Sequence analysis

Prediction of translation initiation site for microbial genomes with TriTISA

Gang-Qing Hu1,2, Xiaobin Zheng1,2, Huai-Qiu Zhu1,2, Zhen-Su She1,2,*

1State Key Lab for Turbulence and Complex Systems, Department of Biomedical Engineering, College of Engineering and 2Center for Theoretical Biology, Peking University, Beijing 100871, China

Received on May 21, 2008; revised and accepted on November 4, 2008
Advance Access publication November 10, 2008
Associate Editor: Limsoon Wong

ABSTRACT

Summary: We report a new and simple method, TriTISA, for accurate prediction of translation initiation site (TIS) of microbial genomes. TriTISA classifies all candidate TISs into three categories based on evolutionary properties, and characterizes them in terms of Markov models. Then, it employs a Bayesian methodology for the selection of true TIS with a non-supervised, iterative procedure. Assessment on experimentally verified TIS data shows that TriTISA is overall better than all other methods of the state-of-the-art for microbial genome TIS prediction. In particular, TriTISA is shown to have a robust accuracy independent of the quality of initial annotation.

Availability: The C++ source code is freely available under the GNU GPL license via http://mech.ctb.pku.edu.cn/protisa/TriTISA.

Contact: she@pku.edu.cn

Supplementary information: Full documentation of the program, containing installation instructions and other operational details, is available on our website. Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Accurate prediction of translation initiation site (TIS) for microbial genomes has been a challenge to a number of gene finders (Besemer et al., 2001; Delcher et al., 2007; Nielsen and Krogh, 2005; Zhu et al., 2007) and TIS post-processors (Ou et al., 2004; Tech et al., 2005; Zhu et al., 2004). For a good prediction of TIS, a number of TIS-related features, such as the length of regulatory signal, start codon usage, operon structure and coding potential have been invoked in the design of prediction methods. Often, assumptions about TIS-related features are tested only for a few genomes such as Bacillus subtilis and Escherichia coli, and the validity to all genomes may be questioned (Makita et al., 2007; Tech et al., 2005). Recently, Tech et al. (2005) have introduced a tool, TiCo, which utilizes a second-order Markov model with positional smoothing to characterize sequence properties around TIS. However, as an unsupervised tool, TiCo is criticized for potential dependency on the quality of initial annotation (Makita et al., 2007). More importantly, all existing tools seem to work for some genomes but not for others. This is due to the diversity of translation initiation mechanisms (Hu et al., 2008a), which cannot be represented by assumptions. In this note, we propose a new, simple and highly robust TIS predictor for microbial genomes, which is based on a universal principle of evolution and has a better accuracy than all existing methods.

2 METHODS

A candidate TIS refers to any in-frame start-codon-like triplet (namely, ATG, CTG, GTG or TTG) within the open reading frame. TIS prediction is to select the right candidate TIS. According to Darwinian evolution, three categories of candidate TISs may arise: true TISs, false TIS upstream of the true TIS and false TIS downstream of the true TIS. Sequence pattern around the true TISs contains species-specific signals that are conserved under selection pressure. In contrast, upstream false TISs are exposed to neutral evolution with minimal sequence features, while downstream false TISs are surrounded by coding sequence and exhibit period-three oscillations (Hu et al., 2008a). The tri-category of candidate TISs is a universal property of all genomes, and a method based on this principle will be applicable to all genomes.

We characterize statistical properties of candidate TISs from each category by a non-homogeneous n-th-order Markov model. Then, a Bayesian probability that a candidate TIS is a true TIS is

\[ P(T|S) = \frac{\alpha_T \prod_{i=1}^{l-1} P_X(s_i | s_{i+1}, \ldots, s_{i+n-1})}{\sum_{X \in \{U,D\}} \alpha_X \prod_{i=1}^{l-1} P_X(s_i | s_{i+1}, \ldots, s_{i+n-1})} \] (1)

where \( S = s_1s_2 \ldots \) is the sequence around candidate TIS such that \( l = L_u + 3 + L_d \), \( L_u \) and \( L_d \) are the number of nucleotides upstream and downstream of TIS, respectively, \( P_X(s_i | s_{i+1}, \ldots, s_{i+n-1}) \) denotes transition probability from \( s_{i+n-1} \) to \( s_i \) and \( i \) is the nucleotide position in sequence \( S \). \( \alpha_T \) is a prior probability, and \( T, U, D \) refer to true, upstream false and downstream false TISs, respectively.

An iterative, unsupervised procedure is designed to train parameters for Markov models of a certain order (Supplementary Material 1). We begin with an initial set of candidate TISs designated as true TISs (set \( T \)). From which we construct the set \( U \) and \( D \) by selecting other TISs than the candidates. Three Markov models are obtained in a straightforward way with a simple counting on the three sets. The Bayesian probability for true TIS in Equation (1) is calculated and used to re-evaluate each candidate TIS; by selecting the one with the highest score as new candidate TIS, we update the three sets of TISs. The procedure repeats until the change in the true TISs is below 0.1%. In practice, we first use a zeroth-order Markov model for initial refinements, and then use higher order Markov models in the later steps of the iteration. This ensures us to have both a robust and accurate TIS predictor (Supplementary Material 1).

The determination of parameters in Equation (1) involves two technical details. First, to have sufficient data, \( P_X \) is computed from all start-codon-like triplets in non-coding regions, regardless of their phases. Second, we use the ratios of sample sizes of the three TIS sets, automatically obtained at each step of iterations, to estimate the three prior probabilities (Marques de Sa, 2001).

*To whom correspondence should be addressed.

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
that is, $P_U = N_U / N$, $P_T = N_T / N$, $P_D = N_D / N$, where $N_U$, $N_T$, $N_D$ are the set sizes of $U$, $T$, $D$, respectively, and $N = N_U + N_T + N_D$.

### 3 RESULTS AND DISCUSSION

TriTISA is designed to post-process an existing microbial gene annotation program, thus to improve its prediction quality of gene starts. It takes an entire genome and a set of initial TIS annotations as input, and predicts one TIS per gene. Here, we apply the method to post-process the RefSeq whole-genome annotation, and evaluate the predictions with experimentally verified TISs currently available for five genomes, including two GC-rich genomes, namely *Natronomonas pharaonis* (63.1%) and *Halobacterium salinarum* (68.1%). Measures for the prediction performance include sensitivity TP/(TP+FN), and specificity TN/(TN+FP), where TP, TN, FP, and FN denote the numbers of true positives, true negatives, false positives and false negatives, respectively (Li and Jiang, 2005). TriTISA reports an average sensitivity of 95.5% (Table 1). The prediction is input sensitive, with a linear correlation between the prediction accuracy and the quality of the input. TriTISA predicts 10–35% higher sensitivity with slightly higher specificity over TiCo when the initial annotation $\alpha < 75\%$ (see Supplementary Material 4 for more details).

In conclusion, we have found a universal, simple and robust method for achieving highly accurate TIS prediction. The method becomes more useful when new genomes with diverse translation initiative mechanisms need good annotation of TIS. It may also represent a powerful tool for refining the annotation of public databases, such as RefSeq whose annotation quality becomes increasingly questionable (Hu et al., 2008a). The computation time of TriTISA is modest; it takes <2 min to analyze the complete genome of *E. coli* K12 on a personal computer (AMD Athlon 64 processor 1.8G).

**Funding**: The National Natural Science Foundation of China (10225210, 30700499 and 10721403); the National Basic Research Program of China (2003CB715905).

**Conflict of Interest**: none declared.

### REFERENCES


### Table 1. Accuracies of TriTISA and other tools on experimentally verified data

<table>
<thead>
<tr>
<th>Genome</th>
<th>Gene finders</th>
<th>Post-processors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No.</td>
<td>TriTISA</td>
</tr>
<tr>
<td><em>Aeropyrum pernix</em></td>
<td>130</td>
<td>96.9 (99.9)</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>883</td>
<td>95.3 (99.8)</td>
</tr>
<tr>
<td><em>Halobacterium salinarum</em></td>
<td>552</td>
<td>95.3 (99.8)</td>
</tr>
<tr>
<td><em>Natronomonas pharaonis</em></td>
<td>321</td>
<td>97.5 (99.9)</td>
</tr>
<tr>
<td><em>Synechocystis sp.</em></td>
<td>124</td>
<td>91.9 (99.6)</td>
</tr>
<tr>
<td>Average</td>
<td>–</td>
<td>95.5 (99.9)</td>
</tr>
</tbody>
</table>

References for tools: EasyGene (Nielsen and Krogh, 2005); Glimmer (Delcher et al., 2007); GeneMarkS (Besemer et al., 2001); GS_finder (Ou et al., 2004); and MED-StartPlus (Hu et al., 2008b). References for datasets: *A. pernix* (Yamazaki et al., 2006); *E. coli* (Rudd, 2000); *H. salinarum* and *N. pharaonis* (Aivaliotis et al., 2007); and *Synechocystis* sp. PCC6803 (Sazuka et al., 1999). Predictions of EasyGene and Glimmer/GeneMarkS were downloaded from the EasyGene website and RefSeq respectively. The other tools were run with default settings. The weighted averages are calculated with weights proportional to gene sizes. The performances are measured by sensitivity, followed by specificity in parenthesis.