SUPPLEMENTAL MATERIAL

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

The Viterbi algorithm in the log-odds space, principal equations

The number of states in GHMM increased significantly (see GeneMark.hmm in GeneMarkS paper versus GeneMark.hmm in GeneMarkS-2). To simplify the Viterbi algorithm implementation, we moved from the standard dynamic programming implementation with the (log) probability values to a dynamic programming implemented in log-odds scores (i.e. where the probabilities along the given path are divided by the probabilities in the non-coding sequence model). We attempted to match the maximum value path in the log-odds space to the max likelihood path calculated in the probability space. In some cases (e.g. zero order models with species-specific parameters) exact identity was demonstrated. We saw that for higher order models and atypical gene states, the log-odds maximum value path was a close approximation of the maximum likelihood path. Extensive tests have showed that the difference between the two types of implementations was small (a difference of 3 genes out of 1,000). It is not clear which approach is more accurate in prediction of genes in real data, as the possible errors in our test sets of validated genes are comparable with the effect (difference in the approaches) that we want to estimate.

In GeneMarkS-2, the first prediction step does not use any species-specific parameters (i.e. the 'typical' model). For every candidate gene, the GC composition is computed and used to choose the coding (M_{cod}) and non-coding (M_{non}) models from the array of *heuristic* 'atypical' models. Thus selected models are then used to compute the 'content' gene score (Equation 1). Using the GC index for model (state) selection allows us to avoid computations of visiting states related to the 82 'atypical' models.

Still, the log-odds formulation excludes certain paths that could be considered in the full GHMM implementation. For instance, we do not directly compare the log-odds score of the native model to that of the atypical model. Rather, we first select the type (atypical or native) of the protein-coding sequence by comparing the probabilities of the native and atypical coding models, as shown in Equation 2. Here, $gene_{type}$ will take the type (native vs atypical) depending on which of the two coding models in Equation 2 (M_{native_coding} vs $M_{atypical\_coding}$, respectively) yields the larger value. We then use the model of that type (e.g. M_{cod} and M_{non}) to compute the log-odds scores defined in Equations 3 and 4.

Parameters of all the models are defined as described in the main text for all the steps of the algorithm; i.e. initialization and subsequent iterations. For a potential protein-coding sequence $x_1x_2 \ldots x_n$ with start codon $x_1x_2x_3$, stop codon $x_{n-2}x_{n-1}x_n$, and upstream sequence $x_{-20}x_{-19} \ldots x_{-1}$ the gene start score is defined in Equation 2 and the rest of the protein-coding score is defined in Equation 3. Here $y_1y_2 \ldots y_k$ denotes the sequence of RBS or promoter box, k is the site length; ϕ denotes the GC content, and M_a denotes the model for state a. The length distributions of prokaryotic protein-coding and non-coding regions (approximated by a gamma-function and an exponential function, respectively), contribute duration values into Equation 3. The last term in Equation 3 is the log-odds score of the durations; C is a constant depending

on parameters *Dc* and *Dn* of the gamma (protein-coding) and exponential (intergenic) length distributions respectively.

Finally, for all prediction steps, overlapping genes are penalized based on the length of the overlap. In particular, for overlapping genes *a* and *b* with lengths L_a and L_b, respectively, and length of overlap *m*, a penalty S_{ovlp} is added to the score (Equation 5).

Equation 1

$$S_{gene} = \log\frac{P(x_1x_2x_3|M_{start\_codon})}{P(x_1x_2x_3|M_{non})} + \log\frac{P(x_4 \dots x_{n-3} \mid M_{coding})}{P(x_4 \dots x_{n-3}| M_{non})} + \log\frac{P(x_{n-2}x_{n-1}x_n| M_{stop\_codon})}{P(x_{n-2}x_{n-1}x_n| M_{non})} + \log\frac{Cn^2e^{-\frac{n}{D_c}}}{e^{-\frac{n-1}{D_n}}}$$

Equation 2

$$gene_{type} = \text{argmax}\left(P\left(x_{16} \dots x_{n-3}\middle|M_{native\_coding}\right), P\left(x_{16} \dots x_{n-3}\middle|M_{atypical\_coding}\right)\right)$$

Equation 3

$$S_{CDS} = \log\frac{P(x_{16} \dots x_{n-3} \mid M_{coding})}{P(x_{16} \dots x_{n-3}| M_{non})} + \log\frac{P(x_{n-2}x_{n-1}x_n| M_{stop\_codon})}{P(x_{n-2}x_{n-1}x_n| M_{non})} + \log\frac{Cn^2e^{-\frac{n}{D_c}}}{e^{-\frac{n-1}{D_n}}}, \quad \textit{where } M_{coding} \textit{ and } M_{non} \textit{ is selected based on the gene\_type value}$$

Equation 4

$$S_{start} = \log\frac{P(x_1x_2x_3|M_{start\_codon})}{P(x_1x_2x_3|M_{non})} + \log\frac{P(x_4 \dots x_{15}| M_{down\_signal})}{P\ (x_4 \dots x_{15}\ |M_{non})} + \frac{P(x_0x_{-1}x_{-2}| M_{up\_signal})}{P(x_0x_{-1}x_{-2}|M_{non})} + \max\begin{cases} \log\frac{P(y_1 \dots y_k| M_{rbs})}{P(y_1 \dots y_k| M_{non})} + \log\frac{P(l \mid M_{rbs\_spacer})}{e^{-\frac{l+k}{D_n}}} & for\ RBS \\ \log\frac{P(y_1 \dots y_k| M_{promoter})}{P(y_1 \dots y_k| M_{non})} + \log\frac{P(l \mid M_{promoter\_spacer})}{e^{-\frac{l+k}{D_n}}} & for\ promoter \\ \log\frac{P(x_{-20} \dots x_{-3}| M_{extra\_up\_signal})}{P(x_{-20} \dots x_{-3}| M_{non})} & for\ upstream\ signature \end{cases}$$

Equation 5

$$S_{ovlp} = -m\log\left(1 + \frac{m}{2}\left(\frac{1}{L_a} + \frac{1}{L_b}\right)\right)$$

Building the models of sequences around the gene starts

The sequence of steps. The identification of the type(s) of regulatory site model(s) goes in parallel with the genome group assignment (Fig. S1). The candidate groups are tested in a particular order, and the genome is assigned to the first group for which the 'membership' test is successful. The process differs slightly for archaeal and bacterial genomes, where an archaea genome is tested against groups A, C, D, E (in that order, while a bacterial genome is tested against groups B, C, D, E. If a genome's domain (i.e. archaeal or bacterial) is not known, then all groups are tested in the order A, B, C, D, and E.

Type A model. This model assumes a presence of both leadered and leaderless transcription. Therefore, both the promoter and the RBS models are to be determined. There are two ways that an archaeon can be assigned to group A; the first method works well when the percentage of leaderless transcripts in the genome is high, and the second caters specifically to the case when leaderless transcripts are less frequent. The two methods only differ in the way the training sequences for the promoter and RBS models are selected.

In the first method, we select the 40 nt long fragments upstream to *all* FGIOs and run GibbsL to possibly detect a 12 nt long motif of the -26 box of the archaeal promoter (Bell and Jackson 1998). On the other hand, we also run GibbsL on the 20 nt long fragments from all IGIO to find the 6 nt long RBS pattern. If the detected promoter motif is *localized at a distance* larger than 14 nt (with the 10% mode threshold) then this genome is assigned to group *A*.

If this condition does not hold, then the method two is applied to detect a lower percentage of leaderless transcripts. In the second method, we choose a set of 20 nt long fragments located upstream to all the FGIOs, and single out those that show a local similarity to the *extended Shine-Dalgarno sequence TA**AGGAGG**TGA* (by checking for 4 consecutive nucleotide matches, with one possible U-G type substitution). This search divides the set of FGIOs into two sets, ones with the upstream fragments having the SD similarity (set X) and ones with upstream fragments having no SD similarity (set Y). We extend set X by adding the 20 nt fragments located upstream to all the IGIOs (expecting them to contain subsequences similar to SD-RBS). Then, we run GibbsL on set X to find the 6 nt long RBS pattern. In a parallel step, we look into set Y to select the 40 nt long fragments upstream to FGIOs and use GibbsL to find the 12 nt long motif. The rational for this is a search for the B Recognition Element (BRE) that in archaea may be located just upstream to the TATA box (Bell and Jackson 1998; Gehring et al. 2016). Again, the localization distance is checked to be larger than 14 nt at a 10% mode threshold. If the condition is satisfied, then the genome is assigned to Group A. It is possible that set Y is sufficiently large but the training the promoter model may not have conclusive results. This may happen when the set Y contains fragments with non-SD RBS (thus, the search for similarity with the extended SD sequence would not produce the desired result). If so, we proceed to derive the type C model (see below). An example of a genome with the model A is *Halobacterium salinarum* where more than 70% of operons have leaderless transcription (Fig. S3).

Type B model. This model is determined for bacterial genomes under assumption that both leadered and leaderless transcription occur along the genome. The approach is similar to the one used for the derivation of the type A model described above. In bacteria with possible instances of leaderless transcription we model the -10 promoter box (with length ~6 nt) which is the closest promoter site to TSS.

We select a set X of the FGIOs for which the 20 nt long upstream fragments have a local similarity to the extended Shine-Dalgarno sequence *TA**AGGAGG**TGA*. This set is then augmented with the 20 nt long fragments located upstream to all the IGIO genes, thus giving us the input set of fragments on which GibbsL is executed to find the 6 nt long RBS motif. Next, we take the set of 20 nt long fragments located upstream to the genes in set Y and run GibbsL to detect the 6 nt long motif of the bacterial promoter box (-10). If we find a motif with the *localization distance* satisfying the threshold 25%, the search for the type B model is finished. An example of a species with the type B model is *Mycobacterium tuberculosis*. (Fig. S4).

Having examined many bacterial species, we observed that, in some cases, the two motifs derived from sets X and Y could be very similar. Since set Y could not produce an SD-RBS motif (given the way set Y was selected) and that set X could not produce a bacterial promoter motif (given the prevalence of IGIOs in set X), it is unlikely that the motifs from set X and Y (when similar) constitute either an SD-RBS or a

promoter. Therefore, in that case, we proceed with the search for a model of type C, where (all) genes may have a non-SD RBS.

Type C model. This type of model is derived for the genomes with the pattern derived for FGIOs is similar to the pattern derived for IGIOs while the consensus of this pattern differs from the one of Shine-Dalgarno. This outcome is observed, e.g., in *Flavobacteriia*, *Bacteriodia* and *Cytophagia*. Since this pattern is present in IGIOs, thus it cannot be related to a promoter, the type C model could be characterized as a non-SD RBS.

To identify such a case, we compare the consensus sequences of the two motifs (from sets X and Y, as described above) derived in model B. If the two consensus sequences share three or more *consecutive* nucleotides (out of 6), they cannot make the distinct promoter & RBS pair as in model B.

No test is needed for archaeal genomes since the promoter is located more than 15 nt away from gene-start. This distance is large for an RBS (thus eliminating archaea from potentially be detected as a non-SD RBS along the line of this logic).

If the matching condition is satisfied, we derive a single 6 nt long motif by the GibbsL alignment of the 20 nt upstream regions of all the genes. Next, the consensus sequence of this motif is compared to the extended Shine-Dalgarno sequence. A significant match to the extended SD sequence constitutes at least four consecutive nucleotide identities (allowing for U-G type substitutions). If such a similarity is not present, while the motif is well localized (i.e. the peak of its position distribution is more than 15%), then we conclude that the single non-SD RBS motif is in place for this group C genome. Otherwise, we continue with the attempt to derive the type D model (see below). An example of a species with the type C model is *Bacterioides ovatus,* (Fig. S5).

Type D model. This single-motif model describes the translation initiation with the SD type RBS, *the most frequent* case in the prokaryotic genomes we have studied. To derive the type D model, we run GibbsL on the set of 20 nt long upstream regions of all the predicted genes. Next, we compute the fraction of *predicted* RBS sites (among all predicted genes longer than 300 nt) that show a local similarity to the extended Shine-Dalgarno sequence (see above). If such a fraction exceeds 0.5, the model type D derivation is finished, otherwise we proceed with the step described below. An example of a species with the type D model is *E. coli*. (Fig. S6).

Type E model. In genomes with the type E model, most genes have leadered transcription but do not have identifiable RBS. Still, the SD type RBS model could be valid for some genes. To derive this model, we select the genes whose 20 nt upstream regions contain a local similarity to the extended Shine-Dalgarno sequence. The common RBS motif is derived from this set by GibbsL. For all remaining genes, the algorithm derives the 'extended upstream signature' model, a 2nd order positional frequency model generated from the alignment of the upstream sequences with respect to the predicted gene starts.

An example of a species with the type D model is *Synechosystis* (Fig. S7).

Motif search by GibbsL

The GibbsL algorithm works with a set of *N* sequences $\{S^{(1)}, \dots, S^{(N)}\}$ such as DNA sequences located upstream to predicted gene starts. We assume for simplicity that all sequences have the same length *L*. Let $\boldsymbol{a} = \{a_1, a_2, \dots, a_N\}$ be the vector of positions, where $a_n = i$ indicates the start position of the predicted motif of fixed length *W* in sequence $S^{(n)}$. The part of sequence S that does not belong to the motif is the "background" part.

The set of motifs can be used to define patrameters of the positional (non-uniform) Markov model, M_{motif}. The set of parameters of the model M_{motif} is also known as a positional probability matrix (PPM).
Then the probability of the motif sequence in the zero-order case is:

$$P(S_{i\dots i+W-1}|M_{motif}) = \prod_{z=i}^{i+W-1} P(S_z|M_{motif}, z)$$

Here $S_{i\ldots j}$ are the nucleotides in S at positions $z = i, i+1, \ldots, j$. The background parts of the sequences $\{S^{(1)}, \ldots, S^{(N)}\}$ are used to define parameters of the "background" model, a uniform Markov model, M_{bgd}. Finally, we change the standard un-informed assumption made in the Gibbs sampling algorithm; this standard assumption is effectively about absence of any externally defined sequence marker that would change the preference for motif starting positions. In fact, the 3' ends of the sequences are supposed to be located next to the first nucleotide of the start codons. This condition imposes preferences on the distances from the motifs to the 3'end of the upstream sequence. We assume that there is a distribution of probabilities, M_{pos}, for motif to start in a given position defined over *L-W+1* possible starting positions. Collectively, the models are designated by $\lambda = (M_{motif}, M_{bgd}, M_{pos})$.

Now the probability of alignment of all motifs (putative functional sites) along with flanking background sequences can be expressed as follows.

$$P(\boldsymbol{a}\,|\boldsymbol{S}, \lambda) = \prod_{n=1}^{N} P(a_n | S^{(n)}, \lambda)$$

$$= \prod_{n=1}^{N} P(S^{(n)}_{1\ldots a_n - 1} |\, M_{bgd}) P\left(S^{(n)}_{a_n \ldots a_{n+W-1}} \middle| M_{motif}\right) P\left(S^{(n)}_{a_n+W \ldots L} \middle| M_{bgd}\right) P(a_n | M_{pos})$$

At each iteration, the distribution from which a new motif position l in sequence n is sampled, is defined by normalization of values

$$A_l^{(n)} = \frac{P(l | S^{(n)}, \lambda)}{P(S^{(n)} | M_{bgd})} = \frac{P\left(S^{(n)}_{l\ldots l+W-1} \middle| M_{motif}\right)}{P\left(S^{(n)}_{l\ldots l+W-1} \middle| M_{bgd}\right)} * P(l\,|M_{pos})$$

The overall model, $\lambda = (M_{motif}, M_{bgd}, M_{pos})$, is updated each time a new motif position is sampled.

This process is repeated in iterations. This stochastic algorithm is supposed to favor the assignments of motif positions that maximize the alignment score F: the log of probability of aligned sequences with given motif positions computed with M_{motif}, M_{bgd} and M_{pos}models divided by the probability of the same sequences computed by the M_{bgd}model only:

$$F = \log \frac{P(\boldsymbol{a}\,|\boldsymbol{S}, \lambda)}{\prod_{n=1}^{N} P(S^{(n)} | M_{bgd})} = \sum_{i=1}^{W} \sum_{j=1}^{J} c(i,j) \log \frac{M_{motif}(i,j)}{M_{bgd}(j)} + \sum_{l=1}^{L} c(l) \log M_{pos}(l)$$

Here J is the size of the alphabet (e.g. 4 in the case of nucleotidies). The function $c(i,j)$ is the number of times element j appears in position i of the predicted motifs, and $c(l)$ is the number of times motifs are located at position l. Similarly, $M_{motif}(i,j)$, $M_{bgd}(j)$, and $M_{pos}(l)$ are the probabilities of symbol j in the motif in position i, symbol j in the background and the motif start location l, respectively. After each K iterations, it is checked if a *shifted* form of the motif model results in a larger score F. This allows the algorithm to escape local optimums and, thus, construct alignments with higher scores.

Effects of the RBS Motif Width on the Gene Start Accuracy

To support the default motif width (6nt) used for RBS motifs, we tested out varying widths (from 5 to 10), and showed the overall gene-start accuracy on the genomes with experimentally verified starts. As is shown in Table S3, the motif width generally has little effect on the start accuracy, peaking at widths 6 and 9. Furthermore, we showed that in the case of *E. coli*,the RBS motifs wider than 6nt did not seem to capture additional information missed by the 6nt motif model (Fig. S11), which led us to settle on 6nt as the default value.

Effects of Addition of the Length Distribution Terms in GibbsL

A canonical form of the Shine-Dalgarno (SD) ribosomal binding site (RBS) motif is represented as AGGAGG. The abundance of G's in that sequence indicates that random sequences with high GC are more likely to exhibit similar hexamers than low GC sequences.

Consider, then, the task of searching for an RBS motif in upstream intergenic regions of length 40 nt in high GC genome. Suppose that one of these upstream sequences has an RBS motif at a distance 6 nt from the gene start. Also assume that, given it's high GC nature, that sequence has a rich-in-G hexamer located further down at a distance of 32 (way beyond the expected location of an RBS).
If the motif search algorithm does not take distance into account, then it will equally likely choosing between these two candidates, and may choose the (incorrect) farther hexamer over the real RBS motif. In the case of GibbsL, however, the distribution over RBS positions (derived from the remaining upstream sequences) is used to handle these quasi tie-breakers by preferring motifs to localized in the same region. The figure below clearly shows that this effect is growing as the GC of the sequence increases (in the case of RBS search).

Analysis of the distribution of the False Negative predictions of the COG supported genes
After running all gene-finders on the 145 genomes containing the COG supported genes, we compared the genes missed by GeneMarkS-2, Prodigal, and Glimmer. As shown in Fig. S12, all three gene finders miss 350 COG genes. We found that 188 of these genes were characterized as hypothetical proteins. From the remaining 162, 32 are 50S ribosomal proteins (with L36 being the most commonly of these proteins, 17 times), 18 labeled as transposase, and 7 are 30S ribosomal proteins. It is interesting that combination of GeneMarkS-2 and Glimmer3 would miss less number of genes than combination of GeneMarkS-2 and Prodigal. This observation suggests that GeneMarkS-2 and Prodigal have more in common than GeneMarkS-2 and Glimmer.

Supplemental Figures

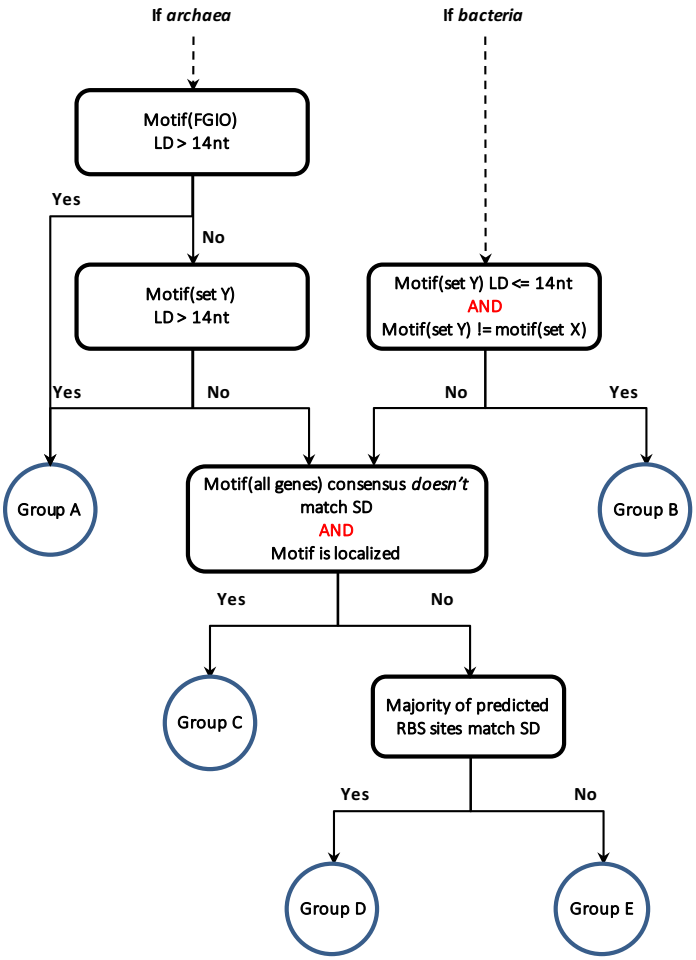

Figure S1. This figure describes the procedure of deriving the type and parameters of the model of a sequence around gene start (models A through E). See Suppl. Methods for details.

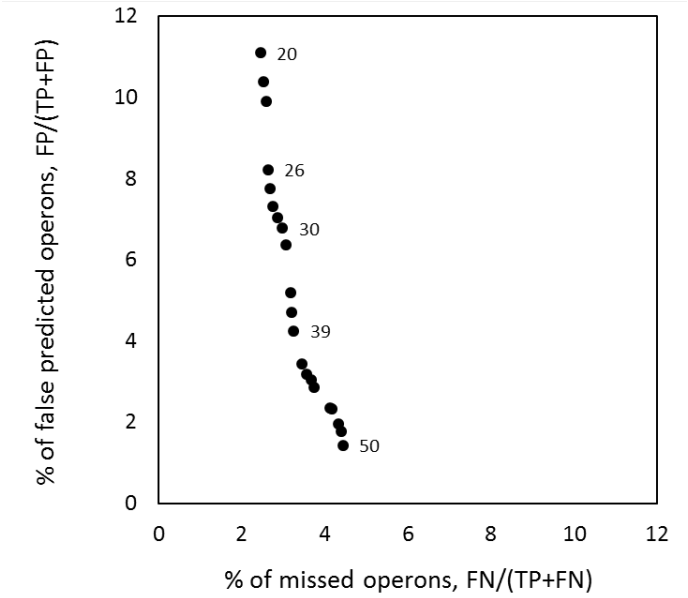

Figure S2. Dependence of false negative and false positive errors made in detection of the *E. coli* operons on the algorithm parameter (the max distance (L) between genes in the same operon, see main text). The computations used the operon coordinates from RegulonDB and gene coordinates from the RefSeq annotation at NCBI. The labels near the dots show the values of the max distance thresholds. The false negative rate is less sensitive to increase of L than the false positive rate. Particularly, increase of L from 25nt to 40nt makes ~6% reduction in false positive rate and ~1% change in false negative rate.

Models of the sequences close to gene starts. Groups A, B, C, D, E

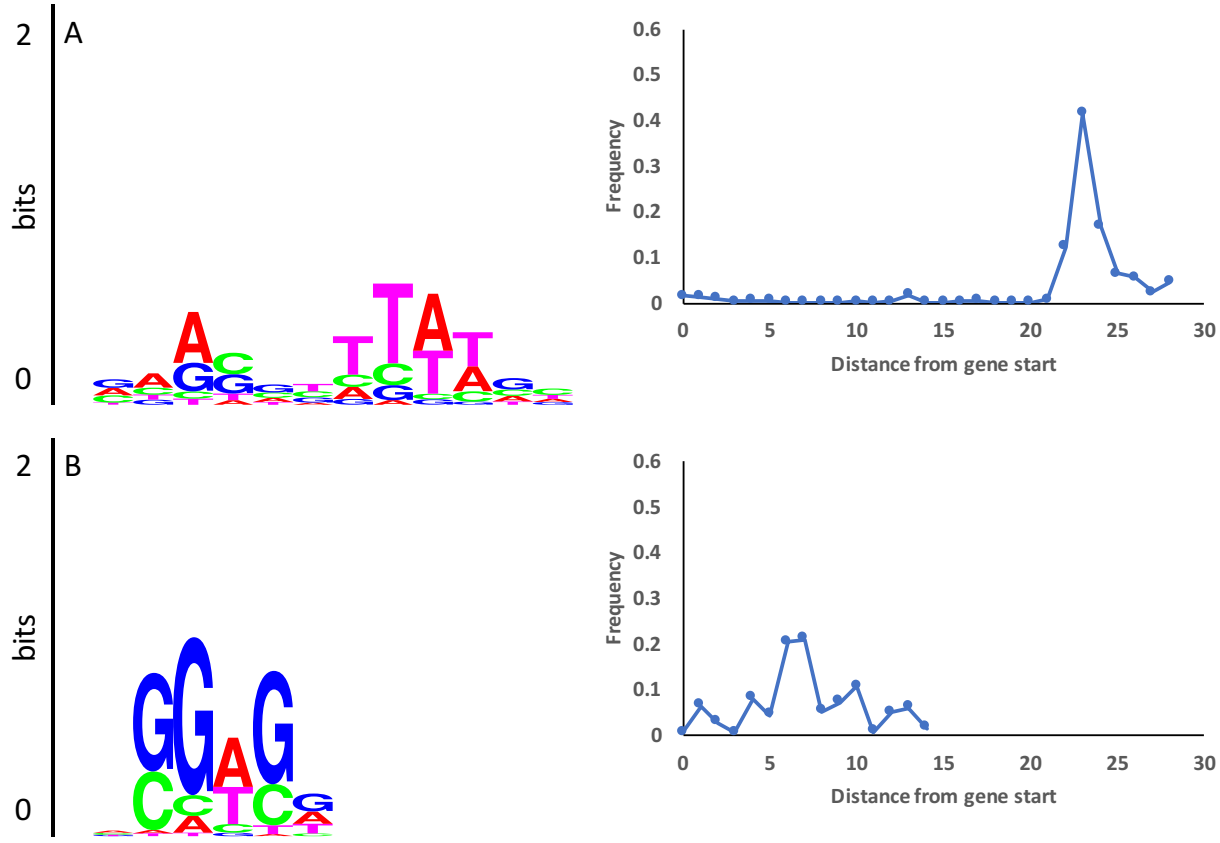

Figure S3. Group A: Archaeal species which genes have both leadered and leaderless transcription. Here for the genome of *H. salinarum* we show A) the promoter motif (leaderless transcription) and spacer length distribution; B) the RBS motif and spacer length distribution.

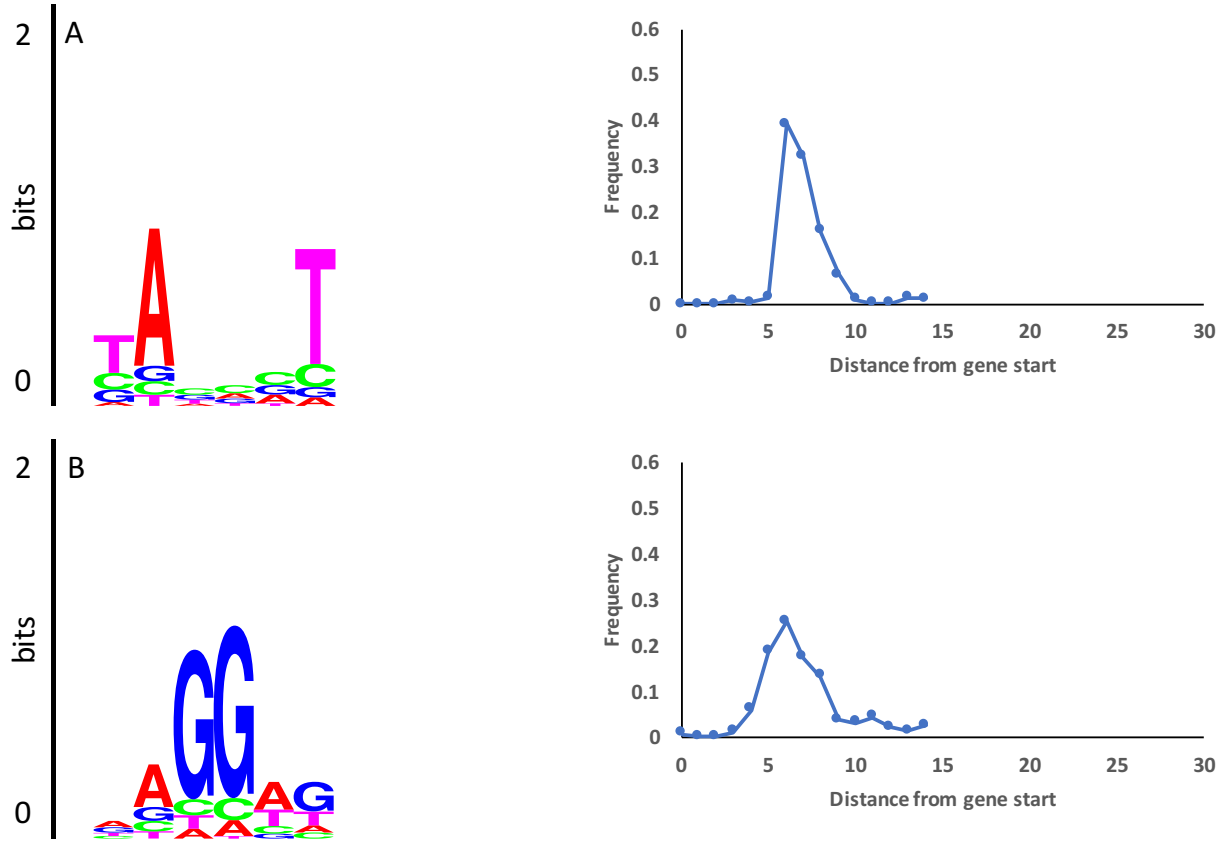

Figure S4. Group B: Bacterial species which genes have both leadered and leaderless transcription. Here for the genome of *M. tuberculosis*, we show A) the promoter motif (leaderless transcription) and spacer length distribution; B) the RBS motif and spacer length distribution.

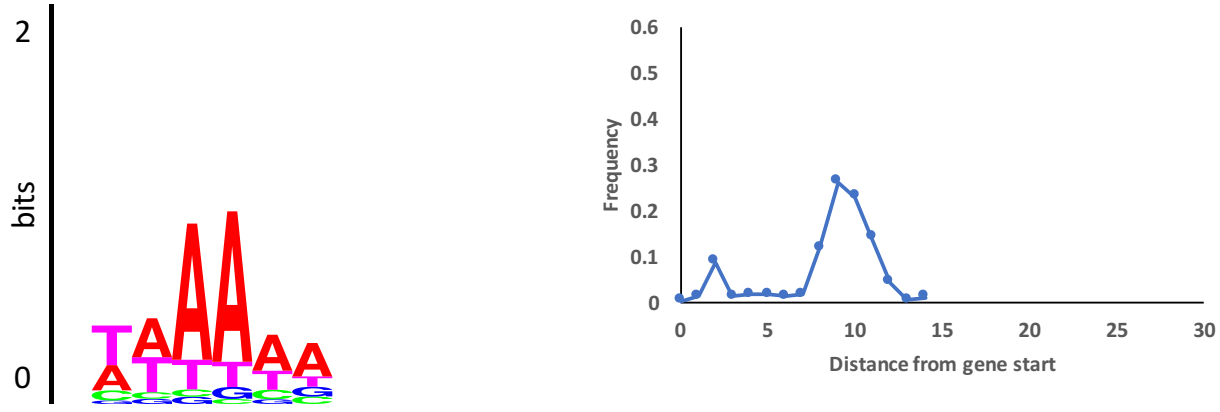

Figure S5. Group C: In the species of this group genes are transcribed in the leadered fashion and the transcripts have non-Shine-Dalgarno RBS. We show the non-SD RBS motif (logo) and the spacer length

distribution for *Bacteroides ovatus*. The consensus of the observed localized motif (with localization mode ~ 9nt) does not match the SD sequence.

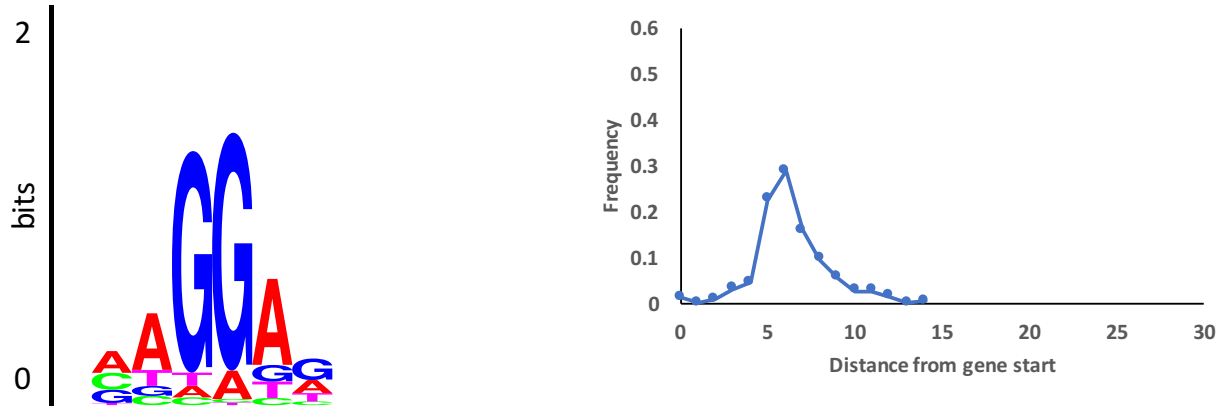

Figure S6. Group D: In species of this group genes are transcribed in the leadered fashion and the transcripts have Shine-Dalgarno RBS. We show here the SD RBS motif (logo) and the spacer length distribution for *Escherichia coli*. The consensus of the observed localized motif (with localization mode ~ 6nt) does match the SD sequence.

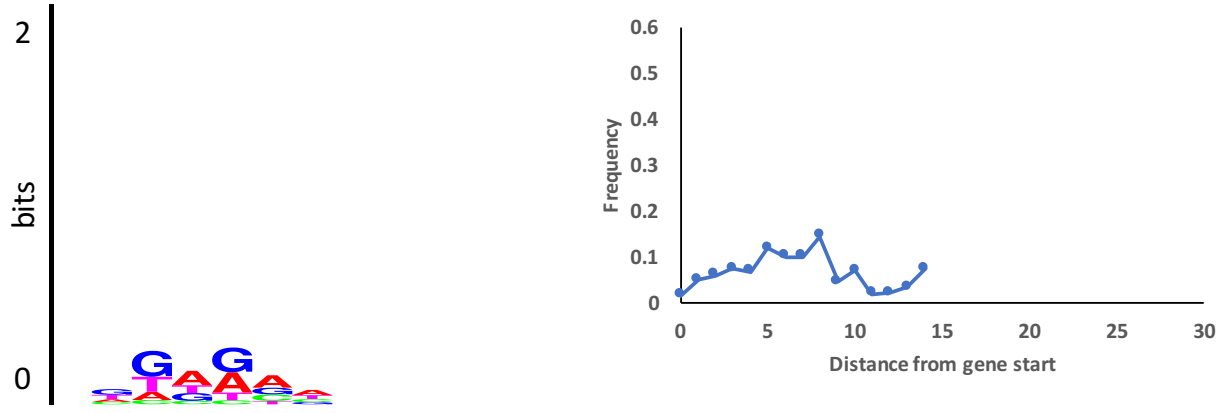

Figure S7. Group E: Genomes of this type have weak RBS signal. We attempted to extract the RBS motif from the gene upstream regions had 4nt match to the extended SD sequence (see the Main text and Suppl. Materials). We show here the RBS motif (logo) and the spacer length distribution for *Synechocystis*.

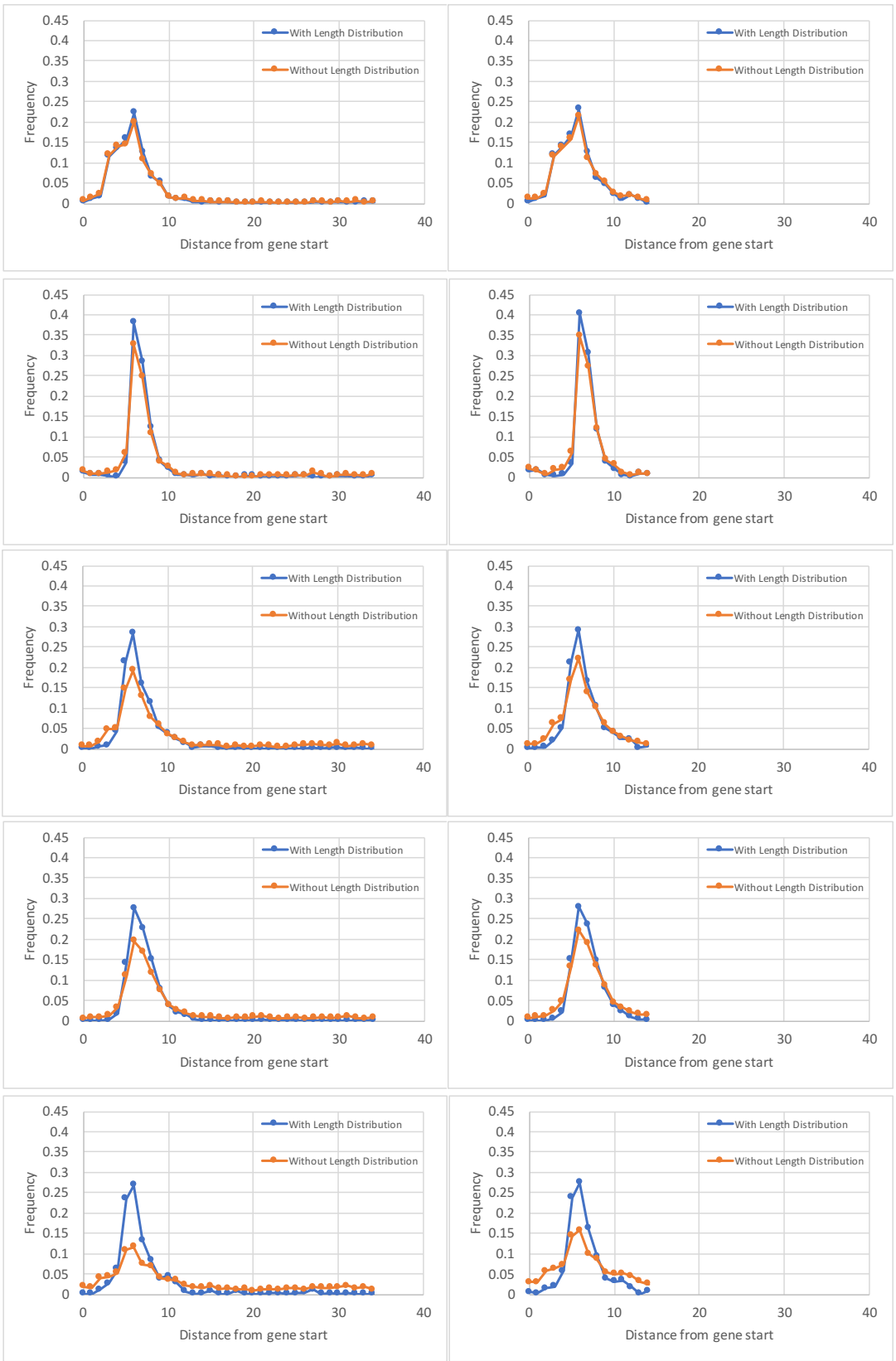

Figure S8: The graphs show distributions of spacer lengths (relative to gene starts) derived by GibbsL (with and without the localization correction) for five genomes with varying GC. For each genome, the

motif search was executed with upstream regions of lengths 40 (left panels) and 20 (right panels). From top to bottom, the species are: *Fusobacterium nucleatum* (GC = 27.2%), *Thermotoga maritima* (GC = 46.2%), Escherichia coli CFT073 (GC = 50.5%), Pseudomonas aeruginosa (GC = 66.6%), *Thermobispora bispora* (GC = 72.4%). As the GC content increases, we see that the GibbsL distance correction feature facilitates more compact localization of the selected motifs.

Reduction of the length of upstream segments from 40nt to 20nt slightly improved the outcomes of GibbsL without localization correction feature as more narrow set of the hexamers were admitted to competition. Thus, GibbsL has better performance both in case of high GC as well as in case of longer upstream regions.

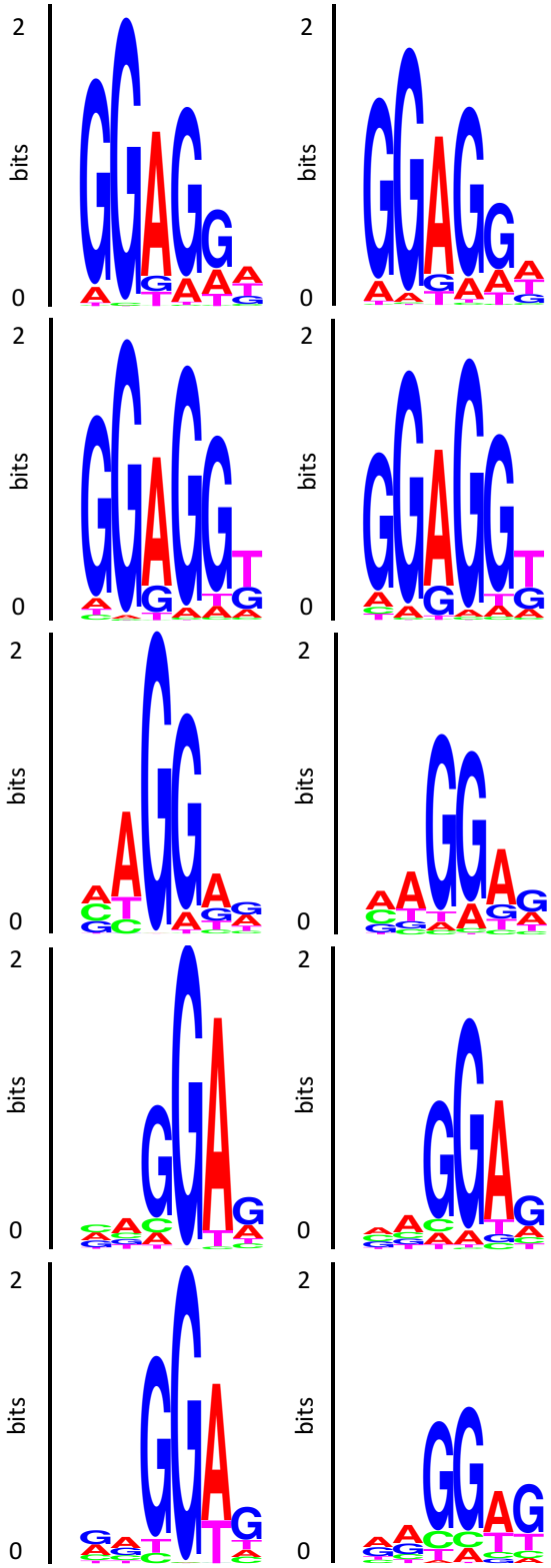

Figure S9: The motif logos that come from the five experiments described in previous Figure. The logos derived when the GibbsL localization correction feature is disabled tend to have stronger G's and A's in

the middle. This happens because the algorithm has more freedom in choosing motifs that can perfectly maximize the frequencies of conserved nucleotides, without taking into account that the motif may deviate too far from the gene start. The length of the upstream regions is 40nt. The left column shows the results when the localization correction is disabled; the right column shows the results with the feature enabled. Each row corresponds to a single species. From top to bottom: *Fusobacterium nucleatum* (GC = 27.2%), *Thermotoga maritima* (GC = 46.2%), *Escherichia coli CFT073* (GC = 50.5%), *Pseudomonas aeruginosa* (GC = 66.6%), *Thermobispora bispora* (GC = 72.4%). See more comments in the legend to Fig. S1.

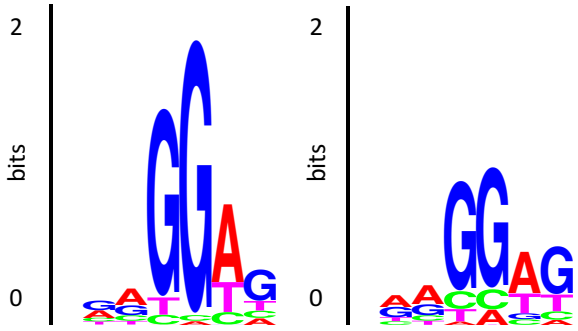

Figure S10: These images show the logos of the RBS motifs found by GibbsL for *Thermobispora bispora* (GC = 72.4%) when the length of the upstream regions is 20nt. The left panel shows the result of GibbsL with the localization correction disabled; the right panel - with this feature enabled.

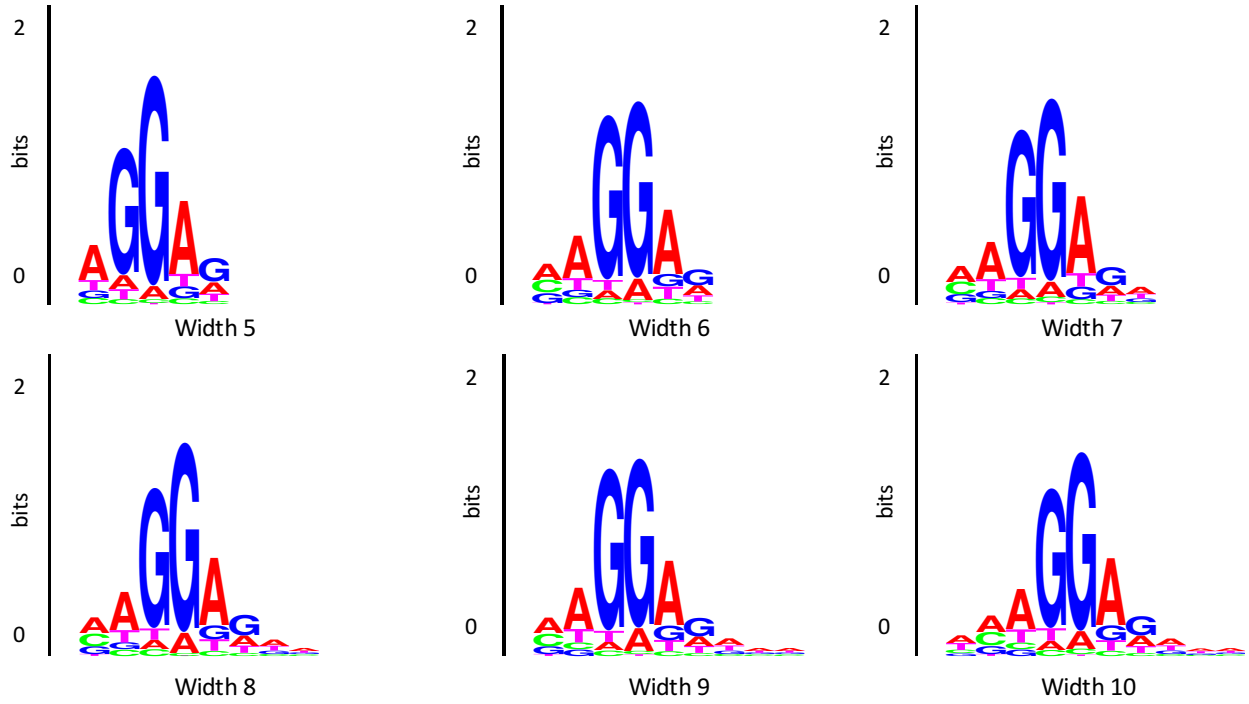

Figure S11. Motif logos for *E. coli* derived in runs of GeneMarkS-2 with differently assigned widths of the RBS motif (5nt through 10nt).

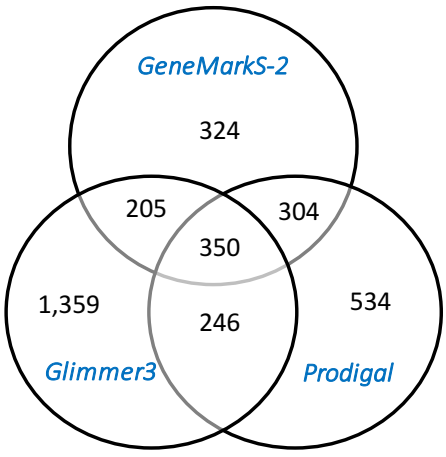

Figure S12. The Venn diagram for the COG genes missed by the three gene finders. Of the 1,605 (1,359 + 246) genes missed only by Glimmer3 and not by GeneMarkS-2, the atypical model in GeneMarkS-2 predicts 42% (679) of them, with the rest predicted by the typical model. Similarly for the 780 (534 + 246) genes missed by Prodigal and not by GeneMarkS-2, the atypical model predicts 30% (233) of them.

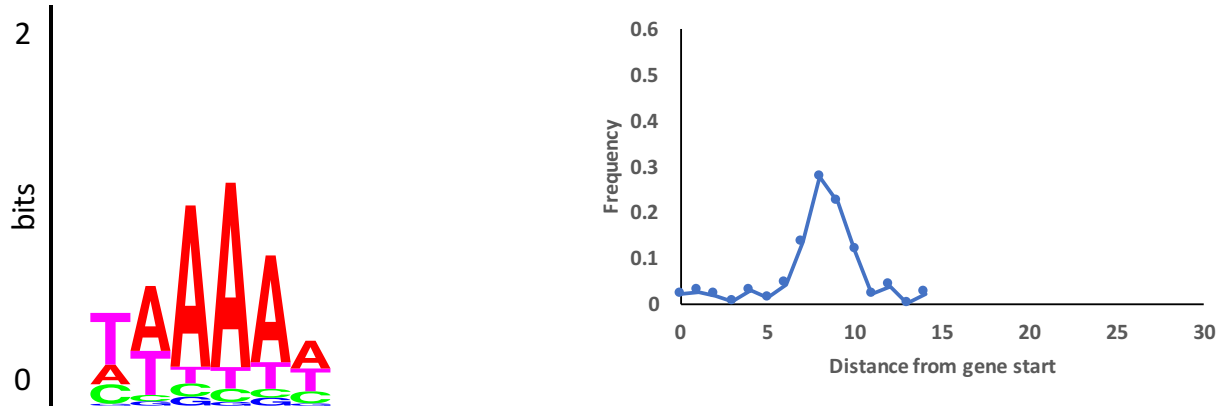

Figure S13. The motif logo and the spacer length distribution of *Flavobacterium frigiarium*, a group C genome.

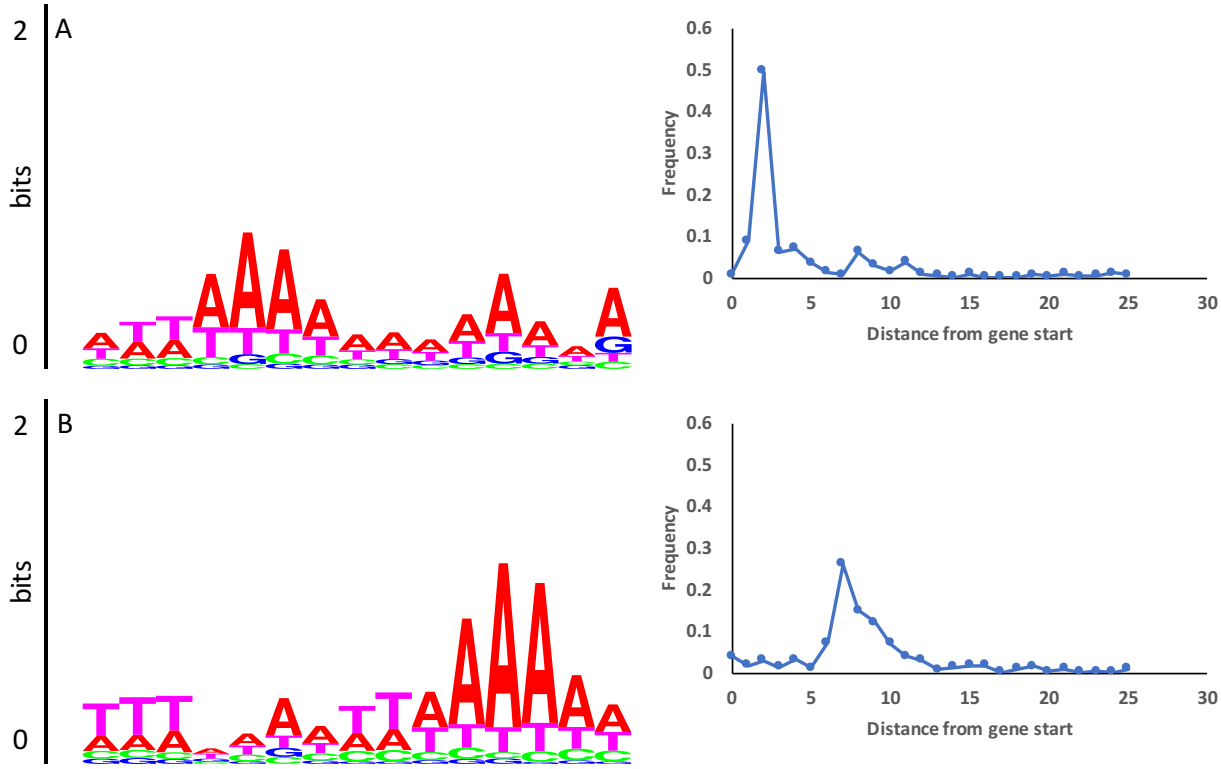

Figure S14. The motif logo and the spacer length distribution for a 15nt motif signal, for two group C genomes: (A) *Bacteroides vulgatus and* (B) *Flavobacterium johnsoniae*.

Supplemental Tables

Table S1. Prokaryotic genomes with proteomic data (*See file Supplemental_Table_S1.xlsx)*

Table S2. Prokaryotic genomes with genes validated by COG annotation (*See file Supplemental_Table_S2.xlsx)*

Species	*Gene-start model type*	*# of verified gene starts*	*Width 5*	*Width 6*	*Width 7*	*Width 8*	*Width 9*	*Width 10*
*A. pernix**	D	130	124	126	125	**127**	126	**127**
D. deserti	B	384	369	369	370	369	**371**	369
E. coli	D	769	742	740	741	743	742	**745**
*H. salinarum**	A	530	**524**	523	522	521	522	523
M. tuberculosis	B	701	632	**635**	631	**635**	634	632
*N. pharaonis**	A	315	**312**	**312**	311	311	**312**	310
Synechocystis	E	96	90	**92**	91	89	90	88
*(*archaea)*	*Total*	2,925	2,793	**2,797**	2,791	2,795	**2,797**	2,794

Table S3. The effect of the RBS motif width on the gene-start accuracy, computed over the set of 7 genomes with experimentally verified starts.

Table S4: Taxonomic representation of ~5000 genomes into five groups, based on the group classification made by GeneMarkS-2 (A through E). (*See file Supplemental_Table_S4.pdf*)
