Sequence analysis

TICO: a tool for improving predictions of prokaryotic translation initiation sites

Maike Tech*, Nico Pfeifer, Burkhard Morgenstern and Peter Meinicke

Abteilung Bioinformatik, Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen, Goldschmidtstrasse 1, 37077 Göttingen, Germany

Received on May 9, 2005; revised on June 24, 2005; accepted on June 27, 2005
Advance Access publication June 30, 2005

ABSTRACT

Summary: We provide the tool ‘TICO’ (Translation Initiation site COrrection) for improving the results of conventional gene finders for prokaryotic genomes with regard to exact localization of the translation initiation site (TIS). At the current state TICO provides an interface for direct post processing of the predictions obtained from the widely used program GLIMMER. Our program is based on a clustering algorithm for completely unsupervised scoring of potential TIS locations.

Availability: Our tool can be freely accessed through a web interface at http://tico.gobics.de/

Contact: maike@gobics.de

For prokaryotes, there are a number of gene-finding tools that can reliably predict the location of genes in a genome under study, for example GLIMMER (Delcher et al., 1999), GS-finder (Ou et al., 2004), MED-Start (Zhu et al., 2004), ZCurve (Guo et al., 2000) and GeneMarkS (Besemer et al., 2001). Essentially, these methods are based on a search for open reading frames with a statistically significant minimal length. In addition, characteristic statistics of sequence content features, such as oligo-nucleotide frequencies, are considered for evaluation of these open reading frames. But, while it is obvious how to identify the end position of a putative gene, it is by no means trivial to determine the corresponding start position as the codons for signaling the initiation of translation may also be used inside genes to code for amino acids. Systematic studies have shown that existing gene finders perform poorly in the prediction of correct translation initiation sites (TIS) (Ou et al., 2004; Zhu et al., 2004; Tech and Merkl, 2003). Consequently, many start positions are incorrectly annotated in databases and, due to the concepts used for gene annotation, these errors tend to be propagated to newly annotated genomes.

We present a tool, TICO (Translation Initiation site COrrection), for improving the results of conventional gene finders by analyzing and relocating prior predictions of prokaryotic TIS. Currently our tool provides an interface for post processing the output of the widely used program GLIMMER. Unlike other programs it is not based on any specific assumptions about prokaryotic TIS. Some existing tools do provide a sequence model and an unsupervised method for optimizing most of the parameters without the need for prior knowledge (Ou et al., 2004; Zhu et al., 2004; Guo et al., 2000; Suzek et al., 2001). However, these models usually include additional TIS related parameters that cannot be adjusted by means of the optimization method. These special parameters can for instance involve the length of a putative RBS motif, the maximal number of RBS motifs considered or the distribution of the start codon usage. Usually these parameters are adjusted to ‘default’ values, which provide good results on genomes like Escherichia coli and Bacillus subtilis. Because with that choice one implicitly makes assumptions about TIS characteristics there exists a certain amount of a risk that the results become suboptimal if the tools are applied to genomes of other species.

Our method is based on the analysis of candidate TIS sequences as obtained from the flanking regions of potential start codons. We implemented a clustering algorithm that performs an unsupervised classification of sequences according to strong-TIS and weak-TIS categories. As potential TIS locations we consider the positions of all admissible start codons in a specified search range (Fig. 1) around the initial TIS, as predicted by a conventional gene finder. In addition, potential start codons have to share the same reading frame of the

Fig. 1. The figure illustrates two parameters, which may be adapted by the user. Above, the search range is shown. This range defines the maximum distance to be searched for alternative start sites around the initially predicted TIS (denoted as initial TIS in the figure). The initial TIS and the alternative start sites are termed candidate TIS. Below, the extract range is shown. This defines the regions around each candidate TIS to be extracted for the scoring based on unsupervised learning. For both parameters, search range and extract range, the length of upstream and downstream regions can be adjusted independently by the user.

*To whom correspondence should be addressed.
Table 1. The accuracy of the TIS prediction obtained by TICO compared with other gene finders and the GenBank annotation (GBK).

<table>
<thead>
<tr>
<th>GBK</th>
<th>Gene finders</th>
<th>Post processors</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLIMMER</td>
<td>ZCURVE</td>
<td>GS-finder</td>
</tr>
<tr>
<td>89.8</td>
<td>63.2</td>
<td>88.6</td>
</tr>
</tbody>
</table>

Among these RBSfinder (Suzek et al., 2001), GS-finder (On et al., 2004) and MED-Start are used as post processors on the same GLIMMER2.02-prediction as TICO. Accuracy was measured in percent of TIS that was predicted correctly with respect to reference annotations for 854 genes from the EcoGene database. The GBK entry refers to EcoGene TIS that coincide with the corresponding GenBank annotation.

**REFERENCES**


TICO: improving predictions of prokaryotic TIS

conflict of interest: none declared.