Deriving Ribosomal Binding Site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N-terminal prediction

William S. Hayes and Mark Borodovsky
School of Biology, Georgia Institute of Technology, Atlanta, Georgia, USA,
{*william,mark*}*@amber.biology.gatech.edu*

Accurate prediction of the position of translation initiation (N-terminal prediction) is a difficult problem. N-terminal prediction from DNA sequence alone is ambiguous if several candidate start sites are close to each other. Protein similarity search is usually unable to indicate the true start of a gene as it would require a strong protein sequence similarity at the N-terminal portion of a protein where conservative regions are rarely situated. With the aid of the GeneMark program for gene identification, we extract DNA sequence fragments presumably containing ribosome binding sites (RBS) from unannotated complete genomic sequences. These DNA segments are aligned to generate the RBS model using the Gibbs' sampling method. N-terminal prediction is then performed by using the RBS model in conjunction with the GeneMark start codon prediction to aid in determining the true N-terminal site.

Abbreviations: CDS - CoDing Sequence; RBS - Ribosomal Binding Site.

1 Introduction

The accurate prediction of gene N-terminals is a difficult problem even in the case when a gene's location as a whole has been predicted. A CDS can be annotated as an ORF with an "open" start, with the true start to be selected from a set of several possible start codons. The difficulty is two-fold. The N-terminal prediction from DNA sequence alone is ambiguous in the cases when alternative starts are close to each other. Gene prediction programs, such as GeneMark[1], which has been used for analysis of several complete small genomes, do not indicate start locations with high confidence. Protein similarity searches are not always helpful since in many cases no orthologous protein is found to define the N-terminal. Even if a similar protein is found, the accuracy of annotation of the N-terminal in the library sequence may be questionable.

We consider this problem in the context of using the gene identifying GeneMark program which may employ several sets of Markov models as defined in the GeneMark method[1]. The method described below can be applied to an already partially studied genome or even to a completely unannotated genome. Training sets for the model derivation can be compiled from experimentally annotated genes. The method can also be used in the case of a new genome

when we apply the model-learning procedure GeneMark-Genesis[2] in the model derivation. To a large extent, the major focus of this research is to improve start codon prediction, though the bonus of generating RBS site predictions is a welcome one.

The procedure proposed in this paper includes several steps to generate an RBS model from unannotated sequence. The Markov models for coding and non-coding sequences are derived by the GeneMark-Genesis program[2] and the coding regions are predicted by GeneMark. Then, the predicted genes are tested for inclusion into the RBS model training sequence set. Finally, the RBS model is generated via multiple alignment through the use of the Gibbs' sampling method[3].

An important note is that the GeneMark-Genesis method produces several Markov models for gene sequences, called Root, Typical and Atypical models. The Root model, based on the largest set (cluster) of genes, is the most general. The Atypical model is trained on the gene sequences (Atypical gene cluster) that are not well predicted by the Root model. The Typical model is close to the Root model[2], and genes that are in the Typical gene cluster are mainly the genes from the Root cluster that do not belong to the Atypical gene cluster. In the case of *E. coli*, it has been shown that Typical and Atypical models roughly correspond to the models derived from *E. coli* Class I and Class III genes respectively[4]. One interesting question to consider will be the level of similarity between the RBS models derived from genes that were predicted by the Typical gene cluster model and the RBS model derived from genes that were predicted by the Atypical gene cluster model.

For each gene, the nucleotide sequence that is selected for multiple alignment is the sequence prior to the start codon, called the "pre-start" sequence. The "pre-start" sequence comprises the nucleotides from position -21 through -4, where -1 is the last base before the start codon. The analysis of information content[5] of the pre-start sequences indicates that position -21 makes a reasonable choice for the 5'-most boundary of the set of RBS sites. Experiments have shown that the RNA sequence segment from position -21 through +12 is protected by the ribosome from RNAase[5] while in a translation initiation complex. The selection of the -4 boundary was determined from a study of the width of an RBS site/start codon spacer using the expression of lacZ as an indicator of complex formation efficiency[6].

2 Materials

The complete genomic sequences of *Escherischia coli*[7] (4.639Mb), *Haemophilus influenzae*[8] (1.830 Mb), *Mycoplasma genitalium*[9] (0.580 Mb), *Methanococcus*

jannaschii[10] (1.665 Mb) and *Synechocystis* PCC6803[11] (3.573 Mb) (GenBank AC numbers: U00096, L42023, L77117, L43967 and SYNECHO respectively) were retrieved from the GenBank database, release 98. Sequence data will be shown using T instead of U in RNA context.

3 Methods

3.1 Deriving RBS model training sets

RBS models were generated by sampling sequences prior to start codons and aligning them using the Gibbs' sampling program[12]. Several different training sets were created using various criteria. The models were checked for bias towards a particular sequence (consensus sequence) and also compared with annotated 16S rRNA's from each particular organism.

One training set was generated through use of the GenBank annotation. Each CDS listed in the GenBank record for the organism was selected to provide a "pre-start" sequence for alignment. The model derived from this training set was considered as a reference model, called the GenBank model, to compare against the other models. GenBank models were derived for each of the five organisms.

For each organism, several other training sets were produced using GeneMark gene predictions. One training set was compiled from the pre-start sequences situated upstream of the 5'-most start codon among the candidate start codons indicated by GeneMark for a given gene. This training set was named the GeneMark set. With regard to the type of model used in GeneMark predictions, the Root, the Typical or the Atypical Markov models[2], we get different GeneMark sets.

Another training set was generated by selecting those GeneMark-predicted genes that had just a single candidate start listed. This set was called the Singlet set. Approximately one out of thirty GeneMark predicted genes were in this group.

To describe one more method for developing a training set, we need to explain how the GeneMark gene Start Score was calculated. The gene Start Score was determined as:

$$Score = P_{non} * P_{cod} \tag{1}$$

with P_{non} designating the probability of the absence of protein-coding properties for a sequence of the length W (GeneMark window width) situated prior to the start codon. P_{cod} is the probability of coding for a sequence of the length W situated immediately after the start codon. Up to ten candidate

starts with a GeneMark gene Start Score greater than or equal to 0.5 are listed in the GeneMark program output for each gene.

The last RBS training set was compiled as follows. Let us designate as set A, the set of the 5'-most candidate starts for each GeneMark predicted gene. Set B, a subset of set A, refers to gene start sites with a GeneMark gene Start Score greater than 0.9 under the condition that all other candidate starts for that particular gene have a GeneMark gene Start Score less than 0.2. Set B is referred to as the Partial Bootstrap set. Set C, a subset of set B, includes only those start sites which are predicted to have 100 nt or more of non-coding sequence prior to their locations (set C is referred to as the Bootstrap set). Pre-start sequences for start sites, sets B and C, were used as training sets. Approximately one in ten GeneMark predicted genes falls into Set C.

3.2 Making RBS model

The alignment of the "pre-start" sequences within windows of 3 to 8 nt widths was performed by the Gibbs sampling program. Those out of these 6 alignments which have relatively high "information per parameter" scores gave us good candidate models of RBS sites in the form of positional frequency matrices. The actual model was chosen out of these candidates based on comparison with the 3' end sequence of the 16S rRNA. The RBS model was used to score putative RBS sites upstream of each candidate start codon indicated by GeneMark.

3.3 Predicting RBS sites

The RBS model was used to predict RBS location and N-terminals. Along the region where the RBS site may be located, the probability of the RBS site starting at a particular position was defined using Bayes' rule:

$$P(RBS_i \mid Seq) = \frac{P(Seq \mid RBS_i)}{P(Seq \mid RBS_i) + P(Seq \mid \overline{RBS_i})}, \; i=-21 \rightarrow (-4+rbs\ \ length) \quad (2)$$

The position $i*$ having the maximum score was accepted as the predicted RBS location.

$$i* = argmax P(RBS_i \mid Seq) \quad (3)$$

4 Results and Discussion

The matrices of *E. coli* RBS models are listed in Tabs. 1-7. The GenBank derived model is given in Tab. 1. This model shows a strong bias towards the consensus sequence AAGGAG. This consensus sequence AAGGAG matches the 16S rRNA portion near its 3' end as seen in Fig. 1. This observation is in good correspondence with the previously proposed mechanism of ribosome binding to mRNA. The GeneMark model with starts of the gene predicted by GeneMark taken as the 5'-most possible start (Tab. 2) is very similar to the GenBank derived model. However, it does not have as strong a bias towards the consensus sequence as the GenBank derived model. The *E. coli* Singlet RBS model shown in Tab. 3 has a even weaker bias to the consensus sequence. In fact, all the Singlet models showed an inconsistent or weaker bias than the GenBank RBS models.

	Position					
	1	2	3	4	5	6
A	**0.38**	**0.55**	0.09	0.10	**0.60**	0.27
C	0.31	0.15	0.07	0.07	0.10	0.07
T	0.07	0.17	0.11	0.05	0.14	0.16
G	0.24	0.14	**0.73**	**0.75**	0.13	**0.46**

Table 1: *E. coli* GenBank derived RBS model

	Position					
	1	2	3	4	5	6
A	0.35	**0.53**	0.09	0.11	**0.60**	0.28
C	0.35	0.17	0.08	0.08	0.12	0.08
T	0.07	0.16	0.13	0.07	0.11	0.16
G	0.23	0.13	**0.70**	**0.69**	0.13	**0.43**

Table 2: *E. coli* GeneMark RBS model derived from 5'-most candidate starts

	Position					
	1	2	3	4	5	6
A	0.16	**0.48**	0.28	0.05	**0.47**	0.39
C	**0.44**	0.28	0.01	0.23	0.23	0.07
T	0.15	0.22	0.13	0.05	0.07	0.10
G	0.26	0.02	**0.58**	**0.59**	0.16	**0.44**

Table 3: *E. coli* GeneMark Singlet RBS model

	Position					
	1	2	3	4	5	6
A	**0.37**	**0.50**	0.10	0.12	**0.56**	0.27
C	0.36	0.21	0.07	0.11	0.18	0.11
T	0.05	0.18	0.08	0.07	0.12	0.13
G	0.23	0.11	**0.75**	**0.67**	0.10	**0.44**

Table 4: *E. coli* Partial Bootstrap RBS model derived from genes predicted by the Root cluster model

	Position					
	1	2	3	4	5	6
A	**0.37**	**0.70**	0.06	0.11	**0.65**	0.32
C	0.34	0.10	0.08	0.05	0.08	0.07
T	0.05	0.12	0.07	0.03	0.09	0.14
G	0.24	0.08	**0.79**	**0.79**	0.16	**0.45**

Table 5: *E. coli* Bootstrap RBS model derived from genes predicted by the Root cluster model

	Position					
	1	2	3	4	5	6
A	**0.36**	**0.58**	0.06	0.08	**0.66**	0.30
C	0.28	0.14	0.07	0.09	0.08	0.09
T	0.05	0.15	0.04	0.02	0.14	0.15
G	0.32	0.13	**0.83**	**0.78**	0.09	**0.44**

Table 6: *E. coli* Bootstrap RBS model derived from genes predicted by the Typical cluster model

	Position					
	1	2	3	4	5	6
A	**0.41**	**0.66**	0.03	0.07	**0.72**	0.28
C	0.23	0.14	0.01	0.10	0.02	0.05
T	0.06	0.07	0.07	0.04	0.12	0.20
G	0.30	0.12	**0.89**	**0.79**	0.14	**0.47**

Table 7: *E. coli* Bootstrap RBS model derived from genes predicted by the Atypical cluster model

```
Org: E.coli
  223771   225312 R 3' ATTCCTCCACTAGGTTGGCGTCCAA 5'
                  C 5' TAAGGAGGTGATCCAACCGCAGGTT 3'
Org: H.influenzae
  128125   126587 R 3' ATTCCTCCACTAGGTTGGCGTCCAA 5'
                  C 5' TAAGGAGGTGATCCAACCGCAGGTT 3'
  569195   570058 R 3' AATGAGAAGACTTAGATTCTAAAGT 5'
                  C 5' TTACTCTTCTGAATCTAAGATTTCA 3'
  770620   772158 R 3' CATTCCTCCACTAGGTTGGCGTCCA 5'
                  C 5' GTAAGGAGGTGATCCAACCGCAGGT 3'
Org: M.genitalium
  170009   171527 R 3' CTCCACTAGGTGGGGGTGCAAGAGC 5'
                  C 5' GAGGTGATCCACCCCCACGTTCTCG 3'
Org: M.jannaschii
  833908   832431 R 3' TAATATAACGATTTCAAACCTATTT 5'
                  C 5' ATTATATTGCTAAAGTTTGGATAAA 3'
 1312899  1314374 R 3' AGAAGTTATTAAAGAAAAAGAAGAG 5'
                  C 5' TCTTCAATAATTTCTTTTTCTTCTC 3'
Org: Synechocystis
 2453675  2452187 R 3' TTTCCTCCACTAGGTCGGTGTGGAA 5'
                  C 5' AAAGGAGGTGATCCAGCCACACCTT 3'
```

Figure 1: Last 25 bases of the 16S rRNA for organisms studied. R refers to 16S rRNA, C refers to the complementary sequence. The consensus-like sequences are underlined.

The *E. coli* Bootstrap RBS models are similar to the GenBank derived RBS model. The Partial Bootstrap model (Tab. 4) appears to be nearly equivalent to the GenBank derived model. In the *E. coli* case, the full Bootstrap model is more highly biased towards the 16S rRNA derived consensus than the GenBank derived model.

Differences in Bootstrap RBS models derived with the aid of the GeneMark program can be observed if one uses the Root, Typical or Atypical gene cluster models. However, these differences are marginal in the *E. coli* case (Tabs. 5, 6 and 7).

Surprisingly, the RBS model obtained by using the Atypical gene model shows a slightly stronger bias towards the consensus sequence AAGGAG than the RBS models derived with the aid of the Root or Typical gene models. This may be an interesting observation towards understanding evolution of

the genes from the Atypical cluster. Many of these Atypical genes are believed to be horizontally transferred genes.

The Bootstrap RBS models for the GeneMark Root gene cluster for the other organisms are given in Tabs. 8-11. As one can see from these tables, the RBS models for the other organisms do not show as strong a bias towards the 16S rRNA related consensus sequence as the Bootstrap RBS model for *E. coli* . The RBS model for *H. influenzae* has a pronounced bias towards consensus sequence AAGGAA where the consensus matching the 16S rRNA perfectly would be AAGGAG. In the *M. genitalium* RBS model, the consensus sequence is TTAAA. In *M. jannaschii*, the RBS model consensus is GGTGA. As can be seen from Fig. 1, *M. genitalium* and *M. jannaschii* 16S rRNA's do not match well with the consensus sequences which may indicate, as may be the case of *H. influenzae*, that not all 16S rRNA's for these species have been annotated. In the case of *Synechocystis*, although the bias to the consensus sequence {T,G}GAGA is not very strong in the RBS matrix, this consensus reasonably matches a portion of the 16S rRNA. It was suggested that the weakness of the *Synechocystis* RBS pattern was related to the abundance of the repetitive sequences CCCCA(A/G)T, TT(G/T)GTCA and CAACAGT[13]. The representation of these sequences in the pre-start sequences used in the RBS model training set is higher than expected by chance alone, yielding 2 occurences among 470 pre-start sequences for the Bootstrap set and 12 occurences among 3000 GenBank pre-start sequences. However, the fraction of the "pre-start" sequences with repetitive elements is not high enough to be responsible for any significant change in the pattern features.

	Position					
	1	2	3	4	5	6
A	**0.50**	**0.76**	0.09	0.07	**0.73**	**0.48**
C	0.08	0.05	0.04	0.04	0.04	0.05
T	0.19	0.08	0.07	0.02	0.14	0.13
G	0.23	0.11	**0.80**	**0.85**	0.06	0.31

Table 8: *H. influenzae* Bootstrap RBS model derived from genes predicted by the Root cluster model

	Position				
	1	2	3	4	5
A	0.07	0.33	**0.77**	**0.40**	**0.40**
C	0.43	0.00	0.00	0.10	0.07
T	**0.50**	**0.63**	0.03	0.07	0.23
G	0.00	0.03	0.20	0.37	0.23

Table 9: *M. genitalium* Bootstrap RBS model derived from genes predicted by the Root cluster model

Sequence logos[14] of each Bootstrap model are listed in Figs. 2,4,6,8 and 10. These logos are a representation of the information content of each model[15].

The Spacer width, the number of nucleotides between the 3' end of the RBS site and the start codon, may vary. We have analysed the Spacer width distribution using RBS site prediction based on the Bootstrap RBS model and Root gene model for each species. The histograms of the Spacer widths are shown in Figs. 3,5,7,9 and 11.

	Position				
	1	2	3	4	5
A	0.15	0.15	0.11	0.02	**0.57**
C	0.00	0.02	0.01	0.00	0.01
T	0.10	0.00	**0.60**	0.02	0.06
G	**0.74**	**0.82**	0.28	**0.92**	0.33

Table 10: *M. jannaschii* Bootstrap RBS model derived from genes predicted by the Root cluster model

	Position				
	1	2	3	4	5
A	0.08	0.08	**0.54**	0.33	**0.35**
C	0.26	0.28	0.23	0.07	0.19
T	**0.33**	0.25	0.14	0.11	0.15
G	**0.33**	**0.39**	0.09	**0.45**	0.28

Table 11: *Synechocystis* Bootstrap RBS model derived from genes predicted by the Root cluster model

Figure 2: *E. coli* Sequence Logo from Bootstrap training set after Gibbs sampler alignment

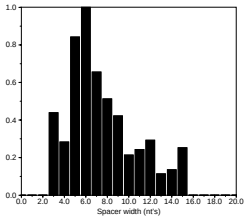

Figure 3: Histogram of *E. coli* RBS Spacer widths

Figure 4: *H. influenzae* Sequence Logo from Bootstrap training set after Gibbs sampler alignment

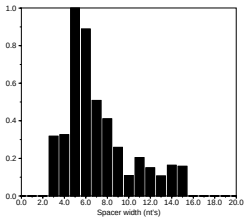

Figure 5: Histogram of *H. influenzae* RBS Spacer widths

As an example, we used the *E. coli* GenBank and Bootstrap RBS models in the GeneMark program for more accurate N-terminal prediction. The sequence from GenBank (Accession number M96795) has two of the only eleven experimentally annotated RBS sites in all of the *E. coli* GenBank records. As one may see from Figs. 12 and 13, both the GenBank and Bootstrap RBS models perform quite well.

Actually, all eleven experimentally annotated RBS sites (Accession num-

Figure 6: *M. genitalium* Sequence Logo from Bootstrap training set after Gibbs sampler alignment

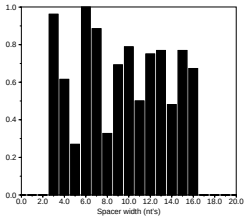

Figure 7: Histogram of *M. genitalium* RBS Spacer widths

Figure 8: *M. jannaschii* Sequence Logo from Bootstrap training set after Gibbs sampler alignment

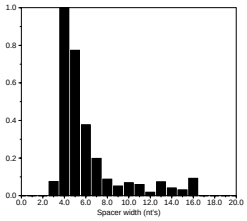

Figure 9: Histogram of *M. jannaschii* RBS Spacer widths

Figure 10: *Synechocystis* Sequence Logo from Bootstrap training set after Gibbs sampler alignment

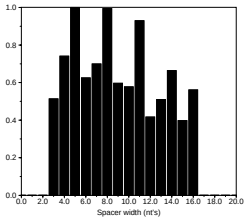

Figure 11: Histogram of *Synechocystis* RBS Spacer widths

bers: Y00720, Z47800, Z11565, X54492, M96795 - two sites, L05381, L14557, D83137, X60699 and U32495) were analyzed using the Bootstrap RBS model. For the "true" start codon in ten out of the eleven genes, the RBS Score was

Left end	Right end	DNA Strand	Coding Frame	Avg Prob	Start Score	RBS Prob	RBS Start	RBS Site
1	198	complement	fr 3	0.75	0.87	0.80	215	TATGAG
1	183	complement	fr 3	0.81	1.00	0.96	195	AATGAG
1	162	complement	fr 3	0.97	0.60	0.89	174	AAAGAT
1	99	complement	fr 3	0.97	0.34	0.81	116	CTGGAC
1	93	complement	fr 3	0.97	....	0.58	110	GCGGTT
1591	2349	direct	fr 1	0.85	0.89	0.92	1578	CAGGGA
1843	2349	direct	fr 1	0.89	0.00	0.39	1827	GCGCAT
1936	2349	direct	fr 1	0.87	0.00	0.28	1924	AATCAC
1975	2349	direct	fr 1	0.86	0.00	0.55	1960	AACGTT

Figure 12: The GeneMark output for *E. coli* sequence (GenBank, M96795) using GenBank RBS model. There are two RBS sites experimentally determined in this sequence. The first one is in complementary strand RBS: 189 to 194 with a CDS start of 183. The second one is in the direct strand, RBS: 1579 to 1583 with a CDS start of 1591. The predicted RBS sequences listed in the rightmost column are given in the 5' to 3' direction.

Left end	Right end	DNA Strand	Coding Frame	Avg Prob	Start Score	RBS Prob	RBS Start	RBS Site
1	198	complement	fr 3	0.75	0.87	0.72	215	TATGAG
1	183	complement	fr 3	0.81	1.00	0.95	195	AATGAG
1	162	complement	fr 3	0.97	0.60	0.87	174	AAAGAT
1	99	complement	fr 3	0.97	0.34	0.81	116	CTGGAC
1	93	complement	fr 3	0.97	....	0.37	110	GCGGTT
1591	2349	direct	fr 1	0.85	0.89	0.96	1578	CAGGGA
1843	2349	direct	fr 1	0.89	0.00	0.24	1827	GCGCAT
1936	2349	direct	fr 1	0.87	0.00	0.19	1924	AATCAC
1975	2349	direct	fr 1	0.86	0.00	0.50	1960	AACGTT

Figure 13: GeneMark output using Bootstrap RBS model for the same sequence as Fig. 12

found to be maximal from among all other candidate start sites. This is a clear improvement in comparison with using just the GeneMark gene Start Score which was maximal for only five "true" start codons out of eleven. The predicted RBS site significantly overlapped the experimental RBS site in nine of the eleven genes; however, one of the experimentally annotated RBS sites was located from $-5 \rightarrow +1$. This annotation indicating that the last nucleotide of the experimental RBS site is the first base of the start codon presents a special case if this annotation is correct.

Although the described method for deriving RBS models from unannotated genomic sequence seems to provide helpful information for predicting N-terminals, we still have a major problem in quantifying the N-terminal prediction performance in a more statistically stringent manner. This is due to the lack of a good control set for prediction accuracy testing. The GenBank

gene annotations often cite the longest possible ORF as a gene. There is a strong argument in favor of this tendency since one can expect about 75% of the true prokaryotic genes to correspond to the longest ORFs (Borodovsky, unpublished). However, rigorously speaking, no annotated gene can be included into a control set without experimental confirmation of the N-terminal.

The results of our experiments with the RBS models appear to be quite promising. Sufficient correlation between the models and the relevant 16S rRNA's for each organism was observed. The control prediction of eleven experimentally annotated RBS sites provided by the *E. coli* GenBank records yields good results from the use of the RBS Bootstrap model, thus, aiding in prediction of N-terminal sites.

Some open questions still exist. Obviously, the weak correlation between the 16S rRNA and RBS consensus sequence for *M. genitalium* and *M. jannaschii* poses a question as to whether some 16S rRNA's in these species are yet to be detected. A much weaker bias to a consensus sequence in *Synechocystis* than in other species may indicate either some unknown features of ribosome-mRNA complex formation, or perhaps, the presence of other signal elements in the RBS mechanism.

Acknowledgements

We wish to thank Jean-Francois Tomb and Owen White for helpful discussions. We gratefully acknowledge useful discussions with John Besemer and James McIninch who also provided valuable software programming help. We also wish to thank Paul J. Turner for providing Xmgr, a presentation quality graphing program and Steven Brenner for making the WebLogo (http://www.bio.cam.ac.uk/seqlogo) service available. The work has been supported in part by the National Institutes of Health.

References

1. Borodovsky, M. and McIninch, J.D., GeneMark: Parallel gene recognition for both DNA strands. *Comp. Chem.* **17**, 123 (1993).
2. Hayes, W.S. and Borodovsky, M., How to interpret an anonymous genome? Learning Markov models for gene identification in parallel with the sequence annotation. *Genome Research* submitted.
3. Liu, J.S., Neuwald, A. and Lawrence, C.E., Bayesian Model for Multiple Local Sequence Alignment and Gibbs Sampling Strategies. *Protein Science* **4**, 1618 (1995).

4. Borodovsky, M., McIninch, J., Koonin, E., Rudd, K., Medigue, C. and Danchin, A. Detection of New Genes in the Bacterial Genome Using Markov Models for Three Gene Classes. *Nucleic Acids Research* **23**, 3554 (1995).
5. Stormo, G.D., Schneider, T.D. and Gold, L.M., Characterization of translational initiation sites in *Escherischia coli* . *Nucleic Acids Research* **10**, 2971 (1982).
6. Ringquist, S., Shinedling, S., Barrick, D., Green, L., Binkley, J., Stormo, G. and Gold, L. Translation initiation in *Escherischia coli* : sequences within the ribosome-binding site. *Mol. Microbiology* **6**, 1219 (1992).
7. Blattner,F.R., Plunkett,G., III, Block, C.A., *et al.*, The Complete Genome Sequence of *Escherischia coli* K-12. *Science* **277**, 1453 (1997).
8. Fleischmann,R.D., Adams,M.D., White,O., *et al.*, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496 (1995).
9. Fraser,C.M., Gocayne,J.D., White,O., *et al.*, The minimal gene complement of *Mycoplasma genitalium* . *Science* **270**, 397 (1995).
10. Bult,C.J., White,O., Olsen,G.J., *et al.*, Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii* . *Science* **273**, 1058 (1996).
11. Kaneko,T., Tanaka,A., Sato,S., Kotani,H., Sazuka,T., Miyajima,N., Sugiura,M. and Tabata,S., Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. I. Sequence features in the 1 Mb region from map positions 64 to 92 percent of the genome. *DNA Research* **2**, 153 (1995).
12. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C., Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science* **262**, 208 (1993).
13. Mazel, D., Houmard, J., Castets, A. M. and Tandeau de Marsac, N., Highly Repetitive DNA Sequences in Cyanobacterial Genomes. *J. of Bact.* **172**, 2755 (1990).
14. Schneider, T.D. and Stephens, R.M., Sequence Logos: A New Way to Display Consensus Sequences. *Nuc. Acids Res.* **18**, 6097 (1990).
15. Schneider, T.D., Stormo, G.D. and Gold, L.M., Information Content of Binding Sites on Nucleotide Sequences. *Journal Mol. Biol.* **188**, 415 (1986).