Improved Prokaryotic Gene Prediction Yields Insights into Transcription and Translation Mechanisms on Whole Genome Scale

Alexandre Lomsadze[1^], Karl Gemayel[2^], Shiyuyun Tang[3^] and Mark Borodovsky[1,2,3,4*]
^ joint first authors, *corresponding author, borodovsky@gatech.edu

[1]Wallace H. Coulter Department of Biomedical Engineering, [2]School of Computational Science and Engineering, [3]School of Biological Sciences, Georgia Tech, Atlanta, Georgia, 30332, USA, [4]Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Moscow, Russia

Abstract
While computational gene finders for prokaryotic genomes have reached a high level of accuracy, there is room for improvement. GeneMarkS-2, a new *ab initio* algorithm, aims to improve prediction of species-specific (native) genes, as well as difficult-to-detect genes that differ in composition from the native genes. We introduce an array of pre-computed heuristic models that compete with the iteratively learned native model for the best fit within genomic neighborhoods that deviate in nucleotide composition from the genomic mainstream. Also, in the process of self-training, GeneMarkS-2 identifies distinct sequence patterns controlling transcription and translation. We assessed the accuracy of current state-of-the-art gene prediction tools along with GeneMarkS-2 on test sets of genes validated by proteomics experiments, by COG annotation, as well as by protein N-terminal sequencing. We observed that, on average, GeneMarkS-2 shows a higher precision in all accuracy measures. Screening of ~5,000 representative prokaryotic genomes reveals frequent leaderless transcription, not only in archaea where it was originally discovered, but in bacteria as well. Furthermore, species with prevalent leadered transcription do not necessarily use RBS sites with the Shine-Dalgarno consensus. The effort to distinguish leaderless and leadered transcription, depending on prevalence of one or the other, leads to classifying prokaryotic genomes into five groups with distinct sequence patterns around gene starts. Some of the observed patterns are apparently related to poorly characterized mechanisms of translation initiation.

[Supplemental material is available for this article].

Key words: gene finding, atypical genes, leaderless transcription, non-SD ribosomal binding site

Running title: Improved Gene Prediction Yields Biological Insights

Introduction

The number of microbial species on Earth was estimated to be of the order of 10^{12} (Locey and Lennon 2016). Therefore, the exponential growth of the number of sequenced prokaryotic genomes, currently ~10^5, is likely to continue for quite a while. To generate structural annotation of a new genome, one could find intervals containing mapped footprints of known proteins and fill in the gaps with predictions from an *ab initio* gene finding algorithm. Given the size of the microbial universe, the search for new microbes will continue to produce genomes with large numbers of genes that are not detectable by mappings protein orthologs. Therefore, improving the accuracy of *ab initio* gene prediction remains an important task.

The proliferation of RNA-Seq presented an opportunity for more accurate inferring of exon-intron structures of eukaryotic genes. Transcriptomes of prokaryotes were thought to be less important for gene finding since the accuracy of *ab initio* prediction of a whole gene (uninterrupted genomic ORF) is significantly higher. Nevertheless, recently introduced modifications of NGS techniques started to generate new kinds of transcript data which impact is yet to be fully appreciated.

For example, the differential RNA sequencing (dRNA-Seq) technique (Sharma et al. 2010; Sharma and Vogel 2014) aims to accurately detect transcription start sites (TSS). Data on TSS locations can be used to verify the annotation of operons, which in turn can help with the detection of the promoter signals as well as translation initiation starts (TIS), particularly in the case of leaderless transcription (Creecy and Conway 2015).

The sequence around a TIS exhibits specific nucleotide patterns that code for effective interactions between mRNA and the translation machinery. In bacteria and archaea, translation initiation is generally thought to occur through the base-pairing interaction between the 3' tail of the 16S rRNA of the 30S ribosomal subunit and the site in the 5' UTR of an mRNA that carries a pattern consistent with the Shine-Dalgarno (SD) consensus (Shine and Dalgarno 1974; Barrick et al. 1994). Frequently in some species and less frequently in others, the transcripts may have very short 5' UTRs (with length < 6 nt) unable to host the ribosome binding site (RBS). Such a mode of transcription, known as *leaderless transcription*, would situate the TSS at or very near to the TIS. In this case, the promoter signal, the TATA box in archaea or the Pribnow box in bacteria with consensus TATAAT, located in close proximity of the TSS, could be used for more accurate TIS identification.

Frequent leaderless transcription was first discovered in the archaea *Pyrobaculum aerophilum* (Slupska et al. 2001). Since then, studies of prokaryotic transcriptomes, including the dRNA-Seq applications detected instances of leaderless transcription in several species of archea and bacteria. Importantly, the fraction of genes with leaderless transcription was observed to vary significantly among species. It was low (<8%) in some bacteria e.g. *Helicobacter pylori* (Sharma et al. 2010), *Bacillus subtilis* (Nicolas et al. 2012), *Salmonella enterica* (Kroger et al. 2013), *Bacillus licheniformis* (Wiegand et al. 2013), *Campylobacter jejuni* (Dugar et al. 2013), *Propionibacterium acnes* (Lin et al. 2013), *Shewanella oneidensis* (Shao et al. 2014), and *Escherichia coli* (Thomason et al. 2015), as well as (<15%) in some archaea e.g. *Methanosarcina mazei* (Jager et al. 2009), *Pyrococcus abyssi* (Toffano-Nioche et al. 2013), *Thermococcus kodakarensis* (Jager et al. 2014), *Methanolobus psychrophilus* (Li et al. 2015), and *Thermococcus onnurineus* (Cho et al. 2017). However, higher frequency of leaderless transcription was observed (> 25%) in bacteria, e.g. *Mycobacterium tuberculosis* (Cortes et al. 2013), *Corynebacterium glutamicum* (Pfeifer-Sancar et al. 2013), *Deinococcus deserti* (de Groot et al. 2014), *Streptomyces coelicolor* (Romero et al. 2014), *Mycobacterium smegmatis* (Shell et al. 2015), and even larger (>60%) in archaea e.g. *Halobacterium salinarum* (Koide et al. 2009), *Sulfolobus solfataricus* (Wurtzel et al. 2010), and *Haloferax volcanii* (Babski et al. 2016).

Current gene finding tools, GeneMarkS, Glimmer3, and Prodigal, have sufficiently high accuracy (>97% on average) in detecting validated protein-coding ORFs (Besemer et al. 2001; Delcher et al. 2007; Hyatt et al. 2010). Though the accuracy in pinpointing the ORFs starts could be lower ~90% (Hyatt et al. 2010) The genes that escape detection altogether are mostly in the atypical category; i.e. genes with sequence patterns not matching the species specific model trained on the bulk of the genome (Borodovsky et al. 1995). The other important accuracy measure, the false positive rate, requires verification that a predicted gene is *not* real; this assessment is difficult to make. In this study, we identify false positives by as (i) predicted genes with an unrealistically large overlap with a confirmed gene located in the opposite strand and (ii) genes predicted in a randomly generated sequence. Separately, the accuracy of gene start prediction was traditionally assessed on sets of genes verified by the protein N-terminal

sequencing. Given the high accuracy of the current tools the task to improve the predictive power of prokaryotic gene finding i.e. to achieve better detection of atypical genes and better gene start prediction is challenging.

In GeneMarkS-2 we utilized and augmented the algorithmic features introduced and implemented in GeneMarkS (Besemer et al. 2001) and MetaGeneMark (Zhu et al. 2010). We expanded the generalized hidden Markov model (GHMM) of GeneMarkS by addition of multiple states accounting for horizontally transferred genes with wide range of GC compositions. We replaced the single heuristic model used in GeneMarkS for recognition of *atypical* genes by an array of heuristic models first introduced in MetaGeneMark and covering the GC range from 30% to 70% (Zhu et al. 2010). In GeneMarkS-2, this set of heuristic (*atypical*) models complements the major model trained on the whole genome and tuned up for the *typical* genes (Besemer et al. 2001).

We revised the GeneMarkS procedure of unsupervised training to account for variability of genome-specific features of transcription and translation that define regulatory sites located close to gene starts. To improve the parameterization of the models of regulatory sites (RBS or promoter boxe), we implemented GibbsL (Gibbs with Localization) a new type of the Gibbs sampler algorithm (Lawrence et al. 1993). We included the distance (the length of the *spacer*) between the *predicted site* and the *anchor point* (e.g. gene start) into the objective function of GibbsL. In a given genome GibbsL operates independently on sets of genes with leadered and leaderless transcription to derive separate models of RBS or promoter boxes.

We assessed the accuracy of several gene finders, including GeneMarkS-2. Since the GenBank genome annotation cannot yet be consider a uniformly reliable (gold) standard, we opted to use subsets of genes validated by one or another external information. Particularly, the genes validated by proteomics experiments, genes validated by the similarity search to the COG annotated genes, on genes with starts determined by N-terminal sequencing. Also, we used sequences simulating genome specific non-coding regions, the sequences that were supposed to contain no genes. The results of the benchmarking experiments demonstrated that, on average, GeneMarkS-2 made more accurate predictions than other existing tools.

Since in process of self-training GeneMarkS-2 had to determine the type of the sites located close to the gene starts, it produced insights into types and frequencies of the transcription and translation related regulatory sites. In many out of ~5,000 representative prokaryotic genomes (Tatusova et al. 2014) we predicted frequent leaderless transcription; transcriptomes of some of these species have already been studied with dRNA-Seq and we compare predictions and observations.

GeneMarkS-2 classified the ~5,000 genomes into five groups, such as archaeal and bacterial genomes with frequent leaderless transcription, genomes with RBS sites of Shine-Dalgarno (SD) type, as well as genomes with non-SD RBS sites. Finally, the there was a group (including e.g. cyanobacteria) that carry very weak patterns in 5' UTR; only a minority of genes in genomes of this group used the SD RBS sites.

Methods

Gene and Genome Modeling

Model of a protein-coding sequence. Protein-coding regions in prokaryotic genomes are known to carry species-specific oligonucleotide (e.g. codon) usage patterns (Grantham et al. 1981; Fickett and Tung 1992). Still, oligonucleotide composition of some genes may deviate from the genome-wide mainstream. To account for genes with some degree of compositional deviation we introduced in GeneMarkS the additional atypical gene model (Besemer et al. 2001).

We estimated the parameters of the 'typical' model by iterative self-training on the given anonymous genome, as it was done in GeneMarkS (Besemer et al. 2001). However, instead of the single atypical model, GeneMarkS-2 uses two sets of atypical models: 41 bacterial and 41 archaeal ones. Parameters of the atypical models, covering GC content range from 30% to 70%, were estimated by the method proposed earlier (Zhu et al. 2010). Each atypical model has an index indicating a narrow (1%) GC content bin it represents; the model parameters are pre-computed and remain the same in any run and any iteration of GeneMarkS-2. Only a subset of atypical models is used in a GeneMarkS-2 run on a given genome, the models with the index value within the GC content range of the genome in question.

In addition to the compositional score, the signal score (see below) and the lengths of candidate genes and intergenic sequences are used in the computing of the gene score.

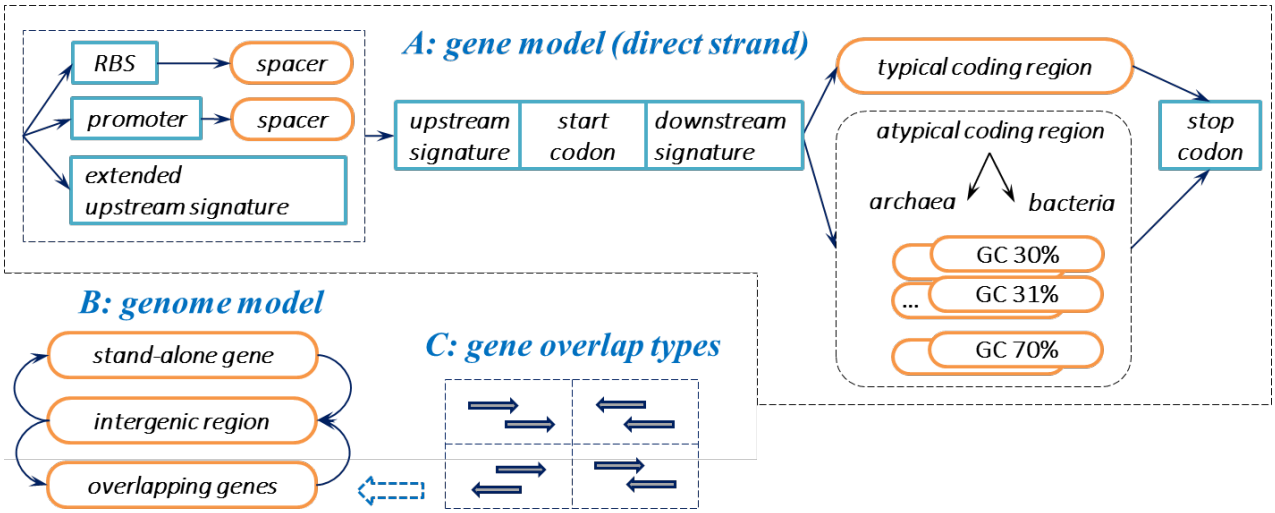

Figure 1. Principal state diagram of the GHHM of prokaryotic genomic sequence. States modeling a gene in the direct strand are shown in 1A. Genes in the reverse strand are modeled by identical set of states with directions of transitions (arrows) reversed. The direct strand states and the reverse strand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

We could interpret the multi-model approach in the following way. If we would disregard for a moment the linear connectivity of genes in a given genome, we could think of this set of 'disjoint' genes as an instance of a small 'metagenome'. The approach developed for metagenome analysis, (Zhu et al. 2010), created a variety of models for analysis of sequence fragments with variety of GC content; we use this library as the set of atypical models. Furthermore, the typical genes making the majority of the whole gene set could be clustered and processed together to derive parameters for the *typical* model. The genes deviated in composition, the minority, still cannot escape detection by the compositionally matching atypical models.

Model of a sequence around gene start. There are characteristic differences in the sequence patterns around gene starts observed in the five just mentioned groups of prokaryotic genomes. Groups A and B exhibit, respectively, the archaeal (-26) and the bacterial (-10) promoter boxes situated upstream to the starts of the genes transcribed in leaderless fashion. Other genes in these genomes may have RBS sites, thus we have *dual gene start models* in the genomes of the A and B groups. Groups C and D where leadered transcription is dominant, carry sequence patterns of a single type; these are the RBS sites with either non-Shine-Dalgarno consensus (group C) or with Shine-Dalgarno consensus (group D). In group E we observed the weak nucleotide patterns that are likely to be related to translation mechanisms not well understood yet.

The sites of promoter boxes as well as the non-Shine-Dalgarno (non-SD) and the Shine-Dalgarno (SD) RBS sites are separated from TIS by variable sequences with variable length (spacers). Therefore, the full model of a regulatory site should include the model of the site-specific sequence pattern, e.g. in a form of the positional Markov chain, as well as the length distribution of the spacer (Fig. 1).

It was also observed that the nucleotide composition of the spacers is not homogeneous, with the part proximal to the translation start exhibiting its own compositional pattern. Therefore, we explicitly model the triplet upstream to the start codon as a 3 nt long site called the *upstream signature*.

Moreover, in the group E genomes we use a 20 nt long *extended upstream signature* (Fig. 1).

Finally, in addition to all the models mentioned above, we also use a *downstream signature* that captures the patterns within the short protein-coding sequence (with length up to 12 nt) located immediately downstream of the start codon (Shmatkov et al. 1999).

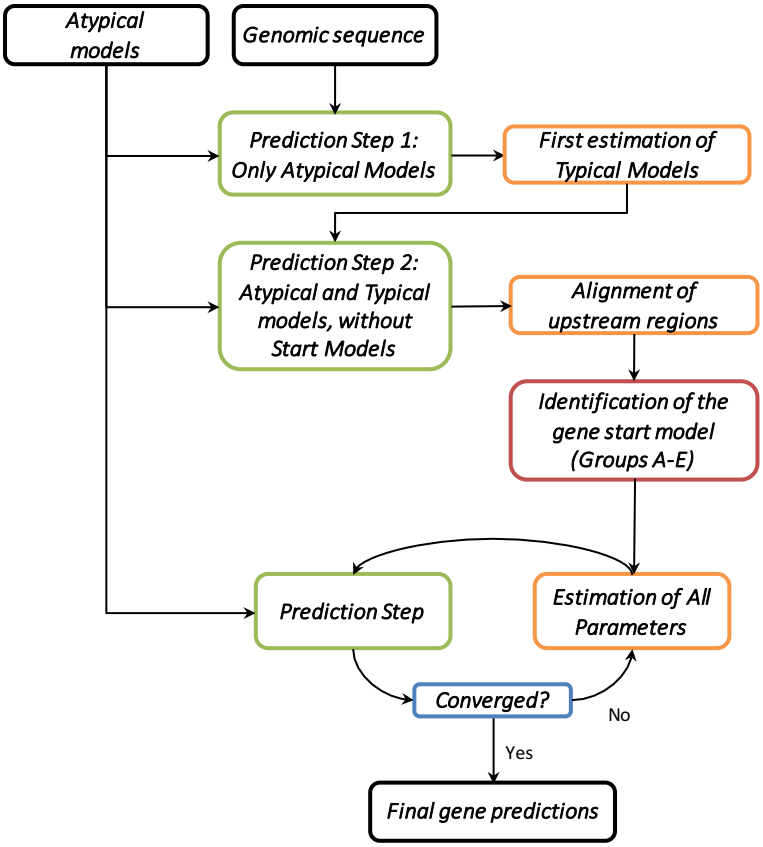

Figure 2. Principal workflow of the unsupervised training.

Unsupervised training

The unsupervised training algorithm makes several two-step iterations (Fig. 2). Each iteration includes i/ genome segmentation into coding and non-coding regions (gene prediction) and ii/ model parameter re-estimation (Besemer et al. 2001).

The first iteration.

Prediction step. The log-odds space Viterbi algorithm (see Suppl. Materials) computes genome segmentation into protein-coding (CDS) and non-coding regions with the maximum value of the log-odds sores and the maximum likelihood. At the first step, the algorithm only uses heuristic (atypical) models. When the Viterbi algorithm runs on a particular segment of the genomic sequence, it utilizes only one of the 82 *atypical* models whose GC index matches the local sequence's GC composition. With this approach, heuristic models are chosen independently for different segments (candidate genes) of the genome. With models in place the algorithm finds the genome segmentation with the maximum value of log-odd scores (the segments could be labeled: coding, reverse coding, gene overlap or non-coding, see Suppl. Materials). Note that, in the first iteration, the Viterbi algorithm does not have means to account for sequence patterns around gene starts since no gene-start model has yet been derived.

Estimation step. After the first run of the Viterbi algorithm, all genomic segments predicted and labeled as 'protein-coding' (CDS) are collected into a training set for the estimation of parameters of the 'typical' model. To decrease the chance of including non-coding ORFs into the training set, predicted genes shorter than 300 nt and those predicted as incomplete CDS are not used in training. From this set of genes, the *typical* model, a 5th order, three-periodic Markov model is derived (Borodovsky and McIninch 1993). Similarly, the genomic segments labeled as 'non-coding' are used to estimate the parameters for the non-coding model, structured as a 2nd order, uniform Markov chain.

The second iteration. At the *prediction step*, the Viterbi algorithm, in addition to the set of 82 *atypical* models, uses the newly derived *typical* model to update the genome segmentation. At the *estimation* step, the updated genome segmentation is used to re-estimate parameters of the *typical* model. Now, having information on initially

predicted gene starts, the algorithm selects sequences situated around the gene starts and derives models of the patterns encoding transcription and/or translation regulation (Fig. S1).

Building the models of sequences around gene starts. The sites near the first genes in operons (FGIOs) are used to regulate transcription and translation (promoters and RBS sites) while the sites located near the interior genes in operons (IGIOs) carry signals related to translation regulation (RBS sites). Given a genome segmentation into protein-coding genes and non-coding regions, we identify FGIOs by the following rule. A gene is an FGIO if the upstream neighboring gene is located either in the complementary strand, or at a distance larger than 25 nt. In the iterations of self-training, the set of FGIOs is updated in each prediction step. We chose the 25 nt threshold after experimenting with values ranging up to 40 nt. Intuitively, larger values offer a more conservative selection of operons and FGIO, since distances between operons tend to be larger than distances between genes within the same operon. An analysis of the annotated operons in the *E. coli* genome (Gama-Castro et al. 2016) shows the effect of this threshold on the accuracy of operon prediction (Fig. S2). The 25 nt cut-off identified 98% of annotated first genes in operons with addition of 8% false positive. Furthermore, while the analysis on *E. coli* suggests that 40 nt might offer a slightly better balance between true and false operon predictions, our experiments show that increasing the value to 40 nt offers an equal (or even slightly worse) prediction accuracy of gene-starts and COGs.

As previously mentioned, the type of gene-start model used is dependent on the group (A through E) to which the genome is assigned.

The two components of a generic site model, the nucleotide frequency matrix and the spacer length distribution are derived by running the modified Gibbs sampler algorithm (GibbsL, described below). In some situations, we decide on the "validity" of the motif by examining the localization (or concentration) of the motifs at some specific distance from the TIS. To do this, we define the *localization distance* as the mode of the spacer length distribution, the most frequent distance between motifs and the gene start. If the frequency at the mode is larger than a threshold Q%, then the motif is said to be localized with Q% mode threshold.

The full details on the derivations of the five model types (A-E) along with the depictions of the positional frequencies patterns of the motifs and associated spacer length distributions (Figs. S3-S7) are given in Suppl. Materials

<u>The third iteration and up</u>. From the third iteration onwards, the *type* of the model of the gene start does not change; however, the model parameters are updated by the type-specific rules. The dual motif models containing RBS and promoter boxes sub-models (defined for groups A and B) compete as the alternative states of the Viterbi algorithm. The typical gene model is updated in this iteration (Fig. 2). GeneMarkS-2 continues the prediction/estimation iterations until convergence (99% identity in gene starts between the iterations). An alternative stopping rule is reaching the maximum number of iteration (the default value is 10). All the genomic segments labeled as coding regions in the last iteration are reported as predicted genes, the output of the algorithm.

A motif finder that accounts for the signal localization pattern

The MCMC motif finder Gibbs3 (Thompson et al. 2003) was designed to learn a probabilistic model of an *a priori* unknown motif present in a set of sequences. We used Gibbs3 for the RBS model delineation in GeneMarkS with reasonable accuracy.

It was observed that the length distributions of sequences separating motifs from gene starts (spacers) indicate a preference for some optimal lengths that facilitate interactions of biomolecules involved in the translation initiation. However, the Gibbs3 algorithm accepts the motif instances with too long or too short spacers equally well in comparison with the motif instances having more optimal spacer length.

We implemented a modification, named GibbsL, of the Gibbs sampler algorithm (see Suppl. Materials). We explicitly included the spacer length into the objective function of GibbsL. At a given iteration of GeneMarkS-2, GibbsL runs a fixed number, N (default N=60), of its own iterations. Furthermore, the instances of GibbsL run are repeated M times (default M=30) and the result with the highest score of the objective function is selected.

Materials

Genes supported by proteomic studies. Mass-spectrometry-determined peptides were obtained in studies of a number of prokaryotic species at the Pacific Northwest National Laboratory (Venter et al. 2011). From all the available genomes, we selected 54, each with more than 250 proteomics validated ORFs (supported by at least two matching peptides). The peptide-supported ORFs (psORFs) annotated in the 54 genomes (Table S1) were used in the assessment of false negative and false positive rates of *ab initio* gene prediction.

A Algorithm	Missed MS confirmed genes (from 89,466)	Missed COG genes (not MS) (from 287,237)
GeneMarkS	376	1,467
Glimmer3	496	1,990
Prodigal	217	1,389
GeneMarkS-2	**181**	**1,147**

B Algorithm	False predictions overlaping MS confirmed genes	False predictions overlaping COG genes (not MS)
GeneMarkS	352	2,046
Glimmer3	921	6,435
Prodigal	211	1,339
GeneMarkS-2	**114**	**932**

Table 1. Statistics of false negative (Panel A) and false positive (Panel B) predictions observed in tests on sequences containing genes validated by proteomics (MS) and by COG annotation.

COG annotated genes. We used 145 genomes (115 bacteria and 30 archaea covering 22 bacterial and archaeal phyla) suggested by our colleagues at DOE Joint Genome Institute (N. Kyrpides, personal communication). The genomes varied in size, type of genetic code, and GC content (Table S2). Among the genes annotated in these genomes we selected genes with the COG characterization, such that orthologous relationships of their protein products were estabilshed with proteins from other species within the Clusters of Orthologous Groups (Tatusov et al. 1997; Tatusov et al. 2003; Galperin et al. 2015). The COG annotation serves as a robust evidence of the functional role of a gene, thus this gene set would be unlikely to include random ORFs. Since 36 out of 145 genomes belonged to the set of 54 genomes with 'proteomics' confirmed genes we removed the redundancy in the actual tests (see below).

Simulated genome specific non-coding regions. Genome specific models (the zero-order Markov chains) for *annotated* intergenic regions of each of 145 genomes (Table S2) were made and used to generate 145 random sequences with length 1 MB. We used the simulated non-coding sequences for assessment of false positive rates of gene predictions. Note that, the zero-order Markov chain models of non-coding regions was not used in Glimmer3, Prodigal or GeneMarkS-2.

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 *Synechocystis sp.* (Sazuka et al. 1999), *E. coli* (Rudd 2000; Zhou and Rudd 2013), *M. tuberculosis* (Lew et al. 2011), and *D. deserti* (de Groot et al. 2014) and the archaea *A. pernix* (Yamazaki et al. 2006), and *H. salinarum*, *N. pharaonis* (Aivaliotis et al. 2007).

Set of representative prokaryotic genomes. The prokaryotic genome collection of NCBI includes a description of 5,007 species as 'representatives' of the whole database of ~100,000 genomes (Tatusova et al. 2014). These include 238 archaeal and 4,769 bacterial species to cover all the genera while leaving out the majority of species of the respective genera along with most of their strains.

Results

Error rates in gene prediction with the focus on the gene 3'end. GeneMarkS, Glimmer3, Prodigal, and GeneMarkS-2 were run with default settings, and their gene predictions were compared (i) with the 'proteomics' validated annotation (54 genomes, Table 1, ~89,500 proteomics supported genes or psORFs) as well as (ii) the COGs validated annotation (145 genomes, Table 2; ~341,486 genes). We determined the frequencies of missed psORFs (false negative) as well as frequencies of false predictions (i.e. those incompatible with psORFs). A predicted gene was judged as false if more than 30% of its length overlapped with a psORF located in a different strand or frame.

In the 54 genomes set, we observed that GeneMarkS-2 missed 181 psORFs out of 89,466, the least number of false negative errors made by the tested tools (Table 1). At the same time, GeneMarkS-2 made the least number of false positive predictions - 114 (Table 1).

The test on 145 genomes with COG validated genes also demonstrated that GeneMarkS-2 made more accurate predictions. The number of missed COG genes, 1,147, was the lowest for GeneMarkS-2, followed by Prodigal with 1,389. The rate of missed COG genes for any gene finders was less than 1% (Table 1). Furthermore, false positives were identified in the same way as above, with GeneMarkS-2 yielding 932 false predictions, a number significantly

A	Bins (nt):	< 150	150-300	300-600	600-900	> 900	Total
Algorithm	COG genes	362	13,985	65,948	83,745	177,446	341,486
		Missed annotated genes (FN)					
GeneMarkS		135	**504**	444	200	305	1,588
Glimmer3		**60**	556	872	347	325	2,160
Prodigal		161	656	442	96	79	1,434
GeneMarkS-2		138	513	**384**	**78**	**70**	**1,183**

B	Bins (nt):	< 150	150-300	300-600	600-900	> 900	Total
Algorithm		False positives (FP) in simulated sequence					
GeneMarkS		4,257	5,865	597	9	0	10,728
Glimmer3		13,590	882	70	35	40	14,617
Prodigal		4,455	7,897	1,669	1,988	889	16,898
GeneMarkS-2		**645**	**345**	**17**	**0**	**0**	**1,007**

Table 2. Panel **A**: Test on 145 genomes containing COG annotated genes. Frequencies of missed genes in groups (bins) of different length. Panel **B**: Test on 145 simulated non-coding sequences (I Mb each). Frequencies of false positives in the same length categories as in Panel A.

smaller than the ones observed for the other gene finders (Table 1). Note that some COG validated genes were identical to the 'proteomics' genes and were excluded in the second test.

On the full set of COG genes, we assessed the effect of gene length on the false negative predictions (Table 2A). Glimmer3 had the lowest number of "missed" short genes (90-150 nt range) as compared to the other tools. However, the cost was a significant increase in the number of false predictions of short genes (Table 2B). We observed that GeneMarkS-2 had the least number of "missed" genes in all the other bins (and in total). At the same time GeneMarkS-2 had the lowest number of false positives in all the bins. We also observed that the GeneMarkS-2 performance was least dependent on genome GC content (data not shown).

Finally, since the significant overlaps between predicted and annotated genes (in the proteomics or COG verified subsets) turned out to be rather rare events (Table 1), we ran additional tests with simulated non-coding sequences to make observations on false positive predictions.

False positive predictions in simulated non-coding sequences. On a sequence simulating non-coding regions of a particular genome **e**ach gene finder was run with the species-specific models trained on the genome in question. We observed that GeneMarkS-2 had lower error rate than the other gene-finders, e.g. more than 50% lower than the second best tool, Prodigal (Table 2B). The increase in the rate of false positive predictions made by Prodigal in high GC sequences was likely related to the tendency of Prodigal to predict longer ORFs as genes; longer ORFs appear more frequently in high GC sequences than in low GC ones.

Glimmer3 had the lowest false positive rate in the length interval 300–600 nt (Table 2B). GeneMarkS-2 demonstrated lower than other tools frequencies of false positives in all the other length intervals. Interestingly, the GeneMarkS-2 improvement over GeneMarkS was mainly in reduction of false positives generated by atypical models (81 vs 9,231 by GeneMarkS). The reduction in false positives attributed to the typical models was much more modest (926 vs 1,642 by GeneMarkS-2).

Accuracy of gene start prediction assessed on the sets of genes with experimentally verified starts. This assessment was done on sets of genes from the genomes of *A. pernix, D. deserti, E. coli, H. salinarum, M. tuberculosis, N. pharaonis*, and *Synechocystis sp.* (Table 3).

Speaking of the whole set of verified genes, we observed that GeneMarkS-2 correctly predicted the largest number of starts among all the gene finders. More precisely, GeneMarkS-2 showed an error rate of 4.4%, followed by Prodigal - 6.1%, GeneMarkS - 10.2% and Glimmer3 - 13.2% (Table 3).

A factor in the GeneMarkS-2 improved start accuracy was the transition to more flexible modeling f the regulatory signals. For instance, the archaea *H. salinarum* characterized as group A had the RBS sequence missing in (most) FGIOs; the detected promoter TATA box was located at a distance 22-24 nt from the TIS the distance indicating the presence of leaderless transcription (Fig. 3B). The RBS sites for the remaining FGIOs as well as for the IGIOs were identified at 6-8 nt distance from the TIS sites (Fig. 3C). The original GeneMarkS with Gibbs3

Species	*Gene-start model type*	*# of verified gene starts*	*GeneMarkS*	*Glimmer3*	*Prodigal*	*GeneMarkS-2*
*A. pernix**	D	130	125	119	**127**	126
D. deserti	B	386	315	314	334	**369**
E. coli	D	769	725	714	**751**	740
*H. salinarum**	A	530	502	454	514	**523**
M. tuberculosis	B	701	572	572	620	**635**
*N. pharaonis**	A	315	309	288	309	**312**
Synechocystis	E	96	81	79	**92**	**92**
*(*archaea)*	*Total*	2,927	2,629	2,540	2,747	**2,797**

Table 3. Numbers of correctly predicted gene starts in the seven test sets of genes verified by N-terminal sequencing.

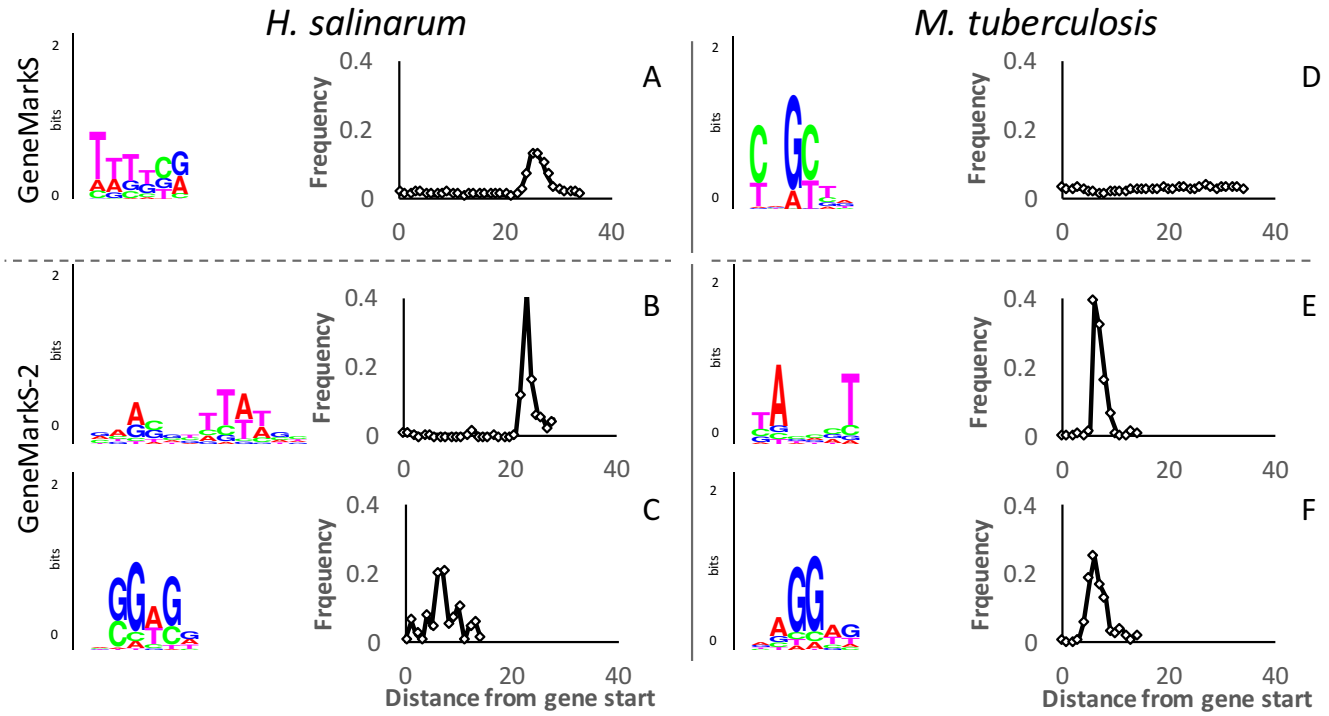

Figure 3. Motif logos and spacer length distributions for genomes of *H. salinarum* (group A) and *M. tuberculosis* (group B). Motifs found by GeneMarkS are shown in 3A and 3D respectively. GeneMarkS-2 detected a promoter signal in *H. salinarum* FGIOs with a better localization than GeneMarkS (3B). In *M. tuberculosis*, the 'mixed' motif found by GeneMarkS has no localization (3D). The motif found by GeneMarkS-2 in FGIOs has a localization typical for bacterial TATA box (3E). In both species GeneMarkS-2 finds the RBS sites for IGIO genes (3C and 3F respectively).

could derive (by design) only a single TATA box model with a less pronounced localization than GeneMarkS-2 with GibbsL (Fig. 3A).

For the bacterial genome *M. tuberculosis* characterized as group *B* the original GeneMarkS did not find sufficiently strong RBS motif (Fig. 3D). GeneMarkS-2 revealed why it a bit difficult feat to accomplish. The new algorithm predicted that 40% of the *M. tuberculosis* operons were transcribed in the leaderless fashion, with the promoter Pribnow box located at a distance of 6-8 nt from the gene starts (Fig. 3F). The remaining ~60% of operons were predicted to have their RBS sites located at the same distance of 6-8 nt from the gene starts (Fig. 3E). Therefore, since both the promoter for genes without 5' UTRs and RBS sites for genes with 5' UTRs were located

at the same distance from the TIS sites, the Gibbs3 algorithm taking all the upstream sequences in one input failed to converge to a single model.

On a general note, the performance of GibbsL in detecting SD RBS motifs in low and mid GC genomes was observed to be somewhat similar to the performance of Gibbs3 (though GibbsL tends to have higher localization peaks). However, in genomes with higher GC content, GibbsL derived the motifs with higher information content and more compact localization (Fig. S8, S9, S10).

We also observed that the GibbsL performance is more robust than one of Gibbs3 when the length of sequences that are supposed to contain common motifs (i.e. the selected gene upstream regions) increased (Fig. S8, S9).

Finally, we have observed that the width of the motif, an assigned parameter of the motif search by GibbsL, did not show significant influence on the results of gene start prediction (if changed between 5 nt to 10 nt, see Table S3). Indeed, if the motifs with larger widths were admitted, the derived RBS models did not show a significant change in information content (Fig. S11).

Predicting leaderless transcripts in ~ 5,000 prokaryotic genomes

GeneMarkS-2 was run on each of the ~5,000 *representative* genomes to generate a genome annotation along with a genome assignment to one of the five groups, *A, B, ... E*, described above. Five separate trees were then created to represent the genomes belonging to each of the groups (Table S4, A-E). These trees show that genomes of similar ancestry tend to belong to the same group.

Species that fell into group *A* were archaeal species with predicted prevalence of leaderless (polycistronic) mRNAs. A promoter model was derived for leaderless FGIO, while an RBS model was derived for the remaining genes. From the 238 archaeal genomes in our dataset, 199 were assigned to group *A*. In particular, some taxa had most (or all) of their members belonging to group A, such as *Halobacteria* (74 out of 74 species, 100%, assigned to group *A*), *Methanomicrobia* (40 out of 42, 95%), *Thermococci* (21 out of 21, 100%), *Thermoplasmata* (11 out of 11, 100%), *Archeoglobi* (7 from 7, 100%), *Thaumarchaeota* (11 from 11, 100%) and *Crenarchaeota* (23 out of 35, 65%) (see Table S4, A). The group *A* characteristic feature was first discovered in the hyperthermophylic archaeon, *Pyrobaculum aerophilum* (Slupska et al. 2001). We have inferred, however, that group D is a home for a significant fraction of the taxon *Crenarchaea*, where *P. aerophilum* belongs. Thus, many members of *Crenarchaea* should have low percentage of leaderless transcripts.

Of a significant interest is group *B* (1028 out of 4769 bacteria) where we included bacterial species predicted to have frequent presence of leaderless mRNAs (Table S4, B). Species of group B are frequent in *Actinobacteria* (773 from 859, 90.0%), in *Deinococcus-Thermus* (37 out of 38, 97.4%) but rare in *Proteobacteria* (104 out of 1854, 5.6%) and *Firmicutes* (36 out of 1064, 3.4%). Particularly high frequency of type *B* species was observed in *Streptomycetales* (129 out of 129, 100%) and in *Corynebacteriales* (197 out of 202, 97.5%) including *Mycobacteriaceae* (56 out of 57, 98.2%).

The group *C* assignments were made for 495 bacteria and no archaea. The characteristic feature of this group is the *presence of the same type signal in both FGIOs and IGIOs*. We interpreted the signal as the *non-SD type RBS* since this signal is present upstream to IGIOs and cannot be a promoter. The species of group C are frequent in the FCB group (409 out of 455, 89.9%), but rare in *Terrabacteria* (1.7%) and *Proteobacteria* (2.0%).

The group D was the largest: 2,935 bacteria and 39 archaea (Table S4, D). The group D species were characterized by the dominance of leadered mRNA with detectable RBS motifs having the Shine-Dalgarno consensus. Among the group D bacteria 39% were Gram-positive and 61% were Gram-negative. However, the majority, 57%, of Gram-positive bacteria were assigned to group D. More than that, if we exclude *Actinobacteria,* that rarely belong to group D (78 out of 859, 9.1%) and mostly appear in group B, we would see that 96% of remaining Gram-positive bacteria belong to group D.

Finally, 311 bacterial species were assigned to Group *E* (Table S4, E) characterized by the absence of pronounced regulatory signals upstream to most genes. Species of this group are relatively frequent in *Cyanobacteria* (90 out of 127, 70.9%) and in *Burkholderiales* (63 out of 166, 37.9%).

The summary list of the distribution of the ~5,000 species among groups A-E is given in Table 4.

	Archaeal Genomes		Bacterial Genomes		Total
	Number	%	Number	%	Number
Group A	199	83.6	-	-	199
Group B	-	-	1,028	21.6	1,028
Group C	0	0.0	495	10.4	495
Group D	39	16.4	2,935	61.5	2,974
Group E	0	0.0	311	6.5	311
Total	238	100	4,769	100	5,007

Table 4. Distribution of archaeal and bacterial genomes among groups A-E.

Discussion

Gene finding accuracy evaluation. We demonstrated in several tests that, on average, GeneMarkS-2 is a more accurate tool than the current frequently used gene finders.

Particularly, GeneMarkS-2 had lower frequencies of *false negative* and *false positive* errors assessed on the sets of genes validated by mass-spectrometry and the COG annotation (Table 1). Also, the numbers of *false positive* predictions made by GeneMarkS-2 in simulated non-coding sequences were significantly smaller than the numbers observed for other tools (Table 2B).

The array of atypical models employed in GeneMarkS-2 improved the prediction of horizontally transferred (atypical) genes. In our observations, the deviation of GC composition of atypical genes from the genome average could be as large as 16% (e.g. 798 nt long *E. coli* gene *b0546* characterized as DLP12 prophage, with GC content 36% vs 52% GC in the bulk of the *E. coli* genes). The GC content of atypical genes frequently is lower than the GC content of the 'typical' ones; this bias is attributed to the possibility that many atypical genes could be transferred from AT rich phage genomes. 'Atypical' genes with large GC deviations are expected to appear more frequently in high GC genomes with the larger space for downward variation.

All over, the atypical genes may constitute a significant fraction of the whole gene complement (e.g. about 15% of genes in the *E. coli* genome (Borodovsky et al. 1995)). In the study of the ~5,000 genomes we found that the distribution of the fraction of predicted atypical genes in prokaryotic genomes are rather similar between archaea and bacteria with the average of about 8-9% (Fig. 4).

Comparison of the COG annotated genes missed by the three gene finders showed that atypical gene constituted 30% among 780 (534+246) genes missed by Prodigal and 42% from 1605 (1359+246) genes missed by Glimmer3 (Fig. S12). Prediction of typical and atypical genes by a single model (as in Glimmer3 and Prodigal) makes prediction of atypical genes more difficult. Improved prediction of atypical genes by GeneMarkS-2 is a compelling argument in favor of the use of atypical models.

One of the positive features of the approach implemented in GeneMarkS-2 is the ability to identify atypical genes as bacterial or archaeal (it is due to the division of the *atypical* models into distinct bacterial and archaeal types (Zhu et al. 2010)). The insights into the possible origin of horizontally transferred genes could be particularly useful for genomes of thermophilic bacteria and mesophilic archaea.

Assessment of the accuracy of gene start prediction was done on the sets of genes verified by the N-terminal sequencing (Table 3). These sets were available for seven genomes that came from all groups except group C. GeneMark-S2, that shows the best performance overall, incorrectly predicted 120 starts out of 2,927 while the second best, Prodigal, made wrong predictions of 170 gene starts. Notably, Prodigal performs better on *E. coli*, however, the set of verified genes from this species was used for supervised training of the Prodigal gene start prediction model. GeneMarkS-2 makes more accurate predictions for genomes of groups A and B where the leaderless transcription is frequent and some genes do not have RBS sites.

Particularly, the experimental study of *D. deserti* identified 384 genes with verified TIS, 262 of which had TSS annotated with dRNA-Seq (de Groot et al. 2014). It was also shown that 167 out of the 262 genes had leaderless transcription. In this genome GeneMarkS-2 correctly predicted 34 more starts than Prodigal. Prodigal detects RBS motifs only; this restriction reduces accuracy of start prediction for genes with leaderless transcription.

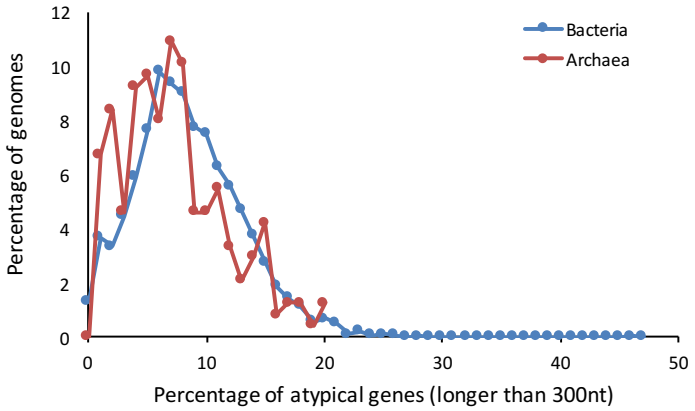

Figure 4. Distributions of the percentage of predicted atypical genes in archaeal and bacterial genomes.

Characterization of the patterns around gene starts. Prediction of the extent of leaderless transcription

GeneMarkS-2, unlike other gene finders, is able to infer the presence of leadered or leaderless transcripts due to the differences in the regulatory signals predicted upstream to the gene starts.

Bacterial and archaeal genomes show different patterns with respect to the frequency of leaderless transcription (Fig. 5). The archaeal species have the bimodal distribution of the frequencies. Large number of studied archaeal genomes are predicted to have 60% to 80% of the operons transcribed in the leaderless fashion. Still some sizable fraction of the group A archaeal genomes had 25-35% of operons transcribed in the leaderless way. Among bacterial genomes of group B, the majority has 25% to 50% of the operons transcribed in the leaderless fashion.

We compared computational predictions of the fractions of the leaderless transcripts with the fractions observed in the dRNA-Seq experiments with the genes expressed in *Deinococcus deserti* (de Groot et al. 2014) *Haloferax volcanii* (Babski et al. 2016), *Sulfolobus solfataricus* (Wurtzel et al. 2010), and *M. tuberculosis* (Cortes et al. 2013). The values of predicted and observed percentages of leaderless transcripts were determined on the sets of genes that had identical starts both in GeneMarkS-2 prediction and annotation. These results were as follows: in archaea *D. deserti*, ~62% vs 62% (1,707 transcripts), in *H. volcanii* 86% vs 82% (1,406 transcripts); *S. solfataricus* 78% vs 76% (859 transcripts), in bacteria *M. tuberculosis* 42% vs 34% (1310 transcripts).
Thus, the predictions were in reasonably good agreement with the experiment.

Genomes with experimentally characterized small numbers of leaderless genes (Sharma et al. 2010; Nicolas et al. 2012; Dugar et al. 2013; Kroger et al. 2013; Lin et al. 2013; Wiegand et al. 2013; Shao et al. 2014; Thomason et al. 2015) (TSS experiments) were all classified as group D. Genomes with large proportion of leaderless transcripts were all classified by GeneMarkS-2 as group B (Cortes et al. 2013; Pfeifer-Sancar et al. 2013; de Groot et al. 2014; Shell et al. 2015) or group A (Koide et al. 2009; Wurtzel et al. 2010; Toffano-Nioche et al. 2013; Jager et al. 2014; Babski et al. 2016).

Experiments on *Synechocystis sp* demonstrated the prevalence of the leadered transcription (Mitschke et al. 2011). However, in less than 15.5% of genes GeneMarkS-2 detected the RBS motif with the SD consensus. The experiments have shown that mutating some "A" rich sequences at 15-45 nt upstream to gene start which was still within long 5' UTR sequence, led to changes in gene expression (Mutsuda and Sugiura 2006). The mechanism used for the TIS recognition in majority of *Synechocystis sp* genes is unknown.

It was observed that translation initiations of the three types, SD-RBS based, non-SD RBS based, and leaderless are present in *E. coli* (Shean and Gottesman 1992; Barrick et al. 1994; Resch et al. 1996). Further observations have shown that the distribution of the numbers of genes controlled by each of the three mechanisms could vary significantly between the species (Gualerzi and Pon 2015). If a particular mode appears rarely in a given species, GeneMarkS-2 training procedure (described below) will not be able to take this mode into account due to the insufficient size of the training set.

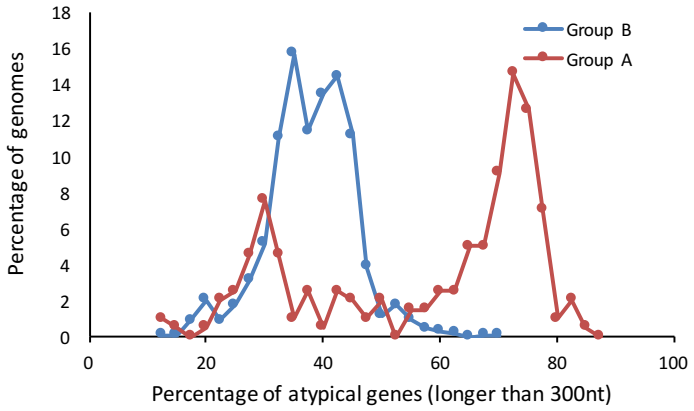

Figure 5. Distributions of the percentage of leaderless transcripts among all transcripts in archaeal group A and bacterial group B.

Regulatory models in Group C

GeneMarkS-2 currently assigns 495 out of 4,769 bacteria (and none out of 238 archaea) to group C. We take as an example the genome of *Bacteroides ovatus*. Experimental evidence shows that, while its 16S rRNA features the *bona fide* anti-Shine-Dalgarno pattern, the SD-matching sequences appear upstream to the TIS sites in only ~3% of genes. A-rich sequences were observed in the upstream regions of the majority of *B. ovatus* genes. It was shown experimentally that mutating these A's reduces the gene expression levels. Thus one would imply that the A-rich sequences are important for TIS recognition (Wegmann et al. 2013). GeneMarkS-2 identified the A-rich non-SD type motif with consistent localization at ~9 nt from TIS (Fig. 6).

GeneMarkS-2 assigned 90% of *Bacteroidetes/Chlorobi* genomes to Group C (408 of 450). While not much is known about the non-SD mechanisms in bacteria, the clustering of the species assigned to group C within parts of the taxonomy tree lends additional credibility to the results (Table S4, C).

For instance, in *Bacteroides* (which includes *B. ovatus*) 21 out of 23 genomes were assigned to group C. In these 21 genomes, GeneMarkS-2 found motifs similar to the ones revealed in *B. ovatus* in conservation pattern and localization distribution.

Also, 30 out of the 30 *Flavobacterium* genomes (a genus from *Bacteroidetes/Chlorobi*) were assigned to group C. The 6 nt wide motifs similar in the conservation pattern to the ones of *B. ovatus* were situated at the same distance from TIS sites (Fig. S13 for *Flavobacterium frigidarium*).

While genomes of the species from these taxa are similar, there are some differences in the derived motifs. In particular, *Bacteroides* tend to have a few strong A nucleotides next to the 'core' motif and close to TIS, while *Flavobacterium* do not (Fig. S14). The 'core' 6 nt motif (TAAAAA) is present in both taxa; it is located at ~ 9 nt from TIS. Consistency of this observation was tested for all 21 *Bacteroides* and 30 *Flavobacteria*. Finally, the 6 nt core motif is not easy to detect in *Prevotella* (a close relative of *Bacteroides*) when the motif width is set to 6 nt. However, a motif with width 15 is easier to detect.

Note that unlike *B. ovatus,* other Group *C* species may have the 16S rRNA with a mutated or truncated tail (Lim et al. 2012).

In the recent publication (Nakagawa et al. 2017) leaderless and non-SD initiation were included into the same class. Here, we made distinction between the leaderless transcripts and, thus, the absence of RBS (groups A and B) and the leadered transcripts with non-SD patterns in 5'UTR (group C).

Overall, GeneMarkS-2 provides the means for revealing genomes that use non-SD RBS as well as for the delineation of the motifs and their use in gene prediction.

Final remarks. Remaining difficulties in automatic genome annotation are concerned about some minor fractions of genes in any given genome.

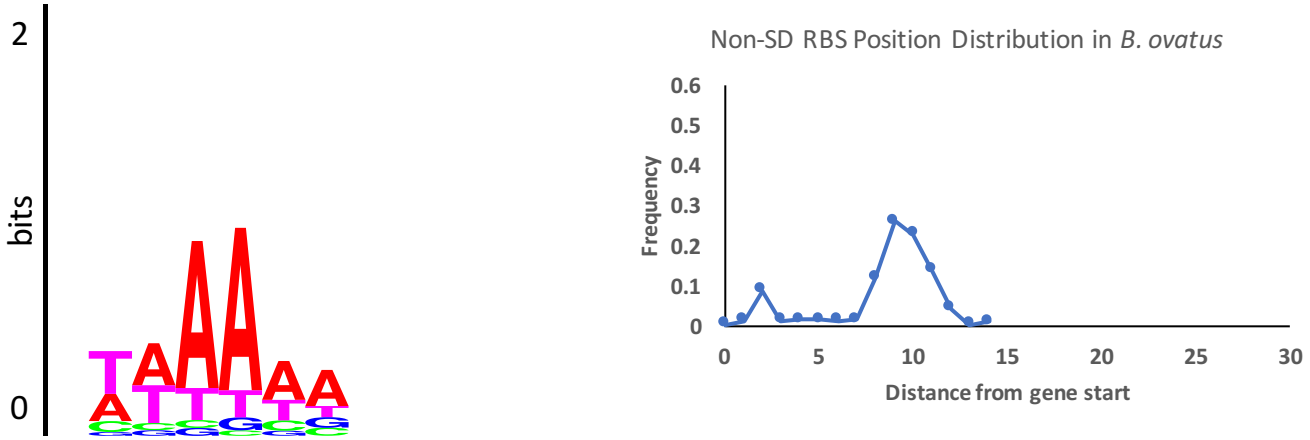

Figure 6. The motif logo and the spacer length distribution of *Bacteroides ovatus*, a group C genome.

Few genes still escape detection by *ab initio* tools, e.g. genes significantly biased in higher order oligonucleotide composition or genes containing frameshifts. When an orthologue of such a gene is present in the database, the frameshift identification can be done rather easily. Some frameshifts, however, are functional and conserved in evolution, such as *prfB* gene, encoding the translation initiation factor (Craigen and Caskey 1986) or genes regulating some mobile elements (Sharma et al. 2011). Still, the fraction of genes with the programmed frameshifts is minor. The pseudogenes, especially expressed pseudogenes, could mislead gene finding tools into generating predictions that eventually would be classified as false positive.

An extension of GeneMarkS (known as GeneMarkS+) integrating external evidence, e.g. protein homology, into *ab initio* gene prediction was developed recently and was used in the latest version of the NCBI prokaryotic genome annotation pipeline as integrator of several types of evidence into genome annotation (Tatusova et al. 2016). Similarly, GeneMarkS-2 has already been extended to the "plus" version (paper in preparation).

Software availability

The software and input files used in this study have been made available through the website
http://topaz.gatech.edu/GeneMark/genemarks2.cgi
Running time of GeneMarkS-2 is currently ~ 3 minutes on genome of the size of *E. coli*.

Acknowledgements

We thank Susan M.E. Smith for useful comments on the manuscript. This work was supported by National Institutes of Health (NIH) [HG000783 to M.B.].

Conflict of interest statement. None declared.

References

Aivaliotis M, Gevaert K, Falb M, Tebbe A, Konstantinidis K, Bisle B, Klein C, Martens L, Staes A, Timmerman E et al. 2007. Large-scale identification of N-terminal peptides in the halophilic archaea Halobacterium salinarum and Natronomonas pharaonis. *J Proteome Res* **6**: 2195-2204.

Babski J, Haas KA, Nather-Schindler D, Pfeiffer F, Forstner KU, Hammelmann M, Hilker R, Becker A, Sharma CM, Marchfelder A et al. 2016. Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). *BMC Genomics* **17**: 629.

Barrick D, Villanueba K, Childs J, Kalil R, Schneider TD, Lawrence CE, Gold L, Stormo GD. 1994. Quantitative analysis of ribosome binding sites in E.coli. *Nucleic Acids Res* **22**: 1287-1295.

Bell SD, Jackson SP. 1998. Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. *Trends Microbiol* **6**: 222-228.

Besemer J, Lomsadze A, 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.

Borodovsky M, McIninch J. 1993. GeneMark: parallel gene recognition for both DNA strands. In *Computers & Chemistry*, **17**: 123-133.

Borodovsky M, McIninch JD, Koonin EV, Rudd KE, Medigue C, Danchin A. 1995. Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res* **23**: 3554-3562.

Cho S, Kim MS, Jeong Y, Lee BR, Lee JH, Kang SG, Cho BK. 2017. Genome-wide primary transcriptome analysis of H2-producing archaeon Thermococcus onnurineus NA1. *Sci Rep* **7**: 43044.

Cortes T, Schubert OT, Rose G, Arnvig KB, Comas I, Aebersold R, Young DB. 2013. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in Mycobacterium tuberculosis. *Cell Rep* **5**: 1121-1131.

Craigen WJ, Caskey CT. 1986. Expression of peptide chain release factor 2 requires high-efficiency frameshift. *Nature* **322**: 273-275.

Creecy JP, Conway T. 2015. Quantitative bacterial transcriptomics with RNA-seq. *Current Opinion in Microbiology* **23**: 133-140.

de Groot A, Roche D, Fernandez B, Ludanyi M, Cruveiller S, Pignol D, Vallenet D, Armengaud J, Blanchard L. 2014. RNA sequencing and proteogenomics reveal the importance of leaderless mRNAs in the radiation-tolerant bacterium Deinococcus deserti. *Genome Biol Evol* **6**: 932-948.

Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**: 673-679.

Dugar G, Herbig A, Forstner KU, Heidrich N, Reinhardt R, Nieselt K, Sharma CM. 2013. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates. *PLoS Genet* **9**: e1003495.

Fickett JW, Tung CS. 1992. Assessment of protein coding measures. *Nucleic Acids Res* **20**: 6441-6450.

Galperin MY, Makarova KS, Wolf YI, Koonin EV. 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* **43**: D261-269.

Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muniz-Rascado L, Garcia-Sotelo JS, Alquicira-Hernandez K, Martinez-Flores I, Pannier L, Castro-Mondragon JA et al. 2016. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Res* **44**: D133-143.

Gehring AM, Walker JE, Santangelo TJ. 2016. Transcription Regulation in Archaea. *J Bacteriol* **198**: 1906-1917.

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. 1981. Codon Catalog Usage Is a Genome Strategy Modulated for Gene Expressivity. *Nucleic Acids Res* **9**: R43-R74.

Gualerzi CO, Pon CL. 2015. Initiation of mRNA translation in bacteria: structural and dynamic aspects. *Cell Mol Life Sci* **72**: 4341-4367.

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.

Jager D, Forstner KU, Sharma CM, Santangelo TJ, Reeve JN. 2014. Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. *BMC Genomics* **15**: 684.

Jager D, Sharma CM, Thomsen J, Ehlers C, Vogel J, Schmitz RA. 2009. Deep sequencing analysis of the Methanosarcina mazei Go1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci U S A* **106**: 21878-21882.

Koide T, Reiss DJ, Bare JC, Pang WL, Facciotti MT, Schmid AK, Pan M, Marzolf B, Van PT, Lo FY et al. 2009. Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol Syst Biol* **5**: 285.

Kroger C, Colgan A, Srikumar S, Handler K, Sivasankaran SK, Hammarlof DL, Canals R, Grissom JE, Conway T, Hokamp K et al. 2013. An infection-relevant transcriptomic compendium for Salmonella enterica Serovar Typhimurium. *Cell Host Microbe* **14**: 683-695.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**: 208-214.

Lew JM, Kapopoulou A, Jones LM, Cole ST. 2011. TubercuList-10 years after. *Tuberculosis* **91**: 1-7.

Li J, Qi L, Guo Y, Yue L, Li Y, Ge W, Wu J, Shi W, Dong X. 2015. Global mapping transcriptional start sites revealed both transcriptional and post-transcriptional regulation of cold adaptation in the methanogenic archaeon Methanolobus psychrophilus. *Sci Rep* **5**: 9209.

Lim K, Furuta Y, Kobayashi I. 2012. Large variations in bacterial ribosomal RNA genes. *Mol Biol Evol* **29**: 2937-2948.

Lin YF, A DR, Guan S, Mamanova L, McDowall KJ. 2013. A combination of improved differential and global RNA-seq reveals pervasive transcription initiation and events in all stages of the life-cycle of functional RNAs in Propionibacterium acnes, a major contributor to wide-spread human disease. *BMC Genomics* **14**: 620.

Locey KJ, Lennon JT. 2016. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci USA* **113**: 5970-5975.

Lukashin AV, Borodovsky M. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res* **26**: 1107-1115.

Mitschke J, Georg J, Scholz I, Sharma CM, Dienst D, Bantscheff J, Voss B, Steglich C, Wilde A, Vogel J et al. 2011. An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp PCC6803. *P Natl Acad Sci USA* **108**: 2124-2129.

Mutsuda M, Sugiura M. 2006. Translation initiation of cyanobacterial rbcS mRNAs requires the 38-kDa ribosomal protein S1 but not the Shine-Dalgarno sequence: development of a cyanobacterial in vitro translation system. *J Biol Chem* **281**: 38314-38321.

Nakagawa S, Niimura Y, Gojobori T. 2017. Comparative genomic analysis of translation initiation mechanisms for genes lacking the Shine-Dalgarno sequence in prokaryotes. *Nucleic Acids Res* **45**: 3922-3931.

Nicolas P, Mader U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, Bidnenko E, Marchadier E, Hoebeke M, Aymerich S et al. 2012. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. *Science* **335**: 1103-1106.

Pfeifer-Sancar K, Mentz A, Ruckert C, Kalinowski J. 2013. Comprehensive analysis of the Corynebacterium glutamicum transcriptome using an improved RNAseq technique. *BMC Genomics* **14**: 888.

Resch A, Tedin K, Grundling A, Mundlein A, Blasi U. 1996. Downstream box-anti-downstream box interactions are dispensable for translation initiation of leaderless mRNAs. *Embo J* **15**: 4740-4748.

Romero DA, Hasan AH, Lin YF, Kime L, Ruiz-Larrabeiti O, Urem M, Bucca G, Mamanova L, Laing EE, van Wezel GP et al. 2014. A comparison of key aspects of gene regulation in Streptomyces coelicolor and Escherichia coli using nucleotide-resolution transcription maps produced in parallel by global and differential RNA sequencing. *Mol Microbiol* **94**: 963-987.

Rudd KE. 2000. EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Res* **28**: 60-64.

Sazuka T, Yamaguchi M, Ohara O. 1999. Cyano2Dbase updated: Linkage of 234 protein spots to corresponding genes through N-terminal microsequencing. *Electrophoresis* **20**: 2160-2171.

Shao W, Price MN, Deutschbauer AM, Romine MF, Arkin AP. 2014. Conservation of transcription start sites within genes across a bacterial genus. *Mbio* **5**: e01398-01314.

Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R et al. 2010. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* **464**: 250-255.

Sharma CM, Vogel J. 2014. Differential RNA-seq: the approach behind and the biological insight gained. *Curr Opin Microbiol* **19**: 97-105.

Sharma V, Firth AE, Antonov I, Fayet O, Atkins JF, Borodovsky M, Baranov PV. 2011. A pilot study of bacterial genes with disrupted ORFs reveals a surprising profusion of protein sequence recoding mediated by ribosomal frameshifting and transcriptional realignment. *Mol Biol Evol* **28**: 3195-3211.

Shean CS, Gottesman ME. 1992. Translation of the prophage lambda cl transcript. *Cell* **70**: 513-522.

Shell SS, Wang J, Lapierre P, Mir M, Chase MR, Pyle MM, Gawande R, Ahmad R, Sarracino DA, Ioerger TR et al. 2015. Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS Genet* **11**: e1005641.

Shine J, Dalgarno L. 1974. The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci USA* **71**: 1342-1346.

Shmatkov AM, Melikyan AA, Chernousko FL, Borodovsky M. 1999. Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes. *Bioinformatics* **15**: 874-886.

Slupska MM, King AG, Fitz-Gibbon S, Besemer J, Borodovsky M, Miller JH. 2001. Leaderless transcripts of the crenarchaeal hyperthermophile Pyrobaculum aerophilum. *J Mol Biol* **309**: 347-360.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN et al. 2003. The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**: 41.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* **278**: 631-637.

Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. 2014. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res* **42**: D553-559.

Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* **44**: 6614-6624.

Thomason MK, Bischler T, Eisenbart SK, Forstner KU, Zhang A, Herbig A, Nieselt K, Sharma CM, Storz G. 2015. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in Escherichia coli. *J Bacteriol* **197**: 18-28.

Thompson W, Rouchka EC, Lawrence CE. 2003. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res* **31**: 3580-3585.

Toffano-Nioche C, Ott A, Crozat E, Nguyen AN, Zytnicki M, Leclerc F, Forterre P, Bouloc P, Gautheret D. 2013. RNA at 92 degrees C: the non-coding transcriptome of the hyperthermophilic archaeon Pyrococcus abyssi. *Rna Biol* **10**: 1211-1220.

Venter E, Smith RD, Payne SH. 2011. Proteogenomic analysis of bacteria and archaea: a 46 organism case study. *PloS one* **6**: e27587.

Wegmann U, Horn N, Carding SR. 2013. Defining the Bacteroides Ribosomal Binding Site. *Appl Environ Microb* **79**: 1980-1989.

Wiegand S, Dietrich S, Hertel R, Bongaerts J, Evers S, Volland S, Daniel R, Liesegang H. 2013. RNA-Seq of Bacillus licheniformis: active regulatory RNA features expressed within a productive fermentation. *BMC Genomics* **14**: 667.

Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R. 2010. A single-base resolution map of an archaeal transcriptome. *Genome Res* **20**: 133-141.

Yamazaki S, Yamazaki J, Nishijima K, Otsuka R, Mise M, Ishikawa H, Sasaki K, Tago S, Isono K. 2006. Proteome analysis of an aerobic hyperthermophilic crenarchaeon, Aeropyrum pernix K1. *Molecular & cellular proteomics : MCP* **5**: 811-823.

Zhou J, Rudd KE. 2013. EcoGene 3.0. *Nucleic Acids Res* **41**: D613-624.

Zhu W, Lomsadze A, Borodovsky M. 2010. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* **38**: e132.