MetaTISA: Metagenomic Translation Initiation Site Annotator for improving gene start prediction

Gang-Qing Hu†, Jiang-Tao Guo†, Yong-Chu Liu†, Huaiqiu Zhu‡
State Key Laboratory for Turbulence and Complex Systems, Department of Biomedical Engineering and Center for Theoretical Biology, Peking University, Beijing 100871, China

Associate Editor Dr. Limsoon Wong

ABSTRACT
Summary: We proposed a tool named MetaTISA with an aim to improve TIS prediction of current gene-finders for metagenomes. The method employs a two-step strategy to predict TISs by first clustering metagenomic fragments into phylogenetic groups and then predicting TISs independently for each group in an unsupervised manner. As evaluated on experimentally verified TISs, MetaTISA greatly improves the accuracies of TIS prediction of current gene-finders.
Availability: The C++ source code is freely available under the GNU GPL license via http://mech.ctb.pku.edu.cn/MetaTISA/.
Contact: hqzhu@pku.edu.cn

1 INTRODUCTION
Accurate gene finding from a large number of shotgun sequences plays an important role in metagenomic sequencing projects. Recently with two noteworthy works, MetaGene (Noguchi et al., 2006, 2008) and Neural Net (Hoff et al., 2008), the current methods show sensitivities and specificities, measured on 3′-ends of genes, that are comparable to those developed for single microbial genome. But, in contrast, accurate prediction of the 5′-ends, namely translation initiation sites (TISs), remains largely unresolved (Hoff et al., 2008; Noguchi et al., 2008). Many of the metagenomics projects is to identify novel proteins followed by experimental characterizations. The need for an accurate prediction of TIS is justified, because expression of genes with incorrect TISs may fail or yield incorrect results (Hoff et al., 2008).

TIS prediction for microbial genomes has achieved high accuracies for a number of tools (Besemer et al., 2001; Delcher et al., 2007; Hu et al., 2008a, 2009; Makita et al., 2007; Nielsen and Krogh, 2005; Tech et al., 2005; Zhu et al., 2004). However, due to the fragmentary nature of metagenomic shotgun sequences and their unknown phylogenetic origins, it is inapplicable to directly apply these methods to metagenomes. In an effort to address this issue, we proposed the MetaTISA (Metagenomic Translation Initiation Site Annotator) tool, which first assigns metagenomic fragments into phylogenetic groups (Sandberg et al., 2001) and then predicts TISs independently for each group with the TriTISA method (Hu et al., 2009). As a post-processor, MetaTISA greatly improves the TIS prediction for current gene-finders such as MetaGeneAnnotator (Noguchi et al., 2008).

2 METHODS
Given a set of predicted coding sequences (CDSs), MetaTISA constitutes two steps in TIS prediction. The first is to assign the fragments into phylogenetic groups. Fragments from the same group are assumed to have closely phylogenetic origin and share a mechanism in translation initiation. We implemented the k-mer method that employs a naïve Bayesian classifier for sequence binning (Sandberg et al., 2001). A set of k-mer frequencies were calculated from current microbial genome annotations for training (one genome per genus to reduce redundancy). Given the k-mer frequencies of a fragment, the classifier compares the likelihood scores calculated from all genera and predicts the fragment’s phylogenetic origin as the one that gives the highest score (Sandberg et al., 2001). Then, the algorithm developed in TriTISA is used to train parameters and predict TIS for each group (Hu et al., 2009). CDSs that are complete in their 5′-ends, thus containing start codons, are selected into the procedure of self-training of parameters, as illustrated in Hu et al. (2009). For each candidate TIS, the algorithm calculates three posterior probabilities: $P_t$ (the candidate as a true TIS), $P_{nc}$ (the candidate from noncoding region), and $P_{co}$ (the candidate from coding region). When CDSs are incomplete in their 5′-ends, it estimates whether the 5′-most candidate TIS is from coding regions (thus the start codon is missing) with the distribution of $P_{co}$ (at a 99% confidence level). For CDS known or predicted to contain a start codon, the algorithm predicts TIS as the one that gives the highest $P_t$ score.

3 RESULTS AND DISCUSSION
As designed to post-process gene predictors in an existing annotation pipeline for metagenomes, MetaTISA aims to improve the quality of TIS predictions. It requires inputs as metagenomic fragments in the FASTA format and a set of gene predictions. At current stage, it accepts gene predictions in formats as specified by MetaGeneAnnotator (Noguchi et al., 2008). As well it accepts our own format called “MED”. The refined predictions are provided in a “MED” format or a general feature format (GFF). Details regarding the formats of input and output are available in our web site.

MetaTISA includes a set of settings for running. The first three parameters specify the sequence region around TIS and the maximal order of the Markov models that characterize the sequences. We have shown that a region that covers 50 bps upstream and 15 bps downstream to TISs and a maximal order of 2 give the best performance, and accuracy improvements brought by a wider region and a higher order are marginal (Hu et al., 2009). The fourth
parameter is the motif length $k$ for the $k$-mer method. Sandberg et al. (2001) have shown that the accuracies of the classifier increases with $k$. We tested MetaTISA with simulation data and found an accuracy improvement of 2-3% for $k = 9$ over $k = 6$, however improvement brought by a larger $k$ is marginal (data not shown).

The last parameter refers to the minimal number of CDSs required for a good estimation of parameters for each group. By testing on the EcoGene data set (Rudd, 2000), the accuracy get saturated when the size is over 200 (the default value). For a group with smaller sample size, MetaTISA predicts TISs by using parameters pre-computed from the genome that is selected to train the $k$-mer method.

Due to the lacking of experimentally verified TISs in metagenome project, the only way to reliably evaluate the prediction performance is to simulate a metagenome based on artificial shotgun sequences from complete microbial genomes (Hoff et al., 2008; Noguchi et al., 2006, 2008). We generated shotgun sequences from 95 randomly selected genomes plus 5 genomes where experimentally verified TISs are available as benchmarks (see Supplementary File 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 (see also Supplementary File 1). Two kinds of simulation were created according to the choice of fragment length: $L = 700$ bps and $L = 400$ bps. We demonstrated the prediction improvements by post-processing the outputs of MetaGeneAnnotator (Table 1) (Noguchi et al., 2008). For $L = 700$ bps, MetaGeneAnnotator shows an average sensitivity of 87.0% and specificity of 99.5%. After being post-processed by MetaTISA, the average accuracy of TISs increases to 93.4% for sensitivity and 99.7% for specificity. Similar improvements are obtained for a more challenging situation of $L = 400$ bps (Table 1). It is interesting to observe that MetaTISA gives generally lower sensitivities for AT-rich genomes than for GC-rich genomes, which is not the case for MetaGeneAnnotator. In fact, this could be an artifact caused by insufficient samples for genome Synechocysis sp. PCC 6803. For another genome *Bacillus subtilis* 168, it should be pointed out that the function characterization of the “non-y” genes by experiments doesn’t guarantee the TISs be 100% correct, and hence the data set is less reliable than others (Yada et al., 2001). Thus, it is likely that figures from the two data sets under-estimate the prediction performances on AT-rich genomes. Moreover, a recent finding that accuracy of TIS annotation in AT-rich genome is generally higher than in GC-rich genomes is in favor of our explanation (Nielsen and Krogh, 2005).

As an indirect way to assess the predictions, it is intriguing to examine the predicted RBS patterns for *Firmicutes*, a major clade in the environmental microbes. Most genes from this phylum contain a Shine-Dalgarno (SD) sequence for translation initiation (Ma et al., 2002). We applied MetaTISA to post-process MetaGeneAnnotator’s predictions for the Human Gut Community Subject 7 (Gill et al., 2006), and visualized for each group the positional weight matrix that summarizes the sequences pattern around predicted TISs. Notably, guanine rich sequences, corresponding to SD signals, are dominant in the sequences upstream to TISs from *Firmicutes* (see Supplementary File 2), highlighting the validity of our method.

Horizontal gene transfer (HGT) is supposed to be involved in environmental bacterial speciation, as implied from an analysis of codon usage bias in *Escherichia coli* (Médigue et al., 1991). Recent HGT may complicate the prediction of a fragment’s origin through composition-based methods like $k$-mer (Diaz et al., 2009), Tamames and Moya (2008) find that recent HGT event for metagenomes occurs at a comparable frequency to that for complete genomes (2-8%). For our method to work on this category of genes, it requires that genes recently transferred may still involve a mechanism for translation initiation that is similar to the donor in the receptor genome. The assumption seems to be plausible due to the universality of the mechanism of translation initiation throughout prokaryotes (Nakamoto, 2009). However, to prove it is a difficult task because of the lacking of donor and receptor data during HGT (Podell and Gaasterland, 2007), and we do not intend to accomplish this in this work. Nevertheless, we collected a total of 305 transfer genes from HGT-DB for *E. coli* K12 (predicted based on compositions) to see if our method makes improvements on this set of genes (Garcia-Vallve et al., 2003). Taken RefSeq’s annotation as benchmarks, test on the simulation data shows that MetaTISA improves the sensitivity by 2.2% (to 87.8%) with a slightly lower

### Table 1. Accuracy improvements of TIS prediction by MetaTISA on genes predicted by MetaGeneAnnotator (MGA)\(^a\).

<table>
<thead>
<tr>
<th>Genome</th>
<th>% GC</th>
<th># genes (^c)</th>
<th>Fragment length = 700 bps</th>
<th>Fragment length = 400 bps</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>MetaGeneAnnotator</td>
<td>MetaTISA</td>
</tr>
<tr>
<td>Bacillus subtilis 168</td>
<td>43.5</td>
<td>1248</td>
<td>3000</td>
<td>87.2 (99.5)</td>
</tr>
<tr>
<td>Synechocysis sp. PCC 6803</td>
<td>47.4</td>
<td>124</td>
<td>197</td>
<td>83.5 (99.3)</td>
</tr>
<tr>
<td>Escherichia coli K12</td>
<td>50.8</td>
<td>883</td>
<td>2123</td>
<td>89.6 (99.8)</td>
</tr>
<tr>
<td>Aeropyrum pernix K1</td>
<td>56.3</td>
<td>130</td>
<td>256</td>
<td>64.4 (98.8)</td>
</tr>
<tr>
<td>Natronomonas pharaonis DSM 2160</td>
<td>63.1</td>
<td>321</td>
<td>651</td>
<td>91.3 (99.6)</td>
</tr>
<tr>
<td>Halobacterium salinarum R1</td>
<td>68.1</td>
<td>552</td>
<td>1124</td>
<td>85.7 (99.5)</td>
</tr>
<tr>
<td>Weighted average(^a)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>87.0 (99.5)</td>
</tr>
</tbody>
</table>

\(^a\)Sensitivity and specificity (in parenthesis) are averaged over 5 simulation replicates. \(^b\)References for datasets: *A. pernix* (Yamazaki et al., 2006); *E. coli* (Rudd, 2000); *H. salinarum* and *N. pharaonis* (Aviaditis et al., 2007); *Synechocysis sp.* PCC6803 (Sazonka et al., 1999). A set of “non-y” genes (experimentally characterized in functions) is proposed by Yada et al. (2001) to evaluation TIS prediction for *Bacillus subtilis*. Number of genes with reliable TISs. \(^c\)Number of true positive predictions over genes with reliable TISs (averaged over the five replicates). Since we focus on accuracy improvements on already predicted genes, sensitivities and specificities are calculated on true positive gene predictions. Note that many of predictions are partial copies from one gene, and the average copy number, as we found, is in general proportional to the mean of gene length (Supplementary File 1). About 30-50% (20-30%) of the predicted genes show true TISs in the fragments for $L = 700$ (400). \(^d\)The weighted averages are calculated with weights proportional to gene sizes.
specificity of 99.2% \((L = 700)\). Note that the figures may be underestimated as well, due to the imperfection of TIS annotations in RefSeq (Hu et al., 2008b; Nielsen and Krogh, 2005).

In conclusion, we presented an accurate method for TIS prediction in metagenomes. One direction of ongoing development is to include more genera to train the \(k\)-mer method and to include other binning methods such as PhyloPythia (McHardy et al., 2007) and TACOA (Diaz et al., 2009). Interestingly, as observed in \textit{E. coli} K12, 17% of the simulated fragments are incorrectly binned \((k = 9)\). However, almost all the misclassified fragments are assigned to close relatives of \textit{Escherichia}, which are expected to share a mechanism of translation initiation. This may explain why MetaTISA still correctly predict most of the TIS even if the fragment is incorrectly binned.

**ACKNOWLEDGEMENTS**

We are grateful to Dr. Hideki Noguchi from Mitsubishi Research Institute and Ms. Katharina J Hoff from University of Goettingen for useful email communications. The work received support by the National Natural Science Foundation of China (30770499 and 10721403) and the Wallace H. Coulter GaTech/Emory-PKU BME Collaborative Research Seed Grant Program.

**Conflict of Interest** none declared.

**REFERENCES**


Ma, J. et al. (2007) Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. \textit{J. Bacteriol.}, 189, 5733-5745.


