Applications of GeneMark in Multispecies Environments

James D. McIninch, William S. Hayes and Mark Borodovsky*

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

james.mcininch@amber.biology.gatech.edu, william.hayes@amber.biology.gatech.edu,
mark.borodovsky@biology.gatech.edu (* author for correspondence)

Abstract

This paper is supposed to bridge the gap between practical experience in using GeneMark for a rapidly widening repertoire of genomes, and the available publications that determine and compare the gene prediction accuracy of the GeneMark method for different genomes. Here we focus on the genome-specific variability of prediction error rates and their sources. DNA sequence inhomogeneity is present both in training and control sets of coding and non-coding regions. Coding region inhomogeneity, caused by differences in sequence composition between "native" and horizontally transferred genes or between genes expressed at different levels, contributes to the false negative error rate. Inhomogeneity of non-coding region may frequently be caused by the presence of unnoticed genes and contributes to the false positive error rate. We have documented such unnoticed genes in GenBank sequences for several species. Some of protein products of these genes have been characterized by similarity search methods. For others, which we call "pioneer genes", no significant similarity has been found at a protein sequence level although the confidence of GeneMark prediction is high. For instance, to date a majority of those pioneer gene predictions made for *E. coli* now show strong similarity to more recently characterized proteins that have been added to protein sequence database. Another practical question is related to genomic sequence inhomogeneity at interspecies level: if GeneMark has not been trained for a particular species, is it possible to apply models derived for phylogenetically close genomes? The answer is, yes. The results of cross-species gene prediction experiments show that cross-species prediction can often be reasonably accurate.

Introduction

Genome sequencing has recently passed the landmark of getting two complete genomes of the free living bacteria *Haemophilus influenzae* and *Mycoplasma genitalium* (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995). Complete sequences of several other genomes that are expected in the near future present a challenging opportunity for computer methods of DNA sequence analysis and gene identification. In prokaryotic genomes many genes may be identified with confidence based solely on the length of an open reading frame (ORF), a DNA segment where the triplet genetic code can be read continuously. Therefore, nontrivial targets of gene identification are short genes, 150-550 nt in length. These genes are difficult to discriminate from randomly occurring ORFs and, on the other hand, 20-40 % of these genes are not detected by protein sequence similarity search methods (for instance, by BLAST type method regardless of PAM or BLOSUM matrix choice).

A number of algorithms to detect coding potential were suggested in the 80's based on the triplet statistical structure of protein coding regions. A comparison of the efficiency of the different types of coding potential statistical measures was made (Fickett & Tung, 1992) and the advantage of in-frame oligonucleotide (k-tuple) statistics was indicated. There are several possible algorithmic ways to use these oligonucleotide statistics (the most popular are trinucleotides and hexamers) for coding potential assessment. For instance, the products of k-tuple frequencies, or sums of their logarithms have been used (Claverie *et al.*, 1990; Snyder & Stormo, 1995). Such expressions, however, do not provide rigorous probabilistic measure for coding potential since adjacent k-tuples are not independent. The inhomogeneous Markov model (IMM) of order k-1, whose parameters are obtained from k-tuple statistics, permits one to obtain a mathematically accurate measure of coding potential: the probability that an observed DNA segment is part of a coding or non-coding region (Borodovsky *et al.* 1986a). IMM models were employed in the GeneMark gene identification algorithm using a Bayesian inference framework (Borodovsky *et al.*, 1986b; Borodovsky & McIninch, 1993). The software implementing this approach has been used for finding prokaryotic genes in several large scale sequencing projects (Burland *et al.*, 1993; Blattner *et al.*, 1993; Fleischmann *et al.*, 1995; Fraser *et al.*, 1995) Currently about 2,900 gene described

in GenBank cite GeneMark as a software tool that was instrumental in their identification.

A Hidden Markov Model (HMM) is an even more flexible tool than IMM. Recently, the fact that prokaryotic gene identification is a non-trivial task has gained the attention of a group working on HMM applications (Krogh *et al.*, 1994a) and a new algorithm has been developed for *E. coli* gene recognition (Krogh *et al.*, 1994b). The program, EcoParse, is supposed to find a maximum likelihood parse of a given DNA sequence into coding and non-coding regions. The advantage of EcoParse is in that it avoids restrictions imposed by sliding window techniques. On the other hand, GeneMark benefits from using the concept of "gene shadow" designating a region of DNA sequence complementary to a true gene residing in the opposite strand. This concept fits well into the Markov model/Bayesian inference formalism and allows for the concurrent analysis of two complementary DNA strands. The EcoParse program processes each strand separately, in two runs. This sort of procedure generates false positive predictions upon entering gene shadows in each strand. The erroneous predictions are related to the self-complementarity of the RNY pattern observed in coding regions (Shepherd, 1981). To cope with the false positive predictions the EcoParse program requires a special non-HMM post-processor. Another problem is to recognize genes that belong to minor classes like Class III in *E. coli* (Médigue *et al.*, 1991).Tests have shown that the EcoParse program is not sensitive to Class III genes from *E. coli* regardless of their length, whereas the GeneMark program has been easily tuned up for finding Class III genes. It appears that merging the Bayesian inference framework of GeneMark with the HMM paradigm of EcoParse could be a promising avenue for future development combining the advantages of both approaches.

In this paper we assess the GeneMark gene prediction performance for several prokaryotic species: *Bacillus subtilis, Salmonella typhimurim, Klebsiella pneumonia, Mycobacterium leprae, Mycobacterium tuberculosis, Mycoplasma capricolum, Sulfolobus solfataricus* and phage T4. Prediction error rates have been defined not only for test sequences from the same genome from which the training set was derived but, also for test sequences from other genomes. As was shown previously, IMM order invariant false positive rates indicate the presence of unnoticed genes (Borodovsky *et al.*, 1995). Upon having such an observation we attempt to characterize putative new genes by using GeneMark in conjunction with sequence similarity search using the program BLAST (Altschul *et al.*, 1990).

Materials and Methods

Algorithm Outline

The key elements of the GeneMark algorithm are the IMMs for protein-coding sequence, coding region shadow sequence and non-coding sequence. Naturally, the accuracy of representing statistical patterns by IMM increases with the increase of the model's order (Borodovsky *et al.*, 1986a). Parameters, initial and transition probabilities of IMM of an order k-1 are determined from k-tuple statistics obtained from training sets of experimentally identified sequences. To identify a given DNA fragment as residing in a region that is either i) protein-coding, ii) non-coding but complementary to a coding sequence (gene shadow), or iii) totally non-coding (see Fig. 1), Bayesian inference is used.

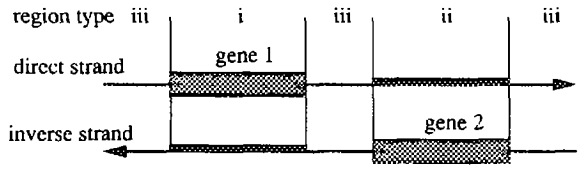


Figure 1. Types of sequence regions shown along the direct strand of DNA.

Each one of the three types of sequence regions (models, j = 1, 2, 3) can be associated with a given DNA fragment, S, by an *a posteriori* probabilistic measure defined by Bayes' theorem:

$$P(\text{model}_j|\text{sequence S}) = \frac{P(\text{sequence S}|\text{model}_j)P(\text{model}_j)}{\sum_j P(\text{sequence S}||\text{model}_j)P(\text{model}_j)}$$

The above expression is generalized to identify possible reading frames in a coding sequence. Then, seven *a posteriori* probability values P_i, i = 1, ..., 7 describe the likelihood of occurrence of seven mutually exclusive events related to the sequence S: sequence S is either directly protein-coding or a shadow of coding region, and the code is to be read in one of six possible frames, or sequence S is non-coding. If one P_i, i = 1, ..., 6 is greater than threshold 0.5, then the DNA fragment S (or its complement) is identified as a protein-coding in its proper reading frame. If P_7 is larger than 0.5 or if no P_i, i = 1, ..., 6 is larger than 0.5 then S (and its complement) is identified as non-coding (for more details see Borodovsky & McIninch, 1993).

Long sequences are analyzed using a sliding window technique. For a given window sequence, each of seven

GeneMark *a posteriori* probability values is interpreted as a probability for a nucleotide in the center of the window to appear in each of the seven states. These probability values, shown graphically as functions of the nucleotide position (see Fig. 4, below), are treated as coding region indicator functions and serve as the basis for parsing the whole sequence into coding and non-coding regions. The protein coding likelihood score for a given sequence region (ORF) is defined as an average of the corresponding likelihood function within the region. All ORFs with likelihood scores larger than a chosen threshold value, for instance 0.5, are included into the list of predicted genes (expressed ORFs).

Sequence Set Compilation

Sets of DNA sequences were extracted from the GenBank database (release #85) using the GENEMAN database search-and-extract utility of the DNAStar software package (DNAStar, Madison, WI). The collection of sequences of a given species was divided into subsets of coding and non-coding sequences according to GenBank feature tables. Sequences that were incomplete or labeled as putative genes were excluded. The sizes of the obtained sequence sets are given in Tab. 1.

	Coding		Non-coding	
Organism	Sequences	Bases	Sequences	Bases
B. subtilis	805	721,459	575	275,744
S. typhimurium	458	436,934	369	173,217
K. pneumonia	164	157,617	116	50,315
Phage T4	259	157,123	140	56,869
S. solfataricus	86	65,761	61	28,813
M. tuberculosis	52	54,685	64	46,529
M. leprae	30	28,230	41	55,511
M. capricolum	29	17,595	61	17,742

Table 1. Size of DNA sequence data sets for the eight species.

Deriving Parameters of IMMs

For each organism, oligonucleotide counts were found for each reading frame of the coding sequence training set and converted into frame-dependent IMM transition probabilities. Oligonucleotide counts were also found for the non-coding training sets and used to compute frame-independent transition probabilities. Matrices of transition probabilities for each training set were derived for several IMM orders. As the order grows, some IMM matrix elements may become zero due to insufficient statistics (the elements other than related to nonsense codons). This fact causes an increase in prediction error rate. Therefore, for a given set of experimentally characterized sequences there is an "optimal" IMM order.

Prediction Accuracy Assessment

The accuracy of the method is characterized by its false positive and false negative error rates. In order to assess these rates, the GeneMark program was used to record the values of likelihood scores in a series of non-overlapping 96 bp DNA fragments. The false negative error rate is defined as the fraction of fragments scoring below 0.5 in the first reading frame when the program is applied to a set of verified protein coding fragment. The false positive error rate is defined as the fraction of fragments scoring above 0.5 in some reading frame when the program is applied to a set of sequences with no known protein-coding genes. (For more detailed discussion see Kleffe et al., 1996). Lower error rates indicate higher accuracy.

To quantify of divergence of species specific patterns in gene sequences we applied IMMs trained for a given species to each of the species specific sequence sets considered in this study. The benchmark error rates for a given species were obtained using a cross-validation procedure. In this procedure, each sequence sample was divided into 7 sub-samples of equal size. The reported error rate is the average of the observed error rates for 7 experiments where each sub-sample was used as the test subject for a model trained on the remaining 6 sub-samples.

Combination With Amino-Acid Sequence Similarity Search

The use of the GeneMark program can be further augmented by subsequent characterization of predicted proteins via similarity search methods. Technically, the GeneMark program is able to forward high-scoring predictions to e-mail servers available through the Internet. In this study, the GeneMark program extracted high-scoring ORFs ($p > 0.5$) and forwarded translated amino-acid sequences to the BLAST e-mail server at NCBI/NIH. This procedure proves to be an efficient way to locate proteins that may belong to specific protein families. Some ORFs with high GeneMark scores produced amino acid sequences with no any significant similarity to known proteins. We call these predicted genes "pioneer genes" and assume that they are worth investigating by all possible means, including repeating the similarity search upon protein sequence database updates.

Implementation

The GeneMark algorithm has been implemented in the C programming language both as a stand alone program and as an e-mail server for UNIX (Solaris 2.3) operating system. The textual output reports predicted coding region locations, their nucleotide sequences and amino-acid sequences of predicted gene products. Postscript graphical output displays the *a posteriori* probability functions pertaining to gene prediction (see Fig. 7 below). There are two e-mail servers located at Georgia Tech and at EBI: genemark@ford.gatech.edu and genemark@ebi.ac.uk. These facilities permit the user to forward amino-acid sequences of predicted proteins to BLAST, BLITZ or FASTA e-mail servers implementing algorithms described in Altschul *et al.*, 1990; Smith & Waterman, 1981; and Pearson & Lipman, 1988 respectively. The protein sequence similarity search results are sent from the remote server directly to the original user.

Results and Discussion

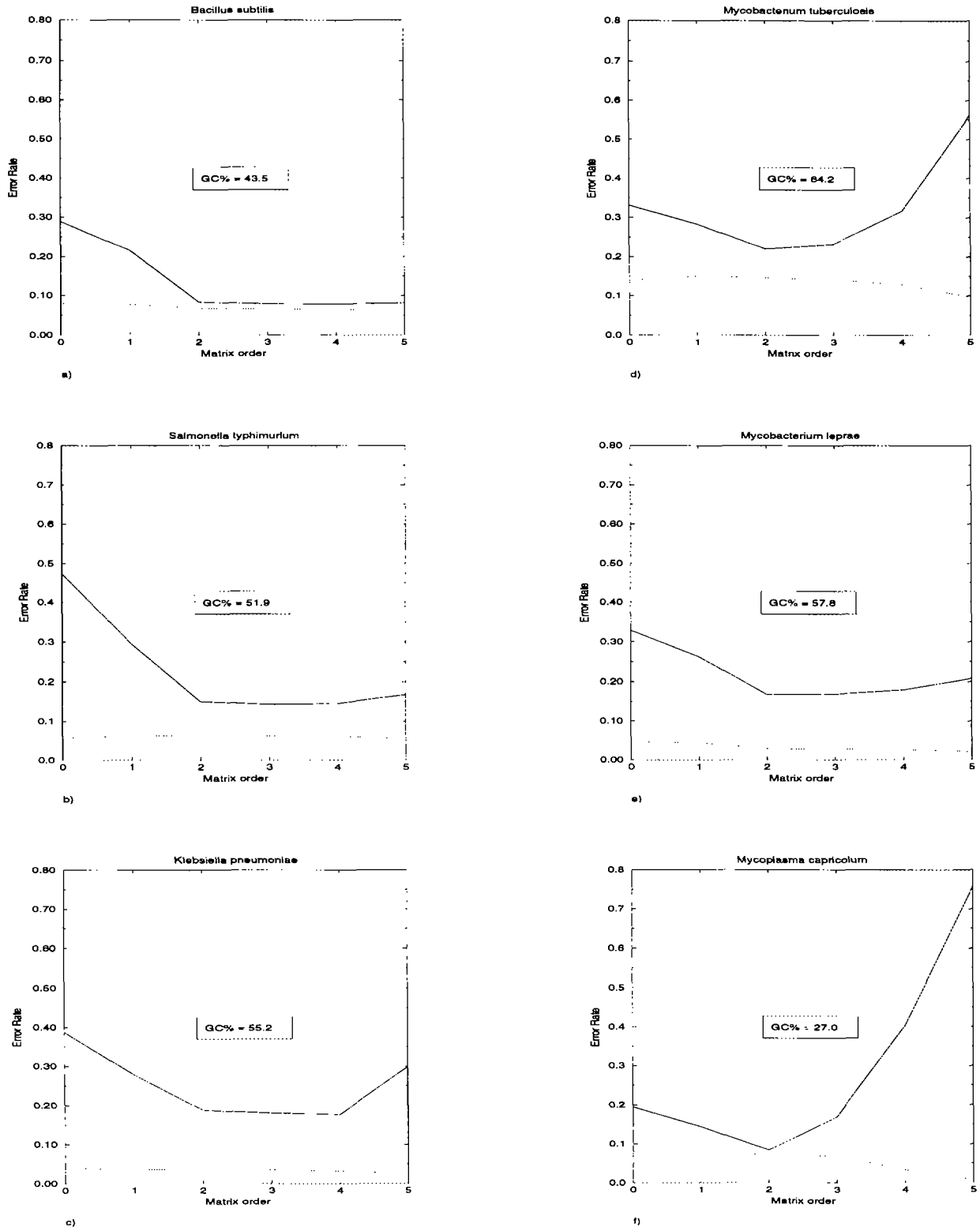
False negative and false positive error rates were assessed, as described in methods, for six species. These values are shown in Fig. 2a-f. It is seen that in case of *M. tuberculosis, M. leprae, M. capricolum*, the false negative error rate achieves the lowest value at IMM order two. Note, that in these cases the size of the coding sequence training set is relatively small. For *B. subtilis, S. typhimurium,* and *K. pneumonia* the nearly lowest error rates are observed for IMM orders two through four. The practical advantage of using higher order models (order four) in this case can be seen from comparison of the whole distribution of *a posteriori* probability scores. Then, it becomes clear that the higher order models produce distributions with lower variance (data not shown). From biological point of view the sharp decrease of false negative error rates for IMM orders from zero through two has a natural explanation since, essentially, parameters of zero order IMM are related to in-phase mononucleotide frequencies, order one IMM carry information on in-phase dinucleotide frequencies and order two IMM carry information on in-phase trinucleotide frequencies (including codon frequencies). The gain of accuracy for higher order models is due to taking into account the bias in frequencies of longer oligonucleotides that cannot be derived from shorter oligonucleotide composition (see also Borodovsky *et al.*, 1995).

The decline of the false positive error rates seen in Fig. 2a-f for higher order models can be partially explained as an artifact related to the simultaneous increase of the false negative error rate. In extreme cases, one may state that "if everything is predicted as non-coding, the false positive rate becomes zero".

There is an important conclusion related to the steadiness of the false positive rates in the range of low IMM orders (up to order three in this study). This tendency contrasts the sharp decrease of the false negative rates observed with the same IMM models (Fig. 2a-f). Such a decrease is expected since the higher order models (but not the highest orders, as discussed above) should better describe statistical patterns of DNA sequences and should produce lower error rates. We assume that this controversy indicates the necessity of re-examination of the set of presumably non-coding sequences in order to detect a presence of unnoticed genes within the coding regions that generate one and the same "false" signal regardless of the IMM order used. The results of this re-examination are given below.

The least false negative error rates observed for particular species vary. This value is a complicated function of the size of the training set, average codon usage bias and the level of homogeneity of codon usage bias among gene sequences of a particular species. For instance, the size of *M. capricolum* training set is small. In this case the observed low false negative error rate for IMM order two is mainly related to the fact that *M. capricolum* genes are AT rich with a strong codon usage bias. *S. typhimurium* and *K. pneumonia* have training sets comparable in size with the one for *B. subtilis*. However, the former species, unlike *B. subtilis* are expected to have classes of genes with an atypical codon usage bias like class III horizontally transferred genes in *E. coli*.

We also investigated the case where IMMs derived from one species are used to analyze sequences of another species (Tab. 2, 3). For example, if the *E. coli* derived IMMs are used for *B. subtilis* sequence analysis it yields a false negative error rate of 63.9%, nearly nine times that seen when the *B. subtilis* models are used. The false positive error rate, however, drops to less than one third that of the *B. subtilis* rate. Naturally, the *E. coli* trained models do not help to recognize *B. subtilis* genes and the majority of *B. subtilis* sequence has been identified as non-coding. On the other hand, models derived from closely related species, such as *E. coli* and *S. typhimurium*, work quite well for each other's sequence analysis. The situation is dramatically different if one takes *E. coli* and *S. solfataricus*. In general, the correlation between



Figures 2a-f. GeneMark prediction error rates as a function of IMM order for *B. subtilis, S. typhimurium, M. tuberculosis, K. pneumonia, M. leprae,* and *M. capricolum.* The false negative rates are shown by solid lines, the false positive rates are shown by dotted lines. Note the general tendency of false negative error rates to decline and then increase at higher orders.

↓Model \ Sequence→	*E. coli*	*S. typhimurium*	*K. pneumonia*	*EcHT*	*B. subtilis*	*Phage T4*	*S. Solfataricus*
E. coli	4.8%	15.8%	18.4%	62.0%	63.9%	81.3%	98.3%
S. typhimurium	7.0%	12.1%	9.3%	57.8%	52.0%	66.6%	92.4%
K. pneumonia	25.9%	30.9%	6.5%	84.2%	84.7%	94.6%	99.5%
EcHT	16.0%	20.4%	32.1%	14.2%	21.7%	11.7%	34.3%
B. subtilis	41.2%	23.9%	31.3%	41.2%	7.0%	24.0%	47.7%
Phage T4	55.0%	60.2%	81.3%	55.0%	37.4%	2.8%	41.6%
S. Solfataricus	92.4%	94.2%	97.6%	81.4%	73.9%	66.1%	14.2%

Table 2. False negative prediction error rates. The forth order IMM were used. The "benchmark" cells are highlighted. The matrix and sequence sets denoted EcHT are horizontally transferred (Class III) *E. coli* genes (Médigue *et al.*, 1991). († IMMs for *S. solfataricus* are of order three).

evolutionary distance and false negative rates is immediately apparent (Tab. 2).

It is worthwhile to note that IMM derived from *S. typhimurium* sequences seem to be relatively good predictors of *K. pneumonia* genes, but the converse does not hold true. The explanation of this phenomenon relates to the fact that *K. pneumonia* and *S. typhimurium* gene sequences share a bias towards the same oligonucleotides (in particular, codons) but the bias in the former species is stronger than in later one. Consequently, when *S. typhimurium* models are used for *K. pneumonia* sequence analysis, the larger transition probabilities related to oligonucleotides frequent in *S. typhimurium* genes are even more frequently used upon scanning *K. pneumonia* genes and produce high coding likelihood scores. However, this does not occur in the inverse situation. A similar relationship exists between *E. coli* native and horizontally transferred genes and between *E. coli* highly expressed and moderately expressed genes [4].

As was stated above, the values and trends of false positive error rates obtained for sets of disjoint sequence fragments may indicate the presence of unnoticed genes in the original sample of presumably non-coding regions. This assumption is supported by the results of GeneMark analysis of continuous sequences where strong predictions of protein-coding regions are seen in locations that are not annotated in the GenBank database. A thorough review of *E. coli* sequences in the EcoSeq6 database (Rudd, 1992) showed a large number of such genes (Borodovsky *et al.*, 1994ab; 1995), and many have since been added to later revisions of this non-redundant *E. coli* sequence database maintained by Kenneth Rudd.

In the present study, in parallel with the GeneMark accuracy assessment stated above, a number of new putative genes were predicted. Predicted protein sequences were analyzed by BLAST local similarity search program. The results are summarized in Tab. 4.

The gene predictions listed above as "pioneer" genes are of particular interest. Not bearing significant similarity to any known protein sequence, these may represent new structural and functional classes of proteins (Fleischmann *et al.*, 1995; Borodovsky *et al.*, 1995; Tatusov *et al.*, 1996). The locations of predicted "pioneer" genes are listed in Tab. 5 below.

One interesting observation made in recent application of GeneMark to *H. influenzae* genomic sequence was that the average length of the pioneer gene is less than an average length of a gene in *H. influenzae* genome (Fig. 3). This difference may contribute to the difficulty in pioneer gene product characterization by protein sequence similarity search.

↓Model \ Sequence→	*E. coli*	*S. typhimurium*	*K. pneumonia*	*EcHT*	*B. subtilis*	*Phage T4*	*S. Solfataricus*
E. coli	15.1%	11.6%	16.7%	15.1%	3.6%	2.5%	1.6%
S. typhimurium	15.4%	11.6%	16.7%	15.4%	6.1%	5.6%	3.5%
K. pneumonia	11.5%	8.6%	13.8%	11.5%	1.3%	13.8%	0.8%
EcHT	17.1%	13.3%	15.1%	17.1%	11.6%	15.8%	12.5%
B. subtilis	16.2%	12.6%	15.8%	16.2%	14.2%	12.6%	12.5%
Phage T4	10.4%	7.4%	6.2%	10.4%	9.1%	16.7%	8.6%
S. Solfataricus	2.0%	1.8%	0.2%	2.0%	4.2%	6.4%	11.5%

Table 3. False positive prediction error rates. The forth order IMM were used. The "benchmark" cells are highlighted. († IMMs for *S. solfataricus* are of order three).

	Pioneer genes	Genes with protein function identified by BLAST	Not annotated in GenBank but present in the protein database	Amount of sequence (kb)
All species	**190**	**92**	**117**	**776.0**
M. leprae	93	26	9	375.0
B. subtilis	51	28	69	229.0
S. typhimurium	15	15	30	97.7
K. pneumonia	2	10	6	30.7
M. tuberculosis	14	6	1	25.8
M. capricolum	15	7	2	17.5

Table 4. Results of a gene search. Pioneer genes are reported if no significant similarity ($P > 10^{-5}$) to any entry in the current protein database is found. BLAST corroborated predictions are reported if the BLAST program was able to find significant similarity to a known protein. 117 predicted proteins have been found in the protein sequence database but their genes are not annotated in GenBank. These proteins could have been reported by researchers other than the authors of the original DNA sequence entry.

Accession	Left end	Right end	AA	d/c	Score
B. subtilis					
A14086	3	155	51	c	0.7322
D14399	3150	3350	67	d	0.6095
D26185	151051	151209	53	d	0.5689
D26185	19392	19631	80	c	0.7607
D26185	7249	7363	38	c	0.5266
D29985	16585	16974	130	d	0.5472
D30689	2	772	257	c	0.6398
D31856	28482	28871	130	d	0.5355
D37799	983	1150	56	c	0.5473
D45242	10475	10609	45	c	0.6770
D45242	20175	20660	162	c	0.5915
D45911	16661	17089	143	d	0.6676
D45911	3232	3621	130	d	0.5338
L03376	1	360	120	d	0.7295
L06664	240	347	36	d	0.5059
L15202	1	375	125	d	0.7627
L17438	1864	2064	67	d	0.5570
L35574	1	243	81	d	0.6086
L42526	130	273	48	d	0.5631
M12620	1	168	56	d	0.5718
M12622	3	164	54	d	0.6587
M16207	1	447	149	d	0.8763
M17445	756	944	63	d	0.5523
M20012	192	440	83	d	0.8305
M27556	3	233	77	c	0.6618
M34826	3	653	217	c	0.7401
M59358	3	287	95	c	0.6679
M73546	1	159	53	d	0.5401
M74538	3860	4066	69	d	0.5037
M85163	860	1147	96	d	0.8427
S70734	374	607	78	c	0.7229
U11039	2046	2270	75	d	0.6472
U20909	175	480	102	d	0.7042
X02150	110	370	87	d	0.7273
X02369	7561	7668	36	c	0.6551
X05680	2	181	60	d	0.6892
X07796	2	412	137	d	0.6600
X56679	1	150	50	d	0.5576
X56680	1	243	81	d	0.6747
X58433	2	136	45	c	0.9972
X73124	38542	38745	68	c	0.5838
X78560	73	186	38	d	0.5085
X87845	1303	1476	58	c	0.5665
S. typhimurium					
D26057	890	1183	98	c	0.6716
D90301	228	569	114	d	0.7077
J01804	2	544	181	c	0.9386
L04307	112	198	29	d	0.5512
L42521	328	768	147	d	0.6057
M84574	2	421	140	d	0.8874
M97752	258	539	94	d	0.8391
U11243	152	301	50	c	0.5941
U11243	3155	3361	69	d	0.6908
U12808	2	205	68	c	0.5537
X63534	1911	2033	41	d	0.6193
X73226	504	863	120	d	0.8171
Z29513	599	970	124	d	0.6332
K. pneumonia					
X53433	20	343	108	c	0.6940
X53433	418	654	79	c	0.5299
M. tuberculosis					
D17369	2	643	214	d	0.6478
L38851	2	154	51	c	0.9117
M62708	2	205	68	c	0.5609
S36714	343	522	60	c	0.5283
U00024	16223	17161	313	d	0.5150
U00024	6319	6504	62	c	0.5898
U27357	2542	2847	102	c	0.8803
U27357	3494	4156	221	c	0.5276
X58485	3	116	38	c	0.9825
X68081	276	1088	271	d	0.5656
X69463	2	178	59	c	0.5750
M. leprae					
L01095	3031	3876	282	d	0.7140
L01095	4263	4382	40	d	0.6259
L01095	5291	5515	75	c	0.7326
L01095	6893	6997	35	d	0.6029
L01095	11651	11800	50	d	0.5342
L01095	13147	13245	33	d	0.5937
L39923	11178	11387	70	d	0.6321
L39923	15683	16156	158	d	0.8693
L39923	3076	3312	79	c	0.6157
L39923	9630	9986	119	d	0.8000
U00011	2756	2947	64	c	0.5497
U00011	14327	14539	71	c	0.5045
U00011	27088	27207	40	d	0.5819
U00011	29861	30133	91	c	0.6810
U00011	30397	30639	81	c	0.5555
U00012	116	532	139	d	0.6597
U00012	20567	21070	168	d	0.7426

U00012	9243	9359	39	c	0.9678
U00013	127	276	50	d	0.5628
U00013	13421	13684	88	d	0.5148
U00013	14985	15122	46	c	0.5690
U00014	5	145	47	c	0.6627
U00014	743	946	68	c	0.5014
U00014	13789	14004	72	d	0.5193
U00014	2029	2196	56	c	0.5396
U00014		18793	63	d	0.5969
U00014	26227	26400	58	d	0.7078
U00015	328	417	30	d	0.5601
U00015	12851	12976	42	d	0.6080
U00015	14961	15191	77	c	0.5382
U00015	23125	23298	58	c	0.5609
U00016	11532	11783	84	c	0.5926
U00016	11883	12020	46	c	0.5390
U00016	13298	13531	78	c	0.5280
U00016	14361	14459	33	d	0.5445
U00016	18858	19124	89	c	0.5436
U00016	20898	21101	68	c	0.7235
U00016	24183	24380	66	d	0.5664
U00016	29667	29753	29	d	0.5709
U00016	31269	31418	50	c	0.5629
U00016	40528	40725	66	c	0.6168
U00016	8879	8974	32	c	0.5910
U00017	19813	19998	62	c	0.6105
U00017	23526	23768	81	d	0.7771
U00017	33215	33409	65	c	0.6790
U00017	39992	40240	83	d	0.7186
U00018	2	175	58	c	0.5622
U00018	1247	1381	45	d	0.6193
U00018	13236	13358	41	d	0.6708
U00018	34890	35066	59	d	0.6244
U00018	35761	35874	38	d	0.5901
U00019	2272	2493	74	c	0.5349
U00020	21506	21658	51	c	0.5462
U00020	25416	25601	62	d	0.6099
U00021	12530	12706	59	c	0.5629
U00021	15971	16117	49	c	0.5795
U00021	37283	37408	42	d	0.5296
U00021	38123	38254	44	d	0.5664
U00021	7388	7540	51	c	0.5046
U00022	27434	27562	59	d	0.5894
U00022	3766	3969	68	c	0.5293
U15180	7740	7820	27	d	0.5451
U15182	28664	28759	32	d	0.5362
U15182	29565	29753	63	c	0.5817
U15182	35985	36131	49	d	0.6008
U15182	8926	9108	61	c	0.6217
U15182	9500	9790	97	d	0.7566
U15183	20364	20486	41	d	0.6297
U15183	7955	8491	179	c	0.5043
U15184	22962	23081	40	c	0.5678
U15184	26160	26324	55	c	0.7768
U15184	32406	32600	65	d	0.5704
U15184	34638	34946	103	d	0.5606
U15187	1477	1614	46	d	0.5863
U15187	15743	15910	56	c	0.6062
U15187	10053	10151	33	d	0.6213
X65546	11	217	69	d	0.6994
X73822	303	518	72	d	0.6551
X77128	1	192	64	d	0.7260
X77128	443	604	201	d	0.6769
Z14314	14430	14729	100	d	0.7764
Z14314	16003	16119	39	c	0.5295
Z14314	16257	16373	39	d	0.5457
Z14314	25748	25849	34	d	0.5291
Z14314	30289	30498	70	d	0.7596
Z14314	33951	34070	40	d	0.6209
Z46257	226	690	155	c	0.6393
Z46257	10162	10287	42	c	0.7343
Z46257	26918	27079	54	d	0.6001
Z46257	32811	32948	46	c	0.6267
Z46257	6151	6933	261	c	0.5109

M. capricolum

D13065	367	597	77	d	0.7918
K02974	3	314	104	d	0.7001
X06414	842	937	32	d	0.5714
Z33008	28	216	63	d	0.5589
Z33015	1	192	64	d	0.6902
Z33030	2	349	116	c	0.8199
Z33032	3	275	91	c	0.7747
Z33044	2	187	62	d	0.7240
Z33047	264	452	63	d	0.6653
Z33059	397	693	99	d	0.7966
Z33060	120	341	74	c	0.6043
Z33102	170	334	55	c	0.6043
Z33195	113	340	76	d	0.6070
Z33235	1	351	117	d	0.7625
Z48956	319	624	102	d	0.7605

Table 5 - Positions of predicted pioneer genes. The length of the protein in amino acids is given in AA column The DNA strand is indicated either as 'd' - direct or 'c' - the reverse complement. The last column presents a protein coding likelihood (an average *a posteriori* probability score).

A reasonable question could be: How do we know that predicted pioneer gene is not just a false positive artifact, and, if not, what is a confidence level of pioneer gene prediction?

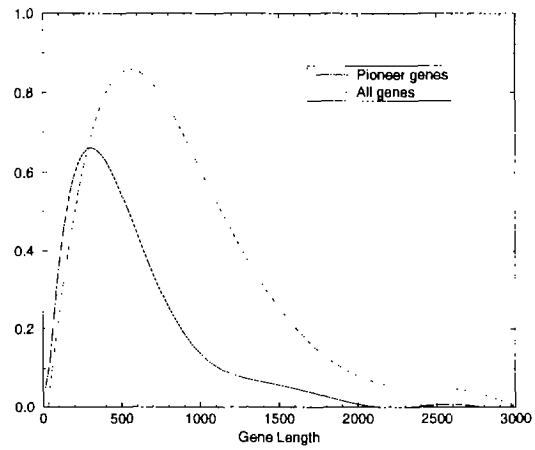


Figure 3. Histograms of gene length and pioneer gene length in *H. influenzae* genome.

This question is difficult to answer precisely, since it amounts to preparing experimentally verified non-coding sequence sets as well as sets of experimentally verified pioneer genes. However, our previous results (Borodovsky *et al.*, 1994ab; 1995) show that pioneer genes are real and GeneMark has an ability to predict these genes (see also Fleischmann *et al.*, 1995; Tatusov *et al.*, 1996). The confidence level can be estimated using the results of computer simulations producing GeneMark score distributions for samples of true genes and non-coding ORFs (of course, "non-coding" in a sense of GenBank annotation). These results indicate that the ORF having a GeneMark score higher than 0.5 has, depending on the given species under consideration, as much as 90-95% of confidence to be a true gene.

In addition to the pioneer genes, GeneMark predicted quite a few new genes whose protein product was corroborated by a BLAST search. These putative proteins have similarities with a wide range of proteins from different organisms and protein families. We made these data available via anonymous FTP from amber.gatech.edu as /pub/GeneMark/new_genes.blast and /pub/GeneMark/new_genes.pioneer or on the world-wide web at http://intron.biology.gatech.edu/new_genes. For example, two putative genes were predicted in GenBank record L13170 (see Fig. 4).

The feature table of the record L13170 does not mention a presence of any genes or gene fragments. The putative proteins derived from GeneMark predictions were used in a BLAST search. It was found that the first protein showed strong similarity to several S1 ribosomal proteins and that the second protein showed similarity to a monoxygenase.

This example is typical for a situation where unnoticed genes are found in the margins of GenBank sequence record. It seems that one strength of GeneMark is to focus the attention of the investigator on putative gene regions which are worthy of efforts to attempt the characterization of a putative protein by various methods available at the protein sequence level. For instance, in addition to conducting a BLAST similarity search, multiple alignment techniques clearly show the presence of an S1 domain in a newly detected protein described above (Fig. 5).

Applications of the GeneMark program to genomic sequence analysis are currently more numerous than we could mention in this short paper. The GeneMark program and e-mail servers have been recently upgraded and now include the refined procedure for prediction of position of translation initiation, as well as the procedure for prediction of location of ribosomal binding site. The new version of the program is able to identify possible sequencing errors that result in frameshifts within protein

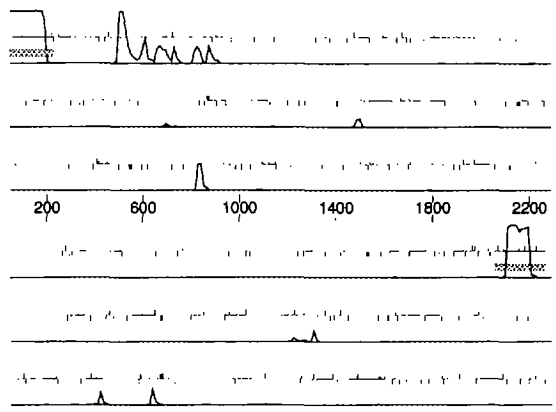


Figure 4 - The GeneMark graph for the GenBank sequence # L13170 (*B. subtilis*). *A posteriori* probability functions are shown by thin lines. ORFs are rendered as horizontal bars at the 0.5 probability level. Regions between subsequent in-frame stop codons in which there is a sustained coding probability higher than the threshold parameter are designated as thick horizontal gray bars. Start and stop codons are indicated as upward and downward ticks respectively.

consensus		....G..U.G.U..U...GA.V.U.....GUU..S.....U.....U..G..V...U..U.....U.L....	
S1H1/BACSU	2	DLKEGMELQGTVRNVVDFGAFVDIGVKQDGLVHISKLSNQFVKHPLDVVSVGDIVTVWVDGVDVQKGRVSLSMVK	L13170
o622/ECOLI	495	DLQPGMILEGAVTNVTNFGAFVDIGVHQDGLVHISSLSNKFVEDPHTVVKAGDIVKVKVLEVDLQRKRIALTMRL	U18997
RS1H_BACSU	268	KVKPGDVLEGTVQRLVSFGAFVEILPGVEGLVHISQISNKHIGTPHEVLEEGQTVKVKVLDVNENEERISLSMRE	P38494
RS1_ECOLI	448	LNKKGAIVTGKVTAVDAKGATVELADGVEGYLRASEASRDRVEDATLVLSVGDEVEAKFTGVDRKNRAISLSVRA	P02349
PNP_ECOLI	617	EIEVGRVYTGKVTRIVDFGAFVAIGGGKEGLVHISQIADKRVEKVTDYLQMGQEVPVKVLEVDRQ-GRIRLSIKE	P05055
YABR_BACSU	1	SIEVGSKLQGKITGITNFGAFVELPGGSTGLVHISEVADNYVKDINDHLKVGDQVEVKVINVEKD-GKIGLSIKK	P37560
RS1H/HUMAN	278	QIAAGSVLEGTVKRVKDFGAFVEILPGIEGLVHVSQISNKRIENPSEVLKSGDKVQVKVLDIKPAEERISLSMKA	U05589

Figure 5. An S1 domain containing protein found in GenBank L13170 - This multiple alignment constructed by the program MACAW (Schuler et al., 1991) includes the C-terminal portion of a newly found putative *B. subtilis* gene product (top row), as well as S1 domains from the *E. coli* and human ribosomal protein S1, and their uncharacterized homologs from *E. coli* and *B. subtilis*. The consensus shows residues conserved in all of the aligned sequences; U - a bulky hydrophobic residue. The distances from the N-termini are indicated.

coding regions. The addresses for GeneMark e-mail servers are given in Materials and Methods section.

Acknowledgments

We are grateful to Dr. Eugene Koonin for help with the analysis of protein sequences from *B. subtilis*. We would also like to thank Dr. Kenneth Rudd for providing the EcoSeq6 database. DNAStar software package was very useful at the stage of sequence sets preparation. This work was made possible by NIH grant HG00783.

References

Altschul, S. F.; Gish, W.; Miller, W.; Myers, E.; and Lipman, D. J. 1990. Basic linear alignment search tool. *J. Mol. Biol.* 215: 403-410.

Blattner, F. R.; Burland, V.; Plunkett, III, G.; Sofia, H. J.; and Daniels, D. L. 1993. *Nucl. Acids Res* 21: 5408-5417.

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

Borodovsky, M.; Koonin, E. V.; and Rudd, K. E. 1994. New genes in old sequence: a strategy for finding genes in the bacterial genome. *TIBS* 19: 309-313.

Borodovsky, M.; McIninch, J. D.; Koonin, E. V.; Rudd, K. D.; Médigue, C.; and Danchin, A. 1995. Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucl. Acids Res.* 23: 3554-3562.

Borodovsky, M.; Rudd, K. E.; and Koonin, E. V. 1994. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucl. Acids Res.* 22: 4756-4767.

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

Borodovsky, M.; Sprizhitsky, Yu. A.; Golovanov, E. I.; and Alexandrov, A. A. 1986b. Statistical features in the *Escherichia coli* genome functional primary structure. III. Computer recognition of protein coding regions. *Molecular Biology*, 20: 1144-1150.

Burland V.; Plunkett, III, G.; Daniels D. L.; and Blattner F. R. 1993. DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: organizational symmetry around the origin of replication. *Genomics* 16: 551-561.

Claverie, J. M.; Sauvaget, I.; and Bougueleret, L. 1990. k-tuple frequency analysis: from intron/exon discrimination to T-cell epitope. *Mapping Methods in Enzymology* 183: 237-253.

Fickett, J. W.; and Tung, C. S. 1992. Assessment of protein coding measures. *Nucl. Acids Res.* 20: 6441-6450.

Fleischmann, R. D.; Adams, M. D.; White, O.; Clayton, R. A.; Kirkness, E. F.; Kerlavage, A. R.; Bult, C. J.; Tomb, J.-F.; Dougherty, B. A.; Merrick, J. M.; McKenney, K.; Sutton, G.; Fitzhugh, W.; Fields, C. A.; Gocayne, J. D.; Scott, J. D.; Shirley, R.; Liu, L.-I.; Glodek, A.; Kelley, J. M.; Weidman, J. F.; Phillips, C. A.; Spriggs, T.; Hedblom, E.; Cotton, M. D.; Utterback, T. R.; Hanna, M. C.; Nguyen, D. T.; Saudek, D. M.; Brandon, R. C.; Fine, L. D.; Fritchman, J. L.; Fuhrmann, J. L.; Geoghagen, N. S. M.; Gnehm, C. L.; McDonald, L. A.; Small, K. V.; Fraser, C. M.; Smith, H. O.; and Venter, J. C. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.

Fraser, C. M.; Gocayne, J. D.; White, O.; Adams, M. D.; Clayton, R. A.; Fleischmann, R. D.; Bult, C. J.; Kerlavage, A. R.; Sutton, G.; Kelley, J. M.; Fritchman, J. L.; Weidman, J. F.; Small, K. V.; Sandusky, M.; Fuhrmann, J. L.; Nguyen, D. T.; Utterback, T. R.; Saudek, D. M.; Phillips, C. A.; Merrick, J. M.; Tomb, J.-F.; Dougherty, B. A.; Bott, K. F.; Hu, P.-C.; Lucier, T. S.; Peterson, S. N.; Smith, H. O.; Hutchison, III, C. A.; and Venter, J. C. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science* 270: 397-403.

Kleffe, J.; Hermann, K.; and Borodovsky, M. 1996. Statistical Analysis of GeneMark Performance by Cross-validation. *Computers & Chemistry* 20: 123-134.

Krogh, A.; Brown, M.; Mian, I. S.; Sjoelander, K.; and Haussler, D. 1994a. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* 235: 1501-1531.

Krogh, A.; Mian, I. S.; and Haussler, D. 1994b. A hidden Markov model that finds genes in E. coli DNA. *Nucl. Acids. Res.* 22: 4768-4778.

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

Pearson, W. R.; and Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* 85: 2444-2448

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. 2.3-2.43 Cold Spring Harbor Press, Cold Spring Harbor, NY.

Schuler, G.; Altschul, S. F.; and Lipman, D. J. 1991. A workbench for multiple alignment construction and analysis. *Protein Struct. Func. Genet.* 9: 180-190.

Shepherd, J. C. W. 1981. Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justifications. *Proc. Nat. Acad. Sci. USA* 78: 1596-1600.

Smith, T. F.; and Waterman, M. S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147: 195-197.

Snyder, E. E.; and Stormo, G. D. 1995b. Identification of protein coding regions in genomic DNA. *J. Mol. Biol.* 248: 1-18.

Tatusov, R. L.; Mushegian, A. R.; Bork, P.; Brown, N. P.; Hayes, W.; Borodovsky, M.; Rudd, K. E.; and Koonin, E. V. 1996. Metabolism and Evolution of *H. influenzae* Deduced from a Whole Genome Comparison to *E. coli*, *Current Biology* 6: 279-291.