PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2008, 35(11): 1254~1262
www.pibb.ac.cn

New Solutions of Translation Initiation Site Prediction for Prokaryotic Genomes*

HU Gang-Qing[1,2], LIU Yong-Chu[1,2], ZHENG Xiao-Bin[1,2],
YANG Yi-Fan[2], SHE Zhen-Su[1,2,3], ZHU Huai-Qiu[1,2,4]**

([1] *Department of Biomedical Engineering, and State Key Laboratory for Turbulence and Complex Systems, Peking University, Beijing* 100871, China; [2] *Center for Theoretical Biology, Peking University, Beijing* 100871, China; [3] *Department of Mathematics, UCLA, Los Angeles, CA* 90095, USA; [4] *The Finnish Genome Center, University of Helsinki*, 00290 *Helsinki*, Finland)

Abstract Accurate prediction of the translation initiation site (TIS) is an important issue for prokaryotic genome annotation. However, it is still a challenge for the existing methods to predict the TIS in the genomes over a wide variety of GC content. Besides, the existing methods have not yet undergone a comprehensive evaluation, leaving prediction reliability as a largely open problem. A new algorithm MED-StartPlus, a tool that predicts TIS in prokaryotic genomes with a wide variety of GC content was presented. It makes several efforts to model the nucleotide composition bias, the regulatory motifs upstream of the TIS, the sequence patterns around the TIS, and the operon structure. Tests on hundreds of reliable data sets, with TISs confirmed by experiments or having annotated functions, show that the new method achieves a totally high accuracy of TIS prediction. Compared with existing TIS predictors, the method reports a totally higher performance, especially for genomes that are GC-rich or have complex initiation mechanisms. The potential application of the method to improve the TIS annotation deposited in the public database was also proposed.

Key words prokaryote, gene prediction, translation initiation site, prediction evaluation

Accurate prediction of the translation initiation site (TIS) of protein coding gene is still one of the substantial issues for prokaryotic genome annotation. Recently, many algorithms have been developed to improve the TIS prediction accuracy [1~6]. For the well-studied genomes with non-biased GC content such as *E. coli* K-12 and *B. subtillis* (with GC content of 50.8% and 43.5% respectively), most of the current algorithms have shown relatively high performance to improve TIS prediction. However, a challenging problem is to capture the subtle and complex motifs in TIS upstream region in the genomes with high or low GC content [5~7]. It is well known that the GC-rich genomes tend to possess weaker ribosomal binding site (RBS) signals, thus complicating the detection of RBS patterns for TISs prediction [1, 7]. In addition, the previous studies using *in silico* and experimental data revealed the diversity of translation initiation mechanisms and components of translation initiation in Archaea [8, 9], where a large proportion of genomes have a high or low GC content. All these may lead to overall lower quality of TIS prediction by current computational methods. A recent investigation showed that in some genomes up to 60% of genes may have been annotated with the wrong start sites, especially in the GC-rich ones [10]. It is generally believed that the genomic GC content is related to bacterial evolution as supported by the phylogenetic tree based on 5 S rRNA sequences[11]. Therefore a method to describe complex TIS-related sequence patterns, then lead to accurate prediction of gene starts would be helpful in facilitating the genome annotation, as well as in understanding full information on protein coding genes of the prokaryotic organisms.

In order to overcome the difficulty of TIS prediction for genomes with high or low GC content, various strategies have been used. For example, the GS-Finder[3] employs a "nine-dimensional super-sphere" method to select initial training sets for genomes with

*This work was supported by grants from The National Natural Science Foundation of China (30770499, 30300071,10721403) and National Basic Research Program of China (2003CB715905).

**Corresponding author. Tel: 86-10-62767261
E-mail: hqzhu@pku.edu.cn, hqzhu@ctb.pku.edu.cn

Received: February 5, 2008 Accepted: June 30, 2008

GC content > 56%. A recent work, Hon-yaku, introduces a Bayesian method by considering elements important in translation initiation[6]. Unlike other methods, Hon-yaku is a supervised learning method. With the training set determined by comparisons between orthologous genes, the algorithm reports good performance in the GC-rich Gamma-Proteobacteria[6].

Here, we present a new algorithm MED-StartPlus, a tool designed to predict TIS in prokaryotic genomes covering a wide variety of GC content. As a result, MED-StartPlus is shown to achieve an overall high performance on hundreds of reliable data sets. We also report a detailed comparison with existing TIS prediction programs such as RBSfinder[2], GS-Finder[3], MED-Start[4], TiCo[5] and Hon-yaku[6]. MED-StartPlus is thus demonstrated with a competitive or better performance, especially on the GC-rich genomes.

1 Materials and methods

1.1 Materials

The complete prokaryotic genomic sequences and their annotations for a total of 300 genomes were downloaded from the RefSeq Release 17 at the time of writing, whereas the genome sequence and annotation for *H. salinarum* were available from the website http://www.halolex.mpg.de/public/. We included two types of data sets as benchmarks.

First, we collected the 14 published TIS data sets available so far for 11 genomes with GC content ranging from 35.6% to 68% . The data sets were verified either by N-terminal protein sequencing (*Aper*[12]; *EcoGene*[13]; *Link*[14]; *Syne*[15]), or through literature review (*Mtub*, *Paer107* and *SolfGene*[16]), or with sequence homology (*Bsub58*[17]), or by comparison between orthologous genes (*Bpse*, *Hars* and *Paer344*[6]), or by analyzing proteomic data (*Npha* and *Hsal*[18]), or with experimentally characterized gene function (*Bsub*[17]).

To make a large-scale evaluation, we then included the RefSeq annotations on function-known genes for each prokaryotic genome, *i.e.*, those with product descriptions excluding any of the key words "-like", "conserved", "hypothetical", "homolog", "probable", "possible", "predicted", "putative", "similarity" and "unknown". A total of 300 data sets containing 450 849 genes for 300 genomes were constructed as the second type of benchmark. Note that we cannot guarantee that the start sites are correctly annotated for all of these genes, but without further experimental evidences they indeed constitute the current best public resources to which the prediction can be compared. For each data set, we have removed genes that contain frameshift, or begin with non-canonical start codon such as CTG and ATT, or are not annotated in RefSeq. The collected data sets can be freely downloaded from http://ctb.pku.edu.cn/main/SheGroup/Software/MEDPWEB/MEDStartPlus.htm.

1.2 Model and algorithm

In developing MED-StartPlus, several efforts were addressed to build a new model of the prokaryotic TIS and the resultant algorithm. First, we considered the effect of nucleotide composition bias to find regulatory motifs and characterize the sequence patterns around the motifs as well as the start sites. Next, the parameters describing operon structure in prokaryotic genomes were added into the model. In addition, we introduced another component to score the coding potential of the context around a candidate start, by referring to the property of the codon positional GC-content. Finally, we introduced and combined two scoring functions that integrated the above parameters to determine the TIS.

1.2.1 Finding motifs upstream of TIS. One of the efforts made in the current algorithm was to construct models to capture the subtle and complex motifs upstream of the TIS in GC-rich genomes. As pointed out previously[1, 5~7], motif finding is challenging in GC-rich genomes, because random nucleotide strings from the background occur even more frequently than regulatory motifs. In MED-StartPlus, to eliminate the effect of nucleotide composition bias, we applied the chi-squared statistic as an over-representation measurement to select motifs.

Let S be the set of all l-mers occurred in the set of TIS upstream region sequences. For each $s \in S$, let Obs(s) denote the number of observed occurrence, and Exp(s) as the expected number of occurrence, which is estimated by the nucleotide composition in the intergenic regions. We regard a motif as being over-represented if Obs (s) > Exp (s). Then $X_s = \frac{(\mathrm{Obs}(s)-\mathrm{Exp}(s))^2}{\mathrm{Exp}(s)}$ characterizes the over-representation of s. Herein, any over-represented motifs with $\xi_s = \frac{X_s}{\max\{X_s\}} > \xi_0 = 0.5$ would be selected as the significant motifs, which are referred to as those associated with the signals of translation initiation. Having the significant motifs determined, a relative positional weight matrix (RPWM) $w_S(b_j, j)$ and a probability of spacer length distribution p_i are then calculated for them. The definition of RPWM $w_S(b_j, j)$ will be described in the following subsection. Here i is position of occurred motif upstream to the start sites, $b_j \in \{A, C, G, T\}$ and j is position within aligned windows around the motifs. In the current algorithm, both $w_S(b_j, j)$ and p_i are inferred based on the predicted start sites with a set of the significant motifs, then lead to convergences by self-training iteration.

We introduced the above-mentioned ξ_0 as a threshold in the motif search. A simulation was performed to test whether the threshold $\xi_0 = 0.5$ could give real motifs. We calculated max $\{X_s\}$ from sequences upstream to the TIS. Meanwhile, we generated random sequences with the same nucleotide composition to calculate max$\{X_r\}$ for random motifs r; this procedure was repeated 1 000 times in order to estimate a distribution of max $\{X_r\}$. The simulation showed that the probability of the motif s with $\frac{X_s}{\max_s\{X_s\}} > \xi_0$ being a random motif r was generally less than 0.000 1.

In this work, motifs were designed as 5-mers. To assure statistical significance, the number of samples demanded increased exponentially over the motif length, whereas the samples were generally insufficient for small genomes when the length was over six. To examine the effect of the selection of motif length, we have tested other options by both 4-mers and 6-mers on the set of published TIS data sets and observed no significant difference in accuracy (data not shown).

As a result, MED-StartPlus reported more than 90% of the 300 genomes recovered with motifs containing tetra-mers from the widely accepted SD sequence "AAGGAGGTGA" [19]. For example, we found motifs in the GC-rich genome *M. tuberculosis* as "AGGAG", "GAGGA" and "AAGGA", which resemble the SD motifs in *E. coli* K-12 and *B. subtilis* [4]. Similar motifs were recovered in two other GC-rich genomes *P. aeruginosa* and *B. pseudomallei* [5, 6].

1.2.2 RPWM for motifs and start site. In prokaryotic genomes, both initiation signal and context around the TIS show content conservation in a certain region. This conservation can be characterized by the standard positional weight matrix (PWM), which has been used in several methods for TIS prediction[2, 4, 6]. To describe the conservation of context around both motif and TIS in the current algorithm, we define a relative positional weight matrix (RPWM) taking into account the contribution of nucleotide composition bias as follows.

We first define the RPWM of motif. For each significant motif, an aligned PWM denoted as $w_{S\_\text{foreground}}(b_j, j)$ can be calculated as in the way used in MED-Start[4], by a multiple alignment of all non-full-matching instances of the motif occurring within the training sequences. We then introduced a background PWM, denoted as $w_{S\_\text{background}}(b_j, j)$, to describe the composition bias by calculating the genomic nucleotide composition. As a result, the RPWM for the motif may be read as

$$w_S(b_j, j) = \frac{w_{S\_\text{foreground}}(b_j, j) / w_{S\_\text{background}}(b_j, j)}{\sum_{b_j} w_{S\_\text{foreground}}(b_j, j) / w_{S\_\text{background}}(b_j, j)} \qquad (1)$$

Similarly for the context around TISs, the aligned PWM denoted as $w_{T\_\text{foreground}}(b_j, j)$ can be calculated, while the background PWM, denoted as $w_{T\_\text{background}}(b_j, j)$ was defined in this way: the upstream region of start sites was calculated with the genomic nucleotide composition, and the downstream region of start sites was calculated using aligned coding sequences. Thus RPWM for the start site is

$$w_T(b_j, j) = \frac{w_{T\_\text{foreground}}(b_j, j) / w_{T\_\text{background}}(b_j, j)}{\sum_{b_j} w_{T\_\text{foreground}}(b_j, j) / w_{T\_\text{background}}(b_j, j)} \qquad (2)$$

1.2.3 Distance distribution between genes. A common feature of prokaryotic genomes is the presence of operons, which results in the feature of genes within an operon having much shorter intergenic distances than others [20]. In the current algorithm, a scoring function was defined to measure the probability of a gene start site with a certain distance to its immediate upstream gene. We referred to any two adjacent genes as neighbors, of which members may locate on either the same or the opposite strand, denoted as type1 and type2 members, respectively. The distance between members from a neighbor, denoted as d, was defined as a gap between them, while a negative distance means the number of overlapping base pairs. The distance between adjacent genes along the genomic sequence may be described by two distributions $f_{\text{type1}}(d)$ and $f_{\text{type2}}(d)$. For each candidate start site m in a gene g, by denoting the immediate upstream gene as g', a score may be calculated as:

$$D_m = \begin{cases} f_{\text{type1}}(d), & \text{if } g \text{ and } g' \text{ are in type1,} \\ f_{\text{type2}}(d), & \text{if } g \text{ and } g' \text{ are in type2.} \end{cases} \qquad (3)$$

Our studies on hundreds of genomes showed that $f_{\text{type1}}(d)$ was rather conserved, with the top two frequent distances as ～4 bps and ～1 bp, while $f_{\text{type2}}(d)$ varied slightly among species (data not shown). In fact, both functions were automatically calculated based on the predicted start sites, then come to certain distributions by self-training iteration in the current algorithm.

1.2.4 Coding potential of regions around the TIS. In GC-rich genomes, protein coding genes usually have a G$\bar{\text{G}}$S pattern at three codon positions, where G, $\bar{\text{G}}$, and S are the bases of G, non-G, and G/C, respectively[21]. Such a pattern would be helpful for a gene prediction method addressed to the GC-rich prokaryotic genomes[16, 21, 22]. In MED-StartPlus, this pattern was used to describe the 5′ end of coding regions.

We defined $GC(i)$ as the G/C occurrence at the ith codon position, where i =1, 2 or 3, then used the parameter $gb_s = \dfrac{GC(1)+GC(3)}{GC(1)+GC(2)+GC(3)}$ to describe the coding property of a given sequence segment s [22]. Using samples from the intergenic and 5′ end coding regions, we noted that the distributions of gb_s, denoted as $f_{\text{c}}\ (gb_s)$ and $f_{\text{nc}}\ (gb_s)$ for coding and noncoding sequences, showed normal distributions and differed significantly in GC-rich genomes. Thus the posterior probability of sequence s being coding, with regard to the codon positional nucleotide usage, can be written as the ratio of $P_{\text{c}}\,(s) = \dfrac{f_{\text{c}}(gb_s)}{f_{\text{c}}(gb_s)+f_{\text{nc}}(gb_s)}$. Similarly, the posterior probability of sequence s being noncoding is $P_{\text{nc}}(s)=\dfrac{f_{\text{nc}}(gb_s)}{f_{\text{c}}(gb_s)+f_{\text{nc}}(gb_s)}$. Then for a candidate start site m, the probability for the upstream sequence as noncoding and the downstream sequence as coding is

$$C_m=P_{\text{nc}}(s_1)\cdot P_{\text{c}}(s_2)\ , \qquad (4)$$

where s_1 and s_2 refer to sequences l bps upstream and l bps downstream from start site m, respectively. The length of l has been tested from 0 to 300 bps at a step of 30. When evaluated on the published data sets, the accuracy fluctuation was generally less than 3% while l > 60, However, the overall best result was obtained under l = 90.

1.2.5 Scoring functions and strategy of self-learning algorithm. With the parameters defined above, we were now able to construct the below scoring function:

$$Score_m=\log\{\max_j[p_i\prod_j w_S(b_j,\ j)]\}+\log\{\prod_k w_T(b_k,\ k)\}$$
$$+\log w_m+\log D_m+\log C_m\ , \qquad (5)$$

where, m means the m-th in-frame candidate start counting from the 5′-most one, $w_S(b_j,\ j)$ and $w_T(b_k,\ k)$ are RPWMs defined by Eq.1 and Eq. 2, and w_m is the probability weight of a relative distance between the stop codon and the candidate start codon m, k means the position within aligned windows around start codons. In addition, j, b_j, b_k, p_i, D_m and C_m have been defined as mentioned above. Omitting components related to motifs in Eq. 5 leads to

$$Score_m=\log\{\prod_k w_T(b_k,\ k)\}+\log w_m+\log D_m+\log C_m. \ (6)$$

Note that the RPWMs $w_T\,(b_k,\ k)$ in Eq. 5 and Eq.6 differ from the window size for calculating the matrix. The former is defined in the range of –20 to 15 bps around the TIS, while the latter in the range of –50 to 15 bps around the TIS to describe the subtle and weak sequence pattern upstream to the TIS.

The algorithm of MED-StartPlus is composed of two sub-modules, the “Signal involved module” and the “Simplified module”, which differ in the scoring function. The former selects Eq. 5, while the latter uses Eq. 6. Each TIS module processes in two stages, the “Self-training stage” and the “Prediction stage” (Figure 1). In the “Self-training stage”, the program begins with selecting ORFs longer than 300 bps as the seed ORFs from the input. The parameters in both Eq. 5 and Eq. 6 are first trained with the input starts of the seed ORFs, and then enable the application of the scoring function to relocate start sites of all the ORFs. The updated starts are used to re-select seed ORFs and re-calculate all parameters. This procedure goes on by iteration until it reaches more than 99.0% unchanged starts or at most 20 rounds. With the converged parameters, both modules score all candidates in each ORF by scoring functions Eq. 5 and Eq. 6 respectively, and then each selects the one with the highest score as the most-likely TIS in its “Prediction stage”.

After independently analyzing the input, each module outputs a list of potential TISs for all ORFs. The system then combines the two modules. Those predicted identically by both modules are defined as certain TISs, which are also the final prediction for this part of ORFs. To determine between the disagreed potential TISs, the program first calculates a standard PWM $w_f(b_k,\ k)$ for sequences around the certain TISs, and another standard PWM $w_b(b_k,\ k)$ around TISs only predicted by the “Signal-involved module”. Then, it employs the equation

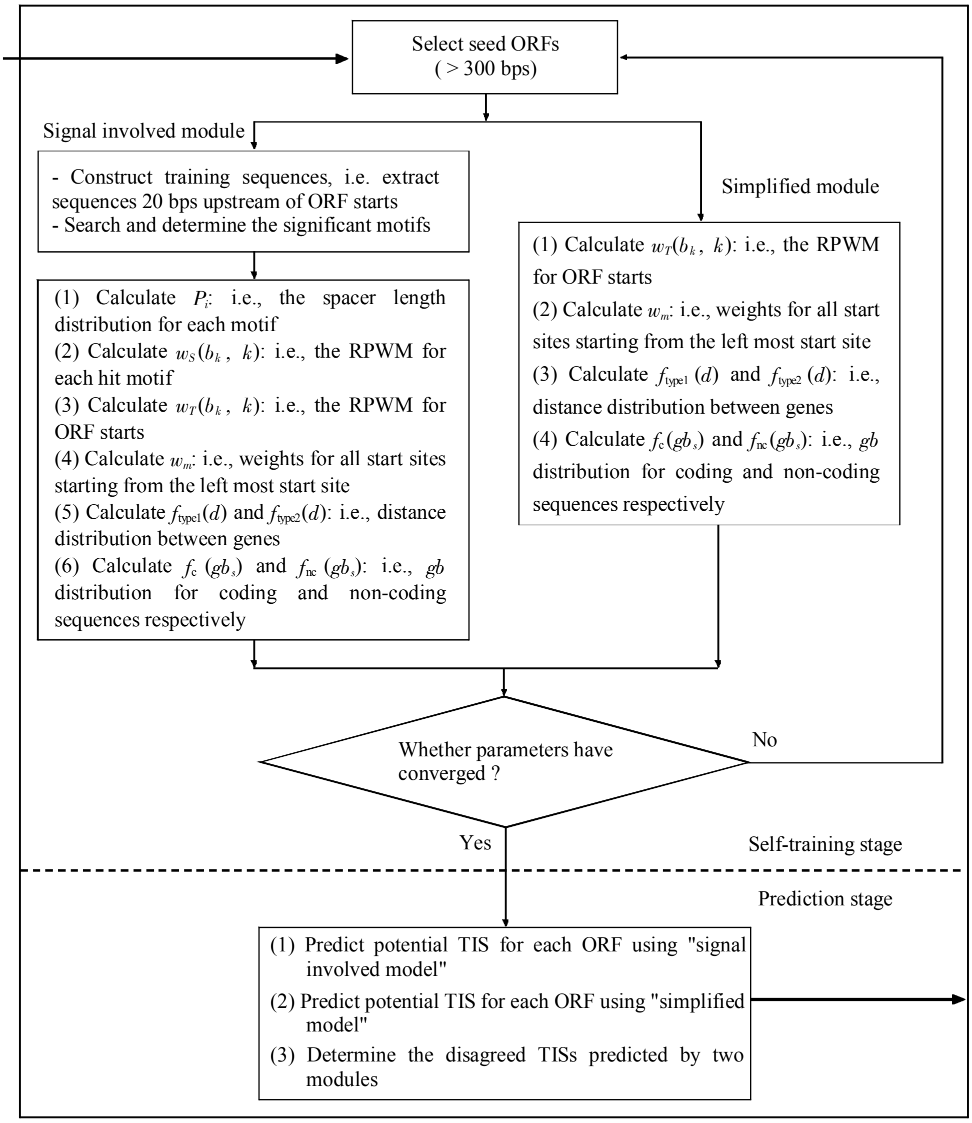
Select seed ORFs
(> 300 bps)
Signal involved module
- Construct training sequences, i.e. extract sequences 20 bps upstream of ORF starts
- Search and determine the significant motifs
(1) Calculate P_l: i.e., the spacer length distribution for each motif
(2) Calculate $w_S(b_k, k)$: i.e., the RPWM for each hit motif
(3) Calculate $w_T(b_k, k)$: i.e., the RPWM for ORF starts
(4) Calculate w_m: i.e., weights for all start sites starting from the left most start site
(5) Calculate $f_{type1}(d)$ and $f_{type2}(d)$: i.e., distance distribution between genes
(6) Calculate $f_c(gb_s)$ and $f_{nc}(gb_s)$: i.e., gb distribution for coding and non-coding sequences respectively
Simplified module
(1) Calculate $w_T(b_k, k)$: i.e., the RPWM for ORF starts
(2) Calculate w_m: i.e., weights for all start sites starting from the left most start site
(3) Calculate $f_{type1}(d)$ and $f_{type2}(d)$: i.e., distance distribution between genes
(4) Calculate $f_c(gb_s)$ and $f_{nc}(gb_s)$: i.e., gb distribution for coding and non-coding sequences respectively
Whether parameters have converged ?
No
Yes
Self-training stage
Prediction stage
(1) Predict potential TIS for each ORF using "signal involved model"
(2) Predict potential TIS for each ORF using "simplified model"
(3) Determine the disagreed TISs predicted by two modules

Fig. 1 Flow chart of "Self-training stage" and "Prediction stage"

$$Score=\prod_k w_f(b_k,\ k)/\prod_k w_b(b_k,\ k) \tag{7}$$

to re-score the disagreed TISs and chooses the one with the higher score as the final prediction.

2 Results and discussion

To benchmark MED-StartPlus, we have tested the five existing TIS predictors referred to as RBSfinder[2], GS-Finder[3], MED-Start[4], TICO[5] and Hon-yaku[6]. Herein for each organism, we took the RefSeq annotation as initial input for a TIS predictor. It has been reported that some methods appear to be sensitive to the input TIS locations[6]. Thus, to make the comparison impartial for each locally executable program, when analyzing a genome, we treated the longest ORFs (assigned with the leftmost start codon) of all genes as the input for a TIS predictor instead of the original TISs given by the full annotation.

2.1 Accuracy of TIS prediction in published data sets

We first used the data sets listed in Table 1 as benchmarks, including 4 074 genes with their TISs confirmed by various evidences. The accuracy of each algorithm is listed in Table 1. The results for the supervised algorithm Hon-yaku were cited from its publication[6]. According to the genomic GC content, the genomes are classified into three groups: AT-rich genomes (GC < 40%), genomes with medium GC content (40% ≤GC ≤56%), and GC-rich genomes (GC > 56%)[3, 23].

Table 1 Performance evaluation against 14 published data sets for MED-StartPlus, MED-Start, RBSfinder, GS-Finder, Hon-yaku and TiCo

Organism	% GC	Data set	MED-StartPlus(%)	MED-Start(%)	RBSfindera$^{1)}$(%)	GS-Finder(%)	Hon-yakub$^{2)}$(%)	TiCo(%)
S. solfataricus	35.8	*SolfGene*	**85.7**	33.9	51.8	80.4	-	82.1
B. subtilis	43.5	*Bsub*	91.7	91.3	82.5	90.3	**92.7**	90.6
		Bsub58	94.7	**96.6**	86.0	**96.6**	**96.6**	93.0
Synechocystis sp.	47.4	*Syne*	88.2	66.4	66.4	80.0	-	**90.9**
E. coli	50.8	*EcoGene*	93.0	93.0	84.6	91.1	93.2	**95.2**
		Link	94.2	**96.9**	90.6	93.7	96.3	**96.9**
H. arsenicoxydans	54.3	*Hars*	**93.2**	87.7	76.5	90.1	92.6	**93.2**
A. pernix	56.3	*Aper*	91.9	**94.3**	30.1	75.6	-	93.5
N. pharaonis	63.1	*Npha*	**95.6**	11.5	59.8	83.5	-	91.3
M. tuberculosis	65.6	*Mtub*	**89.4**	4.5	65.2	72.7	-	83.3
P. aeruginosa	66.6	*Paer344*	**93.6**	65.1	72.7	89.5	92.8	90.4
		Paer107	**98.1**	74.8	86.9	97.2	-	94.4
B. pseudomallei	67.6	*Bpse*	**96.5**	2.5	64.3	87.4	92.6	86.9
H. salinarum	68.0	*Hsal*	**92.0**	4.3	13.9	73.4	-	80.3

$^{1)}$ RBSfinder was run repeatedly to improve performance. $^{2)}$ The results for the supervised algorithm Hon-yaku were cited from ref [6].

2.1.1 Genomes with the medium GC content. For two well-studied genomes *B. subtilis* and *E. coli* K-12, and the genome *H. arsenicoxydans*, the prediction accuracy of MED-StartPlus on the five data sets *Bsub*, *Bsub58*, *EcoGene*, *Link* and *Hars*, showed a similar performance to MED-Start, and matched the overall performance of GS-Finder, TiCo and Hon-yaku, while outperforming RBSfinder by about 10%.

TIS prediction in *Synechocystis* sp. was regarded as a challenge due to the weak RBS information upstream of the start sites [3]. As shown in Table 1, GS-Finder reports an accuracy of 80.0% on the data set Syne, while both MED-Start and RBSfinder report rather lower accuracies. In contrast, MED-StartPlus and TiCo report relatively higher performance with accuracies of 88.2% and 90.9%, respectively.

2.1.2 GC-rich genomes. TIS prediction is believed to be much more difficult in GC-rich genomes. It was reported that GC-rich genomes generally possess less information content in RBS [7]. Moreover, there likely exists more SD-like strings by random and generally more candidate start codons for each ORF in GC-rich genomes[10, 24]. All of these would complicate the designing of a TIS model. Herein we used seven data sets from GC-rich genomes as benchmarks: *Aper*, *Npha*, *Mtub*, *Paer344*, *Paer107*, *Bpse* and *Hsal*.

The archaeal genome *A. pernix* serves as a typical example to show the risk of pre-designed parameters in a TIS predictor. Contrary to the common belief that ATG is the most frequent start codon, a recent analysis on the 130 experimentally confirmed TISs has revealed that over 50% of the start codons in *A. pernix* are TTG, and genome-wide analysis showed that TTG (38%) is the most frequent start codon compared with ATG (33%) and GTG (29%) [12]. Thus, the pre-stored TTG usage in GS-Finder as 7% [3] fails to work on *A. pernix*, resulting in a much lower performance of 75.6% compared with MED-StartPlus (91.9%), MED-Start (94.3%) and TiCo (93.5%) on the data set *Aper*. The rule of selection of ATG prior to TTG as a start codon is applied in RBSfinder [2]. In addition, RBSfinder begins with a known consensus sequence to search motif (“AGGAG” by default). However, “GGGGT” and “GGGTG” are the top two over-represented SD motifs in this genome [9, 12]. Thus, it is easy to understand that RBSfinder gives a surprisingly low performance on *A. pernix*.

N. pharaonis and *H. salinarum* are also archaeal genome studied here. TISs in the data set *Npha* and *Hsal* were confirmed through analyzing the proteomic data with a false positive rate expected to be only 0.2% [18]. On the data set *Npha* with 321 samples, MED-StartPlus had the highest accuracy (95.6%), which was 4% higher than TiCo and at least 10% higher than others. Moreover, the performance of MED-StartPlus on Hsal(552 samples) for *H. salinarum* was much more distinctive: it reported an accuracy of 92.0%, which was nearly 12% higher than the best of the other predictors.

Genomes with the other four data sets, *Mtub*, *Paer344*, *Paer107* and *Bpse*, belong to the Eubacteria. On the data set *Mtub*, MED-StartPlus reached the highest accuracy of 89.4% . For *Paer344*, which possesses the largest size of samples (344), our method also reported the highest accuracy of 93.3%. Similar performance held up on the data set *Paer107*. The accuracy on the data set *Bpse* was also remarkable: our method achieved the highest accuracy of 96.5%, which is about 9% higher than other unsupervised methods such as GS-Finder and TiCo.

Generally speaking, MED-StartPlus successfully overcomes the serious limitation of MED-Start on GC-rich genomes, while outperforming other TIS predictors.

2.1.3 AT-rich genome. For the genome *S. solfataricus* with a high AT content, MED-StartPlus reached the highest prediction accuracy of 85.7% on the data set *SolfGene*. Note that the accuracy on *SolfGene* for all methods are generally lower than those on other data sets, possibly due to the relatively small size of the data set. However, TIS prediction accuracy for AT-rich genomes is higher than that for GC-rich genomes, as will be demonstrated later.

2.2 Comparison against function-known genes in RefSeq

The bottleneck for evaluating TIS prediction is the relatively small number of genes with TISs verified in the laboratory, as well as few genomes have been built even with such small gene data sets listed in Table 1. To make a large-scale evaluation, we then report the results of comparing predictions against the RefSeq annotations on 300 prokaryotic genomes. Since it has long been known that the public RefSeq annotation is not fully accurate, the comparison was performed on totally 450 849 function-known genes for the 300 genomes, of which the starts were regarded as being more reliable than others in the annotations. Here, without further experimental evidence, the data sets indeed constitute the current best public resources for large-scale evaluation. According to the same rule, the genomes were classified into three groups: 94 AT-rich genomes (GC < 40%), 107 genomes with medium GC content (40% ≤GC ≤56%), and 99 GC-rich genomes (GC > 56%).

The average accuracy of prediction herein referred to the rate of TISs consistent with the datasets, by MED-StartPlus, MED-Start, RBS-finder, GS-Finder and TiCo, are summarized in Table 2, while the full details for all 300 genomes are available at the website: http://ctb. pku. edu. cn / main / SheGroup / Software/ MEDPWEB / MEDStartPlus.htm. For Hon-yaku, since it employs a supervised learning method with manual manipulation, we did not include it in the comparison. As we can see from Table 2, MED-StartPlus reached an accuracy of 82.4% on average for 107 genomes with the medium GC content, 0.9% higher than that of TiCo of the best of the others. For the 94 AT-rich genomes, MED-StartPlus had an accuracy of 85.8% on average, 1.0% higher than that of TiCo. With regard to the 99 GC-rich genomes, the averaged accuracy of all programs is lower than 80%. However, the accuracy achieved by MED-StartPlus was 78.1% on average, a remarkable 4.2% higher than that of TiCo with the best performance among others. The results show that MED-StartPlus has a marked advantage in TIS prediction over MED-Start and RBSfinder, especially in GC-rich genomes, where the average improvement was 41.8% and 18.4% respectively.

Table 2 Average percentages of predictions consistent with the RefSeq annotations on function-known genes for MED-StartPlus, MED-Start, RBSfinder, GS-Finder and TiCo

Genome group	# genomes	MED-StartPlus(%)	MED-Start(%)	RBSfinder(%)	GS-Finder(%)	TiCo(%)
AT-rich	94	**85.8**	78.6	78.8	84.9	84.8
Medium GC content	107	**82.4**	72.8	70.9	79.0	81.5
GC-rich	99	**78.1**	36.3	59.7	72.6	73.9

Clearly, MED-StartPlus generally produced high consistent predictions with the function-known gene sets in the RefSeq annotation, and showed a totally higher performance compared with the existing TIS predictors for genomes over a wide variety of GC content.

2.3 Improvement by combining two TIS modules

Finally, we discuss improvement of the strategy by combining the two TIS modules in MED-StartPlus. It has been demonstrated *in vivo* that some genomes such as *S. solfataricus* possesses two different mechanisms of translation initiation[25]. Most genes at the beginning of a transcript unit in *S. solfataricus* are leaderless, which means that the 5′ untranslated region (5′ UTR) is short or missing. In fact, leaderless genes have been reported in several archaeal genomes [9]. Unlike SD-led genes, which possess SD signals in sequence upstream to the starts, only transcription initiation-related signals are found for leaderless genes [9]. Thus, for genomes containing both SD-led genes and leaderless genes, two kinds of signals exist upstream to the starts. However, in most cases, the "Signal involved module" converges to only one kind of signals. For instance, only SD-like motifs like "AGGAG" and "GGAGG" were found in *S. solfataricus*. Such a preference would lead to losing the ability of the "Signal involved module" to predict TISs for leaderless genes. Therefore in the current algorithm, MED-StartPlus deployed the strategy of combination of the "Signal involved module" and the "Simplified module".

To examine the improvement achieved by such a strategy, we compared the prediction performance of each module on the same benchmarks in Table 1. The results showed that the combination had an overall higher performance than the "Simplified module" with a maximum improvement of 3%. While compared with the "Signal involved module", the improvements achieved were even higher in *SolfGene* (by 35.7%), *Syne* (29.1%), *Npha* (15.6%) and *Hsal* (7.9%), all of which are believed to contain leaderless genes in their genomes[9, 16, 25]. Note that the final system of MED-StartPlus reports an overall better performance than any other software for the four genomes. The results herein clearly indicate that our method works efficiently in predicting TIS for genomes with complex initiation mechanism.

3 Conclusion

In this paper, we developed a new algorithm to model the complex statistical patterns associated with translation initiation in prokaryotic genomes over a wide variety of GC content. The resulting method MED-StartPlus has been comprehensively evaluated for the prediction performance on 14 published data sets including 4 074 genes with TISs confirmed by various evidences, and a total of 450 849 function-known genes from 300 genomes with genomics GC content arranging from 24.0% to 74.9%. The results show that MED-StartPlus has a high overall performance. We further report a detailed comparison with the existing methods of TIS prediction. With a performance competitive with the best of them on the AT-rich genomes and genomes with the medium GC content, MED-StartPlus outperformed the current best TIS predictors on the GC-rich genomes and genomes with complex initiation mechanism.

With growing number of completely sequenced prokaryotic genomes, more and more genomes deposited in the public database are exhibiting higher or lower GC content and complex translation initiation mechanisms. We therefore expect the potential application of our method combined with the gene finding tools, which will lead to improving the TIS annotation in the public annotation such as RefSeq, and in ongoing sequencing projects by the improved genome annotation pipeline. Recently, the method has been jointly applied with extrinsic evidences to build the database ProTISA, which collects reliable TIS annotation and predicts motifs associated with TIS for all currently sequenced genomes in prokaryotes[26].

The software and the source code implemented in C++ can be freely downloaded from http://ctb.pku.edu.cn/main/SheGroup/Software/MEDPWEB/MEDStartPlus.htm under the GNU GPL license.

Acknowledgements We thank YAO Xin-Qiu, KANG Hong and SUN Zong-Xiao of Peking University, Philippe Ortet of CEA Cadarache, and Prof. WANG Jin of Nanjing University for beneficial discussions. We also thank Prof. ZHANG Chun-Ting for providing the GS-Finder program, and Dr. Yuko Makita and Dr. Michalis Aivaliotis for providing parts of the data sets. We are very grateful to Dr. Iain C Bruce of Zhejiang University School of Medicine for carefully reading and revising the manuscript.

References

1 Besemer J, Lomsadze A, Borodovsky M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res, 2001, **29**(12): 2607～2618

2 Suzek B E, Ermolaeva M D, Schreiber M, *et al*. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics, 2001, **17**(12): 1123～1130

3 Ou H Y, Guo F B, Zhang C T. GS-Finder: a program to find bacterial gene start sites with a self-training method. Int J Biochem Cell Biol, 2004, **36**(3): 535～544

4 Zhu H Q, Hu G Q, Ouyang Z Q, *et al*. Accuracy improvement for identifying translation initiation sites in microbial genomes. Bioinformatics, 2004, **20**(18): 3308～3317

5 Tech M, Meinicke P. An unsupervised classification scheme for improving predictions of prokaryotic TIS. BMC Bioinformatics, 2006, **7**: 121

6 Makita Y, de Hoon M J, Danchin A. Hon-yaku: a biology-driven Bayesian methodology for identifying translation initiation sites in prokaryotes. BMC Bioinformatics, 2007, **8**: 47

7 Frishman D, Mironov A, Gelfand M. Starts of bacterial genes: estimating the reliability of computer predictions. Gene, 1999, **234** (2): 257～265

8 Londei P. Evolution of translational initiation: new insights from the archaea. FEMS Microbiol Rev, 2005, **29**(2): 185～200

9 Torarinsson E, Klenk H P, Garrett R A. Divergent transcriptional and translational signals in Archaea. Environ Microbiol, 2005, **7** (1): 47～54

10 Nielsen P, Krogh A. Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics, 2005, **21** (24): 4322～4329

11 Hori H, Osawa S. Evolutionary change in 5S rRNA secondary structure and a phylogenic tree of 352 5S rRNA species. Biosystems, 1986, **19**(3): 163～172

12 Yamazaki S, Yamazaki J, Nishijima K, *et al*. Proteome analysis of an aerobic hyperthermophilic crenarchaeon, Aeropyrum pernix K1. Mol Cell Proteomics, 2006, **5**(5): 811～823

13 Rudd K E. EcoGene: a genome sequence database for *Escherichia coli* K-12. Nucleic Acids Res, 2000, **28**(1): 60～64

14 Link A J, Robison K, Church G M. Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. Electrophoresis, 1997, **18**(8): 1259～1313

15 Sazuka T, Yamaguchi M, Ohara O. Cyano2Dbase updated: linkage of 234 protein spots to corresponding genes through N-terminal microsequencing. Electrophoresis, 1999, **20**(11): 2160～2171

16 Zhu H, Hu G Q, Yang Y F, *et al*. MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. BMC Bioinformatics, 2007, **8**: 97

17 Yada T, Totoki Y, Takagi T, *et al*. A novel bacterial gene-finding system with improved accuracy in locating start codons. DNA Res, 2001, **8**(3): 97～106

18 Aivaliotis M, Gevaert K, Falb M, *et al*. Large-scale identification of N-terminal peptides in the halophilic archaea Halobacterium salinarum and *Natronomonas pharaonis*. J Proteome Res, 2007, **6**(6): 2195～2204

19 Ma J, Campbell A, Karlin S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol, 2002, **184**(20): 5733～5745

20 Salgado H, Moreno-Hagelsieb G, Smith T F, *et al*. Operons in *Escherichia coli*: genomic analyses and predictions. Proc Natl Acad Sci USA, 2000, **97**(12): 6652～6657

21 Chen L L, Zhang C T. Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. Biochem Biophys Res Commun, 2003, **306**(1): 310～317

22 Nishi T, Ikemura T, Kanaya S. GeneLook: a novel ab initio gene identification system suitable for automated annotation of prokaryotic sequences. Gene, 2005, **346**: 115～125

23 Guo F B, Ou H Y, Zhang C T. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucleic Acids Res, 2003, **31**(6): 1780～1789

24 Chang B, Halgamuge S, Tang S L. Analysis of SD sequences in completed microbial genomes: Non-SD-led genes are as common as SD-led genes. Gene, 2006, **373**: 90～99

25 Benelli D, Maone E, Londei P. Two different mechanisms for ribosome/mRNA interaction in archaeal translation initiation. Mol Microbiol, 2003, **50**(2): 635～643

26 Hu G Q, Zheng X, Yang Y F, *et al*. ProTISA: a comprehensive resource for translation initiation site annotation in prokaryotic genomes. Nucleic Acids Res, 2008, **36** (database issue): D114～D119

原核基因翻译起始位点预测的新方法 *

胡钢清[1,2] 刘永初[1,2] 郑晓斌[1,2] 杨一帆[2] 佘振苏[1,2,3] 朱怀球[1,2,4]**

([1]北京大学生物医学工程系、湍流与复杂系统国家重点实验室，北京 100871;
[2]北京大学理论生物学中心，北京 100871;
[3] *Department of Mathematics, UCLA, Los Angeles, CA* 90095, USA;
[4]*The Finnish Genome Center, University of Helsinki*, 00290 Helsinki, Finland)

摘要 翻译起始位点(TIS，即基因 5′端)的精确定位是原核生物基因预测的一个关键问题，而基因组 GC 含量和翻译起始机制的多样性是影响当前 TIS 预测水平的重要因素．结合基因组结构的复杂信息(包括 GC 含量、TIS 邻近序列及上游调控信号、序列编码潜能、操纵子结构等)，发展刻画翻译起始机制的数学统计模型，据此设计 TIS 预测的新算法 MED-StartPlus．并将 MED-StartPlus 与同类方法 RBSfinder、GS-Finder、MED-Start、TiCo 和 Hon-yaku 等进行系统地比较和评价．测试针对两种数据集进行：当前 14 个已知的 TIS 被确认的基因数据集，以及 300 个物种中功能已知的基因数据集．测试结果表明，MED-StartPlus 的预测精度在总体上超过同类方法．尤其是对高 GC 含量基因组以及具有复杂翻译起始机制的基因组，MED-StartPlus 具有明显的优势．

关键词 原核生物，基因预测，翻译起始位点，预测评价

学科分类号 Q61，Q93

* 国家自然科学基金(30770499, 30300071, 10721403)和国家重点基础研究发展计划(2003CB715905)资助项目．

** 通讯联系人. Tel: 010-62767261, E-mail: hqzhu@pku.edu.cn, hqzhu@ctb.pku.edu.cn

收稿日期：2008-02-05，接受日期：2008-06-30