28–38.

Shulman, M. J., C. M. Steinbert and N. Westmoreland. 1981. "The Coding Function of Nucleotide Sequences can be Discerned by Statistical Analysis." *J. Theor. Biol.* **88**, 409–420.

Smith, T. F., M. S. Waterman and J. R. Sadler. 1983. "Statistical Characterization of Nucleic Acid Sequence Functional Domains." *Nucleic Acids Res.* **11**, 2205–2220.

Tong, H. 1975. "Determination of the Order of a Markov Chain by Akaike's Information Criterion." *J. Appl. Prob.* **12**, 488–497.

Subba Rao, J., C. P. Geevan and G. Subba Rao. 1982. "Significance of the Information Content of DNA in Mutations and Evolution." *J. Theor. Biol.* **96**, 571–577.

Wilbur, W. J. 1985. "Codon Equilibrium I: Testing for Homogeneous Equilibrium." *J. Molec. Evol.* **21**, 169–181.

Received for publication 1 July 1988