GeneMarkS-2: Raising Standards of Accuracy in Gene Recognition

Alexandre Lomsadze[1^], Shiyuyun Tang[2^], Karl Gemayel[3^] and Mark Borodovsky[1,2,3]
^ joint first authors

[1]Wallace H. Coulter Department of Biomedical Engineering, [2]School of Biological Sciences,
[3]School of Computational Science and Engineering, Georgia Tech, Atlanta, GA 30332, USA

Abstract

Motivation: *Ab initio* gene prediction in prokaryotic genomes is supposed to be so accurate that RNA-Seq data are rarely produced to bring in an additional layer of evidence. In 2016 more than 60,000 prokaryotic genomes were re-annotated by the NCBI pipeline. Given the sheer volume of prokaryotic DNA data flowing from next generation sequencing facilities into genome databases, the annotation accuracy should be at the highest level possible. Still, the prevalence of horizontal gene transfer as well as ubiquitous leaderless transcription observed in prokaryotic species call for introducing more complex models of genes and regulatory regions than it was thought to be sufficient earlier.
Results: We describe a new algorithm and software tool GeneMarkS-2. The new multi-model tool has an option to select parameters best matching local genomic GC content that may vary widely due to horizontal gene transfer. Genomes are automatically classified by the inferred types of organization of gene starts neighborhoods which evolution is directed by species specific transcription and translation mechanisms. A new motif search algorithm, LFinder, introduced to reach higher accuracy in detecting conserved motifs in regulatory regions upstream to predicted gene starts uses objective function depending on motif localization. In performance assessments made on test sets validated by proteomics experiments and other sources of evidence we have demonstrated superior accuracy of GeneMarkS-2 in comparison with other state-of-the-art gene prediction tools including GeneMarkS which "plus" version is currently used by the NCBI prokaryotic genome annotation pipeline.
Availability: http://topaz.gatech.edu/GeneMark/genemarks2.cgi
Contact: borodovsky@gatech.edu

Introduction

The number of microbial species on Earth is estimated to reach beyond 10^{14} (Locey and Lennon 2016), therefore, currently observed exponential growth of prokaryotic genome databases is likely to continue for quite a while. Structural annotation of a new genome relies both on mapping known proteins to genome as well as on *ab initio* gene finding. Undoubtedly, the search for new microbial species will continue to produce genomes with large fractions of genes not showing similarities to previously known ones.

Accurate a*b initio* gene prediction in prokaryotes is hampered by presence of frequent gene overlaps, ubiquitous genes with atypical composition, and multiple candidates for gene starts. Current gene finding tools, GeneMarkS, Glimmer and Prodigal are sufficiently precise (Besemer et al. 2001; Delcher et al. 2007; Hyatt et al. 2010). The gene prediction accuracy in terms of pinpointing the reading frame, and the 3' end is characterized by ~ 3% of errors (the false negative rate estimated on sequences with verified genes). On the other hand, estimation of the false positive rate requires verification that a predicted gene

is not real. These evaluations are harder to get. Finally, the accuracy of translation start site prediction are estimated in the range of 80-95% (Hyatt et al. 2010).

We describe GeneMarkS-2, the new gene finder that combines and develops further machine learning approaches introduced and implemented in the *ab initio* gene finders GeneMarkS (Besemer et al. 2001) and MetaGeneMark (Zhu et al. 2010). The architecture of generalized HMM of GeneMarkS was expanded to account for laterally transferred genes in wide range of GC content (from 30% to 70%). This expansion was made by replacement of a single model for 'atypical' genes by a whole host of 'atypical' gene models validated earlier in MetaGeneMark for bacterial and archaeal genomes (Zhu et al. 2010). Also, we have addressed the issue of species specific variability in organization of the gene start neighborhoods that play critical role in control of gene expression. Iterative unsupervised estimation of model parameters is now combined with inference of the type of organization of regulatory regions. A new objective function for motif detection algorithm was introduced. The new algorithm of the Gibbs sampler type, LFinder, was shown to make a more precise identification of RBS and promoter motifs. The accuracy of the new tool was assessed on sequences containing genes validated by proteomics as well as by other types of external evidence that indicate either gene presence or gene absence. The results show that the performance of the new tool is superior in comparison with other state-of-the-art gene finders.

Methods

Genome modeling

We observed significant differences between prokaryotic species in organization of sequences around gene starts; these differences reflect variations in transcription and translation mechanisms. Majority of prokaryotic species rarely use leaderless transcription; thus each gene is preceded by a ribosome binding site (RBS). In another group of species transcription of the first-genes-in-operons or stand-alone genes is leaderless; transcripts of these genes do not have space for RBS sites. There are also species (e.g. *M. tuberculosis*) where only a subset of first-genes-in-operons have leaderless transcription. In genomes of all types we observed RBS sites in front of *internal* genes of operons.

Genomic sequence can be efficiently described by the generalized hidden Markov models (GHMM). In GeneMarkS-2 we expand the model introduced for GeneMarkS (Besemer et al. 2001; Lukashin and Borodovsky 1998). A new design of the model of a gene (Fig.1) allows for two alternative signals situated closely to gene start: a ribosome binding site (RBS) and a promoter box (in case of leaderless transcription). The species specific models of these two signals include a positional Markov model of fixed length and a variable length spacer. Also, we detected a species specific frequency pattern, the *upstream signature,* in the short sequence proximal to the translation start (sometimes as short as 3nt). Further on, the gene start model includes the species specific model of the start codon. Finally, there is the *downstream signature*, the species specific model of a sequence (that could be as long as 12nt) located immediately after the start codon (Fig. 1).

The three-periodic Markov model of a protein-coding region in GeneMarkS-2 has multiple types, one 'typical' (species specific) and eighty-two 'atypical'. Parameters of the 'typical' model describing the majority of genes with species specific oligonucleotide composition are estimated in iterative self-training. The fifth order 'atypical' models are pre-designed for forty-one equal intervals of GC content in the range from 30% to 70% separately for bacterial and archaeal domains (Zhu et al. 2010). These 'atypical' models

are able to describe minority fraction of the genes assumed to be horizontally transferred to the genome in question along the path of the species evolution. The method of generating the GC dependent models of protein-coding and non-coding sequences was described earlier (Zhu et al. 2010). This method was demonstrated to deliver effective parameters for gene finding in anonymous metagenomic sequences with a wide range of GC content.

Finally, the score of a potential gene depends on the ORF length. The frequency of a length of protein coding sequence obeys a species independent distribution known for prokaryotic genomes; this distribution is conventionally approximated by the gamma function (Lukashin and Borodovsky 1998).

Unsupervised training

The principle of the iterative training procedure is to conduct rounds of sequence labeling into coding and non-coding regions and to estimate parameters necessary for running the labeling algorithm from the sets of sequences with uniform labels assigned in the previous iteration. The difficult issue of starting this cycle is addressed by introducing special 'heuristic' parameters that are in fact our sets of pre-designed 'atypical' models of the fifth order. At the *initialization* step these models are used in the Viterbi algorithm implemented in the log-odds space. As a result of the run, the sections of genomic sequence are labelled as protein-coding and non-coding (intergenic) regions.

The labeling determined in the *initialization* step starts the main cycle of iterative training (Fig. 2). Particularly, the first labeling allows to select a set of genes that appear to be the first-genes-in-operons. This category is assigned to a predicted gene "X" if the adjacent upstream gene is either located in the complementary strand or ends upstream of "X" at a distance larger than 40nt. The upstream sequences of thus selected genes are used to detect the common conserved regulatory motif. Predicted instances of this motif located in front of the genes are compared with the sequence of conserved 16S RNA tail. If more than 50% of the alignments produce sufficiently good scores then the genome is classified as the one with predominant presence of RBS sites (e.g. *Escherichia coli*), otherwise the search for promoter motif may bring in the conclusion that the leaderless mode of transcription is dominant, the leaderless "A" case (e.g. *Halobacterium salinarum*) or that the leaderless mode of transcription is present in a moderate degree (e.g. *Mycobacterium tuberculosis*).

After the *initialization* step, the training procedure continues as follows. All the regions labeled as protein-coding (but those shorter than 300nt) are used in training to estimate parameters of the 'typical' model (Fig. 2). The next run of the log-odds Viterbi algorithm (see details in Suppl. Materials) uses the newly defined 'typical' model in parallel with the whole set of 'atypical' ones. If a given ORF is predicted as coding by some 'atypical' model (with the score S being above the threshold) and not predicted by a 'typical' model, the ORF is excluded from the training of the 'typical' model or a model of intergenic region. All ORFs predicted by the 'typical' model are included into the training of the next version of 'typical' model regardless of being predicted by any 'atypical' model as well. Note that the ORF score includes the gene start score with upstream and downstream signatures as well as the CDS score; the odds-ratio normalization is made with respect to the second order model of non-coding sequence which parameters are estimated on the set of sequences labeled as non-coding.

The *main cycle* of iterative training repeats parameter re-estimation and genome labeling until less than 1% of new labels change in comparison with the labels assigned in the previous iteration (the convergence

condition). The step at which convergence is reached is the last training step. All the segments labeled as coding regions at this step are accepted as the final gene predictions, the output of the algorithm.

A new motif finder

Accuracy of gene start prediction improves if evolutionary conserved sequences located close to gene starts are taken into account. The MCMC motif finder Gibbs3 (Thompson et al. 2003) was designed to learn a probabilistic model of unknown motif present in a set of sequences. It reaches its goal by iterative construction of multiple sequence alignment with respect to predicted motif locations. Gibbs3 was shown to work reasonably well for the RBS model delineation in GeneMarkS. However, its performance deteriorates with increase of the genome GC content (e.g. in *M. tuberculosis*).

Localization of motif with respect to gene start appears to have its own pattern that could be learned. The presence of this pattern is suggested by the mechanism of the action of a ribosome (or RNA polymerase, in the leaderless transcription case). Close proximity of the motif to the gene start facilitates the gene expression. However, the Gibbs3 algorithm does not use information on the motif localization defined by the length distribution of the sequence between the motif and the gene start (the spacer). The new algorithm, LFinder, adds localization measure to the objective function of the Gibbs sampling algorithm. As a result, LFinder is able to avoid motif locations deviating significantly from the presumed gene start (See Suppl. Materials for the details of the LFinder algorithm). LFinder runs a fixed number of iterations N (default N=110). Multiple executions of the algorithm with different starting points improve the performance. The new objective function enables LFinder to outperform Gibbs3 (see Results)

Materials

Sets of genes supported by proteomic data

Mass-spectrometry-determined peptides were mapped to genomes of a number of prokaryotic species by the Pacific Northwest National Laboratory (Venter et al. 2011). In total 1,209,658 peptide spectrums were mapped by the MS-GF+ software tool to 87,417 ORFs. The quality control of the experiments (Venter et al. 2011) included i/ requirement that the score of the match to genome would have P-value better than 1e-10; ii/ removal of low-complexity peptides; iii/ removal of ORFs lacking uniquely mapped peptides, iv/ requirement that an ORF would have at least two matching peptides separated by less than 750nt distance. The peptide-supported ORFs (psORFs) from 58 species (Table S1) made a test set for assessment of accuracy of gene prediction.

Sets of COG annotated genes and simulated non-coding sequences

In addition to tests on peptide supported ORFs, we worked with genomes of 115 bacteria and 30 archaea available at NCBI (the species selection was made together with the DOE Joint Genome Institute; for the names and RefSeq ID see Table S2). This set spanning 22 bacterial and archaeal phyla featured genomes varied in size, type of genetic code, and GC content. To minimize effects of possible annotation errors, we selected only genes whose functional annotation cited a particular COG name from a database of Clusters of Orthologous Groups (Galperin et al. 2015; Tatusov et al. 1997; Tatusov et al. 2003). Functional annotation with COG affilation provides a strong evidence of evolutionary conservation not expected for a random ORF. A missed in prediction ‘COG gene’ was counted as false negative (FN). For

assessment of frequency of false positive (FP) predictions we used simulated non-coding sequences generated by the second-order Markov model. To train this species-specific model we used genomic sequences *not annotated* as protein-coding genes, RNA genes, or pesudogenes. For each species the model generated ten replicas of 1Mb long non-coding sequence to verify that the results do not show unexpected deviations.

Test sets of genes with experimentally verified starts

The N-terminal protein sequencing is a standard but not frequently used technique to validate sites of translation initiation (protein N-terminals and gene starts). Relatively large sets of genes with validated starts are known for the bacteria *E. coli* (Rudd 2000; Zhou and Rudd 2013), *M. tuberculosis* (Lew et al. 2011), *Synechocystis sp.* strain PCC6803 (Sazuka et al. 1999) and the archaea *H. salinarum*, *Natronomonas pharaonis*, and *Aeropyrum pernix* (Aivaliotis et al. 2007; Yamazaki et al. 2006).

Results of accuracy assessment

Genes supported by proteomics

To assess performance of GeneMarkS, Glimmer3, Prodigal, and GeneMarkS-2 we did run the gene finders with default parameters on the 58 genomes (Table S1) with ~87,000 ORFs supported by proteomics (psORFs). The frequencies of the three types of errors were recorded: i/ missed psORFs (false negative); ii/ false positives, the predictions incompatible with psORFs (fully overlapped with a psORF located in different strand or frame); iii/ errors in start prediction (prediction of genes shorter than they had to be given the evidence from a peptide mapped upstream to predicted start).

We observed that GeneMarkS-2 missed the least number of psORFs; this new tool also generated the least number of false positives (Table 1). Notably, the required full overlap seems to be a rare event in this test setting. Therefore, we complement the test for false positives with the test on simulated sequences (see below). Prodigal produced a lower number of "shorter" predictions. However, one should note that the "longer" predictions cannot be detected by proteomics, thus, the start prediction accuracy assessment made only with regard to shorter genes is biased. The tool that has a tendency to predict longer genes (even one making a blunt prediction of the longest ORFs) would get a better result in the category of "shorter" genes without getting any negative points for predicting too long genes. The test on genes with validated starts (see below) is a more balanced and objective test of performance in gene start prediction.

Finding COG genes as well as "genes" in simulated non-coding sequences

In another round of performance assessment, we focused on COG genes (those longer than 89 nt) annotated in 115 bacterial and 30 archaeal genomes (Table S2). The overall rate of missed COG genes for all the gene finders was less than 2% (Table 2A). The total number of missed genes, 867, was the lowest for GeneMarkS-2, followed by Prodigal with 1350. The GeneMarkS-2 performance was least dependent on genome GC content (data not shown).

We also assessed how the performance depends on gene length. The COG genes were divided into groups with length in five intervals starting with minimal length of 90 nt (Table 2A). Glimmer made lower number of "missed" calls for short genes (90nt-150nt range) as compared to the other tools. Still, this

effect came at a cost of significant increase in numbers of false predictions in simulated sequences (Table 2B). GeneMarkS-2 made the least number of "missed" calls for the COG genes in all the other bins. At the same time it demonstrated quite robust performance in terms of false positives generated in simulated sequences.

In the tests with simulated sequences each gene finder was run with its species specific parameters. The error rate was defined as the ratio of the number of predicted "genes" to the total number of random ORFs (longer than 90nt). The test was repeatede ten times for each species. Notably, the numbers of ORFs in a simulated sequence depends on its GC content and is lower for low and high GC composition while reaching maximum at about 58-65% GC. Also, a simulated sequence with high GC (up to 65%) contains more long ORFs than sequences with lower GC. All over, GeneMarkS-2 was observed to have a significantly lower error rate, e.g. more than 50% lower than the second best tool, Prodigal (Table 2B). The increase of false positive rate of Prodigal in high GC sequences which carry longer ORFs more frequently, is likely to be related to the tendency for predicting longer ORFs as genes.

Similarly to analysis done with COG genes, we grouped ORFs predicted in simulated non-coding sequences into five groups depending on length (Table 2B). GeneMarkS-2 demonstrated consistantly better performance than Prodigal in all the length intervals. Glimmer perfomed best in the range 300 – 600 nt, but did make large number of errors in other length intervals especially in the range below 150 nt. Prodigal, in contrast, has shown an increase in false prediction of ORFs longer than 300nt.

The gene start prediction

Assessment of performance was done on the earlier described sets of genes with experimentally verified starts available for *A. pernix, E. coli, H. salinarum, M. tuberculosis, N. pharaonic* and *Synechocystis sp.* (Table 3). We observed improvement in GeneMarkS-2 gene start prediction in comparison with GeneMarkS. As the result GeneMarkS-2 made correct predictions for the largest number of starts in the verified set among all the gene finders.

Here we want to describe a few instructive cases (Fig. 3). The archaeal *H. salinarum* genome, was classified as leaderless class "A" with RBS signal missing in the first-genes-in-operons and promoter for leaderless transcription located at a distance 22-24nt from the translation start (Fig. 3B). Still the RBS sites for the internal genes in operons are present and located in the distance 6-8 nt (Fig. 3C). Original GeneMarkS identified only the promoter signal with less pronounced localization (Fig. 3A). The bacterial *M. tuberculosis* genome was classified as leaderless class "B". Previously, in GeneMarkS Gibbs3 attempted but failed to find a correct RBS motif for *M. tuberculosis* (Fig. 3D). Application of GeneMarkS-2 shows that in *M. tuberculosis* some first-genes-in-operon are transcribed in leaderless fashion, with promoter signals located at 6-8 nt from the gene starts (Fig. 3F), while transcripts of other genes have sufficient space for the RBS sites located at 6-8nt distance from the gene starts (Fig. 3E)

Discussion

Improvement in performance demonstrated in several tests justifies introduction of the new features in the GeneMarkS-2 algorithm. The array of 'atypical' models improves prediction of horizontally transferred genes that appear in a given genome in a rather small number (e.g. less than 15% of genes in *E. coli* K12 genome). Importantly, the difference in GC content between 'typical' and 'atypical' genes could be as

high as 20%. Most of the time, the GC contents of 'atypical' genes were observed to be lower than GC content of 'typical' ones; the difference is less pronounced in low GC genomes than in high GC genomes where the space for downward variation is larger. Existence of several gene classes in terms of differences in codon usage has been known for a while (Borodovsky et al. 1995). Nevertheless, the two gene finders considered here, Glimmer and Prodigal, use single (typical) model for predicting all the genes.

The idea of the multi-model approach could be illustrated as follows. With some stretch of imagination, by disregarding linear connectivity of genes in a given genome, one could think about these 'disjoint' genes as sequences of a small 'metagenome'. A tool we have developed earlier for metagenome analysis, MetaGeneMark (Zhu et al. 2010), appears to be a natural source of models for analysis of such a collection of sequences showing some variability in GC content. The sequences of 'typical' genes could be clustered and processed together to derive more accurate 'typical' model. The remaining small number of genes deviated in composition from the bulk of the genome are targets of the best matching 'atypical' models. A value added feature of this approach is classification of atypical genes into bacterial and archaeal (such a separation of models was described in the MetaGeneMark publication). The insight into possible origin of a horizontally transferred gene is particularly useful for genomes of thermophilic bacteria and mesophilic archaea.

The concept of 'heuristic' modeling as an approach to parameterization of a model suitable for analysis of a short sequence segment isolated from genomic context was proposed in 1999 (Besemer and Borodovsky 1999); with arrival of many more genomes this method was extended to GC specific high order models in MetaGeneMark (Zhu et al. 2010) in 2010.

Necessity of automatic classification of genomes with respect to the types of gene start organization was articulated already in the GeneMarkS publication (Besemer et al. 2001). GeneMarkS was used to identify genes in *Pyrobaculum aerophilum* the species with ubiquitous leaderless transcription (Slupska et al. 2001). Improvement in automatic gene start modeling along with more accurate gene start predictions makes possible to generate useful predictions on structure and evolution of elements of transcriptome. Notably, there has been a good agreement between sets of the *M. tuberculosis* genes predicted to be leaderless by GeneMarkS-2 and detected to be leaderless in RNA-Seq experiment (Cortes et al. 2013).

Now, after one more time pushing the limits of *ab initio* gene prediction, we still have to recognize that the accuracy did not reach 100%. There are genes that still escape recognition, e.g. genes significantly biased in higher order oligonucleotide composition or genes that are corrupted by frameshifts. Genes with frameshifts present a challenge in terms of a need of annotation of all their fragments even those that are *not ending* at a standard stop codon. There is a gray area of pseudogenes, especially expressed pseudogenes, that would distract gene finding tools to generate predictions that have to be appropriately classified. When an orthologue of such a gene is present in the database, the frameshift identification can be done rather easily.

A version of GeneMarkS known as GeneMarkS+ is used in the latest version of the NCBI pipeline as integrator of several types of evidence into genome annotation (Tatusova et al. 2016). To extend GeneMarkS-2 to the "plus" version is the next obvious step. Running time of GeneMarkS-2 is currently ~6 minutes on genome of the *E. coli* size. Further optimization of the speed is another task on the agenda of this group.

SUPPLEMENTARY DATA

Supplementary Data are available at:
http://topaz.gatech.edu/GeneMark/genemarks2.cgi

FUNDING

Wallace H. Coulter Department of Biomedical Engineering, School of Biological Sciences, School of Computational Science and Engineering at Georgia Tech.

Conflict of interest statement. None declared.

REFERENCES

Aivaliotis, M., et al. (2007) 'Large-scale identification of N-terminal peptides in the halophilic archaea Halobacterium salinarum and Natronomonas pharaonis', *Journal of Proteome Research,* 6 (6), 2195-204.

Besemer, J. and Borodovsky, M. (1999) 'Heuristic approach to deriving models for gene finding', *Nucleic Acids Res,* 27 (19), 3911-20.

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 (12), 2607-18.

Borodovsky, M., et al. (1995) 'Detection of new genes in a bacterial genome using Markov models for three gene classes', *Nucleic Acids Res,* 23 (17), 3554-62.

Cortes, T., et al. (2013) 'Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*', *Cell Rep,* 5 (4), 1121-31.

Delcher, A. L., et al. (2007) 'Identifying bacterial genes and endosymbiont DNA with Glimmer', *Bioinformatics,* 23 (6), 673-9.

Galperin, M. Y., et al. (2015) 'Expanded microbial genome coverage and improved protein family annotation in the COG database', *Nucleic Acids Res,* 43 (Database issue), D261-9.

Hyatt, D., et al. (2010) 'Prodigal: prokaryotic gene recognition and translation initiation site identification', *BMC Bioinformatics,* 11, 119.

Lew, J. M., et al. (2011), 'TubercuList-10 years after', *Tuberculosis,* 91 (1), 1-7.

Locey, K. J. and Lennon, J. T. (2016) 'Scaling laws predict global microbial diversity', *Proc Natl Acad Sci U S A,* 113 (21), 5970-5.

Lukashin, A. V. and Borodovsky, M. (1998) 'GeneMark.hmm: new solutions for gene finding', *Nucleic Acids Res,* 26 (4), 1107-15.

Rudd, K. E. (2000) 'EcoGene: a genome sequence database for *Escherichia coli* K-12', *Nucleic acids research,* 28 (1), 60-64.

Sazuka, T., Yamaguchi, M., and Ohara, O. (1999) 'Cyano2Dbase updated: Linkage of 234 protein spots to corresponding genes through N-terminal microsequencing', *Electrophoresis,* 20 (11), 2160-71.

Slupska, M. M., et al. (2001) 'Leaderless transcripts of the crenarchaeal hyperthermophile *Pyrobaculum aerophilum*', *J Mol Biol,* 309 (2), 347-60.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997) 'A genomic perspective on protein families', *Science,* 278 (5338), 631-7.

Tatusov, R. L., et al. (2003) 'The COG database: an updated version includes eukaryotes', *BMC Bioinformatics,* 4, 41.

Tatusova, T., et al. (2016) 'NCBI prokaryotic genome annotation pipeline', *Nucleic Acids Res,* 44 (14), 6614-24.

Thompson, W., Rouchka, E. C., and Lawrence, C. E. (2003) 'Gibbs Recursive Sampler: finding transcription factor binding sites', *Nucleic Acids Res,* 31 (13), 3580-5.

Venter, E., Smith, R. D., and Payne, S. H. (2011) 'Proteogenomic analysis of bacteria and archaea: a 46 organism case study', *PLoS One,* 6 (11), e27587.

Yamazaki, S., et al. (2006) 'Proteome analysis of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum perni*x K1', *Mol Cell Proteomics,* 5 (5), 811-23.

Zhou, J. D. and Rudd, K. E. (2013) 'EcoGene 3.0', *Nucleic Acids Research,* 41 (D1), D613-D24.

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010) 'Ab initio gene identification in metagenomic sequences', *Nucleic Acids Res,* 38 (12), e132.

FIGURES and TABLES

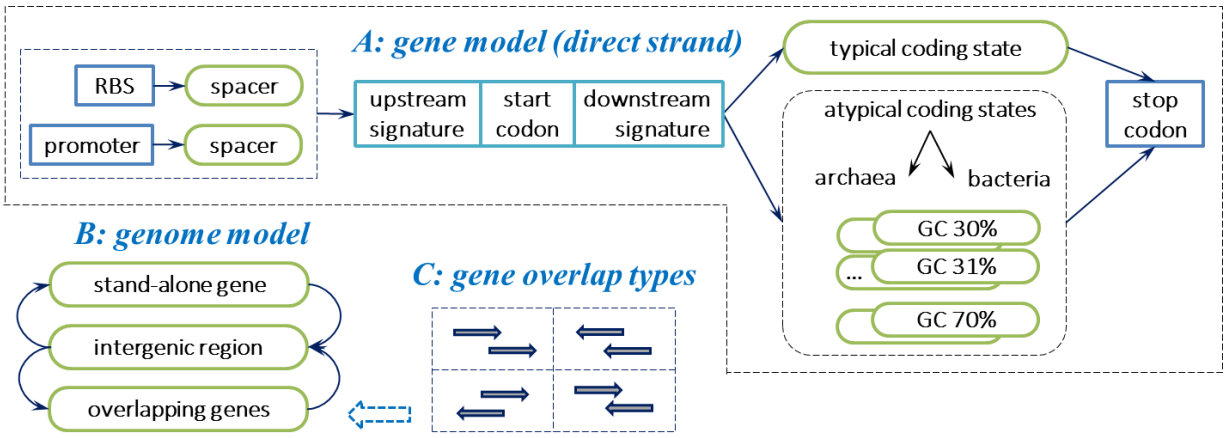

Figure 1. Principal state diagram of the GHHM of prokaryotic genomic sequence. States modeling a gene in direct strand are shown in **1A**. The arrows designate transitions between the states. Genes in reverse strand are modeled by identical set of states with directions of arrows reversed. The reverse strand states are connected to the direct strand states through the intergenic region state and states for genes overlaps in opposite strands (**1B** and **1C**).

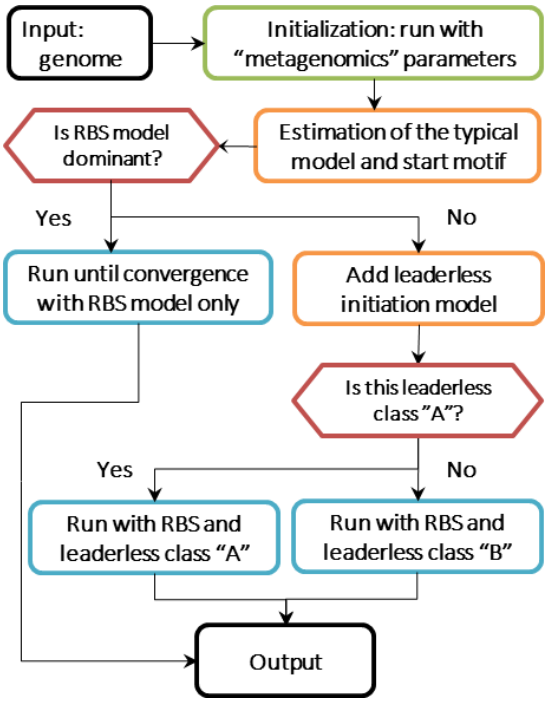

Figure 2. Principal workflow of the unsupervised training.

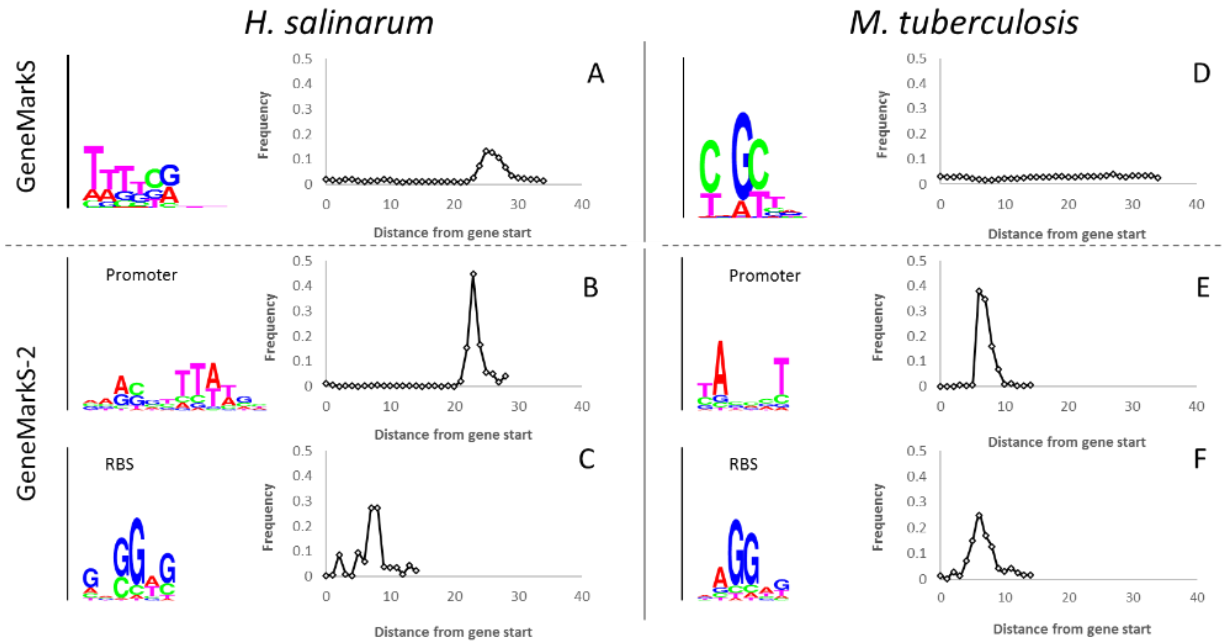

Figure 3. Visualizations of the motif models (logos) and spacer length distributions for genomes of *H. salinarum* (leaderless class "A") and *M. tuberculosis* (leaderless class "B"). Motifs found by GeneMarkS are shown in **3A** and **3D** respectively. GeneMarkS-2 first selects first-in-operons genes and detects the type of signal (non-RBS in both genomes). In *H. salinarum* first-in-operons genes have a promoter signal at a large distance, with no RBS due to leaderless transcription (**3B**). In *M. tuberculosis* only a fraction of first-in-operons genes have leaderless transcription with promoter signal on much closer distance to gene starts (**3E**). In both genomes genes located inside operons have RBS sites (**3C** and **3F** respectively).

Type of error	GeneMarkS	Glimmer	Prodigal	GeneMarkS-2
Missed ORF	375	522	230	**146**
False prediction	250	568	148	**102**
Short prediction	3,413	3,401	**1,338**	1,576

Table 1 Results of gene finding performance assessment on the 58 genomes with annotated peptide-supported ORFs (Venter et al, 2011).

A	Bins (in bp):	< 150	150-300	300-600	600-900	> 900
Algorithm	COG genes:	355	13,574	63,986	81,231	171,457
	Total missed	missed				
GeneMarkS	1,468	130	477	410	175	276
Glimmer	2,053	**61**	536	812	340	304
Prodigal	1,350	157	625	409	85	74
GeneMarkS-2	**867**	90	**404**	**260**	**55**	**58**

B	Bins (in bp):	< 150	150-300	300-600	600-900	> 900
Algorithm	Total FP	false positives (FP) in simulated sequence				
GeneMarkS	39,321	15,782	20,548	2,947	43	**1**
Glimmer	74,061	62,239	10,928	**428**	163	303
Prodigal	23,919	4,518	9,530	3,237	4,706	1,928
GeneMarkS-2	**9,541**	**4,067**	**4,355**	1,096	**21**	2

Table 2. Results of gene finding performance assessment. Panel **A**: Statistics of observed false negatives (missed genes) for the 145 prokaryotic genomes with gene annotation showing COG affiliation. Panel **B**: Statistics of observed false positives for 1Mb non-coding sequences generated by species-specific model of non-coding region (145 simulated sequences).

Species (*archaea)	Genome class	# of verified starts	GeneMarkS	Glimmer	Prodigal	GeneMarkS-2
*A. pernix**	majority RBS	130	125	119	**127**	125
E. coli	majority RBS	769	725	714	**751**	744
*H. salinarum**	leaderless & RBS	530	501	457	514	**522**
M. tuberculosis	leaderless & RBS	701	572	572	**620**	**620**
*N. pharaonis**	leaderless & RBS	315	308	293	309	**312**
Synechocystis	minority RBS	96	82	79	**92**	91
	Total:	2,541	2,313	2,234	2,413	**2,414**

Table 3 Numbers of gene starts predicted correctly in the sets of genes verified by N-terminal sequencing. The classes of genomes are defined with respect to the types of signals for the first-in-operon genes. "Majority RBS" indicates that the RBS in ubiquitously present in upstream regions of such genes. "Leaderless & RBS" indicates significant present of leaderless transcription. "Minority RBS" in *Synechocystis* presents the special case when in absence of leaderless transcription, the RBS signal is detected in less than 50% of all genes.