How to Interpret an Anonymous Bacterial Genome: Machine Learning Approach to Gene Identification

William S. Hayes and Mark Borodovsky

Genome Res. 1998 8: 1154-1171
Access the most recent version at doi:10.1101/gr.8.11.1154

RESEARCH

How to Interpret an Anonymous Bacterial Genome: Machine Learning Approach to Gene Identification

William S. Hayes and Mark Borodovsky[1]

School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332-0230 USA

In this report we address the problem of accurate statistical modeling of DNA sequences, either coding or noncoding, for a bacterial species whose genome (or a large portion) was sequenced but not yet characterized experimentally. Availability of these models is critical for successful solution of the genome annotation task by statistical methods of gene finding. We present the method, GeneMark-Genesis, which learns the parameters of Markov models of protein-coding and noncoding regions from anonymous bacterial genomic sequence. These models are subsequently used in the GeneMark and GeneMark.hmm gene-finding programs. Although there is basically one model of a noncoding region for a given genome, several models of protein-coding region are automatically obtained by GeneMark-Genesis. The diversity of protein-coding models reflects the diversity of oligonucleotide compositions, particularly the diversity of codon usage strategies observed in genes from one and the same genome. In the simplest and the most important case, there are just two gene models—typical and atypical ones. We show that the atypical model allows one to predict genes that escape identification by the typical model. Many genes predicted by the atypical model appear to be horizontally transferred genes. The early versions of GeneMark-Genesis were used for annotating the genomes of *Methanoccocus jannaschii* and *Helicobacter pylori*. We report the results of accuracy testing of the full-scale version of GeneMark-Genesis on 10 completely sequenced bacterial genomes. Interestingly, the GeneMark.hmm program that employed the typical and atypical models defined by GeneMark-Genesis was able to predict 683 new atypical genes with 176 of them confirmed by similarity search.

Pioneer methods for statistical identification of protein-coding regions in DNA sequence were developed in the early 1980s (Fickett 1982; Gribskov 1984; Staden 1984). These methods exploited the statistically significant differences in compositional features of continuous protein-coding and noncoding DNA sequences. The search for formal mathematical tools that would express these differences in the most efficient way led to the introduction of inhomogeneous, three-periodic Markov chain models of protein-coding regions (Borodovsky et al. 1986a). These models, along with ordinary Markov models of noncoding DNA sequence, were incorporated into a Bayesian algorithm, GeneMark, analyzing DNA sequence locally within a sliding window (Borodovsky et al. 1986b, Borodovsky and McIninch 1993). Later on these models were used in a global maximum likelihood algorithm, GeneMark.hmm, analyzing the whole DNA sequence at once (Lukashin and Borodovsky 1998). The combination of parameters of three-periodic Markov models of different orders, the interpolated model, was used in the GLIMMER program for gene finding in bacterial genomes (Salzberg et al. 1998). Three-periodic Markov models have been also used in the programs most often used for eukaryotic gene finding, such as GENSCAN (Burge and Karlin 1997), GRAIL (Xu et al. 1994), and HMMgene (Krogh 1997). Sound application of the Markov model-based approach to gene finding have been performed recently in annotating the first complete bacterial genomes *Haemophilus influenzae*, *Mycoplasma genitalium*, and others (see Methods). Difficulties soon surfaced. Unlike the pioneer genome sequencing project of *H. influenzae*, many genomes completed later were swiftly sequenced starting from the zero point. Therefore, there was an absence of previously experimentally annotated DNA sequence necessary to determine the parameters of Markov models. This difficulty was aggravated by the fact that protein-coding sequences in a given bacterial genome were not homogeneous in their compositional features (Medigue et al. 1991). Therefore, the program using models trained on the bulk

[1]Corresponding author.
E-MAIL mark@amber.gatech.edu; FAX (404) 894-0519.

 8:1154–1171 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

set of protein-coding sequences happened to be insensitive in finding genes of minor inhomogeneity classes. In cases when preliminary information on gene classes is available, class-specific models of protein-coding regions could improve the performance of the gene-finding method (Borodovsky et al. 1995). Therefore, the problem was to derive accurate models of coding and noncoding regions, including specific models for several gene classes, from the anonymous sequence.

The GeneMark-Genesis method described below addressed this problem. The method worked in two main steps. A set of ''long'' open reading frames (ORFs) identified in bacterial genomic sequence was used to start a process of obtaining parameters of Markov models of protein-coding and noncoding regions. The initial models were used in the GeneMark program to score the putative gene sequences and to form the cluster seeds for the class-specific training sets. Then the clusterization procedure using relative entropy (Cover and Thomas 1991) as a distance function was run until convergence and several sets of presumably coding sequences with more homogeneous compositional features were obtained. In the simplest and the most important case, there were just two clusters that gave rise to so-called typical and atypical models of protein-coding region.

The early versions of the GeneMark-Genesis program were used in genome sequencing projects of *M. jannaschii* (Bult et al. 1996) and *H. pylori* (Tomb et al. 1997); however, complete publication of the method was delayed until comprehensive data on the accuracy of various algorithm options were obtained.

We tested the GeneMark-Genesis program on 10 complete bacterial genomes (see Methods). The results of the tests were presented in terms of local ''short fragment prediction accuracy'' characterized by false-negative and false-positive error rates and in terms of global ''whole gene prediction accuracy'' characterized by sensitivity and specificity parameters. The intriguing possibility to predict genes of different evolutionary origins at an early stage of anonymous genome analysis did not escape our attention. For the *Escherichia coli* genome, we have analyzed the relationship between compositionally atypical genes and horizontally transferred genes as identified in previous studies (Medigue et al. 1991; Lawrence and Ochman 1997). We showed that using the atypical gene models in the GeneMark.hmm program led to prediction of >400 new genes in the 10 genomes. From these predictions, 176 were corroborated by protein sequence similarity search by gapped BLAST (Altschul et al. 1997).

RESULTS AND DISCUSSION

Root Models

The Root models, of orders zero to five, were generated, as described in Methods, for the 10 complete bacterial genomic sequences. For each genome, the predictive accuracy of the Root model was evaluated by the ''short DNA fragment identification'' procedure with cross-validation. The accuracy of the Root model was compared with the predictive accuracy of the GenBank model. A summary of these comparisons is given in Table 1. The ''optimal'' order of the Root (GenBank) model was defined as the order

Table 1. False-Negative Prediction Error Rates for Optimal Orders of the Root and GenBank Models for 10 Species

	Root model		GenBank model	
Species name	**optimal order**	**error rate**	**optimal order**	**error rate**
Archaeolglobus fulgidus	3	0.096	5	0.087
Bacillus subtilis	4	0.133	5	0.119
Escherichia coli	4	0.137	5	0.113
Haemophilus influenzae	3	0.073	4	0.077
Helicobacter pylori	3	0.077	5	0.082
Mycoplasma genitalium	3	0.122	4	0.105
Methanococcus jannaschii	3	0.067	4	0.060
Mycoplasma pneumoniae	3	0.112	5	0.104
Methanobacterium thermoautotrophicum	4	0.098	5	0.084
Synechocystis	4	0.122	5	0.120

of the model that produced the lowest false-negative error rate. The error rates produced by the Root (GenBank) models of the optimal order are shown in Table 1. The slightly higher false-negative error rates produced by the Root models are within 2% of the error rates generated by the GenBank models.

More detailed comparison of the Root and GenBank models performance is presented in Figure 1 for the *E. coli* case. The observed relatively high false-positive error rate of the Root model (Fig. 1) is partly attributable to the higher contamination of the noncoding test set for the Root model with true protein-coding regions as compared with the noncoding test set for GenBank model. Note that although the Root model prediction accuracy in terms of the false-negative error rate is close to the GenBank model for each given order (Fig. 1), there is still room for improvement for both the Root and GenBank model performance. For each genome, the accuracy of the Root model and the GenBank model was also characterized in terms of sensitivity (Sn) and specificity (Sp) values (see Methods). The GenBank model performed slightly better than the Root model, for instance, the Sn and Sp average for *B. subtilis* was 0.941 for the GenBank model and 0.921 for the Root model. For the more homogeneous genome of *H. pylori*, the Sn and Sp average was 0.952 and 0.946 for the GenBank model and the Root model, respectively. For all of the species studied in this paper, the difference in predictive accuracy between the two models, measured by the average of Sn and Sp, was less than three percentage points. The upper limit of averaged Sn and Sp was 0.943 for the GenBank model and 0.919 for the Root model.

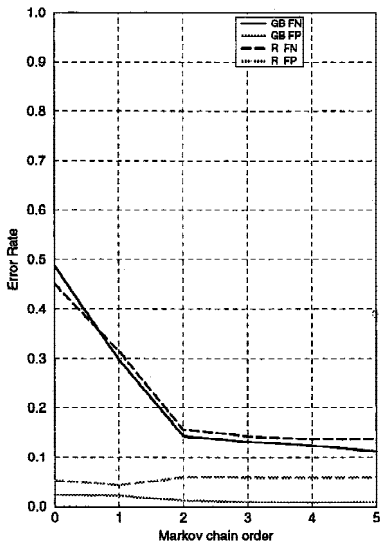

Figure 1 False-negative (FN) and false-positive (FP) prediction error rates observed in the short fragment identification procedure implemented the *E. coli* genome. The Root and GenBank models were employed in the GeneMark program making the predictions. Similar graphs are called accuracy graphs (see Figs. 2 and 3, below).

Clustering

We used the clustering procedures described in the Methods section to reduce the level of inhomogeneity in the sets of coding regions used for training models of several gene classes. The results of ORF clusterization and the characteristics of the performance of these cluster-specific models are presented and discussed in this section.

It was argued that the GeneMark-type algorithm reliably identifies the protein-coding regions whose oligonucleotide composition conforms to the Markov model that GeneMark is using. The Markov model, however, derived from a large enough set of genes represents the composition features common to the majority of genes in this set. Therefore, even using this model in the GeneMark program to analyze the same set of genes, some coding regions with atypical composition could be misidentified as noncoding. Our attempt to divide the initial set of long ORFs into several homogeneous subsets (clusters) pursued the goal of forming at least, and preferably at most, two clusters that would be used as training sets for typical and atypical models. As shown below, for all bacterial genomes in the current study, the two-cluster division was indeed possible and worked sufficiently well for gene-finding purposes. To follow the previous studies of *E. coli* genome (Medigue et al. 1991), however, we also considered the three-cluster case with the third cluster called the highly typical cluster.

The two-means clustering algorithm, which produced the type A clusters, was initially tested on a ''toy'' sequence, the hybrid DNA sequence combining *E. coli* and *H. influenzae* genomic sequences. The long ORF clusters, typical and atypical, obtained by the clustering procedure were named as *EC* and *HI* clusters after the name of the species representing a majority of the ORFs in a particular cluster. The typical cluster turned out to be the *EC* cluster, as there are almost three times as many *E. coli* ORFs as *H. influenzae* ORFs. Table 2 shows the sizes of the overlaps between the sets of true *E. coli* and *H. influenzae* ORFs and the resulting *EC* and *HI* clusters obtained from the hybrid sequence.

The models derived from the *EC* and *HI* clusters

Table 2. Composition of Typical and Atypical ORF clusters Obtained for Combined Sequence in Terms of Native ***E. coli*** and ***H. influenzae*** ORFs

	Cluster type	
	atypical	**typical**
Set of *E. coli* ORFs	265	2270
Set of *H. influenzae* ORFs	925	30

Majority of *E. coli* long ORFs fell into the typical cluster; majority of *H. influenzae* long ORFs fell into the atypical cluster.

of long ORFs were evaluated using the ''short DNA fragment identification'' procedure. In Figure 2, the top panel shows the false-negative error rate produced by the Combined (Root) model of different orders. This model was used to identify coding function in short fragments of long ORFs from the Combined set (solid line), from the *EC* cluster (dotted line), and from the *HI* cluster (dashed line). Obviously, the Combined model more reliably recognized as coding the sequences from the larger *EC* cluster than the sequences from the smaller *HI* cluster (the majority of the training set effect). The *EC* cluster derived model, the middle panel, identified quite well the coding property of short sequences from the *EC* cluster, was less precise, on average, in identifying sequences from the Combined set, and was a poor coding function predictor for the sequences from the *HI* cluster. The *HI* cluster model, the bottom panel, predicted the coding property of sequences from the *HI* cluster quite well, moderately well for those from the Combined set and rather poorly for the sequences from *EC* cluster. As is seen in Table 2 and Figure 2, the toy example demonstrated the ability of the algorithm to detect the presence of two types of protein-coding sequences in one anonymous DNA sequence.

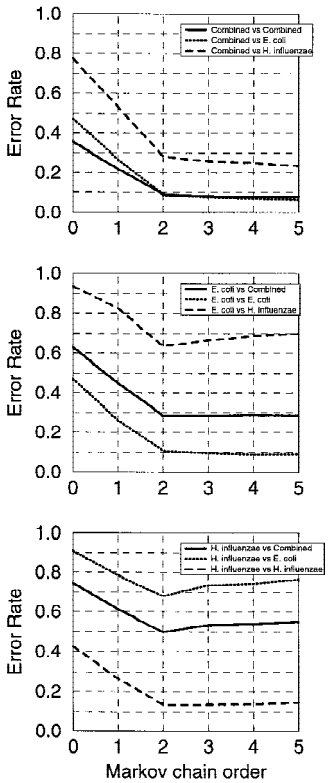

Figure 2 Accuracy graph of the analysis of the combined sequence of *E. coli* and *H. influenzae* (see text and Fig. 3 legend).

The two-means clusterization procedure was applied to the 10 complete bacterial genomes, assuming no annotation was known, to obtain the typical and atypical clusters of long ORFs (A, B, and C types of clusters). The cluster-specific models of protein-coding sequences obtained from these clusters were designated as 2A-, 2B-, and 2C-type models. The three-means ($k = 3$) clustering procedure was also applied to the 10 genomes and produced the typical, atypical, and highly typical ORF clusters (of types A, B, and C). The cluster-specific models obtained from these clusters were designated as 3A-, 3B-, and 3C-type models. The sizes of all clusters are given in Table 3. Interestingly, the largest atypical clusters of type A were found in the two *Mycoplasma* genomes.

Long ORFs from the atypical cluster defined either by the two- or three-means clustering possessed an atypical oligonucleotide composition. Many of these ORFs could be descendants of genes horizontally transferred into the bacterial genome in the course of evolution. For a given genome, one might expect to see two types of horizontally transferred genes, the ones with GC content lower than average and ones with GC content higher than average. Therefore, one may assume that the atypical cluster should contain two subsets, one with higher-than-average GC content and another one with lower than average GC content. The clustering results, however, did not quite meet this expectation. Atypical long ORFs in all bacterial genomes, besides the two genomes of *M. genitalium* and *M. pneumonia*, yielded bell-shaped unimodal GC content distributions. Also, as shown in Table 4, in all but the two *Mycoplasma* genomes, the GC content of the atypical cluster was 3%–5% lower than the typical cluster.

As an exception to the rule, among the *M. geni-*

Table 3. Numbers of Long ORFs in the Clusters of Type A, B, and C Obtained by ***k***-Means Clustering Procedure (***k*** = 2, 3)

Species name	Total no. of long ORFs	Cluster types	Number of ORFs in clusters: 2-means (AT)	2-means (T)	3-means (AT)	3-means (HT)	3-means (T)
A. fulgidus	1222	A,B,C	148	1071	137	263	819
B. subtilis	2234	A	687	1547	600	392	1242
		B,C	335	1899	335	392	1507
E. coli	2553	A	691	1844	628	534	1373
		B,C	380	2155	380	534	1621
H. influenzae	987	A	203	782	172	240	573
		B,C	147	838	147	240	598
H. pylori	907	A	194	706	166	264	470
		B,C	134	766	134	264	502
M. genitalium	309	A	88	221	76	54	179
		B,C	46	263	46	54	209
M. jannaschii	887	A	151	736	136	273	478
		B,C	133	754	133	273	481
M. pneumoniae	416	A	188	228	178	82	156
		B,C	62	354	62	82	272
M. thermoautotrophicum	957	A,B,C	99	858	95	194	668
Synechocystis	1837	A	352	1479	337	438	1056
		B,C	274	1557	274	438	1119

talium long ORFs assigned to type A or C atypical clusters by two-means clustering procedure there were nine ORFs highly deviant in GC content. Only one of the nine ORFs, the one encoding a cell envelope protein of unknown function, was annotated in GenBank. Among the *M. pneumonia* ORFs assigned to type A or C atypical clusters there were 22 ORFs with deviated GC content. Of these 22 ORFs, 20 were annotated in GenBank. These annotated genes were either adhesin genes or genes of hypothetical proteins. Two *M. jannaschii* ORFs with highly deviant GC content were assigned to type A or C atypical clusters.

Incidentally, those of the GC-deviant ORFs observed in these three species that were annotated in GenBank were encoding cell-envelope or surface-structure proteins. The other atypical ORFs, predicted as genes but not annotated, were also likely to belong to the same family. Interestingly, for the three species mentioned above, all GC deviants did not belong to the atypical clusters of type B.

The cross-validation tests have shown (Table 5), that the models derived from atypical clusters were sufficient to identify, almost all ORFs in these clusters as protein-coding sequences.Therefore, typical and atypical models together covered an overwhelming majority of long ORFs in a given bacterial genome. Therefore, assuming no compositional difference between sets of longer and shorter genes there was no practical need, from the gene finding standpoint, to continue sub-clustering any of the atypical clusters.

Note that the outcome of the clustering procedure could vary depending on the procedure parameters. Because clustering is essentially an optimization process, the final division of the initial set into clusters corresponds to a local minimum of the distance function. To prove that the minimum is the global one is rarely computationally feasible. Nevertheless, it was observed that significant variations of the ORF cluster seeds did not influence the result of the clustering. For instance, for a genome such as *E. coli* a random assignment of long ORFs into ''seed'' clusters for two- or three-means clustering still resulted in essentially the same set of clusters.

For each genome, the accuracy of models derived from ORF clusters was assessed by the ''short fragment identification method''. This analysis is illustrated in Figure 5, below for the *E. coli* genome case. The three left panels of Figure 3 show the false-negative error rates observed in application of the Root model (top), the typical model (middle), and the atypical model (bottom) obtained by two-means clustering. The three curves in each panel corre-

Table 4. GC Content for the Typical and Atypical Clusters of Types A, B, and C

Species name	Cluster type	2-means		3-means		
		AT	T	AT	T	HT
A. fulgidus	A	43.8	50.6	43.4	50.4	51.3
	B	43.8	50.6	43.4	50.4	51.3
	C	43.8	50.6	43.4	50.4	51.3
B. subtilis	A	41.7	46.1	41.3	46.2	45.6
	B	42.7	45.2	42.3	45.1	45.6
	C	40.5	45.6	40.4	45.7	45.6
E. coli	A	47.7	53.8	47.3	53.6	54.0
	B	48.6	52.8	48.0	52.5	54.0
	C	46.2	53.4	46.1	53.2	54.0
H. influenzae	A	36.0	39.8	35.9	39.2	40.7
	B	36.6	39.4	36.3	38.9	40.7
	C	35.9	39.6	35.9	39.1	40.7
H. pylori	A	37.2	40.5	37.5	39.4	42.0
	B	37.2	40.2	37.6	39.2	42.0
	C	37.3	40.3	37.6	39.3	42.0
M. genitalium	A	32.2	31.7	33.0	30.7	33.6
	B	32.1	31.8	32.2	31.2	33.6
	C	32.7	31.6	33.8	30.8	33.6
M. jannaschii	A	29.6	32.9	29.8	31.3	35.6
	B	29.2	32.8	29.4	31.3	35.6
	C	29.6	32.8	29.8	31.2	35.6
M. pneumoniae	A	40.9	41.2	41.2	39.9	43.0
	B	39.5	41.2	39.6	40.6	43.0
	C	44.1	40.5	44.7	39.5	43.0
M. thermoautotrophicum	A	44.0	52.1	43.9	51.9	52.5
	B	44.0	52.1	43.9	51.9	52.5
	C	44.0	52.1	43.9	51.9	52.5
Synechocystis	A	42.0	50.7	41.9	50.1	51.8
	B	42.2	49.9	41.9	49.3	51.8
	C	41.6	50.4	41.5	49.8	51.8

spond to the three control sets of short protein-coding fragments derived from the Root cluster (solid line), typical cluster (dotted line), and atypical cluster (broken line), respectively. The error rates observed in application of the models derived by three-means clustering are shown in the right panels (Fig. 3) devoted to typical (top), highly typical (middle), and atypical model (bottom). The three curves in each panel correspond to the control sets of short sequences derived from typical cluster (solid line), highly typical cluster (dotted line), and atypical cluster (broken line). A lighter gray curve is seen in all panels except for the top left one. This curve, which is attributable to the graphical design might occlude the solid line underneath it, represents the cross-validation error rate obtained for the model derived from the preclustering procedure. This curve appears only in the tests where the test set could overlap with the training set and where cross-validation was used, such as the typical model versus typical ORF cluster test, etc. It was observed that in most instances the preclustering cross-validation false-negative error rates are practically the same as the postclustering cross-validation error rates.

The graphs of the error rates produced by the models obtained by the three-means clustering (Fig. 3) well resemble the results of the previous work (Fig. 1 in Borodovsky et al. 1995). In that study, the protein-coding sequence models were derived from the three classes of *E. coli* genes described by Medigue et al. (1991). Classification presented in that study was based on the clustering of 61 dimensional vectors of codon frequencies by correspon-

Table 5. Fraction of Atypical Long ORFs Predicted by the GeneMark Program using the Atypical Model

	Type of AT cluster used as a training set					
Species name	**2A**	**2B**	**2C**	**3A**	**3B**	**3C**
A. fulgidus	0.892			0.883		
B. subtilis	0.978	0.991	0.967	0.980	0.982	0.967
E. coli	0.907	0.982	0.863	0.904	0.982	0.863
H. influenzae	0.990	0.993	0.986	0.988	0.980	0.980
H. pylori	0.985	0.985	0.978	0.982	0.985	0.978
M. genitalium	1.000	1.000	0.978	1.000	1.000	0.957
M. jannaschii	0.980	1.000	0.977	0.978	0.992	0.977
M. pneumoniae	0.963	1.000	0.855	0.949	1.000	0.855
M. thermoautotrophicum	0.949			0.947		
Synechocystis	0.977	0.982	0.974	0.979	0.982	0.974

dence analysis. Genes assigned to class I included the majority of the *E. coli* genes with neither an optimal codon usage pattern nor an unusual one. Class II included genes highly expressed under exponential growth conditions. DNA sequences of class II genes show a strong bias toward the ''optimal'' *E. coli* codons. Class III genes mainly included genes horizontally transferred into the *E. coli* genome during the course of evolution. These gene sequences might keep remnants of the codon usage strategies of their prehistoric hosts. In the current study, we observed clear correlations (Figs. 4 and 5) between typical, highly typical, and atypical clusters of long ORFs and the updated sets of class I, class II, and class III sets of genes (A. Danchin, pers. comm.). Correlation was also observed between codon usage patterns of the class I, class II, and class III genes and the ones of typical, highly typical, and atypical ORFs (data not shown). These observations gave the motivation to name the ORF clusters as typical, atypical, and highly typical.

The atypical gene clusters derived for the *E. coli* genome were also compared with the set of 756 *E. coli* horizontally transferred genes classified as such in the earlier work (Lawrence and Ochman 1997) and kindly provided by J.G. Lawrence. Only 410 of 756 genes were longer than 500 nucleotides. These

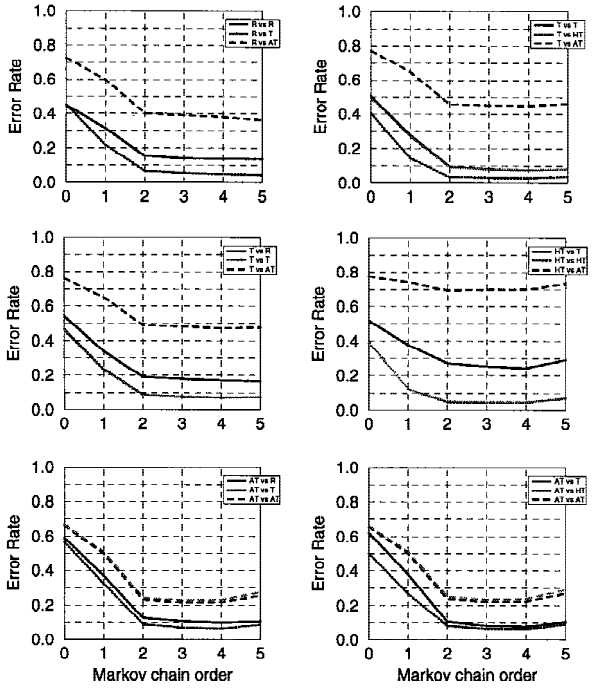

Figure 3 The accuracy graphs for the *E. coli* genome analysis(see text). Average false-positive error rates were 0.055 for the Root model; 0.43 for the typical and 0.060 for the atypical models, respectively, obtained by two-means clustering; 0.055 for the typical, 0.036 for the highly typical, and 0.071 for the atypical model obtained by three-means clustering.

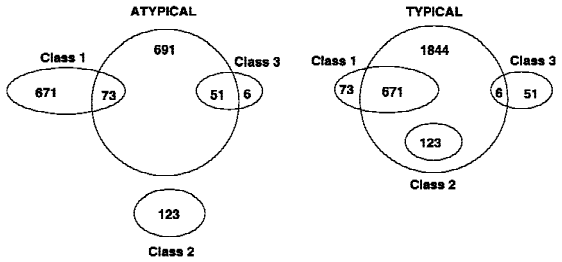

Figure 4 Venn diagrams representing overlaps between the *E. coli* gene classes (Medigue et al. 1991) and ORF clusters obtained by two-means clustering.

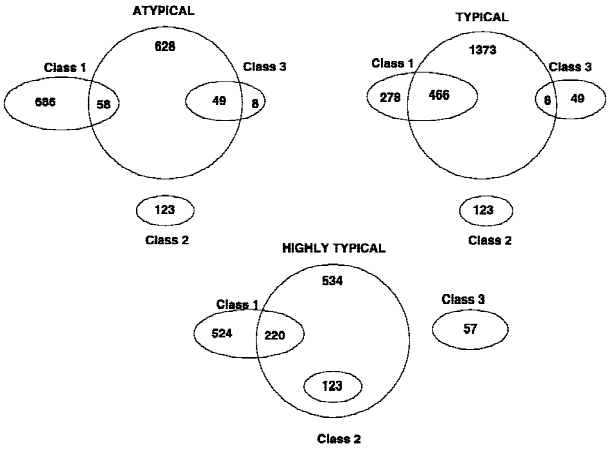

Figure 5 Venn diagrams representing overlaps between the *E. coli* gene classes (Medingue et al. 1991) and ORF clusters obtained by three-means clustering.

410 genes were compared with the A-, B-, and C-type clusters of ORFs obtained by two- or three-means clustering procedure applied to the *E. coli* set of ORFs longer than 500 nucleotides. It is seen (Table 5) that the type-A cluster is more than twice as large as the designated set of 410 genes and includes either 324 or 321 of them, clusters 2A and 3A, respectively. Cluster C with 470 ORFs demonstrated a much stronger relative overlap with the set of 410 genes, 257 to 260 genes in 2C and 3C clusters, respectively. Cluster B with 470 ORFs had only 132 and 152 overlapping items for 2C and 3C, respectively. Interestingly, the cluster B-type construction rule somehow discriminated against longer ORFs. The vast majority of the discussed above *E. coli* A- and C-type clusters, from 92% to 99%, were the ORFs >700 nucleotides (Table 6), whereas in cluster C the ORFs >700 nucleotides constituted only ~78%.

The relatively small overlap of >500 nucleotides, genes of class III with clusters A, B, and C (Table 6) is explained by the fact that the class III list, updated as of 1995, did not include full information on the *E. coli* genome completed in 1997. Classification of genes in genomes other than *E. coli* was not available in terms of the names horizontally transferred genes. It seems plausible, however, that many long ORFs assigned to the atypical gene clusters obtained for the other nine genomes, particularly to the clusters of type C, are the genes horizontally transferred into these genomes in the course of evolution. Application of the GeneMark.hmm program has demonstrated that some of these genes are still absent in the 10 publicly available genome annotations (see below).

Many long ORFs used as initial material in GeneMark-Genesis have been annotated as genes in GenBank and their properties were indicated. This allowed us to analyze the features of the typical and atypical ORFs as follows. Figure 4 shows distribution of the *E. coli* typical and atypical long ORFs with regard to the functional categories defined by Monica Riley (1993) and annotated in the GenBank record (Blattner et al. 1997). The Atypical ORFs have a majority among ''phage, transposon, or plasmid'' category and constitute a significant fraction of cell structure genes, genes of putative regulatory proteins, and genes of hypothetical proteins and genes of proteins with unknown function.

In Figure 3 that when the *E. coli* typical model was used, protein-coding function was easily identified in the highly typical protein-coding sequences (dotted line in the top right panel showing low error rate) but not in the atypical protein-coding sequences (dashed line in the same panel). Actually, the typical model recognized the highly typical segments even better than the typical ones.

Table 6. Compositional Features of Atypical Clusters 2A and 3A of ***E. coli*** Long (⩾500 Nucleotide) ORFs

Clustering		**A**	**B**	**C**
2-means	no. of AT ORFs	952	470	470
	no. of horizontally transferred genes in the overlap	324	132	257
	no. of class III genes in the overlap	68	32	42
3-means	no. of AT ORFs	863	470	470
	no. of horizontally transferred genes in the overlap	321	152	260
	no. of class III genes in the overlap	65	36	42

Note that the number of horizontally transferred genes with a length ⩾ 500 nucleotides is 410 (Lawrence and Ochman 1998). The number of class III genes with a length ⩾ 500 nucleotides is 99 (Medigue et al. 1991).

This is the reason why the highly typical gene model appeared to be redundant for gene-finding purposes as far as the *E. coli* genome is concerned. On the other hand, the *E. coli* atypical gene model, which worked fairly well for all types of genes, seems to be a necessary tool for identifying atypical genes. The similar pattern of error rate dependence on the model type and the model order was observed for the other nine genomes (data not shown). These results support the idea that three gene classes exist in the other nine bacterial genomes. This possibility had already been proven true in the case of the *B. subtilis* genome (Kunst et al. 1997).

Sn and Sp Characteristics

The Sn and Sp values, as defined in Methods, indicate both the ability of GeneMark employing a particular model to predict true (annotated) genes and to avoid false-positive predictions. Predictions that do not match the GenBank annotation reduce the value of specificity. Some newly predicted genes, however, might be correct. There have been many instances where genes predicted by a computer method in presumably noncoding sequence, as determined from the GenBank annotation, were experimentally confirmed at a later time (Robison et al. 1994; Borodovsky et al. 1995; McIninch et al. 1996).

In terms of Sn and Sp, none of the cluster types, A, B, or C, was found to be superior in terms of producing better models. For the 10 genomes, the performance of GeneMark, measured by Sn and Sp, was studied for various combinations of the models. For the *E. coli* case, the comparison of predictive accuracy of single models and combinations of models is presented in Table 7. It is seen that the combination of typical and atypical cluster models generally performed very well. A concise list of the averaged Sn and Sp values obtained in similar experiments for all 10 genomes is given in Table 8. Here, the averaged Sn and Sp values for preclustering cross validation are displayed in italics and parentheses. Karlin et al. (1998) reported differences in codon usage pattern between short (100–300 codons) and long genes (>500 codons). It was indicated that up to 5% increase in G+C content of codon site 3 was observed for long genes of the *E. coli* genome, whereas in the *H. influenzae* genome this increase was <1%. Would the observed difference in codon usage between long and short genes affect the prediction accuray? The results of the accuray assessment (Table 8) showed that Avg(Sn,Sp) values for *E. coli* and *H. influenzae* were as close as 0.1%. This result provides indirect evidence that the difference in codon usage for long versus short genes, changing G+C content of codon site 3, should not contribute significantly to prediction error rate.

Gene Clusters in the Kullback–Liebler Distance Space—Accuracy of the Gene Prediction

The observation (Fig. 3, top right) that the typical gene model was able to correctly identify function of a highly typical gene fragment more reliably than the function of a typical gene fragment is intriguing. The quantitative nature of this observation could be explained in terms of the Kullback–Liebler (KL) distance that measures the ''contrast'' between two competing statistical models involved in the pattern recognition procedure. KL distance, or relative entropy, is a convenient distance measure for clustering procedure dealing with sequence objects that are targets of some pattern recognition process. In such a case, it is desirable that the distance measure is proportional to the ''ease'' to discriminate between two objects. Classical determination procedure is based on likelihood ratio. Relative entropy is an expected logarithm of the likelihood ratio (Cover and Thomas 1991). On the intuitive level, the larger the contrast between models, the KL distance, the easier the task to classify an object, a sequence fragment, into one of two classes, coding or noncoding. For ordinary first-order Markov models P and Q with initial and transitional probabilities designated as p_i and p_{ij}, q_i, and q_{ij}, respectively, the KL distance is defined as

$$D(P\|Q) = \sum p_i p_{ij} \log \frac{p_{ij}}{q_{ij}} \qquad (1)$$

This definition is applied under the assumption that model P actually fits the real sequence of events. This assumption, however, may not always be a true one. For instance, let us consider the short DNA fragment identification procedure where competing models of coding and noncoding regions are involved. In this case, the KL distance is defined by the slightly different equation (Borodovsky et al. 1986a):

$$D(P\|Q) = \frac{1}{3}\sum_{k=1}^{3}\sum_{i,j=1}^{m} p_i^k\, p_{ij}^k \log \frac{p_{ij}^k}{q_{ij}} \qquad (2)$$

Here P and Q represent coding and noncoding models, respectively. Index k defines the phase (frame) of the three-periodic model, indices i and j define nucleotides in adjacent positions, $p^k{}_i$ and $p^{k}{}_{ij}$ are ini-

Table 7. Characteristics of Gene Prediction Accuracy for the ***E. coli*** Genome with 4289 Annotated Genes

Models	AT cluster type	Sn	Sp	Average (Sn, Sp)	No. of correctly predicted genes	No. of predictions
GB		0.908	0.978	0.943	3895	3981
R		0.851	0.988	0.919	3651	3696
AT	2A	0.925 *(0.923)*	0.982 *(0.983)*	0.954	3968	4039
AT,T	2A	0.931 *(0.929)*	0.981 *(0.982)*	0.956	3995	4073
R,AT,T	2A	0.932	0.978	0.955	3999	4087
T	2A	0.835 *(0.834)*	0.991*(0.991)*	0.913	3580	3614
AT	2B	0.916 *(0.911)*	0.985 *(0.985)*	0.951	3930	3991
AT,T	2B	0.922 *(0.917)*	0.983 *(0.983)*	0.953	3954	4023
R,AT,T	2B	0.922	0.983	0.953	3954	4023
T	2B	0.859 *(0.857)*	0.990 *(0.990)*	0.924	3683	3719
AT	2C	0.924 *(0.924)*	0.980 *(0.981)*	0.952	3964	4043
AT,T	2C	0.939 *(0.938)*	0.979 *(0.979)*	0.959	4027	4113
R,AT,T	2C	0.939	0.979	0.959	4027	4113
T	2C	0.848 *(0.845)*	0.990 *(0.990)*	0.919	3639	3674
AT	3A	0.927 *(0.924)*	0.982 *(0.983)*	0.955	3975	4048
AT,HT,T	3A	0.936	0.978	0.957	4014	4103
AT,T	3A	0.934 *(0.930)*	0.980 *(0.981)*	0.957	4004	4085
HT	3A	0.716 *(0.715)*	0.989 *(0.991)*	0.853	3070	3103
R,AT,T	3A	0.934	0.980	0.957	4004	4085
R,AT,T,HT	3A	0.936	0.978	0.957	4014	4103
T	3A	0.852 *(0.850)*	0.990 *(0.990)*	0.921	3653	3689
AT	3B	0.922 *(0.916)*	0.984 *(0.984)*	0.953	3956	4022
AT,HT,T	3B	0.931	0.980	0.956	3993	4076
AT,T	3B	0.929 *(0.922)*	0.982 *(0.982)*	0.956	3983	4058
HT	3B	0.713 *(0.715)*	0.990 *(0.991)*	0.851	3057	3089
R,AT,T	3B	0.929	0.982	0.956	3983	4058
R,AT,T,HT	3B	0.931	0.980	0.956	3993	4076
T	3B	0.872 *(0.869)*	0.989 *(0.989)*	0.930	3741	3781
AT	3C	0.926 *(0.922)*	0.980 *(0.980)*	0.953	3971	4052
AT,HT,T	3C	0.943	0.976	0.960	4046	4146
AT,T	3C	0.941 *(0.938)*	0.977 *(0.978)*	0.959	4035	4128
HT	3C	0.713 *(0.717)*	0.990 *(0.992)*	0.851	3057	3089
R,AT,T	3C	0.941	0.977	0.959	4035	4128
R,AT,T,HT	3C	0.943	0.976	0.960	4046	4146
T	3C	0.863 *(0.860)*	0.989 *(0.990)*	0.926	3701	3741

The results obtained by using preclustering cross validation are shown in parentheses (italics). Boldface numbers show the maximum postclustering value of Avg(Sn, Sp) for a given species.

tial and transitional probabilities of the three-periodic model, q_{ij} are transitional probabilities for the ordinary first order Markov model of noncoding region.

It was observed for several genomes (Table 9), that when the model P fits the tested sequence, such as a true gene sequence, the likelihood of misidentification of a protein-coding fragment correlates

Table 8. Gene Prediction Accuracy in Terms of Avg(Sn,Sp) Observed for the 10 Genomes

	Type of atypical model (cluster)					
Species name	**2A**	**2B**	**2C**	**3A**	**3B**	**3C**
A. fulgidus	0.953 *(0.952)*			0.953 *(0.954*		
B. subtilis	0.949 *(0.942)*	0.943 *(0.936)*	0.951 *(0.945)*	0.950 *(0.943)*	0.943 *(0.933)*	0.951 *(0.947)*
E. coli	0.956 *(0.956)*	0.953 *(0.950)*	0.959 *(0.958)*	0.957 *(0.956)*	0.956 *(0.952)*	0.959 *(0.058)*
H. influenzae	0.955 *(0.954)*	0.953 *(0.953)*	0.953 *(0.954)*	0.955 *(0.956)*	0.954 *(0.954)*	0.955 *(0.955)*
H. pylori	0.950 *(0.950)*	0.950 *(0.951)*	0.949 *(0.948)*	0.948 *(0.950)*	0.950 *(0.950)*	0.948 *(0.948)*
M. genitalium	0.924 *(0.931)*	0.927 *(0.931)*	0.924 *(0.927)*	0.925 *(0.929)*	0.928 *(0.932)*	0.922 *(0.925)*
M. jannaschii	0.974 *(0.974)*	0.974 *(0.974)*	0.974 *(0.974)*	0.975 *(0.973)*	0.974 *(0.974)*	0.975 *(0.972)*
M. pneumoniae	0.925 *(0.921)*	0.925 *(0.927)*	0.917 *(0.913)*	0.923 *(0.921)*	0.927 *(0.927)*	0.924 *(0.922)*
M. thermoautotrophicum	0.972 *(0.969)*			0.970 *(0.970)*		
Synechocystis	0.968 *(0.967)*	0.965 *(0.966)*	0.968 *(0.968)*	0.968 *(0.968)*	0.966 *(0.966)*	0.969 *(0.969)*

The gene-finding program GeneMark included models derived from typical and atypical clusters. The results obtained by using preclustering cross-validation are shown in parenetheses (in italics). All other data were obtained by using postclustering cross-validation. See Table 7 for explanation of boldface numbers.

negatively with the KL distance between the models P and Q as determined by equation 2.

It may happen, however, that the model P does not fit the sequence data. Under the assumption that the input sequence fits another model, P^*, it was shown (M. Borodovsky, unpubl.), that the misidentification error rate should correlate negatively with the value of the ''efficient'' KL distance determined by the formula

$$D(P\&P^*\|Q) = D(P^*\|Q) - D(P^*\|P) \qquad (3)$$

Here $D(P^*\|P)$ is defined as

$$D(P^*\|P) = \frac{1}{3}\sum_{l=1}^{3}\sum_{i,j=1}^{m} p_i^{*l}\, p_{ij}^{*l} \log \frac{p_{ij}^{*l}}{p_{ij}^{l}} \qquad (4)$$

Particularly, if the typical model P is used to identify a highly typical gene segment that actually fits the highly typical model P^*, the ''efficient'' KL distance, computed by equation 3 is larger than the KL distance between the models P and Q determined by

Table 9. Negative Correlation of KL distances and False-Negative Prediction Error Rates Produced by GeneMark in Short Fragment Identification

Model (order 3)	***B. subtilis***	***E. coli***	***M. genitalium***	***M. jannaschii***	***M. pneumoniae***
			False-negative error rate		
AT	0.186	0.223	0.171	0.118	0.165
T	0.098	0.083	0.093	0.044	0.078
HT	0.054	0.045	0.067	0.017	0.052
			KL distance (coding vs. noncoding model)		
AT	0.0648	0.0619	0.0837	0.0999	0.0883
T	0.0930	0.1102	0.0893	0.1299	0.1149
HT	0.1318	0.1687	0.1289	0.1909	0.1587
R =	−0.961	−0.931	−0.770	−0.896	−0.905

For five genomes the error rates are shown along with the KL distance between corresponding coding and noncoding models of the third order. The correlation coefficients (*bottom* row) for *A. fulgidus, H. influenzae, H. pylori, M. thermoautotrophicum,* and *Synechocystis,* averaged with regard to A-, B-, and C-type models, were equal to −0.97, −0.91, −0.96, −0.96, and −0.99, respectively.

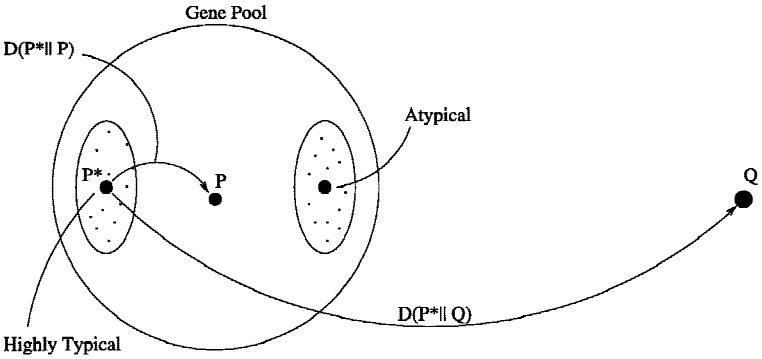

Figure 6 Schematic representation of the genomic sequence space with the KL distance metrics. The elements of the space are protein-coding and noncoding sequences.

equation 2. This leads to the expectation that the typical model will identify the function of a highly typical segment with a misidentification rate lower than one produced by the same model for a typical gene segment. This effect has indeed been observed.

The geometry of the KL distance space of long ORF sequences and noncoding sequences is illustrated in Figure 6. This figure emphasizes the idea that the set of long ORFs is inhomogeneous and should be represented rather by a cloud of points, contrary to the set of noncoding sequences that should be represented by one point. The noncoding point also represents the model of the noncoding region. The sequences from the atypical cluster, located within the cloud (big circle), are shown to be closer to the point representing the noncoding region than the sequences from highly typical cluster. The centers of the clusters (circles) represent the typical, atypical, and highly typical models. Note that the points are connected by curves reminding one that the KL distance is not a regular Euclidean distance and does not satisfy the triangle inequality.

A more elaborate picture, Figure 7, shows the cluster relationships in terms of KL distances for the typical, highly typical, and atypical clusters, type A, and class I, class II, and class III gene sets. The B-type atypical cluster was slightly closer to the other clusters and classes, with the exception of class III. The C-type atypical cluster was located further away from the other sets with the exception of class III.

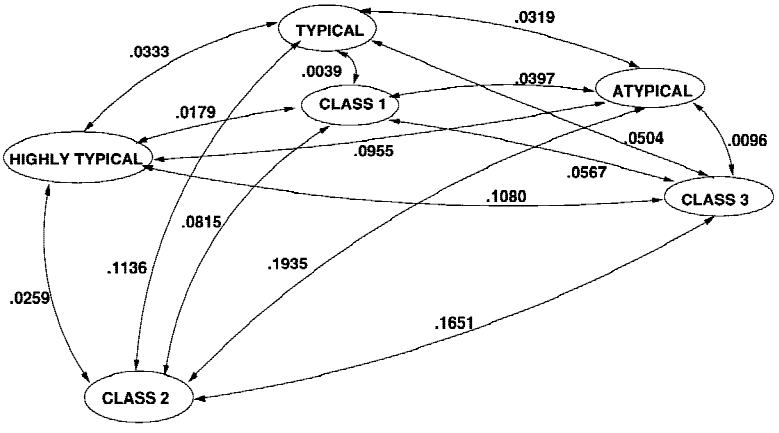

Figure 7 Schematic representation of actual KL distances, as defined by the order 3 Markov models, between the *E. coli* gene classes (Medigue et al. 1991) and ORF clusters obtained by three-means clustering.

New Gene Predictions by Atypical Models

The GeneMark.hmm analysis of the 10 genomes using 2A cluster derived typical and atypical models allowed the identification of 683 new putative genes, each longer than 96 nucleotides. These genes were identified as atypical model predictions, therefore, indicating the atypical model fits better to the composition of these sequences than the typical model. These findings deserve special attention because the genes with atypical composition are harder to detect and therefore thus predicted genes are good candidates for being horizontally transferred genes delivering special adaptive advantages for microorganisms that carry these genes. For 176 predictions, the gapped BLAST analysis (Altschul et al. 1997) corroborated the predictions with statistically significant similarity to known proteins in the NCBI nonredundant sequence database (Table 10). In a majority of cases, the similarity was with hypothetical proteins whereas in 11 cases the similarity was detected with proteins related to insertion elements. Full lists of the results can be found at ftp://genmark.biology.gatech.edu/pub/gmgen.ghmAT.blast ftp://genmark.biology.gatech.edu/pub/gmgen.ghmAT.best.

The differences in distribution of proteins tranlated from typical and atypical long ORFs into functional categories (Riley 1993) can be seen in Figure 8. The proteins translated from atypical ORFs were represented more strongly than the transposon, or plasmid categories. The relative abundance of atypical ORF-derived proteins was observed among hypothetical and unknown pro-

Table 10. New Atypical Gene Predictions for the 10 Genomes Made by the GeneMark.hmm program

Species name	Total	No. of GeneMark.hmm new predictions characterized as atypical [corroborated by gapped BLASP search ($P < 1e-5$)]
A. fulgidus	111	25
B. subtilis	86	14
E. coli	135	22
H.influenzae	44	24
H. pylori	16	5
M. genitalium	37	25
M. jannaschii	80	26
M. pneumoniae	20	14
M. thermoautotrophicum	39	4
Synechocystis	115	17

The GeneMark.hmm program (Lukashin and Borodovsky 1998) shows typical and atypical models in parallel. Numbers of new predictions, as compared with the GenBank records, are shown along with the numbers of cases when the gene prediction was corroborated by similarity search by gapped BLASTP (Altschul et. al. 1997).

teins as well as in cell structure proteins, putative enzymes, and putative regulatory proteins.

METHODS

Materials

To test the GeneMark-Genesis program, we used the complete genomic sequences of the following bacterial species: *Archaeoglobus fulgidus* (Klenk et al. 1997; GenBank accession no. AE000782), *B. subtilis* (Kunst et al. 1997; AL009126), *E. coli* (Blattner et al. 1997; U00096), *H. influenzae* (Fleischmann et al. 1995; L42023), *H. pylori* (Tomb et al. 1997; AE000511), *M. genitalium* (Fraser et al. 1995; L43967), *M. jannacshii* (Bult et al. 1996; L77117), *M. pneumoniae* (Himmelreich et al. 1996; U00089), *M. thermoautotrophicum* (Smith et al. 1997; AE000782), and *Synechocystis* PCC6803 (Kaneko et al. 1996; SYNECO). The size of each sequence, the number of annotated genes, and the average GC content are given in Table 11. The GeneMark-Genesis clustering results were compared with the results of earlier studies by Medigue et al. (1991) and by Lawrence and Ochman (1997). The updated versions of the three *E. coli* gene classes were kindly provided by A. Danchin (Institute Pasteur, Paris, France). Dr. J. Lawrence (University of Pittsburgh, PA) provided us with the DNA sequences of 756 *E. coli* genes, classified as horizontally transferred genes.

Generating the Root Model

The GeneMark-Genesis algorithm presented here makes substantial use of the GeneMark algorithm (Borodovsky and McIninch 1993) and can be viewed as an extension of GeneMark. GeneMark-Genesis does not require experimentally validated sets of training sequences to obtain the crucial set of parameters used in GeneMark, initial and transition probabilities of Markov models of coding and noncoding regions. The Markov model of DNA gene sequence is a machine-learned model that reflects the pattern of correlation between adjacent nucleotides, the pattern evolutionary developed under restrictions intrinsic to coding region (amino acid composition, codon usage pattern, etc.) or noncoding region (dinucleotide composition, etc.).

Given an anonymous bacterial sequence, the parameters of the initial model of protein-coding region were obtained from the set of ORFs identified in the sequence, such as one of the 10 complete genomic sequences with annotation assumed unknown. An ORF starts with a start codon and ends with the in frame stop codon. Whereas the ORF's 3′ end is unique, there is an ambiguity in the position of its 5′ end. This ambiguity is eliminated by specifying an ORF as the longest possible ORF (with the start codon being the furthest possible from the given stop codon).

As Figure 9 shows that an ORF >500 nucleotides found in the *B. subtilis* genome is unlikely to be a noncoding random ORF. This is also true for the nine other bacterial genomes. Therefore, by selecting ORFs >500 nucleotides in anonymous bacterial DNA sequence one has a high likelihood of selecting only true coding regions.

A note of caution should be made concerning the initial part of a long ORF. If the true start codon is located inside the ORF, the sequence upstream to the start codon is noncoding. The average length of this noncoding sequence is small, however. For instance, according to the GenBank annotation of the *B. subtilis* genome this length is equal to 18 nucleotides. Therefore, if the selected ORFs are long enough, the bias in estimated model parameters should be negligibly small. In this work, a threshold of 700 nucleotides was used. This choice made the relative amount of erroneously included noncoding region even smaller than the choice of a 500-nucleotide threshold. If two long ORFs overlapped by >30

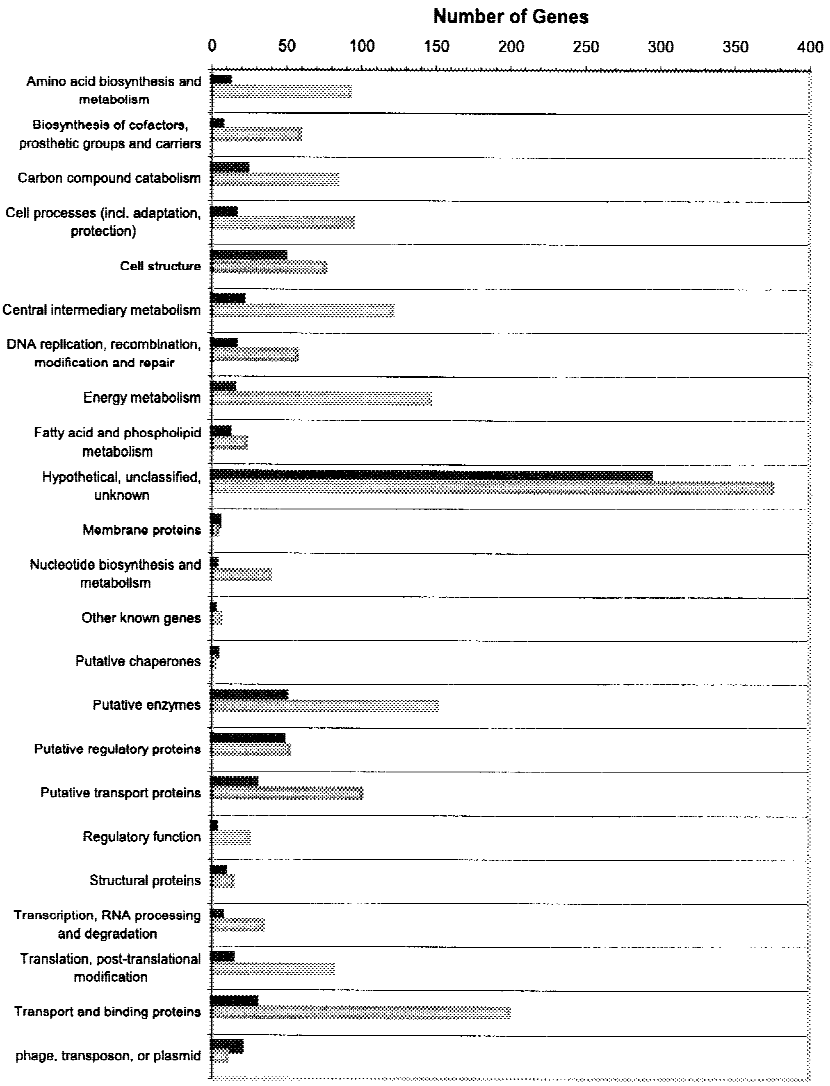

Figure 8 Distribution of functional categories of the *E. coli* genes in the typical and atypical (type A) ORF clusters obtained by two-means clustering. (solid bars) Atypical; (shaded bars) typical.

nucleotides, the longer ORF was kept in the training set. By combining a number of long ORFs detected in bacterial genomic DNA one could obtain an amount of presumably protein-coding sequence sufficient to derive Markov models of high order. For instance, it was estimated that 1029 Mb of combined protein-coding sequence is generally sufficient to derive accurately enough parameters for the three-periodic Markov model of order five (M. Borodovsky, unpubl.).

The set of long ORFs identified in a given complete genomic sequence was used to determine parameters of the Markov model of the protein-coding region, the ''pre-Root'' model of orders zero to five. For a pre-Root Markov model of a noncoding sequence, we used the zero order model with four probability parameters estimated by genome specific frequencies of mononucleotides. The pre-Root models were employed in the GeneMark program for predicting coding regions in the original genomic sequence. Predicted protein-coding regions were combined into a new training sequence. The rest of the sequence, considered noncoding, was included into the noncoding training set. The Markov models, coding and noncoding, whose parameters were defined using these training sets, were called the Root models. For comparison purposes, the GenBank annotation of a given genome was used to compile training sets of coding/noncoding sequence. The models defined by training on these sets were called the GenBank models.

Note that the Root models were generated as described above for complete genomic sequences of *M. jannaschii* and *H. pylori* and used in the GeneMark program to identify protein-coding regions in these genomes (Bult et al. 1996; Tomb et al. 1997). Recently Salzberg et al. (1998) have used sets of ORFs >500 nucleotides to derive the interpolated Markov models of bacterial protein-coding regions that were used in the GLIMMER gene-finding program.

Generating ORF Clusters

The general goal of clustering was to obtain several compositionally homogeneous training sets for class-specific gene models that would enhance the gene-finding ability of the GeneMark program. This goal dictated the choice of seed clusters as well as the choice of the distance definition in the space of ORF sequences. The *k*-means clustering algorithm, with *k* equal to two or three (see below), was employed:

1. Sort *n* objects (ORFs) into c clusters.
2. Compute $m_1 \ldots m_c$ cluster centers.
3. Classify the *n* objects by assigning them to the cluster with the closest center cluster.
4. If any object was assigned to a new cluster, go to 2; otherwise stop.

The initial step of the algorithm was to prepare the ORF cluster ''seeds''. The GeneMark program employing the Root models was used to score all long ORFs identified in a given genomic sequence. The long ORFs with a GeneMark score of <0.5 were assigned to the atypical seed cluster. All other long ORFs, in the two-means clustering procedure, were assigned to the typical seed cluster. In the case of the three-means clustering, the atypical seed cluster was selected as described above, then 15% of the long ORFs with the highest scores were assigned to the highly typical seed cluster, and all other ORFs were assigned to the typical seed cluster.

Each ORF was characterized by a vector of 61 codon frequencies. The ORF cluster center was defined by the vector of cumulative codon frequencies observed in all ORFs in the cluster. The distance between codon frequency vectors

Table 11. Information on Genomic Sequences Used in this Study

Species name	Accession no.	Genome length (nucleotides)	No. of annotated genes	GC (%)
A. fulgidus	AE000782	2,178,400	2407	48.6
B. subtilis	AL009126	4,214,814	4100	43.5
E. coli	U00096	4,639,221	4290	50.8
H. influenzae	L42023	1,830,419	1717	38.1
H. pylori	AE000511	1,667,877	1566	38.9
M. genitalium	L43967	580,073	476	31.7
M. jannaschii	L77117	1,664,987	1680	31.4
M. pneumoniae	U00089	816,394	678	40.0
M. thermoautotrophicum	AE000666	1,751,377	1839	49.5
Synechocystis	AB001339	3,573,470	3169	47.7

could be defined by a symmetrical Kullback–Liebler-type formula

$$D(x\|c) = \frac{1}{2}\sum\left(f_i \log\frac{f_i}{q_i} + q_i \log\frac{q_i}{f_i}\right) \quad (5)$$

Here x is an ORF's codon frequency vector, and c is a cluster's cumulative codon frequency vector. f_i, represents the frequency of the ith codon in the ORF, and q_i represents the frequency of the ith codon in the ORF cluster. In the actual formula used in the clustering procedure, the codon frequency values f_i (q_i) were normalized by the frequency of the encoded amino acid and multiplied by the size of the group of synonymous codons

$$f_i = \frac{C_i + 1}{T_{\text{codons}}} * \frac{N_{i\text{–syn}}}{f_{aa\text{–}i}} \quad (6)$$

Here, T_{codons} is the count of all codons for a particular ORF or cluster. $N_{i\text{-syn}}$ is the number of synonymous codons for the ith codon. $f_{aa\text{-}i}$ is the frequency of the amino acid in an ORF, or ORF cluster, corresponding to the ith codon. C_i is the count of all codons of the ith type. To ward off the zero frequency numerical problem, the codon counts were initialized to one. Substitution of defined frequencies into formula (5) transformed it into a weighted sum of symmetricized KL-distances $D_k(x||c)$ defined for each group of synonymous codons.

$$D(x\|c) = \Sigma\, N_{k\text{–syn}}\, D_k(x\|c) \quad (7)$$

The weights $N_{i\text{-syn}}$ were equal to the size of each synonymous group, and $D_k(x||c)$ was defined as

$$D_k\,(x\|c) = \frac{1}{2}\sum\left(u_i^k \log\frac{u_i^k}{v_i^k} + v_i^k \log\frac{v_i^k}{u_i^k}\right) \quad (8)$$

where u^{ki} and v^{ki} are relative codon usage frequencies in the kth group of synonymous codons. Note, that in formula 7 the terms relating to the methionine and tryptophan amino acids have disappeared.

The final clusters of ORFs to which the k-means clustering algorithm converged were used as training sets for the specific Markov models of protein-coding regions. The regions that were predicted as coding by at least one cluster specific model were eliminated from the set of noncoding sequences used for training the Root model. The sequence left was used for training the Markov model of noncoding sequence. The differences in noncoding models obtained by slightly different clustering processes were not found to be critical for the gene prediction results. Therefore, in what follows, we do not discuss the minor differences between the models of noncoding sequence.

The training set for the atypical gene model could be obtained in several ways. As was indicated earlier, the set of horizontally transferred *E. coli* genes constitutes about 15% of the whole gene pool (Medigue et al. 1991; Lawrence and Ochman 1997). The clustering method described above, however, when applied to a complete bacterial genomic sequence produced the atypical gene cluster, cluster A as it is called below, which usually consisted of ~30% of the total number of genes. We modified the atypical gene cluster content to make its size closer to the size of the horizontally transferred gene set as was estimated in previous studies. For instance, only 15% of all long ORFs that were closest to the center of

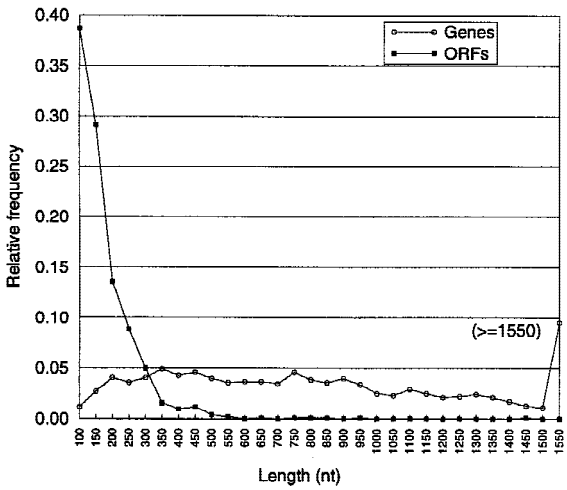

Figure 9 Length distribution for *B. subtilis* annotated genes and random ORFs. Data from the University of Wisconsin, Madison (http://www.genetics.wisc.edu).

the cluster produced by method A, were included into another version of the atypical cluster, cluster B. All other ORFs were included into the typical cluster. In yet another modification, only 15% of all of the long ORFs, those produced by method A that were furthest from the typical cluster center were assigned to the atypical cluster, cluster C.

Note that a different approach to obtain homogeneous sets of protein coding regions and eventually to derive class-specific GeneMark models for anonymous bacterial genomic sequences was suggested by Hirosawa et al. (1997). In this approach, tested on the *Synechocystis* PCC6803 genome, iterative runs of GeneMark were used instead of iterative *k*-means clustering. Recently, clustering of *Arabidopsis thaliana* genes was performed (Peresetsky et al. 1998). In this project, several types of gene sequence characterization were used—by parameters of the Markov models of zero and first order as well as by codon usage vector. The variant of *k*-means clusterization procedure employed as a distance function the Euclidean distance between parameter vectors.

Prediction Accuracy Evaluation using Sets of Short DNA Fragments

The GeneMark program prediction accuracy was characterized by the false-negative and false-positive error rates observed in identification of short DNA fragments, either coding or noncoding. To obtain the test set of presumably coding short sequences, the long ORFs from a designated set were cut into a number, N_c, of nonoverlapping 96-nucleotide sequences. Each individual fragment was then identified as coding or noncoding. If the output GeneMark score was lower than the threshold, 0.5, the fragment was interpreted as noncoding and, therefore, misidentified. The total number of false-negative predictions, FN, was recorded and the ratio FN/N_c defined the observed false-negative error rate. Note that the uncertainty in the position of the true start codon in a long ORF might slightly elevate the false-negative error rate as a few true noncoding fragments could be generated by chopping a long ORF. To define the false-positive error rate, presumably noncoding sequences were cut into a number, N_n, of short 96-nucleotide sequences. Now if the output GeneMark score for a given fragment was higher than the threshold, 0.5, the fragment was interpreted as protein-coding, therefore misidentified. The observed false-positive error rate was defined as the ratio FP/N_n, where FP was the total number of false-positive predictions. When a Markov model of protein-coding regions was tested on protein-coding DNA sequence from the same gene class as one the model was derived from, the above accuracy evaluation procedure was used in cross-validation mode.

Sensitivity and Specificity

The sensitivity, Sn, and specificity, Sp, parameters, formulas 9 and 10, were used to characterize the accuracy of the models and the algorithm in terms of whole gene prediction. Here, as a control, we used the GenBank annotation, assuming it is correct on the whole gene location but might be off in terms of the start codon position. Sn is defined as the ratio of the number of correctly located genes to the number of genes annotated in the GenBank record for this sequence. Sp is the ratio of the number of correctly located genes to the total number of genes predicted by the algorithm in the given sequence.

$$\mathrm{Sn} = \frac{\mathrm{no.\_of\_correct\_predictions}}{\mathrm{total\_no.\_of\_annotated\_genes}} \tag{9}$$

$$\mathrm{Sp} = \frac{\mathrm{no.\_of\_correct\_predictions}}{\mathrm{total\_no.\_of\_predictions}} \tag{10}$$

When the performance of a combination of models was evaluated, the GeneMark program was used to analyze a given sequence once for each model being considered. An ORF was counted as a predicted gene if the ORF was predicted as a coding one by at least one model. If the same ORF was predicted as protein-coding one by several models, however, these coinciding predictions were considered as one item for counting the total numbers of predictions.

Postclustering and Preclustering Accuracy Evaluation Procedures

When the basic cross validation procedure was used for an accuracy evaluation of the Markov model of a protein-coding region, this model was derived from 6/7 of a particular long ORF cluster and the other 1/7 of the ORF cluster was used as a test set. Such an evaluation was done 7 times by testing each 1/7 of the ORF cluster with the model trained on the other 6/7 part of the cluster. The resulting false-negative rate was an average of the seven observed rates. Notably, the accuracy evaluation using the whole set of long ORFs was performed posterior to the clustering of long ORFs identified in a given sequence. This procedure was called postclustering cross-validation.

Additional accuracy evaluation was made to make sure that clustering does not introduce a systematic bias into the model derivation. To perform this preclustering cross-validation procedure, the initial sequence was divided into seven parts. The 6/7 of the whole sequence was used for the *k*-means clustering (see above) and deriving models. The 1/7 of the sequence was used to test the models on the long ORFs assigned to specific, typical or atypical, ORF clusters by the initial *k*-means clustering (with the same parameter k) that had been performed initially with the whole sequence. The resulting false negative error rate was obtained as an average of the seven observed outcomes. Both the postclustering and the preclustering cross-validation procedures were performed only for testing the particular model on the sequence cluster with the same name (i.e., typical model vs. typical ORFs, etc.). The postclustering and the preclustering Markov models were also assessed using the Sn and Sp parameters.

Gene Prediction by Atypical Model

Contrary to using different models in GeneMark in a consecutive manner, the GeneMark.hmm program uses the models, particularly typical and atypical, in parallel; therefore, for each identified gene the program indicated the model that fits best in terms of maximum likelihood to the gene sequence (Lukashin and Borodovsky 1998). Genes predicted by the atypical model were of special interest since these genes were potentially horizontally transferred genes as well as genes that are difficult to spot by the typical model. We analyzed all 10 bacterial genomes using the GeneMark.hmm program, with the models derived from A-type clusters obtained by *k*-means ($k = 2$) clustering. The regions predicted as genes by the atypi-

cal model, >96 nucleotides, were selected, translated, and searched against the nonredundant protein sequence database supported by NCBI using the gapped BLAST program (Altschul et al. 1997). The BLAST hits with *P* values smaller than $1e^{-5}$ were considered as statistically significant evidence for nonrandom similarity to an earlier discovered protein. These findings indicated that the current predictions are most likely functional genes rather than the false positives.

ACKNOWLEDGMENTS

We thank Antoine Danchin, Ashok Kolaskar, Owen White, Jean-Francois Tomb, and Jefferey Lawrence for helpful discussions. We are grateful to James McIninch and John Besemer for discussions and valuable software programming assistance. This work has been supported in part by the National Institutes of Health.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ''advertisement'' in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., T.L. Madden, A.A. Schdffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

Blattner, F.R., G. Plunkett III, C.A. Bloch, N.T. Perna, V. Burland, M. Riley, J. Collado-Vides, J.D. Glassner, C.K. Rode, G.F. Mayhew et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277: 1453–1462.

Borodovsky, M. and J.D. McIninch. 1993. GeneMark: Parallel gene recognition for both DNA strands. *Comp. Chem.* 17: 123–133.

Borodovsky, M.Yu., A. Sprizhitsky, E.I. Golovanov, and A.A. Alexandrov. 1986a. Statistical features in the *Escherichia coli* genome functional primary structure. II. Non-homogeneous Markov chains. *Mol. Biol.* 20: 833–840.

———. 1986b. Statistical features in the *E. coli* genome functional primary structure. III. Computer recognition of protein coding regions. *Mol. Biol.* 20: 1140–1150.

Borodovsky, M. Yu., J. McIninch, E. Koonin, K. Rudd, C. Medigue, and A. Danchin. 1995. Detection of new genes in the bacterial genome using Markov models for three gene classes. *Nucleic Acids Res.* 23: 3554–3562.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, and J.A. Blake. 1996. Complete genome sequence of the methanogenic archeon, *Methanococcus jannaschii*. *Science* 273: 1058–1073.

Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268: 78–94.

Cover, T.M. and J.A. Thomas. 1991. *Elements of information theory*. John Wiley & Sons, Inc., New York, NY.

Fickett, J. 1982. Recognition of protein-coding regions in DNA sequences. *Nucleic Acids Res.* 10: 5303–5318.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J. Tomb, B.A. Dougherty, J.M. Merrick et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496–512.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelly et al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397–403.

Gribskov, M., J. Devereux, and R.R. Burgess. 1984. The codon preference plot: Graphic analysis of protein coding sequences and prediction of gene expression. *Nucleic Acids Res.* 12: 539–549.

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B.-C. Li, and R. Herrmann, 1996. Complete sequence analysis of the genome of the bacterium *Mycomplasma pneumoniae. Nucleic Acids Res.* 24: 4420–4449.

Hirosawa, M., K. Isono, W. Hayes, and M. Borodovsky. 1997. Gene identification and classification in the *Synechocystis* genomic sequence by recursive GeneMark analysis. *DNA Sequence* 8: 17–29.

Kaneko, K., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sigiura, S. Sasamoto et al. 1996. Sequence analysis of the genome of the unicellular cyanobacterium Synechosistis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 3: 109–136.

Karlin, S., J. Mrazek, and A.M. Campbell. 1998. Codon usages in different gene classes of the Escherichia coli genome. *Mol. Microbiol.* 29: 1341–1355.

Klenk, H.P., R.A. Clayton, J.F. Tomb, O. White, K.E. Nelson, K.A. Ketchum, R.J. Dodson, M. Gwinn, E.K. Hickey, J.D. Peterson et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeglobus fulgidus. *Nature* 390: 364–370.

Krogh, A. 1997. Two methods of improving performance of an HMM and their application for gene finding. *Proc. ISMB-1997* 5: 179–186.

Kunst, F., N. Ogasawara, I. Moszer, A.M., Albertini, G. Alloni, V. Azevedo, M.G. Berteron, P. Bessieres, A. Bolofin, S. Borchert et al. 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis. Science* 390: 249–256.

Lawrence, J.G. and H. Ochman. 1997. Amelioration of bacterial genomes: Rates of change and exchange. *J. Mol. Evol.* 44: 383–397.

———. 1998. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci.* (in press).

Lukashin, A.V. and M. Borodovsky. 1998. GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Res.* 26: 1107–1115.

McIninch, J.D., W.S. Hayes, and M. Borodovsky. 1996. Applications of GeneMark in multispecies environment. *Proc. ISMB-1996* 4: 165–175.

Medigue, C., T. Rouxel, A. Vigier, A. Henaut, and A. Danchin. 1991. Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* 222: 851–856.

Peresetsky, A., C. Mathe, P. Dehais, M. Van Montagu, and P. Rouze. 1998. Classification of *Arabidopsis thaliana* gene sequences: Coding sequences clustering into two groups according to codon usage. *J. Mol. Biol.* (in press).

Riley, M. 1993. Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* 57: 862–952.

Robison, K., W. Gilbert, and G. Church. 1994. Large-scale bacterial gene discovery by similarity search. *Nature Genet.* 7: 205–214.

Salzberg, S.L. A.L. Delcher, S. Kasif, and O. White. 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 26: 544–548.

Smith, D.R., L.A. Doucette-Stamm, C. Deloughery, H.-M. Lee, J. Dubois, T. Aldredge, R. Bashirzadeh, D. Blakely, R. Cook, K. Gilbert et al.1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* delta H: Functional analysis. Comparative genomics. *J. Bacteriol.* 179: 7135–7155.

Staden, R. 1984. Measurements of the effect that coding for a protein has on DNA sequence and their use for finding genes. *Nucleic Acids Res.* 12: 551–567.

Tomb, J-F., O. White., A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann, K.A. Ketchum, H.P. Klenk, S. Gill, B.A. Dougherty et al. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* 388: 539–547.

Xu, Y., J.R. Einstein, R.J. Mural, M. Shah, and E.C. Uberbacher. 1994. Constructing gene models from accurately predicted exons: An application of dynamic programming. *Comp. Appl. Biosci.* 10: 613–623.

Received August 17, 1998; accepted in revised form October 23, 1998.