BIOINFORMATICS

Vol. 20 no. 18 2004, pages 3308–3317
doi:10.1093/bioinformatics/bth390

# Accuracy improvement for identifying translation initiation sites in microbial genomes

*Huai-Qiu Zhu[1], Gang-Qing Hu[1], Zheng-Qing Ouyang[1], Jin Wang[3] and Zhen-Su She[1,2,*]*

*[1]State Key Lab for Turbulence and Complex Systems and Center for Theoretical Biology, Peking University, Beijing 100871, China, [2]Department of Mathematics, UCLA, Los Angeles, CA 90095, USA and [3]State Key Lab of Pharmaceutical Biotechnology, Nanjing University, Nanjing 210093, China*

Received on March 26, 2004; revised on June 4, 2004; accepted on June 29, 2004
Advance Access publication July 9, 2004

**ABSTRACT**

**Motivation:** At present the computational gene identification methods in microbial genomes have a high prediction accuracy of verified translation termination site (3′ end), but a much lower accuracy of the translation initiation site (TIS, 5′ end). The latter is important to the analysis and the understanding of the putative protein of a gene and the regulatory machinery of the translation. Improving the accuracy of prediction of TIS is one of the remaining open problems.

**Results:** In this paper, we develop a four-component statistical model to describe the TIS of prokaryotic genes. The model incorporates several features with biological meanings, including the correlation between translation termination site and TIS of genes, the sequence content around the start codon; the sequence content of the consensus signal related to ribosomal binding sites (RBSs), and the correlation between TIS and the upstream consensus signal. An entirely non-supervised training system is constructed, which takes as input a set of annotated coding open reading frames (ORFs) by any gene finder, and gives as output a set of organism-specific parameters (without any prior knowledge or empirical constants and formulas). The novel algorithm is tested on a set of reliable datasets of genes from *Escherichia coli* and *Bacillus subtillis*. MED-Start may correctly predict 95.4% of the start sites of 195 experimentally confirmed *E.coli* genes, 96.6% of 58 reliable *B.subtillis* genes. Moreover, the test results indicate that the algorithm gives higher accuracy for more reliable datasets, and is robust to the variation of gene length. MED-Start may be used as a postprocessor for a gene finder. After processing by our program, the improvement of gene start prediction of gene finder system is remarkable, e.g. the accuracy of TIS predicted by MED 1.0 increases from 61.7 to 91.5% for 854 *E.coli* verified genes, while that by GLIMMER 2.02 increases from 63.2 to 92.0% for the same dataset. These results show that our algorithm is one of the most accurate methods to identify TIS of prokaryotic genomes.

**Availability:** The program MED-Start can be accessed through the website of CTB at Peking University: http://ctb.pku.edu.cn/main/SheGroup/MED_Start.htm

**Contact:** she@pku.edu.cn;she@math.ucla.edu

## INTRODUCTION

Driven by the progress of genome sequencing technology, the need for accurate gene prediction continues to grow. At present, computational gene identification methods for prokaryote genomes have up to sensitivities of 98–99% or higher (Delcher *et al*., 1999; Suzek *et al*., 2001). But their accuracy scores are commonly estimated by comparing the locations of verified translation termination codon of protein coding genes; in other words, the detected genes are open reading frames (ORFs) with an 'open' start. Such annotation does not provide full information for protein coding genes. In order to analyze the putative protein of a gene, it is important to know accurately TIS, the prediction of which is one of the remaining open problems for prokaryotic gene finding. Although many methods for identifying translation initiation site (TIS) have been developed, validation of their actual accuracy has not been carried out thoroughly because of insufficient experimentally confirmed gene start data and a shortage of reliable data for training and testing. In addition, the absence of relatively strong sequence patterns of true gene start sites aggravates the difficulty of identifying gene start codons. Without reliable computational methods, the rule of the 'longest ORF' was often applied to annotate complete microbial genomes with the assignment of TIS to the leftmost start codon (Besemer *et al*., 2001).

In bacterial mRNAs, during the process of protein synthesis, the ribosome binds around initiation site to protect this region from being degraded while ribonuclease is added to the blocked initiation complex. Typically, the ribosome binds

*To whom correspondence should be addressed.

to a region near the 5′ end of the mRNA known as ribosomal binding site (RBS). In order to recognize the RBS, two common features are involved: the start codon and a conserved stretch 7 or 13 bp upstream of the start codon, which forms base pairings with the 16S rRNA in the small ribosomal subunit. This stretch is known as Shine–Dalgarno (SD) sequence, its content may vary among prokaryotes, but is generally highly conserved, a reflection of the high conservation of the 16S rRNA sequences (Shine and Dalgarno, 1974; Mikonnen *et al*., 1994; Lewin, 2000).

There are some pioneer works in the studies of TIS. Stormo *et al*. (1982) presented their results in the early 1980s on the computational characterization of gene starts in prokaryotes. An algorithm (Schurr *et al*., 1993) has been developed to calculate the optimal binding energy between the 16S rRNA of *Escherichia coli* and the upstream regions of a potential start codon, and the difference in the binding energy distribution for upstream regions of true start sites and spurious, gene internal, start codons. In this approach, the binding energy function needs to be provided with prior knowledge about the 16S rRNA sequence, or derived from closely related organism's 16S rRNA. Many algorithms have also been developed to detect various DNA functional sites, including RBS (Hayes and Borodovsky, 1998; Hannenhalli *et al*., 1999; Tompa, 1999). Currently, most methods for TIS prediction employ some RBS model which is either derived by a supervised training or inferred from prior knowledge of the organism-specific 16S rRNA sequence (Besemer *et al*., 2001). The GeneMark and GeneMark.hmm programs (Borodovsky and MaIninch, 1993; Lukashin and Borodovsky, 1998; Hayes and Borodovsky, 1998) use a RBS model in a form of positional nucleotide frequency matrix whose parameters are derived by Gibbs sampling multiple alignment of DNA sequences located at the upstream regions of annotated TIS. GeneMark.hmm 2.0 additionally uses a probability distribution of the length of a spacer, the sequence between the RBS motif and start sites. The latest implementation of GeneMarkS (Besemer *et al*., 2001) utilizes a non-supervised training procedure incorporating GeneMark.hmm to find prokaryotic TIS. The Frame-by-frame program (Shmatkov *et al*., 1999) employs a hidden Markov model (HMM) with several hidden states modeling the trinucleotide frequency pattern specific for upstream regions and downstream sequences of TIS. The RBS model is also used in the ORPHEUS program (Frishman *et al*., 1998) with a weight matrix form of positional frequencies normalized by the frequency of the most probable nucleotide in a given position. In addition, the ORPHEUS model takes into account the pattern of the spacer length variation.

Two of the most recent works are noteworthy: RBSfinder (Suzek *et al*., 2001) and GS-Finder (Ou *et al*., 2004) to which comparisons with our present algorithm are made below. GLIMMER (Delcher *et al*., 1999) assigns, by default, the predicted gene start to the start codon of the longest ORF containing predicted coding region. The RBSfinder (Suzek *et al*., 2001), developed to postprocess the annotation by GLIMMER and other gene finders, uses the entire genomic and first-pass annotation to train a probabilistic model that scores candidate RBS surrounding originally annotated start codons. When a better RBS is found either upstream or downstream the originally predicted start site, then the system replaces the originally annotated start site by the new one. The GS-Finder program developed by Zhang *et al*. (Ou *et al*., 2004) has also designed a self-training method in which six variables are introduced to describe consensus signals (e.g. the SD sequences) in the vicinity of gene starts, a coding potential of DNA sequences near the start codon, the content of the start codon itself and the distance between the leftmost start codon and the candidate start codon, respectively. The former four variables were derived based on the *Z*-curve method (Guo *et al*., 2003; Ou *et al*., 2004), while the latter two variables were given as empirical constants or formulas.

In this paper, we develop a new algorithm based on a statistical model that integrates multiple sources of information about the RBS sequence and the start codons, and then design an iterative self-training system, MED-Start, to improve the accuracy of the gene start prediction. We begin by finding a set of 'candidate motifs' from the upstream regions of pre-predicted coding ORFs (e.g. with leftmost start codons supplied by an automatic gene finder such as GLIMMER 2.02). Part of the candidate motifs are selected to be 'hit motifs' based on a statistical calculation of their position-dependent property; they are considered to be the most significant consensus signals like the SD sequence. For prokaryote genomes with known 16S rRNA, the hit motifs correspond to motifs in the 16S rRNA sequence. It is important that our algorithm does not rely on the prior knowledge of the consensus sequence. We then iteratively characterize true TIS using the weight matrix of the nucleotide alphabets around the chosen start codon and the probability weight of a relative distance between the stop site and the multiple start codons in the same reading frame. Combining them with the weight matrix of the hit motifs, and the probability of spacer length distribution for each hit motif, we finally build a four-component statistical model to describe the TIS of prokaryotic genes which may perform self-training, and in turn to relocate the most likely gene start sites in a predicted coding ORF.

For any prokaryotic genome, the model quickly converges to a set of four statistical parameters which are organism-specific. We believe that our model is simple and the parameters may have significant connection with the functional and evolutionary properties of the genes. The MED-Start system is tested on a validated set of genes from *E.coli*, for which it improves the accuracy of the TIS predicted by a common computational gene finder from 66 to 68% to more than 95%. The MED-Start is developed to refine an initial prediction of the gene starts, and is designed as a postprocessor for MED 1.0, a gene prediction system based on multivariate

entropy distance (MED) method (Ouyang *et al*., 2004). As RBSfinder (Suzek *et al*., 2001), the MED-Start must be run after a gene prediction program, and may also be used to process the output of other programs for microbial gene finding such as GeneMark and GLIMMER.

## MATERIALS AND ALGORITHMS

### Reliable genome sequences data as benchmark

Sequence data used in this paper include the following genomes available in the GenBank database: *E.coli* (Blattner *et al*., 1997), *Bacillus subtillis* (Kunst *et al*., 1997), *Archaeoglobus fulgidus* (Klenk *et al*., 1997), *Haemophilus influenzae* (Fleischmann *et al*., 1995), *Methanococcus jannaschii* (Bult *et al*., 1996) and *Thermotoga maritime* (Nelson *et al*., 1999).

It has been known that the GenBank database annotation of complete microbial genomes has a systematic bias towards TIS annotation by the longest ORF rule (Besemer *et al*., 2001). Since most of genes in bacterial genomes, including those in *E.coli*, have had their start sites predicted computationally rather than experimentally, we have selected only the datasets with experimentally confirmed location of genes. For *E.coli*, we have used two datasets, the EcoGene and Link. The EcoGene database (Rudd, 2000) contains 854 proteins that have been confirmed by N-terminal protein sequences (downloaded from EcoGene website: http://bmb.med.miami.edu/EcoGene/EcoWeb/CESSPages/VerifiedProts.htm, the newest dataset gives 862 genes, but 8 of them are excluded here, in which the length is not a multiple of three, or at least one stop codon exists in the same reading frame of the true stop codon, or the start codon is not one of the canonical start codons ATG, GTG, TTG and CTG). The Link dataset (Link *et al*., 1997) contains 195 N-terminally confirmed genes, which is a subset of EcoGene and contains only genes that either have a processed leader sequence of a single amino acid or do not have a processed leader sequence and therefore do not require an estimation of the correct TISs based on a putative leader sequence. Thus, the Link dataset is usually assumed to be slightly more reliable than the EcoGene dataset. For *B.subtillis*, we used Bsub58 dataset (Yada *et al*., 2001) with 58 genes confirmed by comparison with homologous sequences of *B.halodurans*, as well as the Bsub1248 dataset (Yada *et al*., 2001) with 1248 'non-y' (i.e. experimentally characterized) genes.

In order to examine the performance of the TIS prediction programs (including ours) for short genes, we have also used three sets of short genes with length of 300 bp or shorter, i.e. Bsub123, Bsub72 and Bsub51 dataset, that are selected from the *B.subtillis* genomic sequence and verified by protein similarity search (Besemer *et al*., 2001). The first set Bsub123 includes 123 genes whose protein products possess at least one significant sequence similarity with known protein. The Bsub72 set comprises 72 genes with at least two strong similarities at a protein level. The Bsub51 set includes 51 genes whose protein products have at least 10 strong similarities to known proteins. Note that the start sites of Bsub1248 dataset sequences are not always verified experimentally (Yada *et al*., 2001), and the computational prediction of short gene is currently far from satisfactory, thus Bsub58 dataset can be regarded as the most reliable set of *B.subtillis*.

### Longest ORF rule

With a lack of reliable computational methods for prokaryotic gene start prediction, the rule of the 'longest ORF' was frequently used to annotate complete genomes with gene start assigned to the leftmost start codon. However, there is compelling evidence that systematic bias exists by the use of the longest ORF rule (Besemer *et al*., 2001). Indeed, there is a significant portion of the leftmost ATG (less often GTG or TTG) that are not true start codons. A simple statistical analysis on the accuracy of this rule may be obtained using confirmed gene datasets. For example, 62.9 and 63.0% of 854 EcoGene and 1248 Bsub1248 genes start with the leftmost start codons. Both the GLIMMER gene finder (Delcher *et al*., 1999) and MED 1.0 (Ouyang *et al*., 2004) gene prediction program were designed to locate the leftmost start codon as the predicted start site of gene. As a postprocessor for MED 1.0, our MED-Start system aims at correcting the prediction of TIS with respect to the longest ORFs prediction by MED 1.0 or GLIMMER. If the MED-Start processes a gene predicted by another gene finder (for instance GeneMark), which does not apply the longest ORF, it first automatically extend the ORF to a longest one, and then goes on.

### MED-Start algorithm outline

The MED-Start new algorithm uses a four-component statistical model to describe TIS of prokaryotic genes. The parameters of the model represent important features of biological significance, which include the correlation between translation termination site and TIS in the same reading frame of the gene, the sequence content around a TIS, the sequence content of the consensus signal related to RBS; and the correlation between a TIS and the upstream consensus signal. The parameters are determined by an unsupervised learning. The final model with convergent parameters was then used to select the most likely start codon for each gene predicted. We now describe in detail the steps of the new algorithm.

*Finding candidate motifs in upstream regions of predicted coding ORFs* Motif usually means a subsequence that is well preserved over several sequences, and the occurrences of the motif in those sequences are called instances (Keich and Pevzner, 2002). The motifs in DNA or protein sequences may indicate functional connections, such as RBS in prokaryotes. Usually, the motif is described by some

model, which represents the similarities among the different instances. In this paper, we use the term, $(l, d)$ motif, to refer to the situation where a consensus string of length $l$, without wildcards, and the instances must differ in at most $d$ positions from the consensus (Keich and Pevzner, 2002).

Since the SD signal tends to be a preserved feature in the upstream regions of bacterial gene starts and actually most of the start codons of the longest ORF are real gene starts, it is natural to assume that the SD signal should be often found in the upstream regions of the leftmost start codons. Previous reports indicate that this assumption is correct (Delcher *et al*., 1999; Ouyang *et al*., 2004).

We first search for $(l, d)$ string (as defined above) within $L$ bps upstream of the start codon of the longest ORF in the original annotation (the default values are $l = 5$, $d = 0$, $L = 20$). In order to remove many false positive cases, the initial search is restricted to ORFs longer than 300 bp. For instance, a (5, 0) string is a word of five letters with zero variation that appears in many sequences within 20 bp upstream of the start codons. We select several strings with the highest frequency of occurrence as the candidate motifs. Currently, the system lists the top five ($l$, $d$) strings as the default candidate motifs, part of which may be potential substrings of the SD sequence. Note that, in the next iteration step, the search for candidate motifs will be conducted within $L$ bps upstream regions of the adjusted start sites that may not be the start codon of the longest ORFs. This means that the training sequences, i.e. $L$ bps long upstream regions of start sites of all the training ORFs are updated constantly until the iteration reaches convergence. Note that although the size $l$ of motif string is provided with optional values, our test shows that the default option $l = 5$ appears to generate the best overall performance for our algorithm. The choice is also consistent with that used in the RBSfinder system (Suzek *et al*., 2001).

*Determining hit motifs and their alignment weight matrix* Once the candidate motifs are selected, one goes back to the same set of training sequences again. For each candidate motif, the algorithm searches for its relatives, i.e. $(l, 1)$ instances which differ from the candidate motif by at most one letter. These instances are regarded as candidates for SD signal-like substring. We then calculate the distribution of the location of the last nucleotide of the occurred instance to the first nucleotide of the start codon, which will be referred to as the spacer distribution. As reported earlier (Suzek *et al*., 2001), the RBS-related motifs occur with a strongly position-biased property. We expect that a true motif has a characteristic distribution. We use a deviation $\sigma$ of spacer distribution to characterize each candidate motif. At the $k$-th iterative step, let $p_i^{(k)}$ be the occurrence probability for the candidate motif [including all its $(l, 1)$ instances] at upstream position $i$ from the end of the training sequence for this iterative step, $\sigma$ is calculated by using the formula

$$\sigma = \sqrt{\frac{\sum_{i=l}^{L}\left(p_i^{(k)} - \overline{p}^{(k)}\right)^2}{L - l + 1}},$$

where

$$\overline{p}^{(k)} = \frac{1}{L - l + 1}\sum_{i=l}^{L} p_i^{(k)}. \quad (1)$$

Higher value of $\sigma$ means that the candidate motif occurs in the training sequences with stronger position-biased property, associated with the most likely position of a RBS in the upstream regions of start sites. Therefore, it is reasonable to select those candidate motifs with higher $\sigma$. Usually the algorithm chooses the one having highest $\sigma$, to be so-called 'hit motif', meaning the most significant motif associated with the binding sites upstream the real TIS of genes. In addition, if there exists more than one candidate motif having nearly the same $\sigma$ to the highest one, the algorithm will select all of them, but at most three motifs, as the hit motifs. Therefore, by default option, in each iterative step, MED-Start system will provide at least one hit motif as potential SD signals for a given prokaryotes.

After hit motifs are determined, we compute the positional weight matrix of each hit motif, by a multiple alignment of all its ($l$, 1) instances occurred within training sequences. By the assumption that the hit motifs should be similar to a substring of SD sequence, the algorithm calculates the alignment weight matrix of $[3 + l + 2]$ bp size of window around the hit motif. Our test also demonstrates that the size of window around motif leads to the best overall performance. Therefore the alignment weight matrix may be written as $w_{SD}^{(k)}(b_i, i)$ for $b_i \in$ {A, C, G, T}, where $i$ means position within these alignment windows and $(3 + l + 2) \geq i \geq 1$.

*Weight matrix for start codon context* For prokaryotes, the sites where protein synthesis is initiated may be recognized by binding the ribosome, so that an isolated fragment of protected sequence is formed. This protected sequence has two consensus elements, both SD signal and context around start codon ATG (or less often, GTG and TTG) (Lewin, 2000). To detect the context feature of start codon, under the condition that the true start codons are unknown, fragments around start point of whole training ORFs are taken into account. We calculate their positional probability within the alignment windows around start codon with length of $(4 + 3 + 15)$ bp, where the number 3 is the length of start codon ATG or GTG and TTG, the numbers 4 and 15 correspond to the length of sequences upstream and downstream of start codon. The choice of these numbers is made as the result of our test to obtain the overall best performance of the algorithm. Note that the position weight matrix around the start codon is related to the sequences with frequency pattern around TIS, which

actually may be an indicator for the strength of the SD signal upstream (Shmatkov *et al*., 1999). Thus, a weight matrix of start codon for the predicted coding longest ORFs (longer than 300 bp) as training set is obtained, and the usage of ATG, GTG and TTG may be reflected in this weight matrix automatically. We may represent the weight matrix by $w^{(k)}_{\text{Start}}(b_i, i)$ for $b_i \in${A, C, G, T}, where $(k)$ means the $k$-th iterative step and $i$ means position within these alignment windows and $(4 + 3 + 15) \geq i \geq 1$.

Despite the difficulty of unknown true start codons, we may reach an approximation through this weight matrix, because nucleotides occur more randomly around the false start codons. As the iteration goes, we observe that the weight matrix indeed converge to a final weight matrix which is very similar to that obtained from the analysis of verified gene starts.

*Weights for potential start codons behind the leftmost start codon* The fourth component of our model is a probability that describes the likelihood for a start codon of order $m$ counting from the left most one to be a true start site. Although the true start codons are usually the leftmost ones, many ORF with multiple candidate start codons in the same reading frame may assign the true start site to be the second, third or other start codon downstream from the leftmost one. This implies that start codons in the same ORF starting from the left most one are not equally likely to be a true start site. For instance, there are 62.9% of 854 EcoGene genes starting with the leftmost start codons, while 20.1, 10.2, 3.0 and 1.9% of them start with the second, third, fourth and fifth ones respectively. Thus, different weights should be assigned to different start codons when they are investigated whether to be TIS (Ou *et al*., 2004). Suppose there are more than one start codon (ATG, GTG or TTG), $m$ is the index of start codons, we define $w^{(k)}_m$ as the weight of the $m$-th start codon being true gene start site, $k$ is the iterative step. For instance, $w^{(k)}_1$ is the weight of the leftmost start codon obtained at the $k$-th iteration step. For $k = 1$, i.e. in the initial condition in the iteration, we set an equal weight 1.0 to each $w^{(k)}_m$, i.e. $w^{(1)}_1 = w^{(1)}_2 = \cdots = 1.0$. After the relocation of the start site, $w^{(k)}_m$ is calculated by counting the number of the $m$-th start codon being actually chosen as the start site. This iteration rapidly converges (within four steps at most, see below).

*RBS score for start codon and the most-likely start codon* For each start codon in the same reading frame of a predicted ORF, MED-Start makes a combined RBS score based on the four statistical parameters described above. Each of the above four measurements translates to a probability measure, then the combined score reads:

$$\Phi_i = \log(P_1 \cdot P_2 \cdot P_3 \cdot P_4). \tag{2}$$

The equation above is calculated for each $l$mer occurred at position $i$ from the start codon within the $L$ bp upstream. Where

$$P_1 = p^{(k)}_i, \quad l \leq i \leq L \tag{3}$$

is the occurrence probability as the hit motif at upstream position $i$ from the start codon;

$$P_2 = \prod_{j=1}^{2+l+3} w^{(k)}_{SD}(b_j, j) \tag{4}$$

is given by standard positional weight matrix of the alignment windows described in the previous section for the hit motif, and $j$ is the position within the aligned windows;

$$P_3 = \prod_{j=1}^{4+3+15} w^{(k)}_{\text{Start}}(b_j, j) \tag{5}$$

is given by standard positional weight matrix of the aligned windows described in the previous section around the start codon of ORF, and $j$ is the position within these alignment windows; and finally

$$P_4 = w^{(k)}_m \tag{6}$$

is given by the weight for the $m$-th start codon from the leftmost start codon of the predicted longest ORF. The superscript $(k)$ in Equation (3)–(6) is the number of the iteration.

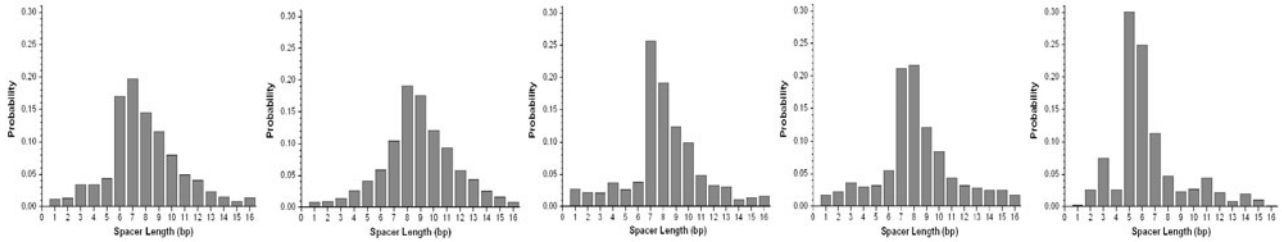
A simple procedure for the iteration may be defined. At each step with a set of given candidate TISs (e.g. beginning with the leftmost start codon), we check the scores $\{\Phi_i\}$ $(l \leq i \leq L)$ for all $l$mers occurred within the $L$ bp upstream regions for each start codon, and select the maximum of $\{\Phi_i\}$ as the RBS score for this start codon, i.e.

$$S^{(k)}_m = \max_{l \leq i \leq L} \{\Phi_i\}, \tag{7}$$

where the subscript $m$ is the order of the start codon starting from the leftmost start codon of the predicted longest ORF. If there exist several hit motifs, the system calculates the scores $\{\Phi_i\}$ of the $i$-th $l$mer according to different hit motifs, and select the maximum one as the score $\{\Phi_i\}$ of that $l$mer.

We then compare the RBS score $S^{(k)}_m$ of different start codon and choose one with the highest score as the most likely candidate for the TIS. The $k$-th iteration completes when all candidate start sites are tested and updated. We then repeat the calculation of candidate motifs and hit motifs and all other probability measures with reference to the newly updated candidate TIS. The iteration begins at the next step.

*Convergence of self-trained model and the final parameters* The rules described above determine the most likely start codons based on RBS score. At each step, after the most likely start codons were relocated, the $L$ bp long upstream regions of those start codons were used as new training sequences to detect SD-like signals for the next step, also the revised parameters, such as $p^{(k)}_i$, $w^{(k)}_{SD}(b_i, i)$, $w^{(k)}_{\text{Start}}(b_i, i)$ and $w^{(k)}_m$ were




(A) (B) (C) (D) (E)

**Fig. 1.** Spacer distribution of the final hit motif with the highest $\sigma$ for various prokaryotes. (**A**) motif 'AGGAG' for *E.coli*; (**B**) motif 'GGAGG' for *B.subtillis*; (**C**) motif 'GAGGT' for *T.maritima*; (**D**) motif 'AAGGA' for *H.influenzae*; (**E**) motif 'AGGTG' for *M.jannaschii*.

**Table 1.** Final hit motifs founded by MED-Start as potential 16S rRNA binding sites of various prokaryotes

| Genome | 16S rRNA | Hit motifs No. 1 | No. 2 | No. 3 |
|---|---|---|---|---|
| *E.coli* | TAAGGAGGTGA | AGGAG | CAGGA | GGAGA |
| *B.subtillis* | TAGAAAGGAGG | GGAGG | AAAGG | AGGAG |
| *T.maritima* | GAAAGGAGGTG | GAGGT | – | – |
| *H.influenzae* | TAAGGAGGTGA | AAGGA | – | – |
| *M.jannaschii* | GGAGGTGATCC | AGGTG | GGTGA | – |

The column labeled '16S rRNA' shows the reverse complement of the 3′ end of the organism's 16S rRNA, the true SD signals should be similar to a substring of this sequence.

taken into the next iterative step. The iterations were repeated until the parameters were at least 99% identical to that of the previous iteration. In this case, the model is regarded as having reached a convergent state. Our calculation indicates that the iteration, as well as the parameters used in Equations (3)–(6), quickly reach the convergent ones within at most four iterative steps.

Table 1 and Figure 1 report the final hit motifs for several prokaryotic genomes founded by the MED-Start and plots of their spacer distribution, applied as a postprocessor for the MED 1.0 gene finding system. The results suggest that the algorithm is rather effective to search the motifs associated with the SD sequences, almost each of the hit motifs is in good agreement with a substring of the reverse complement of the 3′ end of 16S rRNA. The plots of hit motif's spacer distribution versus distance from the start codon reveal that most of the hit motifs are located in upstream of the start codon with highly strong position-biased property, and most of the hit motifs are typically located in a region 5–10 bp upstream of the start codon, it also agrees well with the results of previous reports (Suzek *et al*., 2001; Besemer *et al*., 2001).

Referring to the final results of weights $w_m$ for the genomes studied here, it is concluded that $w_m$ decreases monotonically as $m$ increases (data not shown). Note that these weights are the output of our self-learning algorithm without any prior knowledge. Another interesting feature of our self-learning program is that it takes into account the content of the start codon in its description and the final output automatically gives a distribution of the usage for different start codons (ATG, GTG or TTG). For most prokaryotes, the start codon CTG is rarely used and nearly absent in the annotation. The EcoGene dataset does not have any CTG as the start codon, and the Bsub1248 dataset has only one gene with CTG as the start codon. Even in the GenBank annotation of the complete genome of *E.coli* and *B.subtillis*, there exists only one CTG as the start codon. Thus, codon CTG is excluded from consideration by MED-Start system. The parameters obtained indicate that the general selection preference is such that the start codon ATG is much more favored over GTG and TTG. On the other hand, the result reveals that the ratio for the usage of ATG, GTG and TTG is different for different prokaryotes. For *E.coli* (a Gram-negative bacterium), GTG is favored over TTG; but for *B.subtillis* (a Gram-positive bacterium), TTG is more favored over GTG. Species-dependent variation of the parameters for our algorithm is currently under investigation.

One may find that the iterative strategy of motif detection in our algorithm is similar to the common Gibbs sampler method for solving the same problem. Yet the differences between them should be noted, especially on the sampling criterion of possible motif and the calculation of its occurrence position in all input training sequences. Gibbs sampler usually selects one of the training sequences at random to generate possible instance of the motif, and determine the position of the motif also at random for the next step (Lawrence *et al*., 1993). In contrast, our algorithm never needs to choose them at random, but simply uses a deterministic way in each iterative step, the stability of the parameters representing the SD signals hidden in the training sequences naturally leads to a convergence stage rapidly.

## RESULTS AND DISCUSSION

### Accuracy of the method for reliable datasets

The first test of our self-learning algorithm MED-Start system is performed with several reliable datasets as discussed in the section 'Materials and Algorithms', while the parameters $P_1$, $P_2$, $P_3$ and $P_4$ described above are obtained with the complete

**Table 2.** Prediction accuracy of gene starts by MED-Start with the reliable datasets[a] as test sets

| Species | Test sets | Number of genes in the test set | Percentage of genes as the longest ORF (%) | Accuracy of MED-Start (%) | Accuracy of GS-Finder (%) |
|---|---|---|---|---|---|
| *E.coli* | EcoGene | 854 | 62.9 | 92.9 | 91.1[b] |
| | Link | 195 | 68.2 | 95.4 | 92.3 |
| *B.subtillis* | Bsub1248 | 1248 | 63.0 | 90.1 | – |
| | Bsub58 | 58 | 74.1 | 96.6 | 96.6 |
| | Bsub123 | 123 | 57.7 | 87.8 | 83.7 |
| | Bsub72 | 72 | 56.9 | 93.1 | 90.3 |
| | Bsub51 | 51 | 54.9 | 96.1 | 92.2 |

[a]The reliable datasets have been described in the section 'Materials and algorithm'.
[b]The EcoGene dataset included 838 genes when GS-Finder published its results (Ou *et al*., 2004).

genome annotation from GenBank. (Nearly identical quantities of the parameters may also be trained with the annotation by MED 1.0 or GLIMMER 2.02.) For EcoGene dataset (Rudd, 2000), MED-Start correctly predicts $793/854 = 92.9\%$ of the start sites. While for the Link dataset (Link *et al*., 1997), the accuracy is $186/195 = 95.4\%$. For the *B.subtillis* genome, $1125/1248 = 90.1\%$ of gene start sites in the Bsub1248 dataset (Yada *et al*., 2001) and $56/58 = 96.6\%$ for the Bsub58 dataset (Yada *et al*., 2001) are correctly predicted. For three sets of short genes of *B.subtillis* (Besemer *et al*., 2001), the prediction accuracy of MED-Start is $108/123 = 87.8\%$, $67/72 = 93.1\%$ and $49/51 = 96.1\%$ for the Bsub123, Bsub72 and Bsub51 dataset respectively. Since only the GS-Finder (Ou *et al*., 2004) has recently presented its results for the test of self-training algorithm against these reliable datasets, we have included their results in Table 2 for comparison. In the literature, RBSfinder does not give the same test results, but the highest accuracy of the TIS prediction reported is 88% for the EcoGene dataset (Suzek *et al*., 2001) that includes 717 confirmed genes at that time, and 92% for the Link dataset. As for the GeneMark suit of programs developed by Borodovsky *et al*., their latest version GeneMarkS (Besemer *et al*., 2001) was designed not as a postprocessor for some a gene finder, but as a self-training method for TIS prediction incorporating the GeneMark.hmm, which is one of the best gene finding programs. In terms of prediction of TIS only, the GeneMarkS gives its highest accuracy of 94.4% for the Link dataset, and 82.9, 88.9 and 92.9% for the same three datasets Bsub123, Bsub72 and Bsub51.

In summary, for the most reliable datasets Link from *E.coli* and Bsub58 form *B.subtillis*, the start site prediction accuracy of the MED-Start has achieved higher than 95%. It is worthy mentioning that we does not make special treatment for short genes, but the accuracy of the MED-Start to relocate the start sites of short genes are reasonable, higher than the GS-Finder. This stability of the accuracy with the variation of the gene length is very encouraging as it may suggest possible new algorithm for short gene prediction, which remains challenging for prokaryotic gene prediction. Finally, note that the accuracy of the MED-Start increases a little (from 87.8 to 96.1%) with the degree of similarities with known proteins (from the dataset Bsub123, Bsub72 to Bsub51), see Table 2. Therefore, it is demonstrated that the algorithm gives higher accuracy for more reliable datasets, and is robust to the variation of gene length.

## MED-Start as a postprocessor for gene starts prediction

The MED-Start is an autonomous system that may be used as a postprocessor for MED 1.0 and other prokaryotic gene finding systems, such as GLIMMER and GeneMark, to relocate gene start sites of computational coding ORFs, like RBSfinder (Suzek *et al*., 2001) and GS-Finder (Ou *et al*., 2004). In this section, we demonstrate the performance of MED-Start used to process the outputs of both MED 1.0 and GLIMMER 2.02. Where GLIMMER 2.02 was downloaded from http://www.tigr.org/software/glimmer and run following the instruction given in the distribution file; MED 1.0 is available at our website (http://ctb.pku.edu.cn/main/SheGroup). Both MED 1.0 and GLIMMER 2.02 were run with their default option for each genome.

MED 1.0 detects 5101 genes from the complete *E.coli* genome and 4567 genes from *B.subtillis*. The MED-Start takes the annotation as the input and relocates the position of the TIS. We again choose to compare the prediction of the MED-Start for the subset of confirmed genes whose TIS position is reliable, meanwhile their TIS have been detected by MED 1.0. For *E.coli* genome, MED 1.0 predicts correctly 98.5% of 854 gene termination sites in the EcoGene and all 195 gene termination sites in the Link dataset, while only correctly predicts 61.7% of gene starts in the EcoGene dataset, and 68.2% of gene starts in the Link dataset. After postprocessing by the MED-Start, the prediction accuracy of start sites increases to 91.5% for the EcoGene dataset, and to 95.4% for the Link dataset. For *B.subtillis*, the MED-Start also improves the gene start prediction, the accuracy is raised from 61.4 to 87.7% for the Bsub1248 dataset, from 69.0 to 89.7% for the Bsub58 dataset, from 42.3 to 68.3% for the Bsub123 dataset, from

**Table 3.** Performance comparison of the three gene start predictors, MED-Start, RBSfinder and GS-Finder which correct the gene starts location of coding ORFs predicted by MED 1.0

| Species | Test dataset | Initial prediction by MED 1.0 (%) | | Relocating by postprocessor Both 3′ and 5′ end match (%) | | |
|---|---|---|---|---|---|---|
| | | 3′ end match | Both 3′ and 5′ end match | MED-Start | RBSfinder | GS-Finder |
| *E.coli* | EcoGene (854) | 98.5 | 61.7 | **91.5** | 81.0 | 89.5 |
| | Link (195) | 100.0 | 68.2 | **95.4** | 81.5 | 92.3 |
| *B.subtillis* | Bsub1248 (1248) | 97.4 | 61.4 | **87.7** | 80.1 | 86.6 |
| | Bsub58 (58) | 93.1 | 69.0 | **89.7** | 79.3 | **89.7** |
| | Bsub123 (123) | 78.9 | 42.3 | **68.3** | 60.2 | 63.4 |
| | Bsub72 (72) | 79.2 | 43.1 | **73.6** | 63.9 | 69.4 |
| | Bsub51 (51) | 80.4 | 43.1 | **78.4** | 66.7 | 72.5 |

The used reliable datasets have been described in the section 'Materials and algorithm'. Numbers in parenthesis indicate the number of genes used as test set in the corresponding dataset. Numbers in bold indicate the highest accuracy precisely predicted for each set.

**Table 4.** Performance comparison of the three gene start predictors, MED-Start, RBSfinder and GS-Finder which correct the gene starts location of coding ORFs predicted by GLIMMER 2.02

| Species | Test dataset | Initial prediction by GLIMMER 2.02 (%) | | Relocating by postprocessor Both 3′ and 5′ end match (%) | | |
|---|---|---|---|---|---|---|
| | | 3′ end match | Both 3′ and 5′ end match | MED-Start | RBSfinder | GS-Finder |
| *E.coli* | EcoGene (854) | 99.3 | 63.2 | **92.0** | 81.9 | 90.3 |
| | Link (195) | 100.0 | 68.2 | **95.4** | 80.0 | 92.3 |
| *B.subtillis* | Bsub1248 (1248) | 98.6 | 61.3 | **89.2** | 78.5 | 87.9 |
| | Bsub58 (58) | 98.3 | 69.0 | **94.8** | 82.8 | **94.8** |
| | Bsub123 (123) | 91.1 | 53.7 | **79.7** | 72.4 | 75.6 |
| | Bsub72 (72) | 91.7 | 54.1 | **86.1** | 75.0 | 83.3 |
| | Bsub51 (51) | 88.2 | 47.1 | **86.3** | 70.6 | 82.4 |

The used reliable datasets have been described in the section 'Materials and algorithm'. Numbers in parentheses indicate the number of genes used as test set in the corresponding dataset. Numbers in bold indicate the highest accuracy precisely predicted for each set.

43.1 to 73.6% for the Bsub72 dataset and from 43.1 to 78.4% for the Bsub51 dataset respectively (as shown in Table 3).

Postprocessing the annotation by GLIMMER2.02 has a similar effect. GLIMMER 2.02 predicts 5104 genes in the complete *E.coli* genome and 5068 genes from *B.subtillis*. After the postprocessing by the MED-Start, the accuracy increases from 63.2 to 92.0% for the EcoGene dataset, and from 68.2 to 95.4% for the Link dataset. Improvement in other reliable datasets of *B.subtillis* is given in Table 4. In summary, the MED-Start program greatly improves the accuracy of gene starts prediction by 20–40% for both MED 1.0 and GLIMMER 2.02.

Finally, the performance of the MED-Start is compared to that of the RBSfinder and the GS-Finder. We have downloaded and run these two programs with the default parameters set by their authors as a design decision (Suzek *et al*., 2001; Ou *et al*., 2004) as a postprocessor to MED 1.0 and GLIMMER 2.02, i.e. input the annotation information predicted by MED 1.0 or GLIMMER 2.02, following the instruction given by the documentation with the default option. The results are evaluated in the same way as we described above for testing the MED-Start, and reported in Tables 3 and 4. All three programs have significantly increased the accuracy of the gene start prediction, and the MED-Start is clearly better than the RBSfinder and slightly better than the GS-Finder. In conclusion, the MED-Start is one of the best gene start predictors.

## Comparing the three gene start predictors

It is interesting to discuss the similarities and the differences among the three gene start predictors compared above. The MED-Start uses a set of parameters to describe the TIS, which have similar and also distinct features compared to that by RBSfinder and GS-Finder. The RBSfinder program (Suzek *et al*., 2001) relocates a gene start based on a probabilistic model that scores candidate RBS near annotated start codons predicted by a gene finder; and the program runs in several iteration until it converges. This score is similar to our probabilities $P_1$, $P_2$ and the iteration strategy is also similar. The difference is that prior knowledge is needed in the RBSfinder to specify the statistical properties of the model, and their performance of selecting start sites based on

the RBS score depends on several empirical threshold values. On the other hand, the GS-Finder program (Ou *et al*., 2004) employs an unsupervised learning only on defining the SD signals (without a prior knowledge of 16S rRNA in the genomes), but relies on empirical knowledge for the usage distribution of the start codons (ATG, GTG and TTG) as well as the distance distribution from the leftmost start codon to the true TIS, which hence do not take into account species-dependent variations. The MED-Start algorithm is, on the contrast, based on a more complete unsupervised learning procedure for all the properties describing the TIS, and is thus adaptable to any organisms. The benefit should be more noticeable when one annotates new genomes whose statistical properties somewhat deviate from the existing genomes. In particular, our program should appear to be more suitable to discover new consensus motifs in the upstream regions of the gene starts, or to discover signals that occur in anomalous upstream locations of the start codons, or to discover an anomalous usage distribution of start codons (ATG, GTG and TTG) for a specific genome (see parameters $P_1$, $P_2$, $P_3$ and $P_4$ given in the section 'Materials and algorithms'). After a complete genome sequence and the annotation predicted by a gene finder (such as GLIMMER 2.02 or MED 1.0 program) are available, the MED-Start obtains all parameters described above, without any prior knowledge of these parameters of the organism.

It is thus important to study the complementary property of the three gene start predictors. This may be effectively visualized by a Venn diagram illustrating the overlaps between the three sets of predictions for a common set of data. Figure 2 shows this Venn diagram for the EcoGene dataset (A) and for the Link dataset (B). The diagram indicates that in general the MED-Start shares more common predictions with the GS-Finder than the RBSfinder, yet the MED-Start has its unique predictions. For instance, the MED-Start precisely predicts 6 gene starts in the Link dataset and 22 gene starts in the EcoGene dataset, which neither RBSfinder nor GS-Finder has captured. Hence, the MED-Start algorithm is complement to the two other existing outstanding methods for the prediction of gene start sites in bacterial genomes.

Finally, let us mention a few special merits of the MED-Start. First, the acceleration of microbial genome sequencing has led to an imperative need for entirely unsupervised gene finding methods which may be adapted to any organism and be able to discover species-dependent properties. The MED-Start is more completely autonomous in this regard than any other existing gene start predictors. Second, the set of parameters (weights and probabilities) have easy biological interpretations and its variations over different genomes are interesting to study in relation to the evolutionary properties of the genome. Thirdly, the algorithm achieves a high accuracy in the prediction of the gene starts because of its capability of detecting very subtle properties of the genome through the weights and the probabilities. This makes

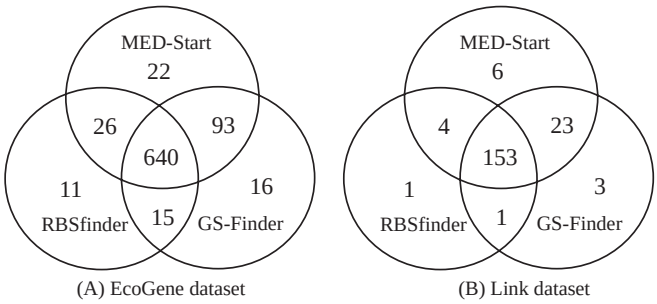



**Fig. 2.** Venn diagram showing group relationship between genes predicted by three individual programs MED-Start, RBSfinder and GS-Finder to process the output of MED 1.0 with analysis of complete genome *E.coli* (**A**) Precisely prediction accuracy for EcoGene dataset. (**B**) Precisely prediction accuracy for Link dataset.

the MED-Start as an effective tool for the study of classification, function and evolution for prokaryote genomes. Finally, the absence of empirical constants and the needs for no prior knowledge make the MED-Start particularly easy to use.

### Iterations convergence of the MED-Start program

We have studied the convergence of the MED-Start program by tracing the accuracy of the prediction versus the iteration number (data not shown). The MED-Start takes as input the annotation by either MED 1.0 or GLIMMER 2.02, and as the iteration goes, the accuracy of the prediction rapidly improves. Using the EcoGene and Link datasets as a standard for benchmark, after one iteration, the accuracy arises to above 90%, that is a more than 25% increase for both datasets. This significant increase is an indication that the statistical model we use reflects the dominant correlations around the gene starts. Within at most four iteration steps, the accuracies converge. At the same time, the rapid convergence is observed for all the parameters $P_1$, $P_2$, $P_3$ and $P_4$ described above. Similar results have been confirmed for other genomes with reliable datasets. Finally, the computation time of MED-Start is modest; it takes about half minute to analyze the complete genome of *E.coli* and relocate all of its TIS on a Pentium-IV personal computer.

## ACKNOWLEDGEMENTS

The work benefits from the research environment at the Center for Theoretical Biology of Peking University. Various discussions with Z.Yang, G.-H.Lan, Z.-X.Sun, Q.Hu and others are acknowledged. The database support from the CBI at Peking University are acknowledged. We thank Prof. C.-T.Zhang at Tianjin University (China) for providing GS-Finder program and part of datasets. The work received support by the National Natural Science Foundation (10225210, 30300071 and 90208021) of China; the present study was also supported

by the 973 Project grant 2003CB715905 founded by MOST of China.

## REFERENCES


Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.

Blattner,F.R., Plunkett,G.III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collodo-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al*. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Borodovsky,M. and MaIninch,J. (1993) GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem.*, **17**, 123–153.

Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton, G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D. *et al*. (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, **273**, 1058–1073.

Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**, 4636–4641.

Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Frishman,D., Mironov,A., Mewes,H.-W. and Gelfand,M. (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes, *Nucleic Acids Res.*, **26**, 2941–2947.

Guo,F.B., Ou,H.Y. and Zhang,C.T. (2003) ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.*, **31**, 1780–1789.

Hannenhalli,S.S., Hayes,W.S., Hatzigeorgiou,A.G. and Fickett,J.W. (1999) Bacterial start site prediction. *Nucleic Acids Res.*, **27**, 3577–3582.

Hayes,W.S. and Borodovsky,M. (1998) Deriving ribosomal binding site (RBS) statistical models from unannotated DNA sequences and the use of the RBS model for N-terminal prediction, *Pac. Symp. Biocomput.*, 279–290.

Keich,U. and Pevzner,P.A. (2002) Finding motifs in the twilight zone. *Proceeding of the 6th international conference on computational molecular biology* (*RESCOMB 2002*). Washington DC, USA, ACM Press.

Kunst,F., Ogasawara,N., Moszer,L., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S. *et al*. (1997) The complete genome sequence of the Gram-positive bacterium *Bacillus subtillis. Nature*, **390**, 249–256.

Klenk,H.P., Clayton,R.A., Tomb,J.F., White,O., Nelson,K.E., Ketchum,K.A., Dodson,R.J., Gwinn,M., Hickey,E.K., Peterson,J.D. *et al*. (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus. Nature*, **390**, 364–370.

Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.

Lewin,B. (2000) *Genes VII*, Oxford University Press, New York.

Link,A.J., Robison,K. and Church,G.M. (1997) Comparing the predicted and observed properties encoded in the genome of *Escherichia coli. Electrophoresis*, **18**, 1259–1313.

Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solution for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.

Mikonnen,M., Vuoristo,J. and Alatossava,T. (1994) Ribosome binding site consensus sequence of *Lactobacilus delbrueckii. FEMS Microbiol. Lett.*, **116**, 315–320.

Nelson,K.E., Clayton,R.A., Gill,S.R., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Nelson,W.C., Ketchum,K.A. *et al*. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritime. Nature*, **399**, 323–329.

Osada,Y., Saito,R. and Tomita,M. (1999) Analysis of base-paring potentials between 16S rRNA and 5′ UTR for translation initiation in various prokaryotes. *Bioinformatics*, **15**, 578–581.

Ou,H.Y., Guo,F.B. and Zhang,C.T. (2004) GS-Finder: a program to find bacterial gene start sites with a self-training method. *Int. J. Biochem. Cell Biol.*, **36**, 535–544.

Ouyang,Z.Q., Zhu,H.Q., Wang,J. and She,Z.S. (2004) Multivariate entropy distance method for prokaryotic gene identification, *J. Bioinform. Comput. Biol.*, **2**, 353–373.

Rudd,K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60–64.

Salzberg,S.L., Delcher,A.L., Kasif,S. and White,O. (1998) Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.*, **26**, 544–548.

Schurr,T., Nadir,E. and Margalit,H. (1993) Identification and characterization of *E. coli* ribosomal binding sites by free energy computation. *Nucleic Acids Res.*, **21**, 4019–4023.

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

Shmatkov,A.M., Melikyan,A.A., Chernousko,F.L. and Borodovsky,M. (1999) Finding prokaryotic genes by the 'frame-by-frame' algorithm: targeting gene starts and overlapping genes. *Bioinformatics*, **15**, 874–886.

Stormo,G.D., Schneider,T.D., Gold,L. and Ehrenfeucht,A. (1982) Use of the 'Perceptron' algorithm to distinguish translation initiation site in *E. coli. Nucleic Acids Res.*, **10**, 2997–3011.

Suzek,B.E., Ermolaeva,M.D., Schreiber,M. and Salzberg,S.L. (2001) A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, **17**, 1123–1130.

Tompa,M. (1999) An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *ISMB*, 262–271.

Yada,T., Totoki,Y., Takagi,T. and Nakai,K. (2001) A novel bacterial gene-finding system with improved accuracy in locating start codons. *DNA Res.*, **8**, 97–106.