Nucleic Acids Research, 2005, Vol. 33, Web Server issue ***W451–W454***
doi:10.1093/nar/gki487

GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses

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

[1]School of Biology and [2]Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received February 14, 2005; Revised and Accepted April 20, 2005

ABSTRACT

The task of gene identification frequently confronting researchers working with both novel and well studied genomes can be conveniently and reliably solved with the help of the GeneMark web software (http://opal.biology.gatech.edu/GeneMark/). The website provides interfaces to the GeneMark family of programs designed and tuned for gene prediction in prokaryotic, eukaryotic and viral genomic sequences. Currently, the server allows the analysis of nearly 200 prokaryotic and >10 eukaryotic genomes using species-specific versions of the software and precomputed gene models. In addition, genes in prokaryotic sequences from novel genomes can be identified using models derived on the spot upon sequence submission, either by a relatively simple heuristic approach or by the full-fledged self-training program GeneMarkS. A database of reannotations of >1000 viral genomes by the GeneMarkS program is also available from the web site. The GeneMark website is frequently updated to provide the latest versions of the software and gene models.

INTRODUCTION

Computational gene finders can be divided into two classes: intrinsic and extrinsic. Intrinsic, or *ab initio*, gene finders make no explicit use of information about DNAs or proteins outside the sequence being studied. Extrinsic gene finders utilize sequence similarity search methods to identify the locations of protein-coding regions. It is common for gene finders of both types to be used in concert in a gene finding project, owing to their complementary nature.

The programs of the GeneMark family are *ab initio* gene finders. Such programs are the only means to identify genes with no homologues in current databases. As these genes make up a sizeable percentage of the whole gene complement for particular species, the importance of *ab initio* programs will not diminish in the foreseeable future. The GeneMark web software includes two major programs, called GeneMark (1) and GeneMark.hmm (2). Both programs employ inhomogeneous (three-periodic) Markov chain models describing protein-coding DNA and homogeneous Markov chain models describing non-coding DNA. GeneMark uses a Bayesian formalism to calculate the *a posteriori* probability of the presence of the genetic code (in at least one of six possible frames) in a short DNA sequence fragment, thus being a local approach. The GeneMark.hmm program uses a hidden Markov model (HMM) framework and the generalized Viterbi algorithm to determine the most likely sequence of hidden states (which are actually labels designating the coding or non-coding function) based on the whole observed DNA sequence. Additional details about the GeneMark and GeneMark.hmm algorithms can be found in Refs (1–3).

The architecture of the HMM itself can be altered to fit the organization of a particular type of genome under study in a better way. For example, the prokaryotic version of GeneMark.hmm contains hidden states for the characteristic features of genes in prokaryotes, including ribosomal binding sites (RBS), uninterrupted genes and gene overlaps. The eukaryotic version utilizes an extended HMM architecture, including states for splice sites, translation initiation (Kozak) sites and interrupted genes (exons and introns). The web software programs have been extensively used—currently >21 000 nucleotide sequence and 329 000 protein sequence records in GenBank (4) contain references to the GeneMark programs.

As many of the model parameters are species-specific, the accuracy of an *ab initio* gene finder is highly dependent on the selection of adequate training data as well as on the use of sound methods to create the models. The models available at the GeneMark website were constructed using our recently developed self-training methods (3) and were tested locally before being released.

WEB SERVER DESCRIPTION

Both GeneMark (1) and GeneMark.hmm (2) can be used via the GeneMark website for the analysis of prokaryotic DNA,

*To whom correspondence should be addressed. Tel: +1 404 894 8432; Fax: +1 404 894 0519; Email: mark.borodovsky@biology.gatech.edu

with 175 pre-computed species-specific statistical models available. Analysis of DNA from any prokaryotic species is supported by (i) a special version of GeneMark.hmm using a heuristic model calculated from the nucleotide frequencies of an input sequence at least 400 nt long (5) and (ii) a self-training program, GeneMarkS (3), which can be used for longer sequences on the order of 1 Mb in length. Thus, the DNA of any prokaryote can be analysed, via either a pre-computed species-specific model or a model created on the fly.

As many of the programs at the GeneMark website share similar interfaces, we use here the prokaryotic GeneMark.hmm program as an exemplar and discuss program-specific differences below, where appropriate.

The GeneMark.hmm web interface accepts as input a single DNA sequence as an uploaded file or as text pasted into a textbox. If a FASTA description line begins the sequence, all text on the line following the 'greater than' symbol (>) is used as the title. In the remainder of the submission, digits and white space characters are ignored and letters other than T, C, A and G (assumed to appear rarely) are converted to N. The interface requires selection of the species name. Selection of a model for the RBS (in the form of a position-specific weight matrix and a spacer length distribution) is optional. In certain cases, such as the crenarchaeote *Pyrobaculum aerophilum*, the RBS model is replaced by a promoter model, which is the dominant regulatory motif located upstream to gene starts in this species (6). The interface also includes the option of using other types of genetic codes such as the Mycoplasma genetic code.

GeneMark.hmm reports all predicted genes in a format that includes the strand the gene resides on, its boundaries, length in nucleotides and gene class (Table 1). Class indicates which of the two Markov chain models used in GeneMark.hmm, Typical or Atypical gene model, provided the higher likelihood for the gene sequence. Genes of the Typical class exhibit codon usage patterns specific to the majority of genes in the given species, while Atypical class genes may not follow such patterns and frequently contain significant numbers of laterally transferred genes (7,8). The nucleotide sequences of predicted genes and translated protein sequences are available as an output to facilitate further analysis, such as BLAST searching (9). An option to generate GeneMark predictions in parallel with the GeneMark.hmm analysis provides important additional information. In this case, GeneMark is set up to use models derived from the same training data as models for the current run of GeneMark.hmm.

It is worth noting that the GeneMark.hmm and GeneMark algorithms are complementary to each other in the same way as the Viterbi algorithm and the posterior decoding algorithm are. Therefore, though the two algorithms are distinct, they are supposed to generate predictions largely corroborating and validating each other. Differences frequently indicate sequence errors and deviations in gene organization, very short genes, gene fragments, gene overlaps, etc.

Table 1. Gene predictions made by the prokaryotic version of GeneMark.hmm for a fragment of the *Escherichia coli* K12 genome

Gene	Strand	Left end	Right end	Gene length	Class
1	+	61	825	765	1
2	+	846	1112	267	2
3	+	1145	2092	948	1
4	–	2254	4386	2133	1
5	–	4388	520	366	1

In the 'Class' column, 1 and 2 indicate Typical and Atypical, respectively. Direct and reverse complement strands are indicated by '+' and '–', respectively. The graphical output for the first three predictions is shown in Figure 1.

Graphical output of the analysis is available in PDF or PostScript format. A fragment of this output, illustrating the predictions of both GeneMark and GeneMark.hmm, is shown in Figure 1. The graphical output clearly depicts the advantage of using multiple Markov chain models representing different classes of genes. Here, the coding potential graph obtained using the Typical gene model, derived by GeneMarkS, is denoted by a solid black line, and the coding potential graph obtained using the Atypical gene model (derived by a heuristic approach) is denoted by a dotted line. Whereas the first and last genes in Figure 1 could be detected using either of the two models, as both of them produced high enough coding potentials, the gene located in positions from 846 to 1112 was detected only by the Atypical model. Further, Figure 1 demonstrates the ability of the GeneMark programs to detect genes of both the Typical and Atypical gene classes (7). The GeneMark graph also includes indications of frameshift positions (also listed in the text report), which are often sequencing errors but in rare cases are natural and biologically very interesting.

For the GeneMark program, there are several specific options. The window size and step size parameters (96 nt and 12 nt, respectively, by default) define the size of the sliding window and how far this window is moved along the sequence in one step. The threshold parameter determines the minimal average coding potential for an open reading frame (ORF) to be predicted as a gene. There are several options which allow fine-tuning of the GeneMark graphical output. In addition, there are options supporting the analysis of eukaryotic DNA sequences by GeneMark including the ability to provide lists of putative splice sites and protein translations of predicted exons. As might be expected, GeneMark (the posterior decoding algorithm) does not produce high enough resolution for the precise prediction of exon–intron borders. Thus, GeneMark.hmm (the generalized Viterbi algorithm) in its eukaryotic version is the major tool for the identification of exon–intron structures in eukaryotic DNA sequences.

The output of the GeneMark program consists of a list of ORFs predicted as genes, i.e. those with average coding potential above the selected threshold. Although each predicted gene can have more than one potential start, additional data is provided to help the researcher annotate one of the alternatives as the 'true' one. The start probability (abbreviated 'Start Prob') is derived from the sequences in the windows immediately upstream and downstream of each potential start. RBS information is provided in the form of a probability score along with the position and sequence of the potential RBS (abbreviated 'RBS Prob', 'RBS Site' and 'RBS Seq'). In addition to the list of predicted genes, GeneMark provides a list of 'regions of interest', spans of significant length between in-frame stop codons where spikes of coding potential are wide enough and may warrant further analysis even if no genes are predicted therein based on automatic comparison with the threshold.

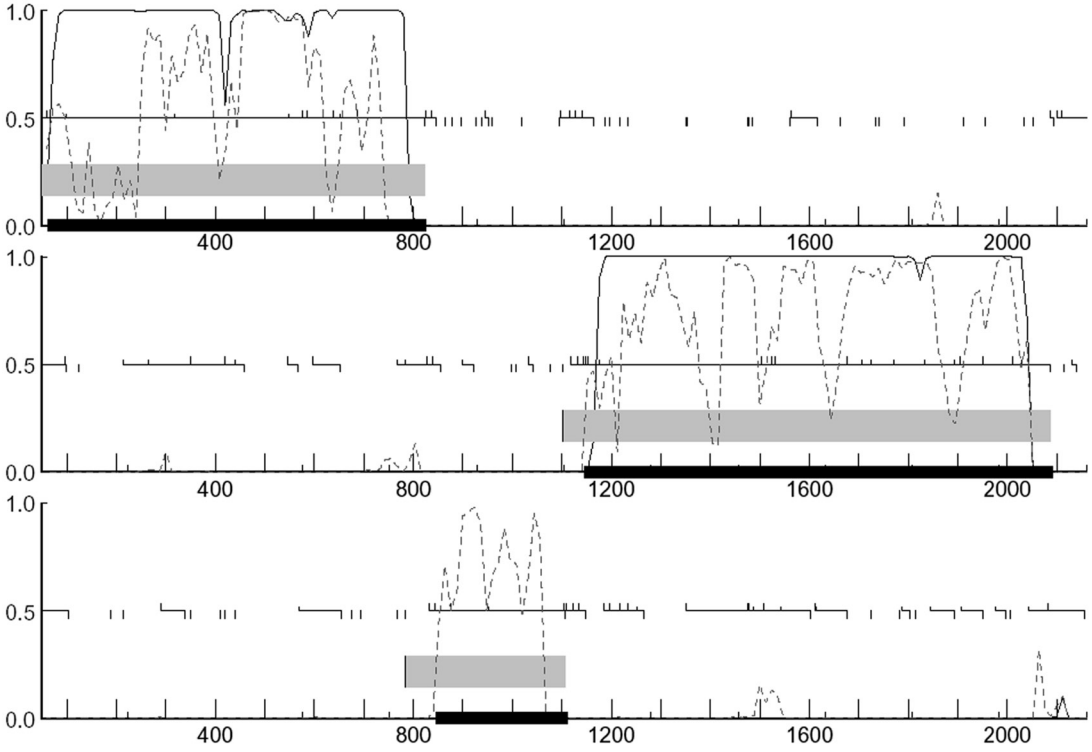

Figure 1. Graphical output from the combination of GeneMark and GeneMark.hmm for a fragment of the *Escherichia coli* K12 genome. The solid black and dashed traces indicate the coding potential calculated by the GeneMark program using the Typical and Atypical Markov chain models of coding DNA, respectively. Only the three reading frames in the direct strand are shown as there are no genes (either predicted or annotated) on the reverse strand in this section of the genome. The thick black horizontal bars indicate the locations of the predictions made by GeneMark.hmm. The thick grey horizontal bars indicate 'regions of interest' provided by the GeneMark program. The thin black horizontal lines indicate (longest) ORFs observed in each reading frame; ticks extending above and below this line indicate potential start and stop codons, respectively.

Analysis of prokaryotic DNA sequences for which there is no pre-computed species-specific model can be carried out using a program version which heuristically derives a model for any input sequence >400 nt (5). This approach has also proven useful for the analysis of inhomogeneous genomes, particularly regions too divergent from the bulk of the genome, such as pathogenicity islands (10,11).

If models (including RBS models) have to be computed *de novo* for an anonymous DNA sequence with length of the order of 1 Mb or longer, the GeneMarkS program can be used (3). This program needs significantly more computational resources; thus, its output is provided via email. A modified version of GeneMarkS tuned for the analysis of viruses of eukaryotic hosts creates a model for the Kozak consensus sequence instead of a two-component RBS model.

The eukaryotic version of GeneMark.hmm is currently available for the analysis of 11 eukaryotic genomes: *Homo sapiens*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Chlamydomonas reinhardtii*, *Drosophila melanogaster*, *Gallus gallus*, *Hordeum vulgare*, *Mus musculus*, *Oryza sativa*, *Triticum aestivum* and *Zea mays*. From the prediction accuracy tables given at the website (http://opal.biology.gatech.edu/GeneMark/plant_accuracy.html), it follows that the latest versions of GeneMark.hmm produce remarkably accurate gene predictions for plant genomes such as rice and *Arabidopsis*. This fact has not escaped the attention of plant genome sequencing consortiums, which have used the program intensively (12,13). The analysis of cDNA and EST sequences from eukaryotes, which typically contain no introns, is facilitated by a special version of GeneMark called GeneMark.SPL. Interestingly, eukaryotic genomes with rare introns present difficulty in terms of collecting enough statistics for the intron and internal exon related models, the important components of a full-fledged eukaryotic gene finder. For this reason, a special interface is available for low eukaryotes such as *Saccharomyces cerevisiae*. Currently, this interface employs versions of prokaryotic GeneMark and GeneMark.hmm augmented with Kozak start site models instead of the prokaryotic RBS model.

The eukaryotic species-specific models are represented by several variants built for distinct G + C% ranges covering the whole scale of G + C inhomogeneity observed in a particular genome. GeneMark.hmm automatically selects the model variant which fits the G + C% of the input sequence. Note that, in the eukaryotic case, the RepeatMasker program (A. M. A. Smit, R. Hubley, and P. Green, www.repeatmasker.org), which is frequently used for pre-processing, can introduce a significant number of 'N' characters. These characters do not influence the selection of the Markov chain model used in prediction.

A sample text output produced by the eukaryotic version of GeneMark.hmm is shown in Table 2. In the graphical output of the eukaryotic version of GeneMark.hmm, the thick horizontal bars (which represent whole genes in the prokaryotic case) indicate predicted exons. Vertical ticks on these bars show

Table 2. Prediction of a single gene (with seven exons) made by the eukaryotic version of GeneMark.hmm for a fragment of the *Arabidopsis thaliana* genome

Gene no.	Exon no.	Strand	Exon type	Exon range	Exon length	Start frame	End frame
1	1	+	Initial	23 525–24 451	927	1	3
1	2	+	Internal	24 752–24 962	211	1	1
1	3	+	Internal	25 041–25 435	395	2	3
1	4	+	Internal	25 524–25 743	220	1	1
1	5	+	Internal	25 939–25 997	59	2	3
1	6	+	Internal	26 292–26 776	485	1	2
1	7	+	Terminal	26 862–27 012	151	3	3

The gene that an exon belongs to and its strand are necessarily the same for all exons in a gene. Each exon is described by a type: 'initial' (begins with ATG), 'internal' or 'terminal' (ends with TAA, TAG or TGA) or 'single' for exons which are both initial and terminal (intronless gene). The start frame and end frame indicate the position of the codon (first, second or third) that the exon begins and ends with, respectively. Notably, all complete gene structures begin in codon position 1 and end in codon position 3.

the starts and ends of predicted initial and terminal exons, respectively.

For the analysis of virus and phage DNA, the heuristic (for short genomes) and GeneMarkS (for long genomes) options, mentioned above, are recommended. In addition, a database called VIOLIN containing pre-computed reannotations of >1000 virus genomes is available (14).

Future directions for GeneMark web software development include detection of several genomic elements currently not predicted by either GeneMark or GeneMark.hmm, such as rRNA and tRNA genes (which can be mis-predicted as protein-coding genes in low G+C% species) and improving the detection of gene 5′ ends. Currently, the server supports the analysis of sequences masked by tRNAscan (15) or similar programs. The GeneMark programs will not find genes in these masked areas (sequences of 'N' characters); thus, the predictions will be compatible with this extrinsic information. The detection of exact gene starts remains a challenging problem in gene finding, as many genes have relatively weak patterns indicating sites of translation and transcription initiation. This problem is made especially difficult by the lack of available data sets containing verified gene start locations to be used for training and evaluation. Refinements in the RBS and Kozak models and the potential inclusion of hidden states representing upstream promoter sequences are currently being explored to address this issue.

ACKNOWLEDGEMENTS

The authors are grateful to Alexandre Lomsadze, Ryan Mills, Vardges Ter-Hovhannisyan and Wenhan Zhu for help in updating and maintaining the GeneMark website. We also appreciate the efforts of the European Bioinformatics Institute, who have set up a server for the GeneMark program at www.ebi.ac.uk/genemark/. Development of the programs available on the GeneMark website has been supported in part by a research grant from the US National Institutes of Health. Funding to pay the Open Access publication charges for this article was provided by NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Borodovsky,M. and Mcininch,J. (1993) GenMark: parallel gene recognition for both DNA strands. *Comput. Chem.*, **17**, 123–133.
2. Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.
3. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res.*, **32**, D23–D26.
5. Besemer,J. and Borodovsky,M. (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.*, **27**, 3911–3920.
6. Slupska,M.M., King,A.G., Fitz-Gibbon,S., Besemer,J., Borodovsky,M. and Miller,J.H. (2001) Leaderless transcripts of the crenarchaeal hyperthermophile *Pyrobaculum aerophilum*. *J. Mol. Biol.*, **309**, 347–360.
7. Borodovsky,M., McIninch,J.D., Koonin,E.V., Rudd,K.E., Medigue,C. and Danchin,A. (1995) Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res.*, **23**, 3554–3562.
8. Hayes,W.S. and Borodovsky,M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res.*, **8**, 1154–1171.
9. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
10. Huang,S.H., Chen,Y.H., Kong,G., Chen,S.H., Besemer,J., Borodovsky,M. and Jong,A. (2001) A novel genetic island of meningitic *Escherichia coli* K1 containing the ibeA invasion gene (GimA): functional annotation and carbon-source-regulated invasion of human brain microvascular endothelial cells. *Funct. Integr. Genomics*, **1**, 312–322.
11. Censini,S., Lange,C., Xiang,Z., Crabtree,J.E., Ghiara,P., Borodovsky,M., Rappuoli,R. and Covacci,A. (1996) cag, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc. Natl Acad. Sci. USA*, **93**, 14648–14653.
12. Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
13. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). *Science*, **296**, 79–92.
14. Mills,R., Rozanov,M., Lomsadze,A., Tatusova,T. and Borodovsky,M. (2003) Improving gene annotation of complete viral genomes. *Nucleic Acids Res.*, **31**, 7041–7055.
15. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.