Nucleic Acids Research, 1999, Vol. 27, No. 19 **3911–3920**

Heuristic approach to deriving models for gene finding

John Besemer[1] and Mark Borodovsky[1,2,*]

[1]School of Biology and [2]School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0230, USA

Received April 30, 1999; Revised and Accepted July 27, 1999

ABSTRACT

Computer methods of accurate gene finding in DNA sequences require models of protein coding and non-coding regions derived either from experimentally validated training sets or from large amounts of anonymous DNA sequence. Here we propose a new, heuristic method producing fairly accurate inhomogeneous Markov models of protein coding regions. The new method needs such a small amount of DNA sequence data that the model can be built 'on the fly' by a web server for any DNA sequence >400 nt. Tests on 10 complete bacterial genomes performed with the GeneMark.hmm program demonstrated the ability of the new models to detect 93.1% of annotated genes on average, while models built by traditional training predict an average of 93.9% of genes. Models built by the heuristic approach could be used to find genes in small fragments of anonymous prokaryotic genomes and in genomes of organelles, viruses, phages and plasmids, as well as in highly inhomogeneous genomes where adjustment of models to local DNA composition is needed. The heuristic method also gives an insight into the mechanism of codon usage pattern evolution.

INTRODUCTION

Computer-aided gene finding frequently employs statistical gene prediction methods based on Markov models (1–3). Parameters of inhomogeneous Markov models for a protein coding DNA sequence could be inferred from training sets of experimentally annotated DNA sequences (1) or from a large enough set of anonymous DNA sequences (2,4–6). In this paper we present a rather simple training procedure that produces fairly efficient Markov models using a minimum amount of training data. The idea of the method is based on two observations made upon analysis of the performance of the prokaryotic versions of the GeneMark and GeneMark.hmm programs (1,3). First, for the *Escherichia coli* genome, whose genes have been divided into three classes that differ in codon usage pattern (7–9), it was noticed that predicting genes of each class by GeneMark did not require carefully tuned up class-specific Markov models. For instance, the genes of Class II, the highly expressed genes possessing the most biased codon usage pattern, could be accurately predicted just by using the models of the genes of Class I, encompassing the majority of *E.coli* genes (10). Secondly, the GeneMark.hmm program (3) was able to detect a vast majority of genes of all three *E.coli* classes using second order Markov models trained on the Class III *E.coli* genes, presumably horizontally transferred genes whose *E.coli*-specific codon usage pattern was the least pronounced. Having realized that practically useful models of protein coding regions may be learned from a rather small amount of genomic sequence, we attempted to avoid the traditional training process. The proposed heuristic procedure for Markov model derivation used a fragment of genomic DNA just long enough to accurately estimate the nucleotide composition. This procedure also used linear functions that related nucleotide frequencies in the three codon positions to the global nucleotide frequencies and linear functions that related amino acid residue frequencies to genome GC content. These functions were obtained by linear regression analysis of DNA sequence data of several completely sequenced prokaryotic genomes. Tests of the new approach were performed on 10 complete bacterial genomes. The heuristically derived models were used with the GeneMark.hmm and GeneMark programs. The tests have shown that the heuristic models worked surprisingly well. Particularly, when GeneMark.hmm was used an average 93.1% of annotated genes were detected, while in comparison, models built by traditional training predicted an average of 93.9% of genes. The heuristic approach to model building will be useful for dealing with prokaryotic species whose genomic sequence information is available in small amounts and for small genomes of organelles, viruses, phages and plasmids, as well as for genomes with highly inhomogeneous DNA composition, when models need adjustment to local DNA composition.

It should be noted that although we have not moved far enough in this direction, the heuristic method produced an interesting by-product, a heuristic codon usage table. Comparison of the experimental and heuristically defined codon frequencies revealed a strong correlation, especially for species with a highly biased nucleotide composition. This correlation indicates the presence of general factors related to genome and proteome composition at the levels of nucleotides and amino acids, respectively, that are involved in shaping the species-specific codon usage patterns. For the species with a balanced nucleotide composition, such as *E.coli*, the heuristic codon usage frequencies may help to single out 'outliers', the instances when codon frequencies deviate from expectation due to some other important factors that become hidden in genomes with a biased composition.

*To whom correspondence should be addressed at: School of Biology, Georgia Institute of Technology, Atlanta, GA 30332-0230, USA. Tel: +1 404 894 8432; Fax: +1 404 894 0519; Email: mark@amber.biology.gatech.edu

MATERIALS AND METHODS

Materials

To obtain the information necessary to generate the heuristic models, we examined the 17 complete bacterial genomes available in GenBank as of November 1998. The species used were *Aquifex aeolicus* (11), *Archaeoglobus fulgidus* (12), *Bacillus subtilis* (13), *Borrelia burgdorferi* (14), *Chlamydia trachomatis* (15), *E.coli* (16), *Haemophilus influenzae* (17), *Helicobacter pylori* (18), *Methanobacterium thermoautotrophicum* (19), *Methanococcus jannaschii* (4), *Mycobacterium tuberculosis* (20), *Mycoplasma genitalium* (21), *Mycoplasma pneumoniae* (22), *Pyrococcus horikoshii* (23), *Rickettsia prowazekii* (24), *Synechocystis* PCC6803 (25) and *Treponema pallidum* (26). Amino acid frequencies and positional nucleotide frequencies were measured using genes annotated in these genomic sequences. For testing purposes, we used 10 of the 17 complete genomes mentioned above as well as the complete genomes for human immunodeficiency virus type I (27) and human T cell lymphotrophic virus type I (28). We also used 650 *Pseudomonas aeruginosa* sequences and 385 *Chlamydomonas reinhardtii* sequences.

The heuristic method of Markov model derivation

This method was designed to build the set of Markov models using a minimal amount of sequence information. When using this approach, only three independent parameters, three of the four nucleotide frequencies specific for the particular genomic sequence, are needed to generate the models necessary to utilize gene-finding programs.

Upon analysis of the 17 complete bacterial genomes we have observed relationships between the positional nucleotide frequencies and the global nucleotide frequencies (Fig. 1A–D) as well as relationships between the amino acid frequencies and the global GC% of the bacterial genomes (Fig. 2A–J). These relationships were approximated by linear functions using standard linear regression. Interestingly, the graphs for positional frequencies of T and G nucleotides (Fig. 1A and C) show a 'Z-pattern' caused by a noticeable difference in frequencies at the first and second codon positions. The frequencies of C and A nucleotides at the first and second codon positions are close to each other and the 'Z-pattern' is absent from the graphs (Fig. 1B and D).

Of the 20 amino acids, the frequency of only 10 were observed to change significantly over the range of GC percentages for the 17 complete genomes examined: 28.6% GC (*B.burgdorferi*) to 65.6% GC (*M.tuberculosis*). Of these 10, four amino acids are coded by SSN type codons (S stands for C or G and designates strong Watson–Crick pairing): alanine (A), glycine (G), proline (P) and arginine (R). Arginine, though, is encoded not only by GCN codons, but also by AGA and AGG. We conventionally considered arginine as an SSN-type amino acid, since four of its six codons are of the SSN type. Frequencies of all four SSN-type amino acids, A, G, P and R, increased as the GC content of the genome increased (Fig. 2A–D).

Five other amino acids whose frequencies significantly changed over the GC% range are coded for by WWN-type codons (W stands for A or T and designates weak Watson–Crick pairing): phenylalanine (F), isoleucine (I), lysine (K), asparagine (N) and tyrosine (Y). As could be expected, the frequency of each of these amino acids decreased as GC% increased (Fig. 2E–I). Methionine, although technically a WWN-type amino acid, occurs very infrequently in bacterial genomes (~1.8%) and changed the least of all 20 amino acids over the GC% range examined.

Amino acids coded by combinations of strong and weak nucleotides in the first two positions of a codon were considered as neutral. Only one neutral amino acid, valine, had a frequency that changed significantly as GC% changed and behaved like a member of the SSN group (Fig. 2J). Valine belongs to the group of aliphatic amino acids, along with isoleucine and leucine. The frequency of isoleucine, classified as a WWN-type amino acid, decreased as GC% increased (Fig. 2F). No significant change in the frequency of leucine, with four of its six codons being of neutral type and the other two of the WWN type, was observed. Perhaps some evolutionary pressure exists to hold near constant the sum of the frequencies of the aliphatic amino acids at the level of the proteome. Thus, as genomic GC% increases, the increase in frequency of valine may be explained as compensation for a deficiency of isoleucine.

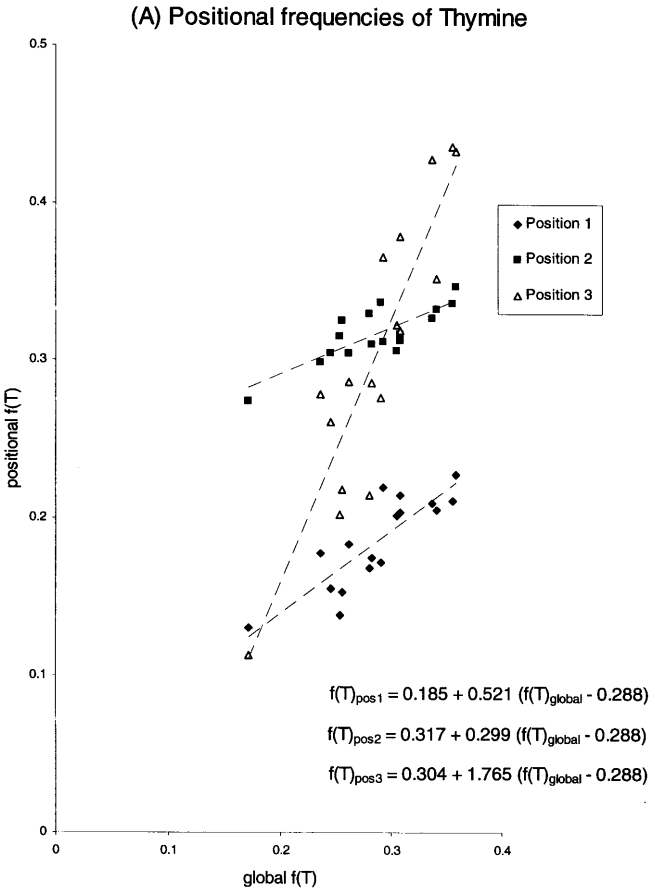
Given these observations, amino acid frequency dependence on global GC% was taken into account only for the four amino acids of SSN type, A, G, P and R, for the five amino acids of WWN type, F, I, K, N and Y, and for the one neutral amino acid, V. For all other amino acid frequencies, the values observed in the *E.coli* proteome were used as constants. For mycoplasma genomes, an additional constant was added to the *E.coli* tryptophan frequency since the codon TGA, usually a stop codon, codes for tryptophan in these species.

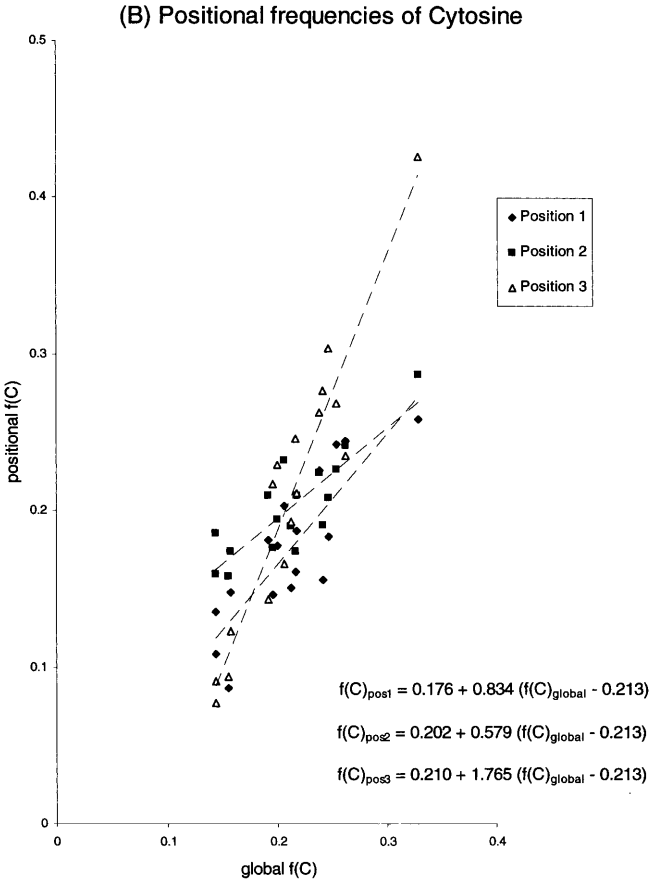
The parameters of the set of Markov models, the three-periodic models for coding sequence of orders zero, one and two and a single zero order model for non-coding sequence, were derived as follows. Learning the global nucleotide frequencies from a given genomic sequence allowed us to determine nucleotide frequencies in each of the three codon positions using the linear relationships shown in Figure 1A–D. Then, the initial values of frequency of occurrence of each of the 61 codons, f_I(XYZ), were obtained as products of the three positional nucleotide frequencies of corresponding nucleotides. The frequency of a particular amino acid was determined for a given GC content and was then used to modify the initial value of codon frequency. For example, the refined frequency of the alanine codon GCT is defined by the formula

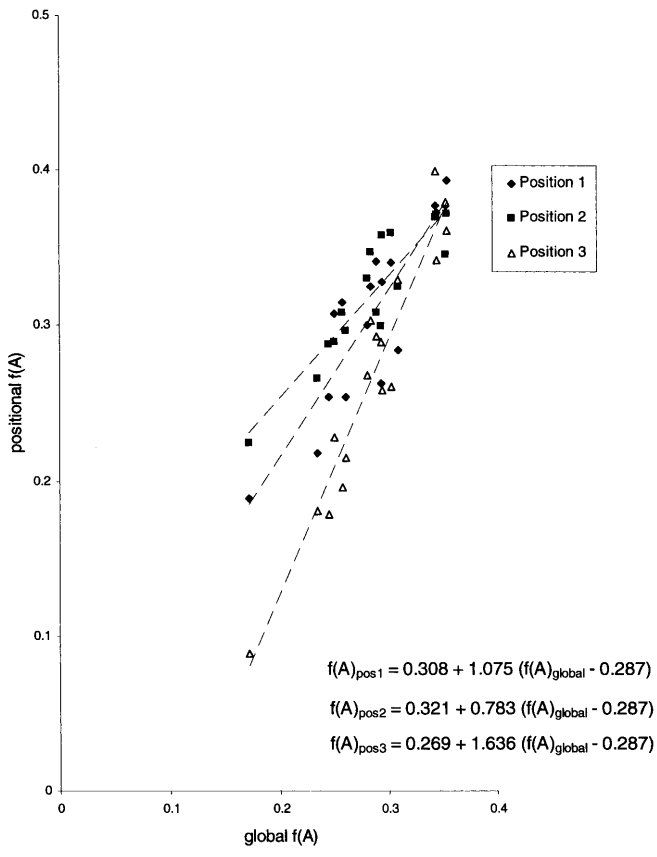
$$f_R(\mathrm{GCT}) = f_{\mathrm{alanine}}(\mathrm{GC\%_{global}}) \times [f_I(\mathrm{GCT})/(f_I(\mathrm{GCC}) + f_I(\mathrm{GCA}) + f_I(\mathrm{GCG}) + f_I(\mathrm{GCT}))] \quad \mathbf{1}$$

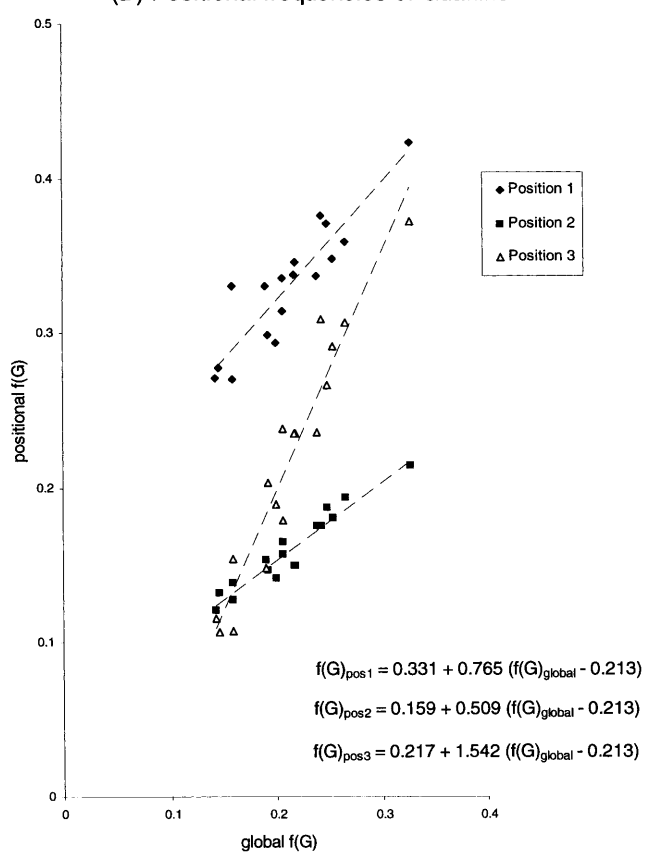
In the case of alanine, encoded by four codons, we could have reached the same result by taking into account only the nucleotide frequencies in the third codon position. Equation **1**,

Figure 1. (Opposite) (**A**) Frequency of nucleotide T in three codon positions observed in 17 bacterial genomes shown as a function of global nucleotide T frequency in a given genome. The equations of the lines approximating the observed data were obtained by linear regression analysis. (**B–D**) As in (A) for positional frequencies of nucleotides C, A and G, respectively.









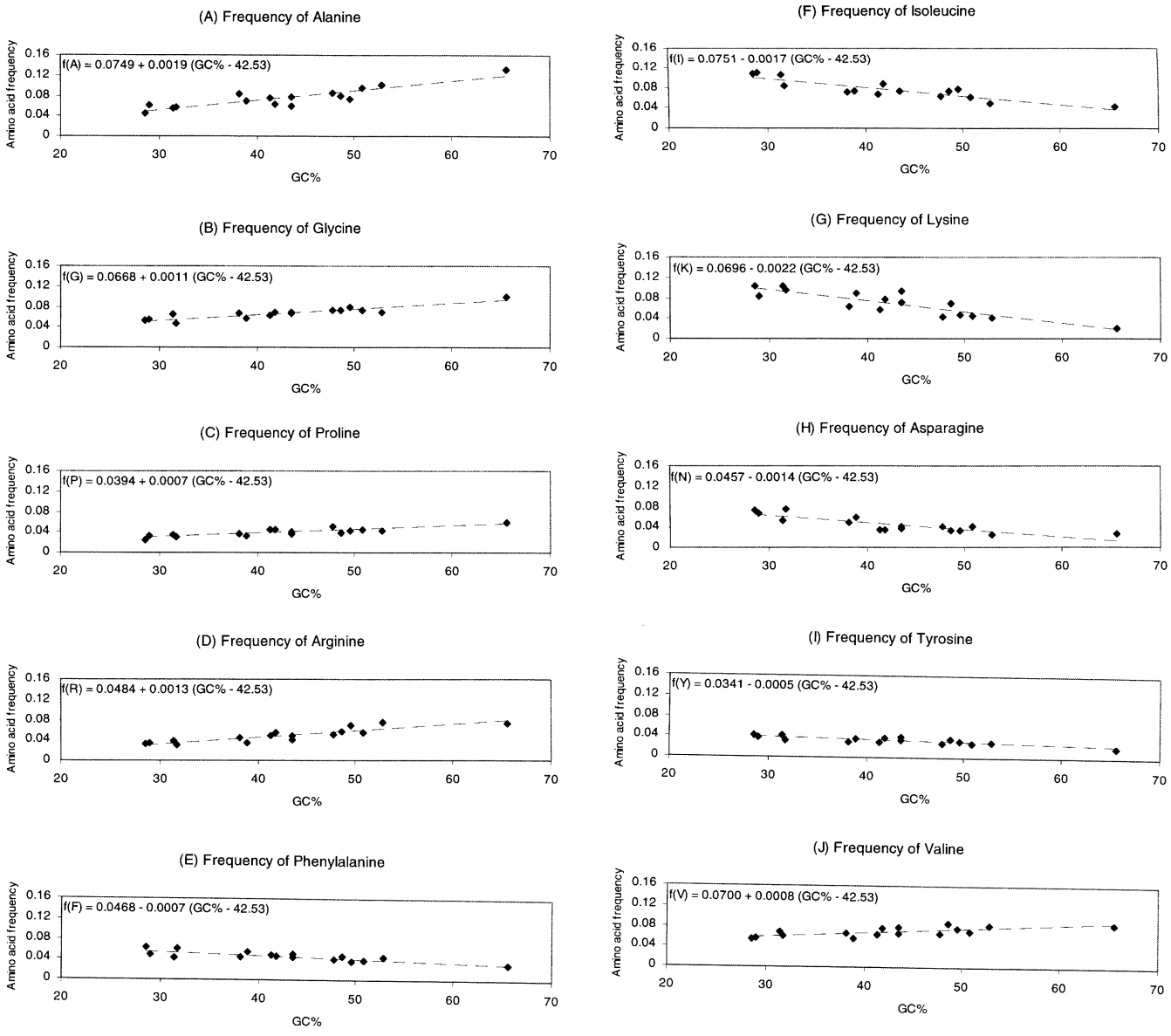


Figure 2. (**A**) Frequency of amino acid alanine shown as a function of GC content of the bacterial genome for 17 genomes. The equation of the line approximating the observed data was obtained by linear regression analysis. (**B–J**) As (A) for glycine, proline, arginine, phenylalanine, isoleucine, lysine, asparagine, tyrosine and valine, respectively.

however, functioned properly independently of the number of synonymous codons, so all amino acids were handled in the same manner. Obviously, this method guarantees that the sum of the refined frequencies of codons is equal to the sum of the frequencies of the amino acids. By completing this computation for all 61 codons, we produced the heuristically built codon usage table for the input genomic sequence.

To construct the three-periodic zero order Markov model of a protein coding region the codon usage table is all that is needed. For example, to determine the probability of A in the first position of a codon, the probabilities of all codons that start with A were added together. In the zero order model of non-coding sequence the global frequencies of the respective nucleotides were used.

For the first order three-periodic Markov model, the codon usage table provides enough data to calculate only two matrices of transition probabilities out of three. To define the values of transition probabilities related to nucleotides occupying the third position of one codon and the first position of the next codon it was assumed that occurrences of adjacent codons are independent events. Indeed, a rather weak correlation exists between nucleotides of adjacent codons. Thus the probability of nucleotide Y in the first position of a codon following a nucleotide X in the third position of the previous codon, $P(X \rightarrow Y)$ for the (..X||Y..) configuration, is equal to the probability of nucleotide Y in the first position of a codon defined previously for the zero order Markov model.

For the second order Markov model, only the transition probabilities for the nucleotide in the third codon position could be produced from the codon usage table.

To find the transition probabilities related to the first and second codon positions, we used the same assumption of

independence of adjacent codons. The second order transition probability $P(XY \rightarrow Z)$ for the (.XY||Z..) configuration was assumed to be equal to the probability of nucleotide Z in the first codon position as defined in the zero order Markov model. The transition probability $P(XY \rightarrow Z)$ for the (..X||YZ.) configuration was equal to the value of probability of Z in the second position following Y in the first position. This value was already defined for the first order Markov model.

Gene finding accuracy estimation

To characterize the gene prediction accuracy of the models, predictions made by the GeneMark and GeneMark.hmm programs were compared with the GenBank annotation. Although it may not always be true, the annotation was assumed to be precisely correct in terms of gene (ORF) location and position of the start codon. We use the term 'close prediction' or mere 'prediction' for the case when the predicted stop codon of the ORF matched the annotated stop, regardless of whether the predicted start codon location matched the annotated start. The term 'exact prediction' describes the case when both the positions of the predicted stop and start codons matched the annotation. Since almost none of the annotated sequence was directly used in building the heuristic models, there was no need to use a cross-validation procedure, which is regularly used to assess the accuracy of the models with a large number of parameters learned from a training set.

RESULTS AND DISCUSSION

Testing on 10 complete bacterial genomes

In order to gauge how well the matrices generated through the heuristic approach performed when used in gene prediction, they were tested on 10 complete bacterial genomes. One test was done with the GeneMark.hmm program only. Another test was performed using a combination of both GeneMark.hmm and GeneMark. Note that the heuristic models could be employed in any gene-finding program using Markov models. Prediction of the 5′-end of a gene was aided by the RBS model built from the *E.coli* sequence data (3).

The complete genomes of the following species were analyzed in the tests: *A.fulgidus*, *B.subtilis*, *E.coli*, *H.influenzae*, *H.pylori*, *M.genitalium*, *M.jannaschii*, *M.pneumoniae*, *M.thermoautotrophicum* and *Synechocystis* PCC6803. By using GeneMark.hmm with heuristic models, an average of 93.1% of the genes present in the GenBank annotations of the above 10 genomes were closely predicted. At the same time, 72.1% of annotated genes were predicted exactly. These results compare favorably with the results obtained using 'native' models derived from the genomic sequences by traditional training (1). When native models were used in GeneMark.hmm, 93.9% of the annotated genes were closely predicted and 77.4% of the genes were exactly predicted (Table 1). This information is broken down into the 10 species tested in Table 2, which shows the results using heuristic models for every species.

A limitation of GeneMark.hmm is that a gene that overlaps at its 3′-end with an adjacent gene in the opposite orientation can be missed. To recover missing genes, the GeneMark program was also run on the sequence. The predictions made by the two programs were then parsed and a single list of predicted genes was produced. In the case where both GeneMark.hmm and GeneMark predicted a gene with the same stop position, the GeneMark.hmm prediction was selected as the representative one. Since we were specifically interested in recovering rather long genes missed by GeneMark.hmm, we only used the GeneMark predictions longer than 500 nt.

Using the combined approach, the heuristic models closely predicted 94.6% of the genes annotated in the 10 bacterial genomes. Native models worked better, predicting on average 97.3% of the genes (Table 1).

When run with the native models, GeneMark.hmm predicted an average of 11.4% genes which were not present in the annotation for the 10 bacterial genomes. These predictions were denoted as potential new genes. The heuristic models predicted an average of 12.4% potential new genes over the same conditions.

Overall, the heuristic models performed similarly to native models in terms of percentage of genes missed, exactly predicted genes and percentage of potential new genes predicted. Only minimal gains were made in terms of gene prediction accuracy by moving from heuristic models to native models. This suggests that heuristic models can be used to accurately predict genes in cases where the amount of sequence data necessary for traditional training of higher order native models is not yet available or cannot be obtained at all. Possible applications include sequencing projects very near their beginning and prediction of genes in small genomes, such as organelles, viruses, phages and plasmids. Heuristic models may also help in analyzing genomes with highly inhomogeneous nucleotide composition.

Table 1. Average gene prediction performance of the heuristic and native models as defined in tests on 10 bacterial genomes

Gene prediction program	Model type	Annotated genes predicted (%)	Annotated genes exactly predicted (%)	Potential new genes (%)
GeneMark.hmm	Native	93.9	77.4	11.4
GeneMark.hmm	Heuristic	93.1	72.1	12.4
GeneMark.hmm and GeneMark	Native	97.3	77.4	11.4
GeneMark.hmm and GeneMark	Heuristic	94.6	73.4	13.4

The figures for annotated genes predicted give the percentage of annotated genes closely predicted (with possible misplacement of the start codon). The combination of GeneMark.hmm with GeneMark described in Lukashin and Borodovsky (3) allows for a 3.4% improvement in the accuracy of native models and 1.5% improvement in heuristic models.

Table 2. Gene prediction accuracy of the GeneMark.hmm program using heuristic models for each of 10 bacterial genomes

	Genes annotated	Genes predicted	Genes predicted (%)	Genes exactly predicted (%)	Potential new genes (%)
A.fulgidus	2407	2516	88.1	70.7	13.2
B.subtilis	4099	4384	96.5	66.4	10.8
E.coli	4289	4426	93.4	75.3	10.7
H.influenzae	1717	1840	95.9	84.9	10.7
H.pylori	1566	1612	93.7	76.8	9.6
M.genitalium	467	509	88.9	66.8	19.3
M.jannaschii	1680	1841	94.5	68.2	13.2
M.pneumoniae	678	734	91.9	65.6	17.0
M.thermoautotrophicum	1869	1944	93.9	65.8	6.9
Synechocystis	3169	3360	94.6	80.5	12.6
Average			93.1	72.1	12.4

The figures in the third column give the percentage of annotated genes closely predicted by the GeneMark.hmm program (with possible misplacement of the start codon). Percentage of potential new genes (false positives) is defined with regard to the number of annotated genes.

Prediction of short genes

Approximately 25% of the 4289 genes found in the *E.coli* genome are shorter than 500 nt in length. Typically, using computational methods these genes have been much more difficult to accurately predict than longer genes. We have done an additional analysis to find out how efficient the models generated through the heuristic approach are in predicting genes of short length in comparison with native models. The percentages of genes that were closely predicted in the *E.coli* genomic sequence using both types of models are shown in Figure 3A. In all length categories other than the shortest one (shorter than 100 nt), which includes 11 annotated genes out of 4289, the heuristic models and native models produced similar results. Figure 3B shows a comparison of the percentage of potential new genes predicted in the *E.coli* genome by both native and heuristic models. The heuristic models and native models predict similar numbers of potential new genes in all length categories.

Testing on bacterial genomes with inhomogeneous composition

Genomic sequences of *P.aeruginosa* were used for this test since they range fairly widely in GC content from 41.6 to 70.3%. Native Markov models based on *P.aeruginosa* sequence were generated using 650 records available in GenBank. The whole GC content range was divided into three bins and the native models were derived for each GC bin. A set of 262 of these sequences was used as a test set for gene prediction accuracy. The GeneMark.hmm program using native models closely predicted 95% of the genes annotated in *P.aeruginosa* sequences and exactly predicted 71% of the annotated genes. The GeneMark.hmm program using the heuristic models, tested on the same 262 genes, also made close predictions for 95% of the annotated genes and exactly predicted 75% of the annotated genes. The program using native models predicted 23% genes not present in the annotation while using heuristic models produced a slightly smaller number of new predictions, 21%. Overall, better results were achieved with heuristically derived models than with native models. Note that the need to use a cross-validation procedure to estimate the accuracy of the heuristic model was virtually abolished since the number of the model parameters learned from the genomic sequences is so small. Using cross-validation for testing the native models could lead to a decrease in the accuracy figure given above by several percentage points.

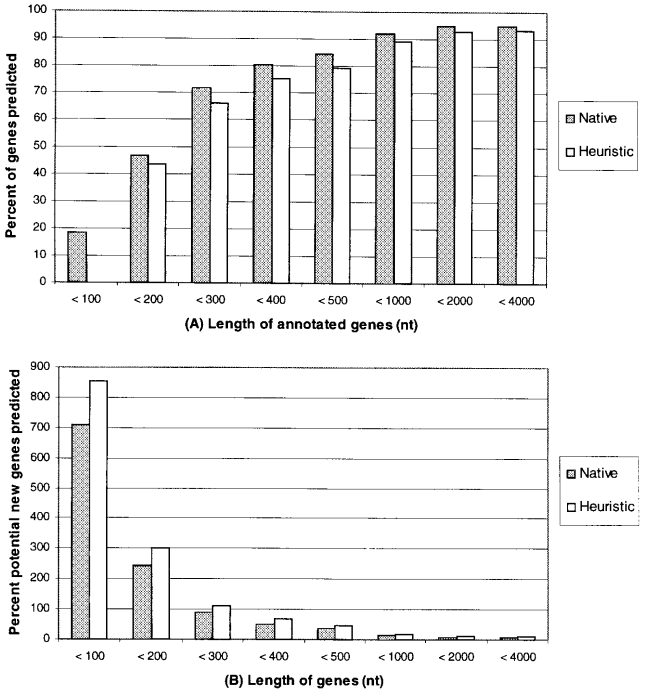


Figure 3. (**A**) Percent of genes predicted by the GeneMark.hmm program out of those annotated in the *E.coli* genome as a function of gene length. (**B**) As (A) for the percent of potential new genes.

Testing on the smallest genomes

Due to the small genome size of phages and viruses, there is not enough sequence data to properly train the Markov models in the traditional way. Since the heuristic approach eliminates the need for large amounts of DNA sequence for model building, genomes of phages and viruses were an ideal test case for this method. The heuristic approach was tested on the complete genomes of human immunodeficiency virus type I and human T cell lymphotropic virus type I.

The HIV type I genome is only 9719 nt long and contains eight annotated genes, two of which contain a single intron. The intron for both genes is in exactly the same position, 6046–9378, although the two genes have different start and stop locations. The prokaryotic version of GeneMark.hmm used here was unable to accurately predict genes with introns. Therefore, we ignored the two genes that contain introns. GeneMark.hmm predicted five of the six annotated genes and no genes that were not annotated. The gene that was not predicted was 237 nt in length and overlapped the preceding gene. The GeneMark program recovered this annotated gene. However, GeneMark also predicted the existence of five other short genes in this sequence that were not present in the annotation. Four of these five predictions were located in the long terminal repeat (LTR) region of the genome as indicated by the annotation. A BLAST search on these four predictions revealed that all four sequences coded for the HIV type I *nef* protein. The *nef* protein has been shown to play a role in virus replication and it has recently been suggested that it may be related to pathogenicity as well (30). The other prediction, interestingly, was located inside the annotated intron. A BLAST search on this sequence revealed that it coded for the Vpu protein. Therefore, all the intronless predictions made by heuristically derived models were supported by either nucleotide sequence annotation in GenBank or by already existing entries in protein sequence databases.

The GenBank record for the human T cell lymphotropic virus type I, 9068 nt, contains only three annotated genes. Using GeneMark.hmm with heuristic models, all three annotated genes were closely predicted. In addition, four genes were predicted that were not present in the annotation. Using GeneMark, another potential new gene was predicted. Of the five potential new genes, two were exactly identical and were found in the LTR regions. Although there were no BLAST hits for this duplicated putative gene, a similarity search against the SCOP database produced a number of matches to known protein structural domains (31). BLAST searches on the other three predicted proteins revealed that two of them were already annotated proteins for this virus. The similarity search for one remaining protein produced no significant homology.

Interpretation of the results in terms of Kullback–Liebler distance

A gene-finding method exploiting the maximum likelihood concept, such as GeneMark.hmm (3), attempts to fit each one of the set of initially defined Markov models, such as models of coding or non-coding regions, to a given DNA sequence eventually parsed into segments where one or another model fits best. Essentially, the algorithm implements 'competition' among the models for the best fit for each given DNA segment. In the simplest case of two competing models, this competition is quantified by a likelihood ratio value that determines which one of two models fits the particular DNA segment better. In making this decision there are two types of possible errors, false positive and false negative. The lower these errors are, the higher is the discrimination power of the method. The Kullback–Liebler (KL) distance or relative entropy, $D(P||Q)$, is a measure of affinity of two statistical (Markov) models P and Q. Rigorous theory shows (29) that false positive (false negative) error rates decrease exponentially with growth of the sequence fragment length, n, with the exponent value proportional to $-nD(P||Q)$ $[-nD(Q||P)]$. This means that the discrimination power should increase as the sequence fragment increases. As shown earlier, this trend was indeed observed in the tests of gene finding as the longer genes were detected with higher accuracy.

The KL distance, as a single parameter, has proven to be a useful characteristic for comparative analysis of performance of gene-finding methods using high order Markov models described by a large number of parameters. Some rather surprising results could be explained in terms of the KL distance (5). Particularly, it was observed that Markov models trained on the set of *E.coli* Class I genes provided higher discrimination power for detecting the *E.coli* Class II genes than for genes from Class I itself (4). This observation agrees with the fact that the effective KL distance, $D^e(P||Q)$, between one model (the *E.coli* Class II gene, P) and another (the *E.coli* non-coding sequence, Q), defined by the formula $D^e(P||Q) = D(P||Q) - D(P||P^*)$, where P^* is the model of the *E.coli* Class I gene, is larger than the regular KL distance, $D(P^*||Q)$, between the *E.coli* Class I gene model and the non-coding sequence model (M.Borodovsky, unpublished data).

Table 3. Values of the Kullback–Liebler distance

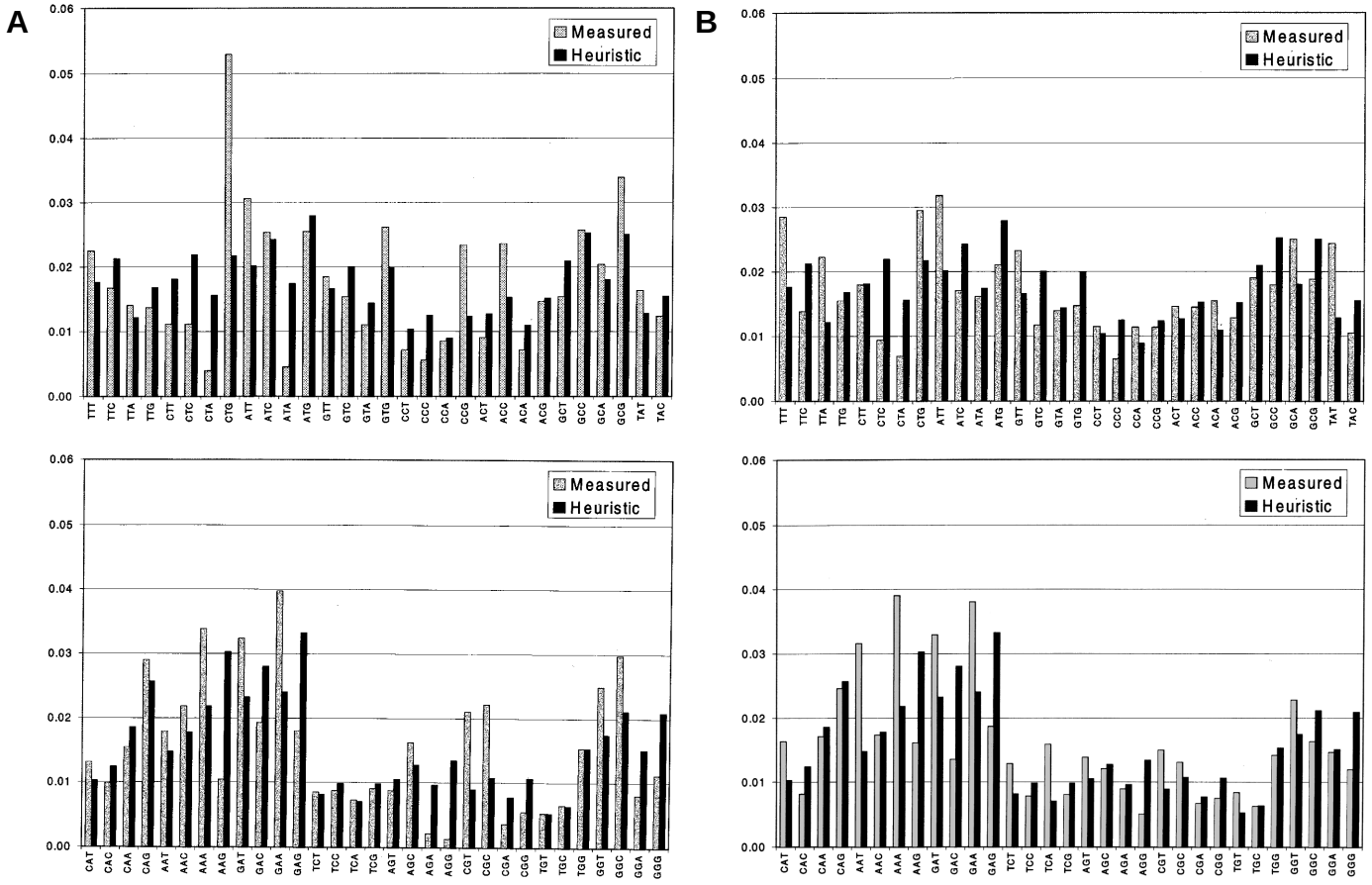
Models	Class 1	Class 2	Class 3	Heuristic	Non-coding
Class 1		0.060	0.063	0.083	0.148
Class 2	0.059		0.145	0.154	0.244
Class 3	0.069	0.183		0.056	0.063
Heuristic	0.097	0.193	0.053		0.083
Non-coding	0.112	0.247	0.024	0.055	

The KL distance values defined by equation **2** are given for both $D(P||Q)$ and $D(Q||P)$, the values related to false positive and false negative error rates (see text).

In our computations we used the standard formula for the KL distance (29) modified to deal with the periodic Markov models P and Q:

$$D(P||Q) = 1/3\sum^3_c\sum^4_i\sum^4_j\sum^4_k p^c_{ijk} \log_2(p^c_{ijk}/q^c_{ijk}) \quad \mathbf{2}$$

Here index c represents the three codon positions and indices i, j and k represent the four types of nucleotides. We computed the KL distance values between several second order models: the three three-periodic Markov models for the *E.coli* genes of Classes I, II and III; the heuristic model of *E.coli* genes; the ordinary Markov model of the *E.coli* non-coding region (Table 3). It is seen that in terms of KL distance the *E.coli* heuristic model lies close to the *E.coli* Class III model. Both of these



models are closer to the model of the non-coding region than the model of *E.coli* Class I and, especially, the model of *E.coli* Class II genes. Let us assume that the *E.coli* Class III model or the *E.coli* heuristic model is used in the gene-finding algorithm. The effective KL distance between the *E.coli* Class I or Class II gene and non-coding models is then large enough (see 5) and the genes of these classes should be predicted with sufficient accuracy. This expectation was confirmed to be a reasonable one for both the heuristic model (current paper) and for the Class III model (M.Borodovsky and A.V.Lukashin, unpublished data). The *E.coli* heuristic model was also able to predict *E.coli* Class III genes, as would be expected from the closeness of these two models in terms of KL distance. However, in comparison with the Class III model the heuristic model has an advantage in that it could be built in a regular way from a small portion of DNA sequence data without dealing with clusterization of the whole gene pool of a given species (5,6,10). Thus, we conclude that the *E.coli* heuristic model is able to predict genes of all *E.coli* classes, while also being easy to build. A similar analysis could be conducted for genomes of other bacterial species with the gene pools divided, if necessary, into sets of typical and atypical genes (5).

Why the method works? Implications for evolution of codon usage

The comparison of actual codon usage frequencies with the codon frequencies defined by the heuristic codon usage table might help understand why the heuristic model is efficient for gene finding. In Figure 4A–B both sets of codon frequencies are shown for the whole *E.coli* genome (Fig. 4A) and the *E.coli* Class III genes (Fig. 4B). These sets are also shown for the genomes of *B.burgdorferi* and *M.tuberculosis* (Fig. 4C and D) having, respectively, the lowest and the highest GC% content in the sample of 17 complete genomes. The correlation coefficient, *R*, between the experimental codon frequencies and the heuristic codon frequencies were calculated. The *R* values for the *E.coli* whole genome and the *E.coli* Class III genes were equal to 0.58 and 0.48, respectively. For *B.burgdorferi* and *M.tuberculosis* the *R* values were equal to 0.94 and 0.87, respectively. The *R* values indicate that the pattern of codon usage frequency was captured by the heuristic procedure to some extent for the case of the *E.coli* genome, with medium GC content. This correlation, as indicated earlier, turned out to be sufficient to generate models providing reasonably accurate gene-finding ability. For the genomes with highly biased GC content the high level of correlation indicates that the codon usage pattern is modeled by the heuristic procedure in great detail. This observation raises the question of to what extent does the simple principle of the heuristic method explain the diversity of codon usage patterns in bacterial and perhaps in higher species. The basic parameters of the method, the global nucleotide frequencies, appear to be the fundamental variables defined by the complex structure of the species-specific biochemical pathways producing the building blocks of the

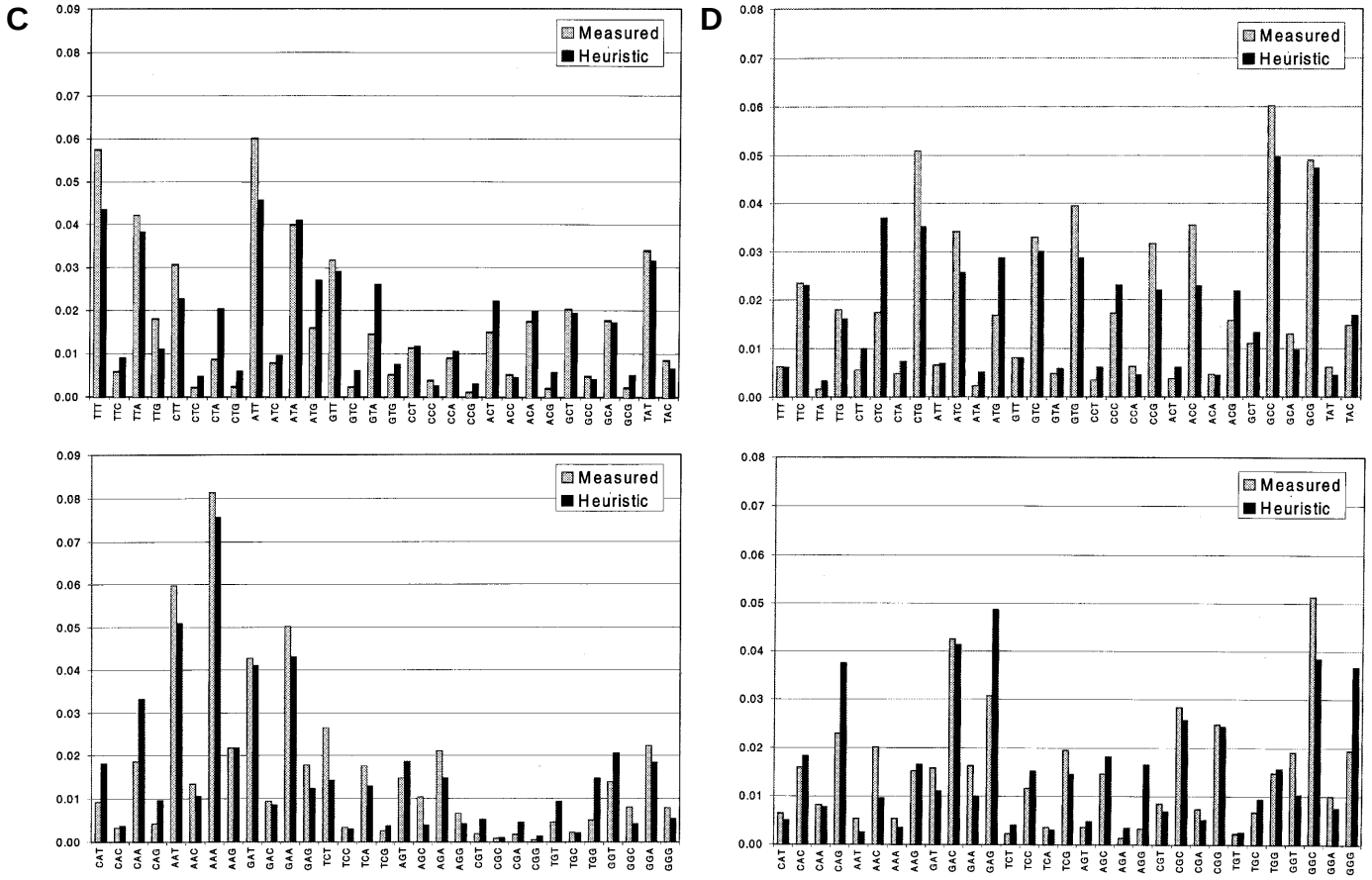


Figure 4. (Opposite and above) (**A**) Frequencies of 61 codons as observed in *E.coli* protein coding regions are shown along with the codon frequencies defined by the heuristic codon usage table. The order of codons follows the order accepted in the table of genetic code with the exception of serine and arginine codons. Although a correlation between natural and model frequencies is seen in general, some significant differences exist in frequencies of particular codons: CTG, CCG, CGT and CGC, as well as CTA, ATA, AAG, GAG, AGA and AGG. (**B**) As in (A) with observed codon frequencies taken from 158 genes of *E.coli* Class III. Again, a general correlation is observed with the noticeable exception of CTC, CTA, AAG, GAC, GAG and AGG, as well as AAT and AAA. (**C**) As in (A) for *B.burgdorferi*. A clear periodic pattern is seen in codon frequencies. This pattern relates to the type of nucleotide occupying the third codon position. The *B.burgdorferi* genome is AT-rich. The codon frequencies for codons ending with A/T are greater in most cases than the frequencies of those ending with C/G. The natural and heuristic frequencies are in good correlation. (**D**) As in (A) for *M.tuberculosis*. Here the periodic pattern similar to that observed in (C) is reversed since the *M.tuberculosis* genome is GC-rich. The natural and heuristic frequencies correlate reasonably well.

DNA double helix. The amino acid composition of a species proteome represents other fundamental variables that are even more conserved among species and vary slightly as the GC% of the whole genome changes. Therefore, the observations, though limited, suggest that a significant component of the codon usage pattern in evolution may be the result of a compromise between restrictions on the nucleotide and amino acid compositions, achieved mainly using elasticity of the silent codon positions. This mechanism of developing the codon usage pattern seems to be more pronounced in the species with highly biased nucleotide composition, such as *B.burgdorferi* and *M.tuberculosis*. In species such as *E.coli* the pressure of compositional restrictions is weaker and other factors come more freely into play to form the codon usage pattern (8,9,32).

Testing on eukaryotic genomes

Although the heuristic approach to model building was designed with bacterial gene prediction in mind, we had reasons to try this approach for eukaryotic DNA sequences as well. The advantage of the heuristic model is the ability of being produced 'on the fly' and of being adjusted to local sequence composition. Thus, the heuristic models under certain circumstances may replace a native protein coding model, which is part of a whole set of models needed for eukaryotic gene prediction. This whole set, as used in the eukaryotic version of GeneMark.hmm (M.Borodovsky and A.V.Lukashin, unpublished data), includes contextual models for start codons, stop codons and splice sites as well as models for length distributions for coding and non-coding regions (exons, introns and intergenic regions). We have estimated the parameters for these additional models from a set of 350 *C.reinhardtii* sequences available in GenBank. Parameters of the 'local' heuristic models for protein coding regions were derived using nucleotide frequencies counted in DNA fragments of ~1000 nt. Although we used the linear functions derived for bacterial genomes, a separate analysis has provided evidence that similar linear functions are valid for the

eukaryotic case (data not shown). When GeneMark.hmm using native models (of order five) was run on a randomly selected test set of 58 *C.reinhardtii* sequences, 88.5% of the annotated exons and 65% of the annotated genes were exactly predicted. Using heuristic models of up to only order two, 82% of the exons and 52% of the genes were predicted exactly. These results suggest that a heuristic approach aimed at eukaryotic genomes has some potential and requires further study. It could be especially useful for genomes with a highly inhomogeneous composition.

Gene finding with heuristic models via a web server

The software program that builds the heuristic model for input sequences is accessible via the Internet at http://dixie.biology.gatech.edu/GeneMark/heuristic.cgi . This program produces heuristic models for a sequence longer than 400 nt. The models are then applied to the analysis of the input sequence by the GeneMark and GeneMark.hmm programs. Output of the web server includes a list of predicted genes in text format and, optionally, a list of predicted protein sequences and a graph of protein coding potentials.

ACKNOWLEDGEMENTS

We are grateful to Dr Jeffrey Lawrence for useful discussion and to Maribeth Norris for help with manuscript preparation. J.B. and M.B. were supported in part by a grant from the US National Institutes of Health. M.B. was also supported in part by a grant from the US Civil Research Development Foundation.

REFERENCES

1. Borodovsky,M.Y. and McIninch,J.D. (1993) *Comput. Chem.*, **17**, 123–133.
2. Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) *Nucleic Acids Res.*, **26**, 544–548.
3. Lukashin,A.V. and Borodovsky,M. (1998) *Nucleic Acids Res.*, **26**, 1107–1115.
4. Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., Fitzgerald,L.M., Clayton,R.A., Gocayne,J.D. *et al*. (1996) *Science*, **273**, 1058–1073.
5. Hayes,W.S. and Borodovsky,M. (1998) *Genome Res.*, **8**, 1154–1171.
6. Audic,S. and Claverie,J.M. (1998) *Proc. Natl Acad. Sci. USA*, **17**, 10026–10031.
7. Grantham,R., Gautier,C., Gouy,M., Jacobzone,A. and Mercier,M. (1981) *Nucleic Acids Res.*, **9**, r43–r74.
8. Gouy,M. and Gautier,C. (1982) *Nucleic Acids Res.*, **10**, 7055–7074.
9. Médigue,C., Rouxel,T., Vigier,P., Henaut,A. and Danchin,A. (1991) *J. Mol. Biol.*, **222**, 851–856.
10. Borodovsky,M., McIninch,J.D., Koonin,E.V., Rudd,K.E., Médigue,C. and Danchin,A. (1995) *Nucleic Acids Res.*, **23**, 3554–3562.
11. Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., Graham,D.E., Overbeek,R., Snead,M.A., Keller,M., Aujay,M. *et al.* (1998) *Nature*, **392**, 353–358.
12. Klenk,H.P., Clayton,R.A., Tomb,J., White,O., Nelson,K.E., Ketchum,K.A., Dodson,R.J., Gwinn,M., Hickey,E.K., Peterson,J.D. *et al*. (1997) *Nature*, **390**, 364–370.
13. Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S. *et al.* (1997) *Science*, **390**, 249–256.
14. Fraser,C.M., Casjens,S., Huang,W.M., Sutton,G.G., Clayton,R.A., Lathigra,R., White,O., Ketchum,K.A., Dodson,R., Hickey,E.K. *et al*. (1997) *Nature*, **390**, 580–586.
15. Stephens,R.S., Kalman,S., Lammel,C.J., Fan,J., Marathe,R., Aravind,L., Mitchell,W.P., Olinger,L., Tatusov,R.L., Zhao,Q. *et al*. (1998) *Science*, **282**, 754–759.
16. Blattner,F.R., Plunkett,G., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al*. (1997) *Science*, **277**, 1453–1462.
17. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J., Dougherty,B.A., Merrick,J.M. *et al.* (1995) *Science*, **269**, 496–512.
18. Tomb,J.-F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleischmann,R.D., Ketchum,K.A., Klenk,H.P., Gill,S., Dougherty,B.A. *et al*. (1997) *Nature*, **388**, 539–547.
19. Smith,D.R., Doucette-Stamm,L.A., Deloughery,C., Lee,H.-M., Dubois,J., Aldredge,T., Bashirzadeh,R., Blakely,D., Cook,R., Gilbert,K. *et al*. (1997) *J. Bacteriol*., **179**, 7135–7155.
20. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E. *et al*. (1998) *Nature*, **393**, 537–544.
21. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G.G., Kelley,J.M. *et al*. (1995) *Science*, **270**, 397–403.
22. Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.-C. and Herrmann,R. (1996) *Nucleic Acids Res.*, **24**, 4420–4449.
23. Kawarabayasi,Y., Sawada,M., Horikawa,H., Haikawa,Y., Hino,Y., Yamamoto,S., Sekine,M., Baba,S., Kosugi,H., Hosoyama,A. *et al*. (1998) *DNA Res.*, **5**, 55–76.
24. Andersson,S.G.E., Zomorodipour,A., Andersson,J.O., Sicheritz-Ponten,T., Alsmark,U.C.M., Podowski,R.M., Naeslund,A.K., Eriksson,A.S., Winkler,H.H. and Kurland,C.G. (1998) *Nature*, **396**, 133–140.
25. Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S. *et al*. (1996) *DNA Res.*, **3**, 109–136.
26. Fraser,C.M., Norris,S.J., Weinstock,G.M., White,O., Sutton,G.G., Dodson,R., Gwinn,M., Hickey,E.K., Clayton,R., Ketchum,K.A. *et al*. (1998) *Science*, **281**, 375–388.
27. Wong-Staal,F., Gallo,R.C., Chang,N.T., Ghrayeb,J., Papas,T.S., Lautenberger,J.A., Pearson,M.L., Petteway,S.R.,Jr, Ivanoff,L., Baumeister,K. *et al*. (1985) *Nature*, **313**, 277–284.
28. Inoue,J., Watanabe,T., Sato,M., Oda,A., Toyoshima,K., Yoshida,M. and Seiki,M. (1986) *Virology*, **150**, 187–195.
29. Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*. John Wiley & Sons, New York, NY.
30. Zaher,H., Kay,D.J., Rebai,N., Guimond,A., Jothy,S. and Jolicoeur,P. (1998) *Cell*, **95**, 163–175.
31. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540.
32. Lawrence,J.G. and Ochman,H. (1997) *J. Mol. Evol.*, **44**, 383–397.