3554–3562 Nucleic Acids Research, 1995, Vol. 23, No. 17

Detection of new genes in a bacterial genome using Markov models for three gene classes

Mark Borodovsky*, James D. McIninch, Eugene V. Koonin[1], Kenneth E. Rudd[1], Claudine Médigue[2,3] and Antoine Danchin[3]

School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA, [1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [2]Institute Curie, 11 rue Pierre et Marie Curie, 75231 Paris Cedex 05, France and [3]Institute Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

Received April 3, 1995; Revised and Accepted July 27, 1995

ABSTRACT

We further investigated the statistical features of the three classes of *Escherichia coli* genes that have been previously delineated by factorial correspondence analysis and dynamic clustering methods. A phased Markov model for a nucleotide sequence of each gene class was developed and employed for gene prediction using the GeneMark program. The protein-coding region prediction accuracy was determined for class-specific Markov models of different orders when the programs implementing these models were applied to gene sequences from the same or other classes. It is shown that at least two training sets and two program versions derived for different classes of *E.coli* genes are necessary in order to achieve a high accuracy of coding region prediction for uncharacterized sequences. Some annotated *E.coli* genes from Class I and Class III are shown to be spurious, whereas many open reading frames (ORFs) that have not been annotated in GenBank as genes are predicted to encode proteins. The amino acid sequences of the putative products of these ORFs initially did not show similarity to already known proteins. However, conserved regions have been identified in several of them by screening the latest entries in protein sequence databases and applying methods for motif search, while some other of these new genes have been identified in independent experiments.

INTRODUCTION

Recent progress in *Escherichia coli* genome sequencing has made possible a more precise characterization and classification of *E.coli* genes and proteins (1–4). In 1991 Médigue and co-workers explored the set of *E.coli* genes using factorial correspondence analysis and dynamic clustering methods (5). This statistical analysis suggested the division of *E.coli* gene sequences into three classes that differ not only in the statistical but in the biological sense as well. Class I genes, with intermediate codon usage bias, maintain a low or intermediate level of expression, although some genes may occasionally be expressed at a very high level in environmentally triggered (rare) conditions. Class II genes, which have a high codon usage bias, are highly expressed under exponential growth conditions. Genes from Class III, with low codon usage bias, mainly belong to plasmids and insertion sequences; this class also includes genes coding for fimbriae, major pili, many membrane proteins, restriction endonucleases and lambdoid phage lysogeny control proteins. Many Class III genes can be expressed at a fairly high level, but their weakly biased codon usage pattern does not reflect the proportions in the distribution of *E.coli* tRNAs under exponential growth conditons (5–8).

The results obtained by Médigue *et al.* (5) show that there is no single variable, like codon adaptation index (9), that would unambiguously indicate to which class a given *E.coli* gene sequence belongs. At least two variables are necessary for this purpose (see Fig. 1 in ref. 5).

In this work, we explore the statistical patterns existing in *E.coli* gene sequences by incorporating gene classification into a broader context of the gene identification problem (see ref. 10 for review). We define a model of a gene sequence for each gene class as an artificial nucleotide sequence with a specific oligonucleotide composition generated by a phased inhomogeneous Markov model.

When these models, together with homogeneous Markov models for non-coding sequences, were used in the GeneMark gene identification program, the accuracy of coding potential prediction by GeneMark was high enough to identify several 'genes' annotated in GenBank as spurious and to predict a number of new genes in unannotated sequences. These findings explain the apparent high error rates previously observed with GeneMark.

MATERIALS AND METHODS

Sequence data

For deriving gene class-specific Markov models, we used *E.coli* DNA sequences of 812 genes from Class I, 281 genes from Class

* To whom correspondence should be addressed

II and 158 genes from Class III (a total length of 731 550, 232 431 and 102 003 nt, respectively) that had been defined by Médigue *et al.* using clustering of the codon frequency vectors in a 61-dimensional space (11–13). The list of gene names and the nucleotide sequences can be retrieved by anonymous FTP from the directory /pub/genemark/ecoli3 at amber.biology.gatech.edu.

For testing the effects of various parameters of the program on gene identification, the sequences were divided into sets of non-overlapping fragments of identical length. As shown previously (14), the reasonable range of window lengths for GeneMark analysis is between 48 and 144 nt. The accuracy of discriminating between coding and non-coding DNA fragments does not sharply depend on their length; therefore we present here the results obtained with 96 nt fragments (96 nt is the default window length used in the GeneMark e-mail server (14). Three sets of protein-coding 96 nt fragments, designated as Cod_i, i = 1,2,3 were compiled from gene sequences of the three classes.

A set of apparently non-coding regions was compiled from the non-redundant *E.coli* sequence database EcoSeq6 (15) by excluding annotated coding regions. All unannotated sequences longer than 100 nt were pooled into a non-coding set, 359 279 nt in length, and subsequently divided into two non-overlapping data sets designated Set I (193 578 nt) and Set II (165 701 nt). The sequences were assigned randomly to Set I or Set II; therefore, statistical properties of these sets were assumed to be identical. Similar to the protein-coding case, we compiled two sets of 96 nt non-coding fragments designated NonCod_i, i = 1,2. Thus, the samples of 96 nt fragments, Cod_i, i = 1,2,3 and NonCod_i, i = 1,2, used for testing GeneMark performance were derived from three non-overlapping sets of genes and two non-overlapping sets of non-coding regions employed for GeneMark training. The accuracy of GeneMark program for various combinations of training and testing sets was assessed (see results below).

The GeneMark method

The GeneMark algorithm has been described in detail previously (14,16,17). The algorithm has been designed to distinguish between three types of DNA sequences, namely: (i) protein-coding sequence (gene); (ii) non-coding sequence that is the complement of a coding sequence (gene shadow); or (iii) non-coding sequence whose complement also is non-coding. Inhomogeneous, phased Markov models are used to describe genes and gene shadows and ordinary Markov models are used for non-coding sequences (18–21). Given the sequence of a fragment S, an *a posteriori* probability for S to belong to one of the above three categories is, according to the Bayes theorem,

$$\mathrm{P(model|sequence)} = \frac{\mathrm{P(sequence|model)P(model)}}{\Sigma\ \mathrm{P(sequence|model)P(model)},} \quad (1)$$

where P(model) is the *a priori* probability of one of the above sequence categories. The formula (1) is used to determine the probability values p_i, i = 1,...,7, ($\Sigma p_i = 1$) for S coding for a protein in each of the six possible reading frames or for S being a non-coding region (14). If one of p_i, i = 1,...,6 is >0.5, the fragment S or its complement is identified as a protein-coding region in the respective reading frame. If $p_7 > 0.5$ or if none of p_i, i = 1,...,6 is >0.5, S and its complement are predicted to be non-coding. Note that the decision rule required that any one of the p_i, i = 1,...,6 was greater than the threshold, rather then some combination of these values, as no significant overlap between genes in the same strand or in complementary strands has so far been detected in *E.coli* (16,17 and unpublished observations) and accordingly, Formula 1 treats events of protein coding in the six possible frames as mutually exclusive.

Accuracy of gene prediction

The parameters of class-specific phased Markov models of different orders (from 0–6) were determined for the three classes of *E.coli* genes, as well as for the three respective sets of gene shadows. The parameters of an ordinary Markov model for a non-coding region (up to the sixth order) were determined from the Set_1 of *E.coli* non-coding sequences.

A version of the GeneMark program is defined by three models, namely those for a gene, a gene shadow and a non-coding region; in each case, three models of the same order were used. We designate different GeneMark versions as GMd_ECOi, where d refers to the order of the models, d = 0,1,2...,6 and i refers to the gene class used for training. The accuracy of prediction as a function of the model parameters was analyzed at a single step of the algorithm dealing with an isolated DNA fragment of a given length. The false negative error rates were determined for each program version using control sets of true coding fragments Cod_i, i = 1,2,3, which, therefore, include the set of fragments derived from the original training set of genes. The error rate obtained for such a 'pseudo-control' set gives a convenient reference point. The false positive error rates were determined using control sets of non-coding fragments NonCod_j, j = 1,2. Upon evaluation of the performance of several program versions, two versions, namely GM5_ECO1 and GM4_ECO3, were chosen and used to score known *E.coli* genes as well as unannotated ORFs (>101 nt) found in intergenic regions.

The GeneMark score for a gene (or ORF) is computed as the average value of *a posteriori* probabilities for each of the 96 nt window covering the gene (ORF) sequence with a step of 12 nt. This choice of scoring function was shown to be robust with respect to variations in the step length. For the classifiers GM5_ECO1 and GM4_ECO3, the score distributions were obtained for known genes and apparently non-coding ORFs. The score threshold for identifying an ORF as a gene was set at 0.4 for each of these two classifiers (see below). If both scores are <0.4, the ORF in question is identified as non-coding.

Database searches for sequence similarity

The ORFs identified by GeneMark were translated into proteins and the resulting amino acid sequences were used to screen the non-redundant protein and nucleotide databases at the National Center for Biotechnology Information (NCBI, NIH) using the programs BLASTP and TBLASTN, respectively (22,23). The SEG program was used to filter the query sequences to remove low complexity (compositionally biased) segments that produce spurious results in database searches (24). The results produced by the BLAST searches were screened for conserved motifs using the programs BLA (25), CAP and MoST (26).

RESULTS AND DISCUSSION

False negative error rates: some gene models work for another gene class better than for their own

Figure 1a–c shows the false negative rates for the GeneMark versions trained on Class I, II and III, respectively, as a function of

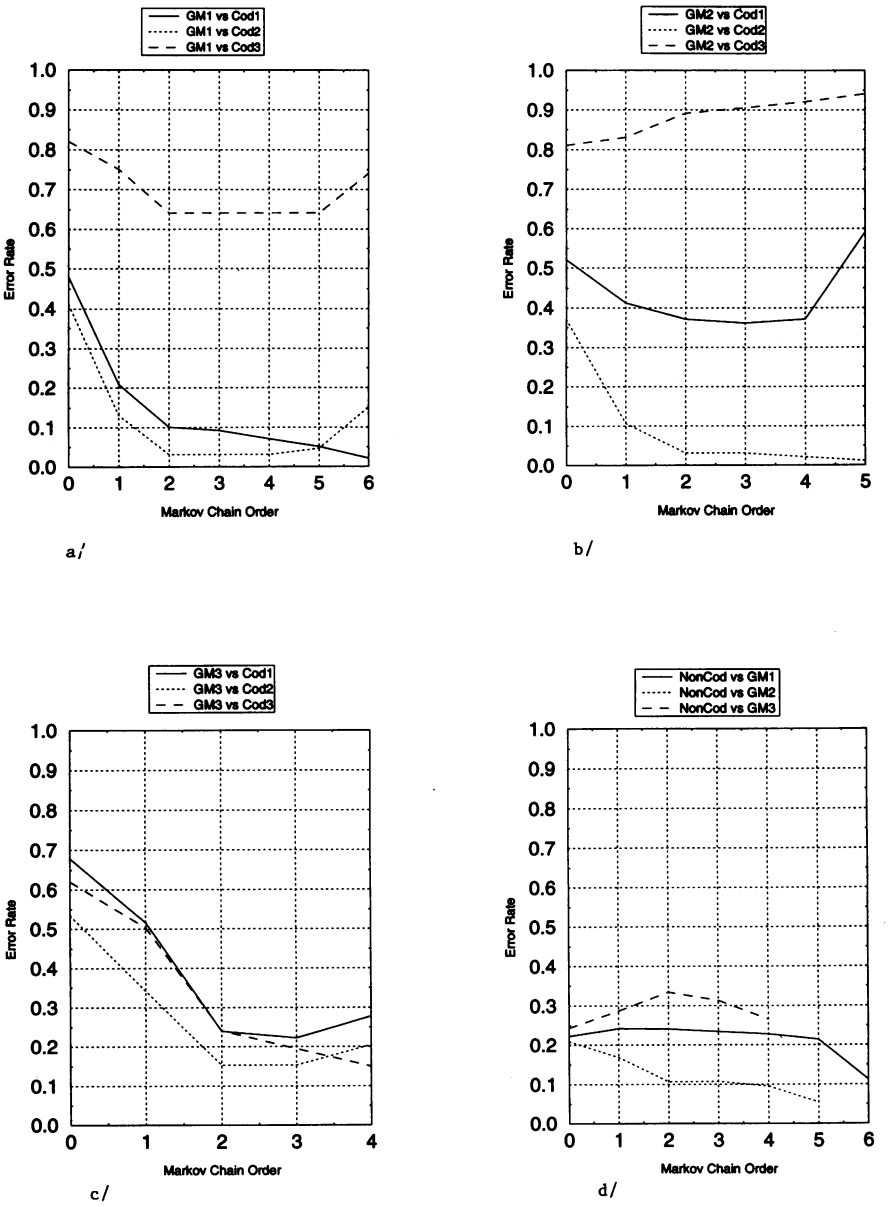

Figure 1. False negative error rates produced by GeneMark programs derived from the three gene classes (in groups related to one and the same training set). (**a**) The error rates computed by GMd_ECO1 programs of different Markov model orders d = 0,1,2,...,6. Three curves correspond to control sets Cod_i , i = 1,2,3, representing the three *E.coli* gene classes. The models derived from the Class I gene failed to recognize gene fragments from Class III but work for Class II genes even better than for the initial training set. (**b**) As in a for GMd_ECO2 programs, d = 0,1,...5. The models derived from Class II genes are good for identification of Class II genes only. (**c**) As in a for GMd_ECO3 programs, d = 0,1,2,...,4. The programs using the Class III gene models of orders 2 and 3 are satisfactory predictors for genes from all classes. Fragments of Class II genes are again predicted even better than fragments from the initial training set. (**d**) The false positive error rates computed by GMd_ECOi programs, i = 1,2,3; d = 0,1,...,max(i), for control set NonCod_2. Three curves correspond to the three training sets of genes. These results are similar to those obtained for fragments from the pseudo-control set NonCod_1 (data not shown).

the Markov model order. Each program version was applied to three control sets Cod_i, i = 1,2,3. Typically, the error rate decreases sharply when the model order increases from 0 to 2, after which it continues to go down slowly. The lowest error rate usually corresponds to the second highest model order as discussed below.

One may expect a model trained on a certain class of genes to be most accurate when applied to a pseudo-control set that contains objects from the original training set. However, the program trained on Class I genes misidentified Class II gene fragments in <4% of the cases for model orders 2–5; these rates are lower than those obtained for the pseudo-control set (Fig. 1a). This observation suggests that programs trained on Class I genes can be used with equally high accuracy to identify both Class I and II genes comprising almost all 'native' *E.coli* genes. As shown in Figure 1a the GM5_ECO1 program correctly identifies 96% of the 96 nt sequence fragments from the 'native' *E.coli* (Class I and Class II) genes.

The program versions trained on Class II genes failed to identify gene fragments from Class III (Fig. 1b). The fragments from Class I genes were recognized poorly as well.

If the program is trained on Class III genes, then (for orders ≤3) gene fragments from Class I are identified with about the same accuracy as gene fragments from Class III (pseudo-control set), yielding 22 and 20% error rates for the orders 2 and 3, respectively (Fig. 1c). The fragments from Class II are again recognized with a better accuracy, namely with an error rate of only 15%, than fragments from the pseudo-control set derived from Class III genes.

Obviously, Class II gene fragments are easy targets for identification by GeneMark program trained on any other set of *E.coli* genes. The Class II genes, however, are not a good choice for a training set for GeneMark (Fig. 1b). In contrast, Class III genes are a difficult target for recognition by the programs trained on other gene classes (the error rates are shown by the dashed curves in Fig. 1a and b) but the program trained on Class III genes performs satisfactorily for genes from other classes (Fig. 1c).

Judging from Figure 1 (a–c), the lowest chance for a true *E.coli* gene fragment to be identified as non-coding is when both GM5_ECO1 and GM4_ECO3 are applied, which is indeed the strategy used in the *E.coli* genome sequencing project (G. Plunkett III, pers. comm.).

A noticeable increase in the error rate is observed in interclass comparisons for the Markov model highest orders. This increase relates to the difference in oligonucleotide composition between different gene classes that is most pronounced for longer oligonucleotides (i.e. higher order models).

It has to be indicated that the steady decrease of the error rate observed for pseudo-control sets does not necessarily mean that the models are getting better and better. For the highest model orders, this tendency appear to be an artifact known as 'overfitting', which is observed when a model is trained and tested at the same set of objects. In such a case, the parameters (transition probabilities) that are determined for the given training set fit ideally the test objects (sequences), while the model would not necessarily fit the objects of the same class that were not included in the training set. Indeed, an independent analysis of the GeneMark accuracy using cross-validation (J. Kleffe, K. Hermann and M.B., unpublished) has shown that the accuracy of each GeneMark program, when it is trained and tested on the same gene class, slowly deteriorates starting from a certain Markov model order, namely, the fifth order for Class I and the fourth order for Class II and Class III. Obviously, the larger the size of the training set the higher the Markov model order that gives the best prediction accuracy.

Surprisingly, the programs of order 2, 3 and 4 trained on Class III genes outperform the same order programs trained on Class II genes for Cod_1 test set (Fig. 1b and c). This observation suggests that the statistical pattern of the highly expressed Class II genes that presumably have evolved from the homogeneous pool of ancestral *E.coli* genes has less in common with 'native' Class I gene sequences than the latter have with the pattern typical of horizontally transferred Class III genes. One may speculate that in the course of evolution, the horizontally transferred Class III genes have converged to the *E.coli* 'native' pattern, whereas Class II sequences have diverged significantly from the ancestral pattern due to the selection for the elevated level of expression.

False positive error rates. What is the actual false positive rate?

As indicated in Materials and Methods, the set of non-coding sequences was compiled from unannotated intergenic regions. This approach obviously is imperfect as a robust set of experimentally validated non-coding regions is needed to define a reliable set for program training as well as for testing and determining the false positive error rate. The difficulty with compiling a sufficiently large set of such sequences prompted the use of the poorly characterized set of unannotated regions which serendipitously resulted in interesting findings.

As shown in Figure 1d, the apparent false positive error rates are quite high, from 10 to 24%, for each program version regardless of the training set and the model order. The significant observation is that, as shown in Figure 1d, false positive error rate computed for the programs trained on Class I sequences increases with the increase of the model order from 0 to 2. This tendency contrasts the decrease of the false negative rates observed with the same programs (Fig. 1a–c). Such a decrease is expected since the higher order models (but not the highest considered in the experiment) should better describe statistical patterns of DNA sequences and should produce lower error rates. This controversy suggested a re-examination of the control sets of presumably non-coding regions and subsequently evidence was obtained that these sequences include unannotated protein-coding regions (discussed below in more detail). The presence of unnoticed coding regions may also explain the clear decrease in the false positive error rates observed for the highest model orders (Fig. 1d). This decrease may be accounted for by the tendency of the highest order programs to classify an increasing fraction of true coding sequences which do not belong to the training set, as non-coding (Fig. 1a–c). Thus, 'false positive' errors due to the presence of unannotated coding sequences will be increasingly suppressed by the high order models.

We attempted to evaluate the lower bound of the false positive error rate under the assumption that all ORFs in unannotated regions predicted by GeneMark as coding are true genes. In this case, the predicted coding sequences should be excluded from training and test sets of non-coding regions. When such a reduced set of non-coding sequences was used, the false negative error rates did not change noticeably, whereas the false positive rate dropped down to 1% for the GM5_ECO1 program and to 5% for the GM4_ECO3 program (data not shown). In order to reliably determine the actual false positive error rate, a systematic experimental testing of the GeneMark predictions is required.

Prediction of complete protein coding regions in *E.coli* DNA sequences

The results of the accuracy testing allowed us to choose the GeneMark versions GM5_ECO1 and GM4_ECO3 as the two complementary versions performing best for the whole set of the *E.coli* sequence data. The above discussion pertains to a single step of the algorithm. This case is rigorously tractable mathematically and the threshold of 0.5 is naturally used. For the case of the sliding window technique, when GeneMark probability functions are calculated at each step and their average (the GeneMark score) is used for prediction, an adequate threshold has to be determined from simulations.

Unannotated ORFs (longer than 101 nt) found in intergenic regions were extracted from 461 non-overlapping contigs comprising the EcoSeq6 database. Figure 2 shows that the distribution of gene lengths and the distribution of lengths of unannotated ORFs overlap significantly in the range from 100 to 450 nt. Thus, statistical analysis of ORFs in this length interval is particularly important.

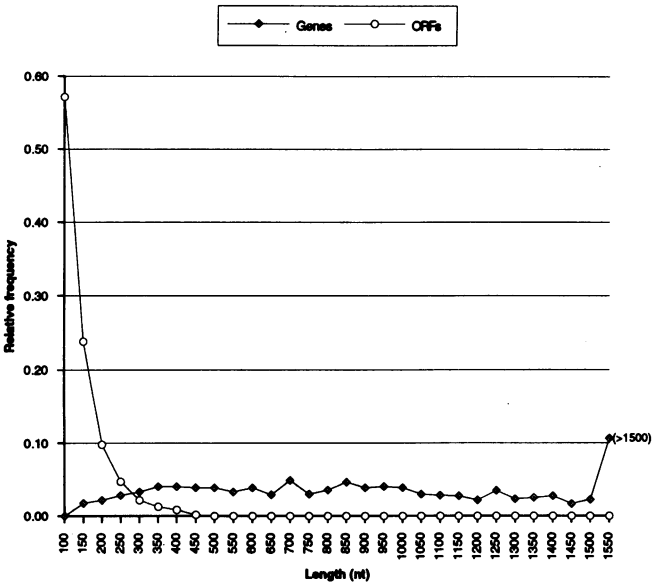

Figure 2. Length distribution for 1305 genes annotated in the EcoSeq6 (◆) and for 3272 ORFs (>101 nt) with no annotation in the EcoSeq6 (○).

The GeneMark scores were calculated (see Materials and Methods) for the complete genes from the three classes and for the set of unannotated ORFs using the two programs mentioned above. Inspection of the score distributions suggests the use of a threshold of 0.4 to identify expressed ORFs. This threshold, if GM5_ECO1 program is used, allows one to identify Class I and Class II genes with false negative errors rates of 0.1 and 0.9%, respectively. In contrast, this program has a high error rate in identifying Class III genes: 44.6% (Fig. 3a). The program GM4_ECO3 has higher error rates for Class I and II genes (2.4 and 2.7%, respectively) but a much lower error rate of 11.9% for Class III genes (Fig. 3b). When the two classifiers are used in parallel, an ORF is considered as non-coding if both scores are < 0.4. It can be shown, assuming that Class III genes constitute 1/5 of *E.coli* genes, that the false negative rate of such a combined strategy is not >3.1%.

The threshold of 0.4 resulted also in 10.1 and 11.3% false positive error rates for GM5_ECO1 and GM4_ECO3 trained programs, respectively. These relatively high error rates triggered a detailed analysis of unannotated ORFs in EcoSeq6 (29). About 350 of these ORFs had a score >0.4 with at least one of the two chosen GeneMark versions. More than one half of these ORFs showed significant sequence similarity to proteins present in sequence databases and for many of these putative proteins, a function could be predicted (16,17; see also 27,28).

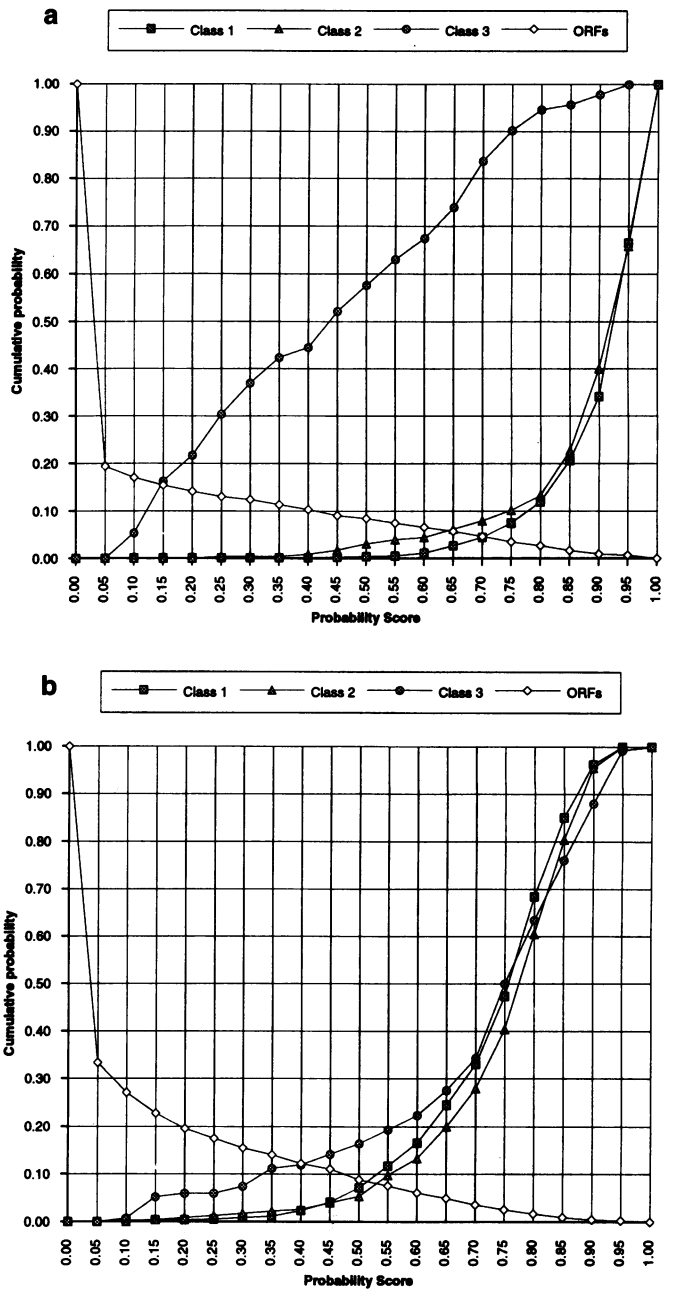

Figure 3. Cumulative histograms of the GeneMark scores for each class of genes and for unannotated ORFs. (**a**) The scores were computed by GM5_ECO1 program. Three rising curves correspond to three gene classes. The descending curve is the cumulative histogram for ORFs' scores (the summation is made from right to left). (**b**) The same as in a with the exception that the scores were computed by GM4_ECO3 program.

Re-evaluation of gene and ORF annotation in GenBank

For 126 ORFs that have been identified as probable coding regions using GeneMark, the initial sequence similarity searches performed in January, 1994, have provided no support. However, our latest analysis performed in January, 1995 indicated that among these 126 predicted expressed ORFs, most of which were relatively short partial gene sequences, 54 have already been identified as putative genes by completion of the ORF sequence and findings of indirect evidence for expression, for example significant ORF length and presence of ribosomal binding sites. For another 14 ORFs, the function of predicted proteins has been demonstrated experimentally; and eight predicted proteins showed significant similarity to proteins that have been added to the sequence databases lately (Table 1). Thus, 50 predicted new genes are still awaiting validation and, given the apparent high accuracy of the GeneMark prediction, they seem to be plausible subjects for direct experimental analysis (Table 2).

Table 1.

REGION/ ACCESSION #/ SWISSPROT NAME	TYPE OF SEGMENT/ ORF SIZE (AA)	GENE TYPE PREDICTED	BEST HIT/ P VALUE	CROSS-PHILUM HIT/ P VALUE	CONSERVED MOTIF	PREDICTED FUNCTION
pabB_sdaA M28695/K02673 YeaB	complete 192	N	GB:M22078 (K.aerogenes) 1.10E-59	none		unknown
end_serS X05017 YcaJ	C 80	F	GB:CBTRXB_4 (C.burnetti) 9.60E-23	YN02_YEAST (S.cerevisae) 8.20E-05	ATP binding site*1	ATPase involved in DNA replication; distantly related to DnaX
hemH_ybaC D90259 YbaC	C/complete 59/319 frameshift	N	GB:AFAGBD_6 (A.eutrophus) 1.70E-22	LIPS_RAT (Rattus sp.) 9.80E-20		lipase
emrB_end M86657 YgaG	C 116	N	GB:HSA07B091*2 ? 2.70E-11	?		unknown
rpoD_end J01687 YgiF	C 32	N	SMAAACGIC (S.marcescens) 1.90E-04	none		unknown
zwf_end M55005 YebK	N 123 frameshift	F	YNGB_CLOPE (C.perfringens) 1.30E-13	none		unknown
end_pyrF M23250 YciM	C 82	N	GB:SCU16783_1 (S.cerevisae) 5.20E-04	GB:SCU16783_1 (S.cerevisae) 5.20E-04	Zn finger	DNA binding
ybdD_end X52904 YbdH	C 81	N	GB:CFU09771_3 (C.freundii) 6.50E-04	none		dehydrogenase
end_nrdA K02672 YfaL	complete 158 frameshift	N	YDEK_ECOLI (E.coli) 8.60E-04	none		unknown
end_ttdA M16194 YgiP	N 40	N	YAFC_ECOLI (E.coli) 1.60E-02	none	HTH motif	transcription regulator
end_mayB X59939 YceK	complete 67	F	YDDQ_ECOLI (E.coli) 1.90E-02	none		unknown

The Tag of the intergenic region where the predicted expressed ORF is located is assigned based on downstream or upstream gene names. If an intergenic region is located at the 5′ end of a EcoSeq6 contig, it is designated end_Gene_name, (Gene_name is the name of the nearest annotated downstrean gene). If an intergenic region is located at the 3′ end of a EcoSeq6 contig, it is designated Gene_name_end (Gene_name is the name of the nearest annotated upstream gene). An ORF is predicted as expressed if at least one of the scores, computed by GM5_ECO1 program (Class I score) or by GM4_ECO3 program (Class III score), is >0.4. This highest score defines the type of the predicted gene, that is native (N) or 'foreign' (F). If both scores are >0.4, N is assigned if Class I score > Class III score and F is assigned if Class III score > Class I score + 0.15. Otherwise, no class designation is assigned (gray area). This is a simplified rule derived from the analysis of the distributions of scores computed for known native and foreign genes (M.B. and K.E.R. unpublished observations). The position in the putative new protein (N-terminal, C-terminal or complete) and the number of residues are indicated for amino acid sequence; extension indicates that predicted ORF has been extended due to the new sequence data with the first number showing the initially predicted length and the second number showing the new total length; the cases of apparent frameshift are noticed. The similarities were found by screening the non-redundant protein sequence database as implemented at NCBI in January 1995.

[1]The conserved motifs comprising an ATP-binding site are predicted to be in the N-terminal portion of YcaJ, upstream of the segment for which the amino acid sequence is available.

[2]The observed similarity is with a human cDNA which however, originates from a collection that is apparently heavily contaminated with bacterial sequences (31), therefore, the origin of this sequence remains uncertain.

The similarities were found by screening the non–redundant protein sequence database as implemented at NCBI in January 1995.

The distribution of GeneMark scores for these 126 ORFs, computed using the GM5_ECO1 and GM4_ECO3 programs, is shown in Figure 4. The 54 genes that already have been completely sequenced and classified as coding for hypothetical

Table 2.

REGION/ ACCESSION #/ SWISSPROT NAME	TYPE OF SEGMENT ORF SIZE (AA)	GENE TYPE PREDICTED
end_act X08035 YcaK	C 64	F
agp_end M33807 YccJ	complete 75	N
end_ansB M34234 YggM	N 45	
ansB_end M34234/M34277 YggN	C 65	N
carB_caiE J01597/V01500 D10483/X73904 YaaV	complete 59	F
end_cysC M74586 YgbE	N 81	N
cysS_folD X56234/X59293 D10588 YbcI	complete 173	N
cysS_folD X56234/X59293 D10588 YbcJ	C/complete 70/frameshift	N
dcm_end X13330 YedJ	C 76	N
dsd_yfdA J011603 YfdD	complete 71	F
end_fdnG M75029 YddG	N 54	N
fepA_fes J04216/M13748 YbdI	complete 57	N
frr_end D13334 YaaM	N 35/49	N
fur_fldA M59426 YbfJ	complete 84	F
end_gnxA M13449/U18655 YbjC	N/complete 41/95	N
end_hisL V00284 YefM	complete 42	F
lap_end M27059 YgbF	complete 116	F
inaA_glpQ K02672 InaA	N/extension 68/216/frameshift	N
inaA_glpQ K02672 YfaH	complete 68	F
ieuS_end X06331 YbeL	N 85	N
end_narL X13360 YchP	C 112	N
end_nrdA K02672 Yfal	N 85	N
end_nrdA K02672 YfaJ	complete 237/frameshift	N
end_nrdA K02672 YfaK	complete 188/frameshift	N
ogt_end Y00495 YdaH	C 57	N
end_ompA J01654 YcbG	C 39	N
pal_lysT X65796 YbgF	complete 258	F
ppsA_ydiA M69116 YdiG	complete 74/frameshift	N
prc_end D00674 YebJ	C 39/frameshift	N
priC_apt D13958 YbaN	complete 116	N
end_priC D13958 YbaM	complete 53	F
pssA_kgtP M58699/X53027 YfiM	complete 90	
purE_end M19657 YbbF	C 90	N
end_recN Y00357 YfjE	C 89	N
end_rimL X15860 YdcG	C 57	N
end_rimL X15860 YdcH	complete 55	N
end_ebmA X54153 YaiH	N 69	N
sohB_topA X04475 YciN	complete 50	N
end_speD J02804/D26562 YacL	C/complete 103/120	N
speF_kdpE M64495 YbfK	complete 85	F
end_tyrR M12114 YcjF	C 102	
uvrY_sdiA X03691 YecF	complete 99/frameshift	F
ysjB_queA M37702 YajH	complete 108	F
end_ybaD X84395 YajI	N 80	N
end_ybhB J01638 YbhC	N 39	N
end_yccA X00547 YccK	N 33	N
ycfA_pin X01805 YcfK	complete 125	F
ycfB_end X59307 YcfB	C/extension 56/231/frameshift	N
end_yedA X13330 YedI	N 136	N
end_yfiB X01818 YfiN	C 79	F
ygfB_ssr D90281 YgfE	complete 109	N

The designations are as in Table 1.

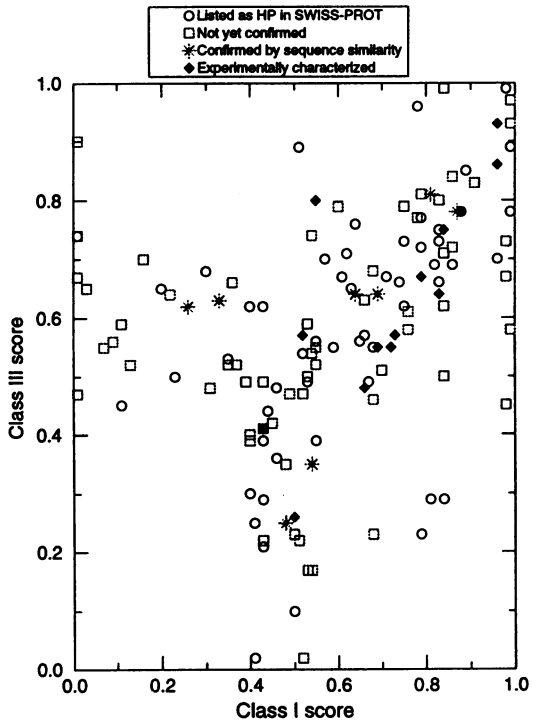

Figure 4. Distribution of GeneMark scores for 126 new genes. The x axis represents the score computed by GM5_ECO1 program, y axis represents the score computed by GM4_ECO3 program. The quadrant x < 0.4, y < 0.4 is empty since a threshold of 0.4 was applied.

```
(a)

YciM'              47 RYRCQKCGFTAYTLYWHCPSC
YRE9_BACSU          7 KFICQSCGYESPKWMGKCPGC   P37572
SMS_ECOLI           8 AFVCNECGADYPRWQGQCSAC   P24554
Zn protein rat    543 REMCDVCDTTIFNLHWVCPRC   S28499 PIR
SOL_DROME         138 RWVCHACGTDNSSVTWHCLIC   P27398
consensus             ...C..C......U...C..C

(b)

                                  HTH  domain

YgiP'              11 LQVLVEIVHSGSFSAAAATLGQTPAFVTKRI
YAFC_ECOLI          8 LAIFVSVVESGSFSRAAEQLGQANSAVSRAV  P30864
TDCA_ECOLI         12 LVVFQEVIRSGSIGSAAKELGLTQPAVSKII  P11036
YDHB_ECOLI          7 LEVVDAVARNGSFSAAAQELHRVPSAVSYTV  P37598
consensus             L.UU..U...GSU..AA..L......V...U
```

Figure 5. Conserved amino acid sequence motifs in putative new proteins predicted by GeneMark. The multiple alignment blocks containing the conserved motifs were generated from BLASTP outputs using the CAP program (26). The position of the first aligned residue in each protein sequence is indicated by a number. The SWISS-PROT or PIR accession numbers are shown in the rightmost column. (**a**) Zn finger motif in the putative protein YciM′. The two conserved pairs of cysteines are highlighted by bold type; U indicates a bulky hydrophobic residue. (**b**) Helix–turn–helix motif in the putative protein YgiP′. Amino acid residues that are conserved in most of the HTH proteins are shown by bold type.

proteins, the eight genes whose existence has been corroborated by sequence similarity and those 50 that remain to be confirmed, are distributed uniformly in this plot. Thus it appears that there is no significant difference in the statistical properties of these three categories of sequences. Interestingly, of the 14 genes whose products have been identified experimentally, 13 have scores >0.4 with both programs. Perhaps this clustering may be due to the correlation between GeneMark scores and the expression level which is reflected in our observation that Class II genes have high scores with programs trained on both Class I and III genes. Obviously, the products of highly expressed genes are more likely to be identified experimentally.

Putative new genes detected by both GeneMark and sequence similarity searches. As discussed previously, combining coding potential prediction methods such as GeneMark with sequence similarity analysis provides an effective strategy for identification of new bacterial genes (16,17). Predictably, in the sequence set studied in this work, the fraction of new GeneMark predictions that could be corroborated through sequence conservation was low. Nevertheless, these cases included interesting new genes and illustrated the problems arising when partial protein sequences are used for database screening (Table 1).

As GeneMark typically identifies parts of gene sequences coding for relatively short protein fragments, the likelihood of detecting statistically highly significant similarity to other proteins is relatively low (17) and using a combination of straightforward database screening with search for conserved motifs (25,26) is particularly important. Two findings of putative new proteins with highly conserved, functionally characterized motifs are illustrated in Figure 5. In both of these cases, the sequence similarity detected using BLAST was not striking (Table 1) but the finding that the conserved regions in YciM′ and YgiP′ contained a Zn finger motif and a helix–turn–helix motif, respectively, still allows one to make functional predictions. Both of these putative new proteins probably are involved in the regulation of gene expression. Prediction of a new Zn finger protein is of particular interest as *E.coli* encodes only a few proteins of this class which is ubiquitous in eukaryotes. Even though YciM′ showed the highest similarity to the *E.coli* protein Sms and its homolog from *B.subtilis*, the new protein is likely to have a different domain organization as the detected partial sequence is from the C-terminus of YciM′, in contrast to Sms which contains the Zn finger at its N-terminus (see Fig. 8 in ref 29).

In some instances, when a partial sequence only shows similarity to uncharacterized proteins from the database and the alignments do not contain any conserved motifs, further rounds of database search with the sequences selected at the initial step still allow one to predict the function. This was the case of the putative protein YcaJ′ whose C-terminal part was similar to uncharacterized putative proteins from *Citrobacter* and yeast (Table 1). An additional database search with these sequences combined with the identification of conserved motifs resulted in a clear prediction that YcaJ is an ATPase distantly related to the *dnaX* gene product and possibly involved in DNA replication (data not shown).

Getting rid of 'very hypothetical' genes. Given the very low false negative rate of GeneMark, we examined those *E.coli* genes that scored unusually low with both GM5_ECO1 and GM4_ECO3 and whose products did not show similarity to other proteins. This analysis revealed that six 'genes' from Class I are complementary to other genes that have been biochemically characterized and have been clearly identified by GeneMark, in contrast to their low scoring complementary counterparts (Table 3). We believe that

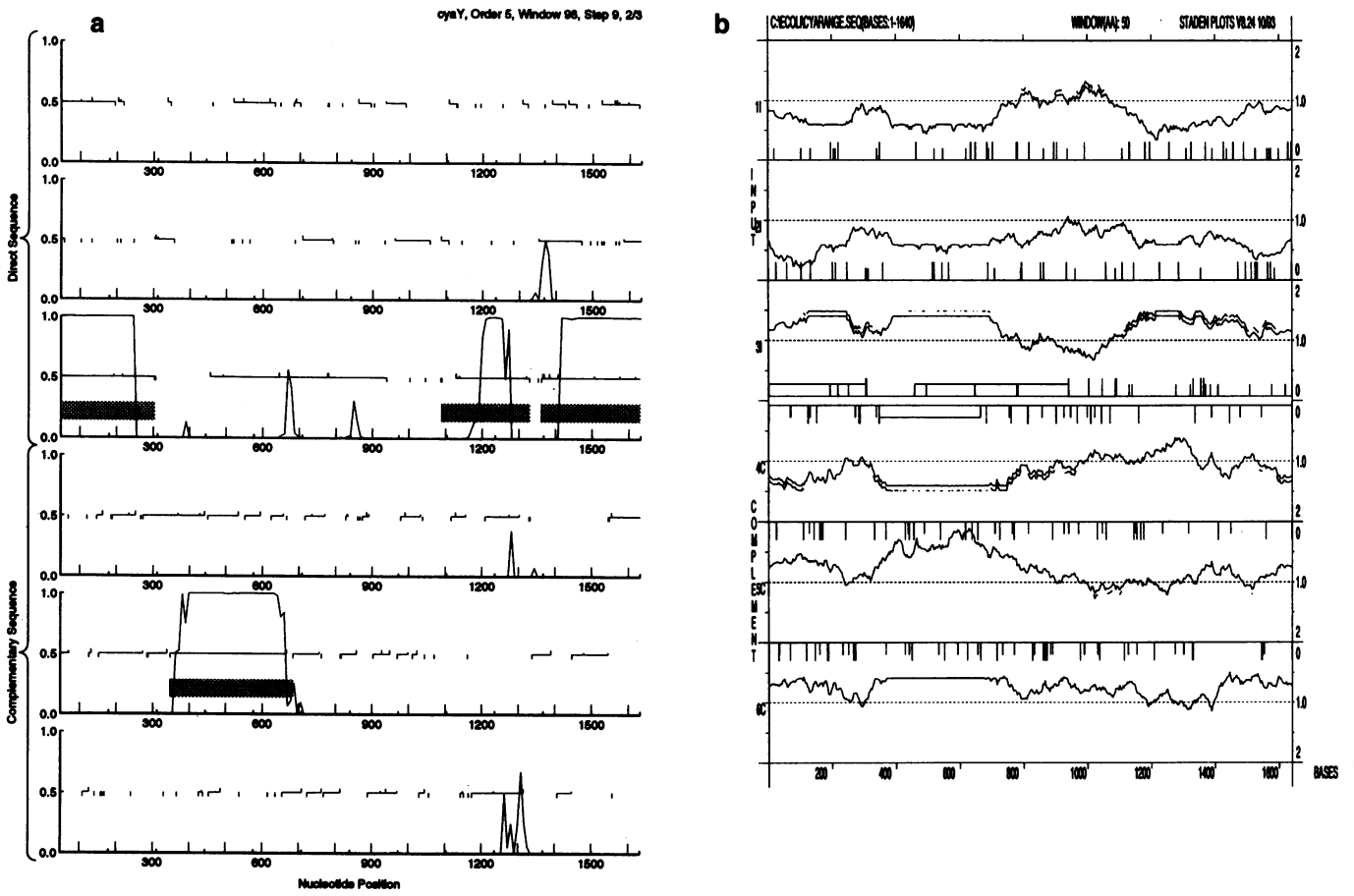

Figure 6. The GeneMark graph for the cyaA downstream region. (**a**) The plot produced by the GM5_ECO1 program as implemented at Georgia Tech E-mail server. Four regions are indicated by gray bars: the 3′ end of *cyaA* gene in the +3 frame (left); the *cyaY* in the –2 frame; the short ORF in the +3 frame; the 5′ region of dapF gene in the +3 frame (right). (**b**) The plot produced by the codon usage based algorithm suggested by Staden (30) and implemented in the DNASTAR software package.

this result indicates that these ORFs residing on complementary strands are not genes.

Table 3.

Proteins encoded by spurious genes	Proteins encoded by genes on complementary strand	GenBank accession nos
VHP in dcm 3′ region	DNA-cytosine methyltransferase	X13330
VHP 13.8 kDa in phn operon	phnP protein	D90227
VHP 12.5 kDa in phn operon	Phosphate transort ATP-binding protein	D90227
VHP encoded by cysX	Protein encoded by cysE	U00039
HP encoded by cyaX	HP encoded by cyaY	X66782
HP FWD1566	HP encoded by yejD	P33918

In particular, a spurious gene designated *cyaX* has appeared in the *E.coli* genomic sequence containing *cyaA* and *dapF* genes and the intergenic region between them (1). There is an ORF located in the –2 frame (*cysY*) and a longer overlapping ORF in the +3 frame (*cysX*) as shown in Figure 6a. Both ORFs have a codon usage characteristic for *E.coli* genes, so it was not easy to discriminate between these competitive ORFs by the method based on codon usage (Fig. 6b). The longer ORF initially has been thought to be the actual gene and has been predicted to encode a hydrophobic protein. However, the GM5_ECO1 program identified the expressed ORF in the –2 frame (Fig. 6a) with the score of 0.90. The score by GM4_ECO3 is 0.72. The respective scores for *CyaX* were 0.05 and 0.09. Strong evidence has been obtained later to support this prediction. First, the sequence comparison with several enterobacterial counterparts has shown that their *cyaX* regions are interrupted by termination codons. Secondly, experiments measuring expression of a fusion of *lacZ* gene with the *cyaA* downstream region indicated that it is from the

complementary strand that a gene, tentatively named *cyaY*, is expressed. Finally, the putative *cyaY* protein had the same length in *Erwinia chrysanthemi* and *Yersinia intermedia*, and in each of these species included ~70% identical amino acids with the *E.coli* protein (P. Glaser, A. Roy and A. D., unpublished observations); more distantly related putative proteins could be identified among uncharacterized nucleotide sequences from *Pasteurella haemolytica* and *Pseudomonas aeruginosa* (E. V. K., unpublished observations).

Among Class III genes, there are 19 that had a score <0.4 with both GM5_ECO1 and GM4_ECO3. For 13 of these genes, biochemical functions have been identified, and five others belong to the HP and VHP categories. In this case, we cannot put forward as strong doubts as for Class I genes since the score calculated by the GM5_ECO1 program for Class III genes is often below 0.4.

CONCLUSIONS

Our results show that there is no single training set which would be suitable for efficient recognition by GeneMark method of all *E.coli* genes. At least two training sets and two program versions derived for different classes of *E.coli* genes are necessary.

Detection of Class III genes is most difficult. These genes can be easily overlooked if inappropriate parameters for the gene-predicting program are used. Class III genes are likely to be recognized with an acceptable accuracy only by a program that has been trained on a representative sample of genes from the same class. This observation substantiates the conclusion by Médigue and co-workers that Class III is mostly comprised of genes that are exchanged horizontally (5). This class represents a significant fraction of the *E.coli* chromosome, perhaps as much as one fifth; at least some of these genes may undergo continuous exchange with other microbial genomes.

This work logically extends our previous reports (16,17) by demonstrating the complementarity of approaches to gene identification by both DNA and protein sequence analysis.

Note

The protein sequences translated from ORFs listed in Tables 1 and 2 can be obtained by FTP from the directory:

/pub/genemark/ecoli3 at amber.biology.gatech.edu.

The programs GM5_ECO1 and GM4_ECO3 can be used via Georgia Tech e-mail server (genemark@ford.gatech.edu). The program versions for *B.subtilis, S.typhimurium, K.pneumonia, M.tuberculosis, M. leprae, M.capricolum* and several other prokaryotic and eukaryotic species are accessible via this server as well.

ACKNOWLEDGEMENTS

We are grateful to William Hayes and Konstantin Derenstein for valuable programming assistance and to Jurgen Kleffe and Michael S. Gelfand for useful discussions. Paul J. Turner kindly provided ACE/gr, a UNIX plotting tool. This work was made possible by NIH grants HG00783 and GM47853 to M.B. and J.D.M.

REFERENCES

1 Daniels, D.L., Plunkett, G.III, Burland, V. and Blattner, F.R. (1992) *Science*, **257**, 771–778.
2 Burland, V., Plunkett, G. III, Daniels, D.L. and Blattner, F.R. (1993) *Genomics*, **16**, 551–561.
3 Plunkett, G.III, Burland, V., Daniels, D.L. and Blattner, F.R. (1993) *Nucleic Acids Res.*, **21**, 3391–3398.
4 Blattner, F.R., Burland, V., Plunkett, G.III., Sofia, H.J. and Daniels D.L. (1993) *Nucleic Acids Res.*, **21**, 5408–5417.
5 Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A. (1991) *J. Mol. Biol.*, **222**, 851–856.
6 Grantham, R., Gautier, C., Gouy, M., Jacobzone, A. and Mercier M. (1981) *Nucleic Acids Res.*, **9**, r43–r74.
7 Gouy, M. and Gautier, C. (1982) *Nucleic Acids Res.*, **10**, 7055–7074.
8 Ikemura, T. (1985) *Mol. Biol. Evol.*, **2**, 13–34.
9 Sharp, P. and Li, W.-H. (1987) *Nucleic Acids Res.*, **15**, 1281–1295.
10 Fickett, J.W. and Tung, C.S. (1992) *Nucleic Acids Res.*, **20**, 6441–6450.
11 Lebart, L., Morineau, A. and Warwick, K.A. (1984) *Multivariate Descriptive Analysis*, John Wiley and Sons, New York.
12 Delorme, M.O. and Henaut A. (1985) *CABIOS*, **4**, 453–458.
13 Hill, M.O. (1974) *Appl. Statist.*, **23**, 340–353.
14 Borodovsky, M. and McIninch, J.D. (1993) *Comput. Chem.*, **17**, 123–133.
15 Rudd, K.E. (1992) In Miller, J. (ed.) *A Short Course in Bacterial Genetics: A Laboratory Manual and Handbook for Escherichia coli and Related Bacteria*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 2.3–2.43.
16 Borodovsky, M., Koonin, E.V. and Rudd, K.E. (1994) *Trends Biochem. Sci.*, **19**, 309–313.
17 Borodovsky, M., Rudd, K.E. and Koonin E.V. (1994) *Nucleic Acids Res.*, **22**, 4756–4767.
18 Borodovsky, M., Sprizhitsky, Yu.A., Golovanov, E.I. and Alexandrov, A.A. (1986) *Mol. Biol.*, **20**, 833–840.
19 Tavare, S. and Song, B. (1989) *Bull. Math. Biol.*, **51**, 95–115.
20 Kleffe, J. and Borodovsky, M. (1992) *CABIOS*, **8**, 433–441.
21 Borodovsky, M., Sprizhitsky, Yu.A., Golovanov, E.I. and Alexandrov, A.A. (1986) *Mol. Biol.*, **20**, 1144–1150.
22 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman D.J. (1990) *J. Mol. Biol.*, **222**, 851–856.
23 Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994) *Nature Genet.*, **6**, 119–129.
24 Wootton, J.C. and Federhen, S. (1993) *Comput. Chem.*, **17**, 149–163.
25 Tatusov, R.L. and Koonin, E.V. (1994) *CABIOS*, **10**, 457–459.
26 Tatusov, R.L., Altschul, S. F. and Koonin, E.V. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 12 091–12 095.
27 Robison, K., Gilbert, W. and Church, G.M. (1994) *Nature Genet.*, **7**, 205–214.
28 Krogh, A., Saira, I. M. and Haussler, D. (1994) *Nucleic Acids Res.*, **22**, 4768–4778.
29 Neuwald, A.F., Berg, D.E. and Stauffer, G.V. (1992) *Gene*, **120**, 1–9.
30 Staden R. (1984) *Nucleic Acids Res.*, **12**, 551–567.
31 Savakis, C. and Doelz, R. (1993) *Science*, **259**, 1677–1678.