JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 7, Numbers 1/2, 2000
Mary Ann Liebert, Inc.
Pp. 249–260

Translation Conditional Models for Protein Coding Sequences

FRANÇOIS RODOLPHE[1] and CATHERINE MATHÉ[2]

ABSTRACT

A coding sequence is defined as a DNA sequence coding the primary structure of a protein (a polypeptide). Such a sequence must satisfy a specific constraint, which consists in coding a functional protein. As the genetic code is degenerated, there exists, for a given polypeptide, a set of synonymous sequences which would code the same polypeptide. Translation conditional models are being defined on such sets. The aim of this paper is to give a common formalism. Besides the codon bias model, a few other conditional models will be defined. Statistical estimators and comparison methods will be briefly presented. These models can be used for gene classification, or to find out, in a real sequence, remarkable features. An example will be presented on *Escherichia coli* genes.

Key words: codon bias, conditional models, coding sequences, gene classification.

INTRODUCTION

CODING SEQUENCES CONSTITUTE a well defined, common and easily recognizable (at least in bacterial genomes) class of DNA sequences.

Coding for a functional protein is certainly a major constraint, but does not specify the sequence. Genetic code degeneracy allows for a huge number of synonymous DNA sequences; even the polypeptide itself is far from being entirely determinated by its function: neutral polymorphism is almost an everyday observation in proteins. Many degrees of freedom are left, and it is a legitimate question to ask how they are used.

As exposed by Trifonov (1989), coding sequences support several signals, and not only the primary structure of proteins. In a slightly different point of view, Hénaut and Vigier (1995) and Karlin and Mràzek (1996) also consider that coding sequences must satisfy several constraints (amino acid requirements, genomic mutational biases, constraints imposed by DNA structure, translational processes, transcriptional processes, mRNA stability and structure); one can add too, the necessary presence of motifs. So, besides neutral polymorphism, genetic code degeneracy is certainly widely "used" to satisfy such constraints.

Neutral polymorphism in proteins is not easily described. On the contrary, synonymy is, and allows the definition of translation conditional models. As these models charge only synonymous sequences, the effect of the primary structure of the protein to be coded is eliminated. Thus, they provide a tool for the study of other constraints. On the basis of these models, it is always possible to separate on a coding sequence features contained in the amino-acid sequence from those due to genetic code degeneracy. For

[1]INRA, Unité MIG, 78352 Jouy en Josas cedex, France.
[2]Vlaams Interuniversitair Instituut voor Biotechnologie, Laboratorium voor Genetika, Universiteit Gent, K.L. Ledeganckstraat 35, B-9000 GENT, België.

instance, if we consider word frequencies, it is interesting to consider separately, for a given word, presence opportunities provided by the polypeptide and effective presence due to a particular synonymous choice. Codon bias studies rely on such models, eliminating the effect of amino acid composition.

CONDITIONAL MODELS

Consider a coding DNA sequence $S = s_1, s_2, \ldots s_n$, of length n, $s_i \in \{A, G, C, T\}$. We will denote by τ the translation operator, i.e., $\tau(S)$ is the amino-acid sequence of length $n/3$ obtained by applying the genetic code to the codons of S in the usual way. Let $\tau^{-1} \circ \tau(S) = \sigma(S)$ denote the set of all synonyms of S.

Let P be a probability distribution on the set of all sequences of length n. We will define the translation conditional probability distribution Q on the same set, using Bayes formula, by:

$$Q[Z] = P[Z \,|\, \tau(Z) = \tau(S)] = \begin{cases} \frac{P[Z]}{\sum_{X \in \sigma(S)} P[X]} & \text{if } \tau(Z) = \tau(S) \\ 0 & \text{if } \tau(Z) \neq \tau(S) \end{cases}$$

Codon bias

Codon bias refers to the fact that synonymous codons are not equally used. Conditional frequencies of codons were analyzed, leading, for instance, to the evidence of different gene classes. In fact, this corresponds to a conditional model, here referred to as the CB model.

Consider any process on the set of amino-acids $P[A_1, A_2, \ldots A_{n/3}]$ and conditional probabilities $\pi[C|A]$ (the probability codon C is used for amino acid A). $P[A_1, A_2, \ldots A_{n/3}] \times \prod_{i=1}^{n/3} \pi[C_i|A_i]$ defines a probability distribution on the set of DNA sequences of length n, and given the amino-acid sequence, the corresponding conditional model is: $Q[C_1, C_2, \ldots C_{n/3}] = \prod_{i=1}^{n/3} \pi[C_i|A_i]$. In this CB model, whatever the amino acid sequence is, given this sequence, codons are supposed to be choosen independently of each other.

Parameters are the conditional probabilities. They belong to a vector space of dimension 41 (if stop codons are discarded). Their maximum likelihood estimators are the relative frequencies of codons with all desirable properties (consistency, asymptotic normality, optimality).

Structure of $\sigma(S)$

Before going on with other models, let us look at $\sigma(S)$. Due to the genetic code structure, synonymous sequences have a lot of common letters. Consider the following example:

S	*ATG*	*ACC*	*TGT*	*GTC*	*CGG*	*CCT*	...
$\tau(S)$	Met	Thr	Cys	Val	Arg	Pro	...
Fixed letters in $\sigma(S)$	*ATG*	*AC.*	*TG.*	*GT.*	*.G.*	*CC.*	...
Admissible values for		*A*		*A*	*C A*	*A*	...
variable letters in the		*G*		*G*	*C G*	*G*	...
codons of sequences		*C*	*C*	*C*	*C C*	*C*	...
in $\sigma(S)$		*T*	*T*	*T*	*C T*	*T*	...
					A A		...
					A G		...

Definition: *Given $\sigma(S)$, we call* bank of order k *any maximal sequence of successive fixed letters, of length at least* k*, which* separates codons*. A sequence between two successive banks of order* k*, is called a* gap of order k*, its size is the number of letters in it.*

In the example above, $ATGAC$, TG, GT, CC, are all banks of order 2. There are four gaps of order 2: 2 of size 1, 1 of size 4, plus an other one starting in the last position. Letters G (respectively T) common to all arginine (respectively leucine) codons in position 2 are *not* banks of order 1, since they do not separate codons.

The notation $B_0, V_1, B_1, \ldots V_l, B_l$ will also be used for a coding sequence, emphasizing its structure as a succession of banks and gaps.

Markov chains

The main characteristic of the CB model is the supposed conditional independence of codons. Alternative models are interesting too. For instance, Markov chains provide flexible models which can introduce a certain degree of dependence between codons, without overincreasing the number of parameters.

Consider a Markov chain of order k, and the sequence $S = B_0, V_1(S), B_1, \ldots V_l(S), B_l$ where banks, and hence gaps, are of order k, and l stands for the number of gaps.

Since banks separate codons, different gaps do not intersect a same codon, and the set of admissible values for $(V_1, \ldots V_l)$ is the product of the sets $\mathcal{E}_i$ $(i = 1, \ldots l)$ of admissible values for gaps. Using the Markov property we have:

$$\sum_{X \in \sigma(S)} P[X] = \sum_{(V_1, \ldots V_l) \in \mathcal{E}_1 \times \ldots \mathcal{E}_l} P[B_0] \prod_{i=1}^{l} P[V_i|B_{i-1}].P[B_i|B_{i-1}, V_i]$$

$$= P[B_0] \prod_{i=1}^{l} \sum_{V \in \mathcal{E}_i} P[V|B_{i-1}].P[B_i|B_{i-1}, V]$$

and the conditional likelihood can be factorized over gaps:

$$Q[S] = \prod_{i=1}^{l} \frac{P[V_i(S)|B_{i-1}].P[B_i|B_{i-1}, V_i(S)]}{\sum_{V \in \mathcal{E}_i} P[V|B_{i-1}]P[B_i|B_{i-1}, V]}.$$

This formula shows that *in a conditional Markov chain model, gaps of suitable order, are independent.*

Moreover, the present factorization makes computation easy. Practically, gaps of order 2 are usually small, and cardinals $\# \mathcal{E}_i$ small too. Hence, for chains of order ≤ 2, this formula makes even direct likelihood computation feasible.

Higher order gaps are too long. Other methods were defined for statistical estimation of higher order Markov chains; they are based on other contrasts; they will not be developed here. In any case, a complete statistical analysis of these models can always be performed.

In a conditional k order Markov chain, probabilities on a k order gap are determined by the set of all $(k+1)$-nucleotides which intersect the gap (all variable $(k+1)$-nucleotides in B_{i-1}, V, B_i), compared with those of all synonymous gap sequences. Thus, in such models, like in unconditioned Markov chains, one can speak of *"preferences"* among $(k+1)$-nucleotides $(a_1, \ldots a_{k+1})$, measured by the transition probabilities $P[a_{k+1}|a_1, \ldots a_k]$.

Self-complementary Markov chains

Complementation is the operation which transforms a sequence into its complementary sequence. Let $\tilde{()}$ represent complementation on nucleotides: $\tilde{A} = T$, $\tilde{G} = C$, $\tilde{C} = G$ and $\tilde{T} = A$. Consider a word $w = (s_{i+1}, s_{i+2}, \ldots s_{i+h})$. Its complementary is $\tilde{w} = (\tilde{s}_{i+h}, \tilde{s}_{i+h-1}, \ldots \tilde{s}_{i+1})$.

Self-complementary Markov chains are defined as Markov chains invariant by complementation, i.e., such that: $\forall w$, $P[w] = P[\tilde{w}]$; hence, if unconditioned, they induce the same statistical structure on both DNA strains. Nevertheless, given the translation, a self-complementary Markov chain does *not* induce the same oligonucleotidic composition on both DNA strains.

We will not develop these models here but will only note that there exist such chains for any order. Obviously, such models have fewer parameters than general chains of the same order. Parameters for self-complementary Markov chains of order 1 and 2 belong to spaces of dimension 7 and 25, respectively.

If we consider that constraints other than the polypeptide to be coded act locally and identically on both strains, they should favour equally complementary oligonucleotides. These models enable one to test this hypothesis.

Filling small gaps of small order and size

Markov chains introduce constraints on the probabilities with which gaps are filled. It is also interesting to consider a generalization of the CB model allowing great flexibility and interaction between adjacent

codons. For clarity, as well as for theoretical and practical reasons, *we now restrict ourselves to gaps of size 1 and order 2*.

Consider such a gap, with its i neighbors on the left and j neighbors on the right (we suppose i and $j \leq 2$: the neighbors are fixed letters). The type of such a gap is defined by its left and right neighbors and its set $\mathcal{E}$ of admissible values.

Definition: *We call* FG$i,j$ *the model in which gaps are filled independently, with an arbitrary probability distribution for each type of gap.*

Obviously, restricted on these gaps, the CB model is FG2,0, and FGi,i contains Markov chains of order i. Such models allow influence by a codon, up to its first two letters, on its immediate neighbor on the left: codon bias can be influenced by the next amino acid, not by the precise codon. This is why codons can be dependent, but, conditionally, not gaps. On a restricted data set (gaps of size 1), these models enable one to test the existence of local interactions between codons.

Anxious persons will ask for some theoretical justification. Consider the following idealized evolution process: suppose a sequence is subject to substitutions only, with substitution probability rates in a given position of the sequence depending on a neighborhood of the position. Suppose that only synonymous sequences survive. This defines a time jump process on a set of synonymous sequences.

If substitution probability rates depend only on i neighbors on the left and j neighbors on the right, $k = \max(i, j)$ order gaps will follow independent trajectories. If the process runs for a sufficiently long time, the probability distribution on a gap is its stationary distribution, which depends on the substitution probability rates, its left and right neighbors, and its set $\mathcal{E}$ of admissible values. In fact, it is the left eigenvector (corresponding to eigenvalue 1) of the infinitesimal generator of the jump process restricted to the gap. Gaps of the same type have the same probability distribution.

Substitution probability rates are unknown (let them remain so!). But one can demonstrate that for any combination of probability distributions on the different gap types of size 1, there exists a set of substitution probability rates that leads to these distributions as stationary ones on gap types of size 1. And there exists such a set of substitution probability rates, depending on the same neighborhood definition, as the one used in the definition of gap types; thus leading to model FG$i,j$. This is not true for large gaps and is a reason for the restriction on gaps of small size.

Statistics for these models is as simple as for the CB model: the probability distribution of a type of gap is estimated by empirical frequencies in the data of that type.

CODON BIAS IN *ESCHERICHIA COLI*

Codon bias has been compared between *E. coli* genes by M'edigue *et al.* (1991), with the CB model, leading to the recognition of three different gene classes. We will present here a comparison of just a few models on the same *E. coli* coding sequences.

Gene classification, data set

A sample of 1572 *E. coli* coding sequences from the ECDC database (February, 1996) were selected using the following criteria:

- sequence must be featured as a gene,
- sequence must have no missing nucleotide,
- sequence of first codon must be a start codon,
- sequence of last codon must be a stop codon TAA,TAG or TGA.

Of these genes, 904 had been assigned to one of the three gene classes by Médigue: 612 to class 1, 223 to class 2, and 69 to class 3.

Gene classification, results

We first looked to see if the structure put in evidence by Médigue could be modified with other conditional models. General and self-complementary homogeneous conditional Markov chains of order 1 and 2 were adjusted to all genes separately. They are called here GM1, GM2, SM1, SM2. Obviously, SMi $\subset$ GMj

($i \leq j$), but no one of these models neither contains the CB model, nor is contained in it. Estimation was done by direct maximization of the likelihood factorized on gaps. Except for a few genes which had long gaps (up to 22 nucleotides), maximization was fast and easy.

These models are characterized by transition matrices (one per gene) or, equivalently, by stationary probability distributions on di- or trinucleotides according to chain order. But as the model is translation conditional, stationary probability distributions on $(k+1)$-nucleotides, for a k order chain, do not represent the actual distribution of $(k+1)$-nucleotides in the gene analyzed. They merely represent the distribution which would be expected in an unconditioned chain.

Parameter space dimensions make visualization difficult. Figures 1, 2, 3, 4 represent these genes in the planes spanned by the two first factors of factorial correspondance analysis on relative codon frequencies for the CB model (41 degrees of freedom), transition probability matrices for CM2 (48 d.f.), CM1 (12 d.f.) models, or alternatively, stationary probability distribution on dinucleotides for the SM1 model (7 d.f.).

These results will not be extensively presented here, but it is worth noting that in all four homogeneous Markov chain models, the three gene classes are quite well discriminated, although less well than in CB model (which served for their definition). Self-complementary homogeneous conditional Markov Chains discriminate as well as general ones of the same order, even SM1, which, as will be seen below, does not fit the data well. This is somewhat remarkable.

Model comparisons, data set

In order to compare all these models, we adjusted them once to the same unique set of genes which could be considered as homogeneous. Indeed, it would make no sense to fit the same model to several genes known, a priori, to be heterogeneous. As FG$i,j$ models were only defined on particular gaps, all model comparisons were done on the set of all gaps of size 1 and order 2 of genes belonging to class 1. We choose the first gene class of Médigue, since it is the most numerous one and it appears quite homogeneous. In these data, among 182638 gaps, there are 140887 gaps of size 1. As a by-product, computations are very fast on these structures.

These structures consist in pentanucleotides where only the central nucleotide is variable. The first three positions form a codon for one out of 15 possible amino acids (excluding amino acids with 1 or 6 codons and stops). The last two positions must be the first two positions of a codon for one out of 17 possible amino acids (those with up to 4 codons). Therefore, right and left banks can take (independently) only 12 different values out of the 16 possible dinucleotides. In total, there are 180 different types of such

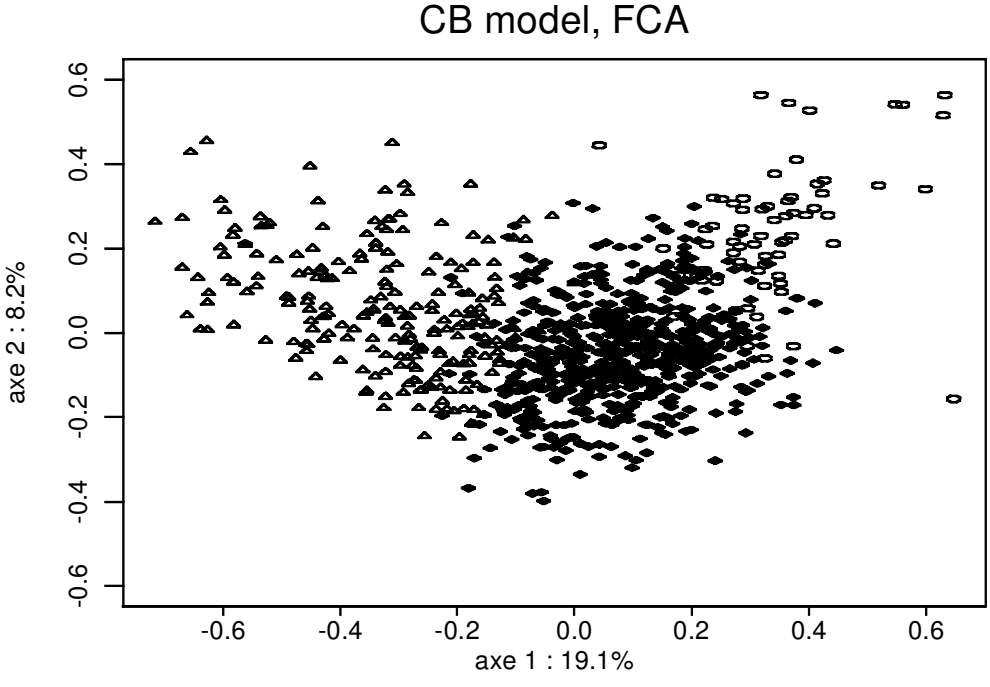

FIG. 1. Gene representation in CB model (first plane of a factorial correspondance analysis of relative codon frequencies). ♦ = class 1; △ = class 2; ○ = class 3

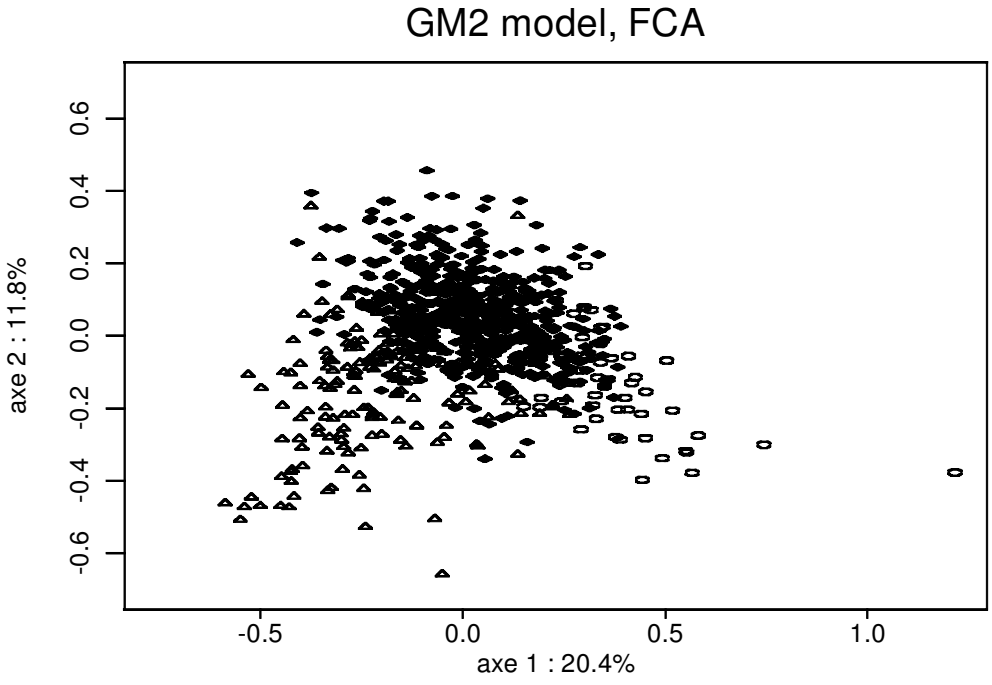

FIG. 2. Gene representation in GM2 model (first plane of a factorial correspondance analysis of transition probability matrices). ♦ = class 1; △ = class 2; ○ = class 3

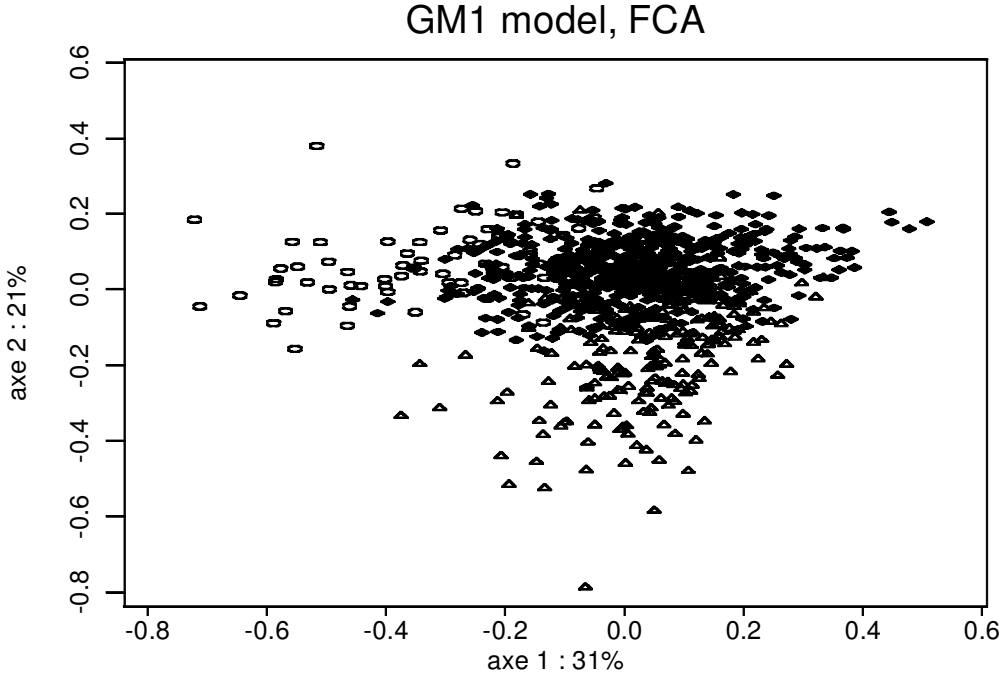

FIG. 3. Gene representation in GM1 model (first plane of a factorial correspondance analysis of transition probability matrices). ♦ = class 1; △ = class 2; ○ = class 3

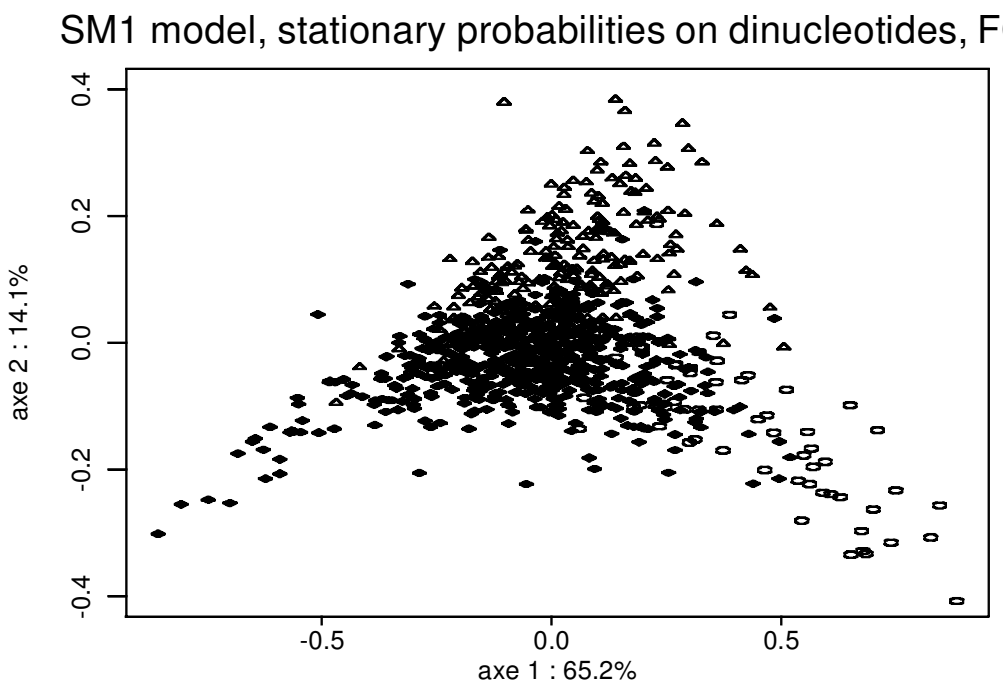

FIG. 4. Gene representation in SM1 model (first plane of a factorial correspondance analysis of stationary probability distributions on dinucleotides). ♦ = class 1; △ = class 2; ○ = class 3

gaps (15×12). Frequencies in each gap type make a sufficient statistics for all these models. There are in total 492 such frequencies ($((64 - (3 + 18 + 2)) \times 12)$, with a maximum of 312 degrees of freedom ($((64 - (3 + 18 + 2 + 15)) \times 12)$.

Model comparisons, results

When only gaps of size 1 and order 2 are considered, the models used here have the following inclusion relationships:

$$\begin{array}{ccccc} FG2,0 & \subset & FG2,1 & \subset & FG2,2 \\ & & \cup & & \cup \\ \cup & & GM1 & \subset & GM2 \\ & & \cup & & \cup \\ M00 & \subset & SM1 & \subset & SM2 \end{array}$$

$M00$ stands for the model in which nucleotides are supposed to be independent and of equal probability. Given the translation, it charges synonymous codons with equal probabilities. It serves here as a reference.

The dimension of parameter space for models FG2,0 (CB restricted on these data), FG2,1, FG2,2 are easily calculated. As homogeneous Markov chains of order ≤ 2 remain identifiable on this restricted data set, their parametric dimensions are unchanged.

Results are summarized in the table below:

Model	*Param. dim.*	*Max. Log likelihood*
$M00$	0	−147063.
$SM1$	7	−141690.
$SM2$	25	−134018.
$GM1$	12	−139795.
$GM2$	48	−131280.
CB	26	−131628.
$FG2,1$	104	−128533.
$FG2,2$	312	−126436.

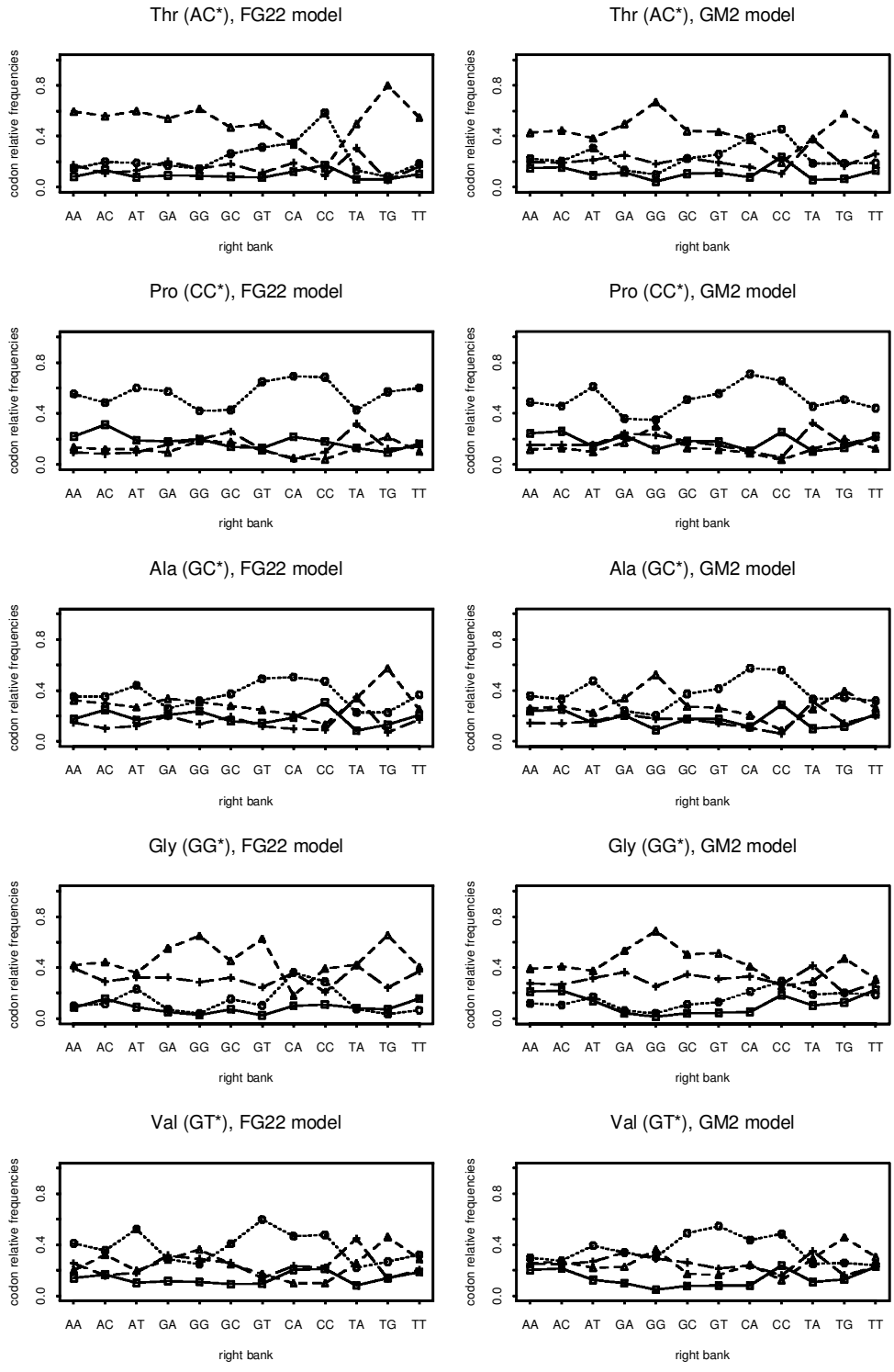

FIG. 5. Empirical frequencies (FG2,2 model) and estimated probabilities in GM2 model, on gaps of size 1, according to the right bank. $\square = A$; $\circ = G$; $\triangle = C$; $+ = T$

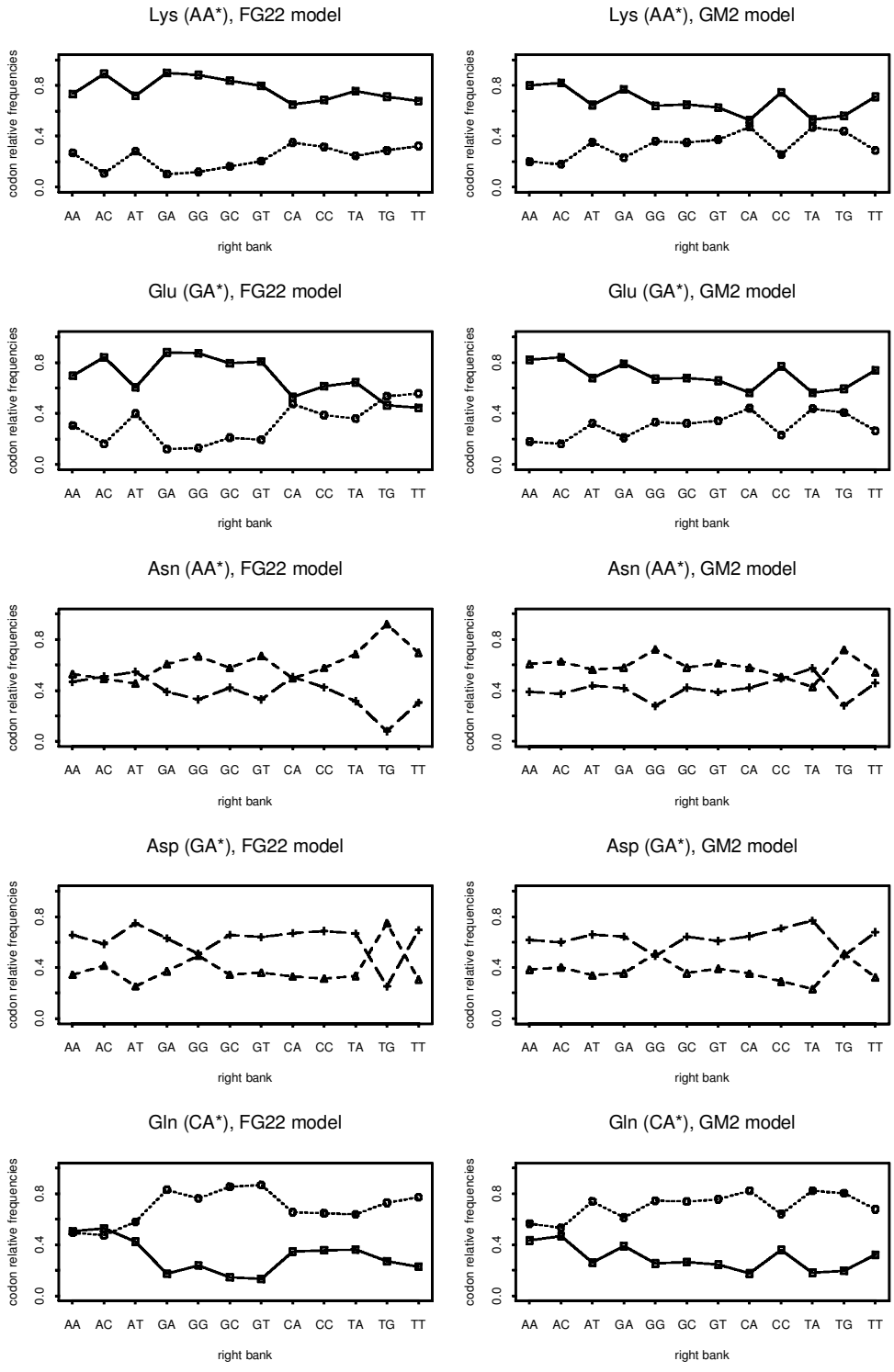

FIG. 6. Empirical frequencies (FG2,2 model) and estimated probabilities in GM2 model, on gaps of size 1, according to the right bank. $\square = A$; $\circ = G$; $\triangle = C$; $+ = T$

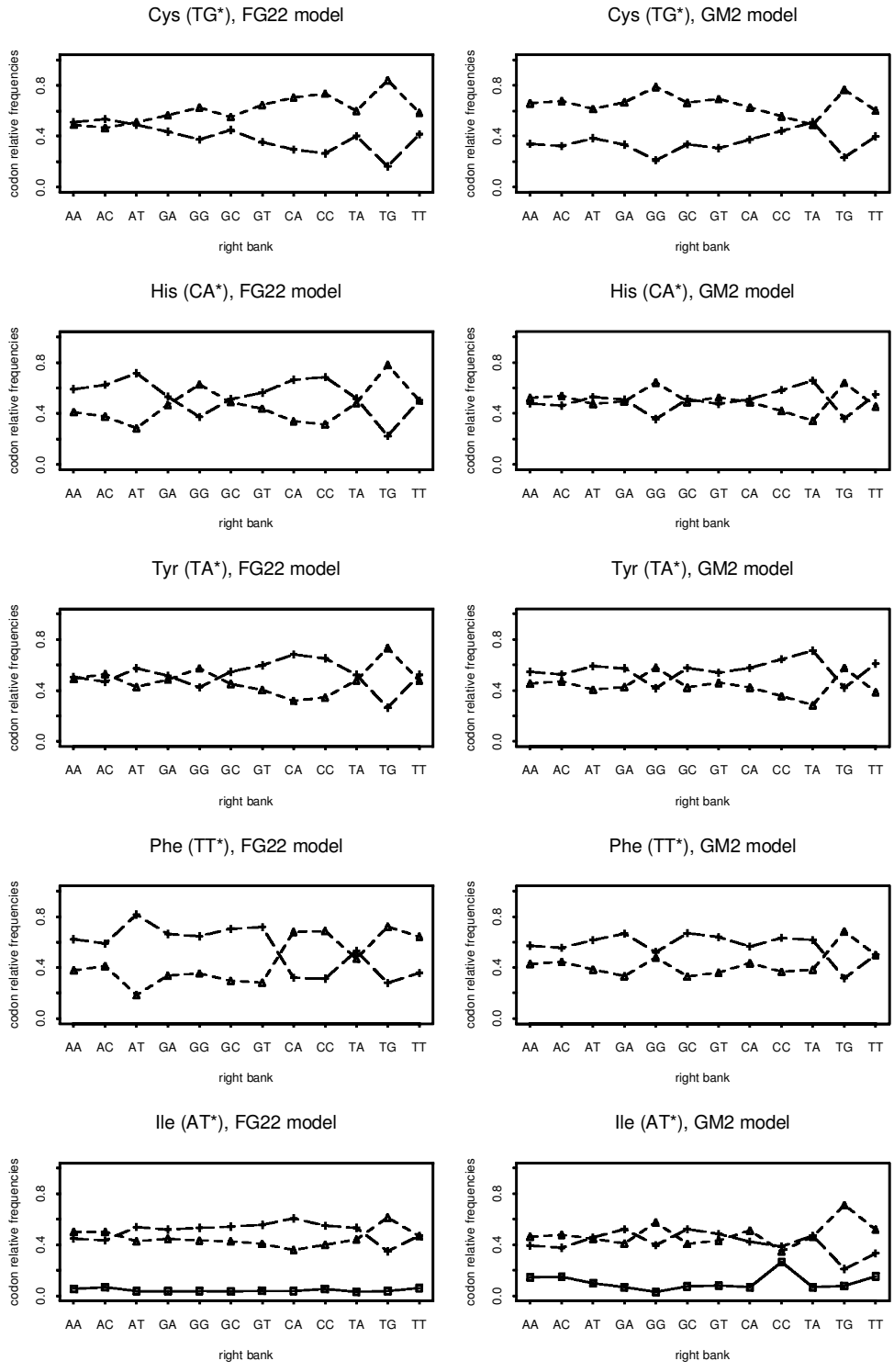

FIG. 7. Empirical frequencies (FG2,2 model) and estimated probabilities in GM2 model, on gaps of size 1, according to the right bank. $\square = A$; $\circ = G$; $\triangle = C$; $+ = T$

Owing to the inclusion relations given above, likelihood ratio tests can be performed to compare inbedded models. For models which are not inbedded, other criteria, such as Akaike's, must be used for comparison purposes. Here, any model is found to be significantly better than any other with fewer parameters. But the huge amount of data makes these tests very powerful, and a significant difference is not necessarily a large one. Most significant and important differences are found between GM1/FG2,1, GM1/GM2, and SM1/SM2 indicating that models which take into account at least some of the trinucleotides present in a gap of length 1 are far better than those limited to dinucleotidic composition.

Figures 5, 6, 7 show, for all gap types considered here, observed relative frequencies (FG2,2 model) and estimated probabilities in model GM2. Codon usage and its variation according to right neighbors are relatively well represented in the GM2 model. This again leads to the idea that local composition in trinucleotides, not only in the coding phase, is an essential feature of coding sequences.

CONCLUSION

Among conditional Markov chains, self-complementary chains do not fit coding sequences as well as general ones. If all constraints, except the primary structure of the polypeptide to be coded, were acting identically on both DNA strains, which is certainly the case for some of them, we would expect preferences on oligonucleotides, as estimated in conditional Markov chains, to be invariant by complementation. Hence, this is evidence for the existence of a major constraint acting differently on both DNA strains.

If codon bias were a by-product of constraints acting on DNA without regard to the phase, we would expect conditional homogeneous Markov chains to be better models than FG2,0. This is not really the case: a second order homogeneous Markov chain is, albeit significantly, not very much better than FG2,0. There must be a major constraint responsible for codon bias that depends on the phase.

These two properties make a very strong argument in favor of the influence of the transcription and translation process on coding sequence's oligonucleotidic composition. This has already been well established, by other means, for translation; for instance, nonuniform usage of synonymous codons has been related to tRNA's abundance (Ikemura, 1981; Bulmer, 1987) and gene expressivity (Gouy and Gautier, 1982; Holm, 1986), has been explained in terms of efficiency requirements by Grosjean and Fiers (1982).

But there is something more. FG2,2 is a significantly better model than FG2,1, itself better than FG2,0. This proves that codon choices are dependent, as was shown by Bulmer (1994). As discussed above, this dependence induces, among oligonucleotides, preferences different from those induced by all of the transcription and translation process. But if these preferences were completely different, GM2 would be worse than FG2,0, which is not the case. So there are constraints other than those related to the transcription and translation process acting everywhere on coding DNA sequences. Due to the importance of phase-related constraints, even in a conditional model other constraints cannot be well analyzed with homogeneous models.

Possibly, second order phased Markov chains would provide interesting models. They would enable one to separate preferences on codons and preferences on noncoding trinucleotides; they would provide better discrimination between constraints related to the transcription and translation processes, from other constraints on DNA structure. In particular, it would be interesting to know at which point these last constraints induce self-complementary preferences on trinucleotides.

REFERENCES

Bulmer M. 1987. Coevolution of codon usage and transfer RNA abundance. *Nature*, 325, 728–730.

Bulmer M. 1994. Synonymous codon usage. in *Informatique et Biologie Moléculaire; Septièmes Entretiens du Centre Jacques Cartier.*

Gouy M., and Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, 10, 7055–7074.

Grosjean H., and Fiers W. 1982. Preferential codon usage in procaryotic gene: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed gene. *Gene*, 18, 199–209.

Hénaut A., and Vigier P. 1995. Etude des contraintes qui s'exercent sur la succession des bases dans un polynucléotide: I. La signification de la dégénérescence du code. *C.R. Académie des Sciences de Paris*, 301(6).

Holm L. 1986. Codon usage and gene expression. *Nucleic Acids Research*, 14(7), 3075–3087.

Ikemura T. 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of Molecular Biology*, 146, 1–21.

Karlin S., and Mràzek J. 1996. What drives codon choices in human genes? *Journal of Molecular Biology*, 252, 459–472.

Médigue C., Rouxel T., Vigier P., Hénaut A., and Danchin A. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *Journal of Molecular Biology*, 222, 851–856.

Trifonov E., N. 1989. The multiple codes of nucleotide sequences. *Bulletin of Mathematical Biology*, 51(4), 417–432.

Address correspondence to:
François Rodolphe
INRA, Unité MIG
78352 Jouy en Jonas Cedex, France

E-mail: fr@jouy.inra.fr

This article has been cited by:

1. Guang Wu, Shaomin Yan. 2005. Determination of mutation trend in proteins by means of translation probability between RNA codes and mutated amino acids. *Biochemical and Biophysical Research Communications* **337**:2, 692-700. [CrossRef]