W588–W590 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl313

TICO: a tool for postprocessing the predictions of prokaryotic translation initiation sites

M. Tech*, B. Morgenstern and P. Meinicke

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

Received February 14, 2006; Revised March 6, 2006; Accepted April 11, 2006

ABSTRACT

Exact localization of the translation initiation sites (TIS) in prokaryotic genomes is difficult to achieve using conventional gene finders. We recently introduced the program TICO for postprocessing TIS predictions based on a completely unsupervised learning algorithm. The program can be utilized through our web interface at http://tico.gobics.de/ and it is also freely available as a commandline version for Linux and Windows. The latest version of our program provides a tool for visualization of the resulting TIS model. Although the underlying method is not based on any specific assumptions about characteristic sequence features of prokaryotic TIS the prediction rates of our tool are competitive on experimentally verified test data.

INTRODUCTION

The accuracy of translation initiation site (TIS) prediction in prokaryotes is still not sufficient. Though existing gene finders may reliably identify coding regions of significant length, they usually show a poor performance in predicting the correct TISs (1–3). For genomes with a high G+C content in particular, the rate of inaccurate TIS predictions is usually high (3). Recently, several postprocessors have been proposed (1,2,4,5) to improve the accuracy of TIS prediction in prokaryotic genomes.

Our tool TICO (for ‘TIs COrrector’) is based on an unsupervised learning scheme for relocation of putative gene starts. The underlying TIS model is very general and does not include any specific assumptions about TIS-related sequence features. In particular, no a priori assumptions about Shine–Dalgarno (SD) motifs (6) and their positional and compositional variation are required. In addition, we avoided any kind of empirical thresholds which might also imply a severe bias towards certain genomes.

Despite the generality of the implemented method, the prediction performance of our program is competitive (7), with good results even on high-G+C genomes. The program is supplemented with a tool for visualizing relevant sequence features which have been learned by the algorithm. This extension provides an effective means for monitoring the resulting model and may also reveal unknown sequence characteristics associated with prokaryotic TISs.

OUTLINE OF THE IMPLEMENTED ALGORITHM

The algorithm implemented in the tool TICO is based on a constrained clustering scheme which we described in detail in (7). The clustering starts with an initial gene annotation given as input. Within a specified *search range* around each annotated start candidate, all possible gene starts are defined as additional TIS candidates. These candidates have to share the reading frame of the annotated open reading frame (ORF) without any in-frame stop codon occurring between a potential start and the annotated stop. At the start of the clustering, the initially annotated gene starts are labelled as *strong* TISs; all other candidates are labelled as *weak* TISs. Based on that labelling, a positional weight matrix (PWM) is estimated from position-dependent trinucleotide frequencies. Positional smoothing is applied to the estimated probabilities in order to prevent vanishing entries. Finally, logarithms of strong and weak probabilities are subtracted to build the PWM. Then the PWM is used to score the candidates and, in turn, the score is used to reassign the candidates according to strong and weak TIS categories. Among all candidates associated with an ORF, the candidate with the maximum positive score is labelled strong; all other candidates of that ORF, are labelled weak. Estimation of the PWM is repeated until the labels no longer change or a maximum of 20 iterations has been reached.

As the clustering algorithm requires a suitable initialization, the resulting prediction to some degree depends on the prior annotation of TIS locations used for the initial labelling. In comparison with other tools our algorithm has proven to be

*To whom correspondence should be addressed. Tel: +49 551 3914927; Fax: +49 551 3914929; E-mail: maike@gobics.de rather robust with respect to low accuracy of the initial annotation (7). Nevertheless, the quality of the initial annotation can be too bad to serve as an appropriate starting point for our algorithm. Also, in cases where no TIS-related signals are actually present in the sequence, our algorithm is unlikely to improve the prediction.

DESCRIPTION OF THE TOOL

TICO can be accessed through our web interface and it is also available as commandline tool for Linux and Windows. The tool requires the input of a genome sequence in the FASTA format and a valid gene annotation, as obtained, for example, using the tool GLIMMER (8). The current version accepts two kinds of input format: GLIMMER2.x and our own format called 'simple coord'. In the download section, several Perl scripts are available for conversion of other formats (e.g. the PTT format provided by GenBank) into the TICO input format. The output of the tool is provided in a GLIMMER-like format, in our own format and in general feature format (GFF), according to the specifications of the Sanger Institute (http://www.sanger.ac.uk/). All formats contain the gene coordinates as predicted by TICO and a score which indicates the quality of the predicted TIS, i.e. a measure of fit with respect to the model resulting from the clustering algorithm. In GFF and the GLIMMER-like format the relocations of the TIS are also represented by means of the distance between the initially annotated position from the input and the new position for the TIS as predicted by TICO.

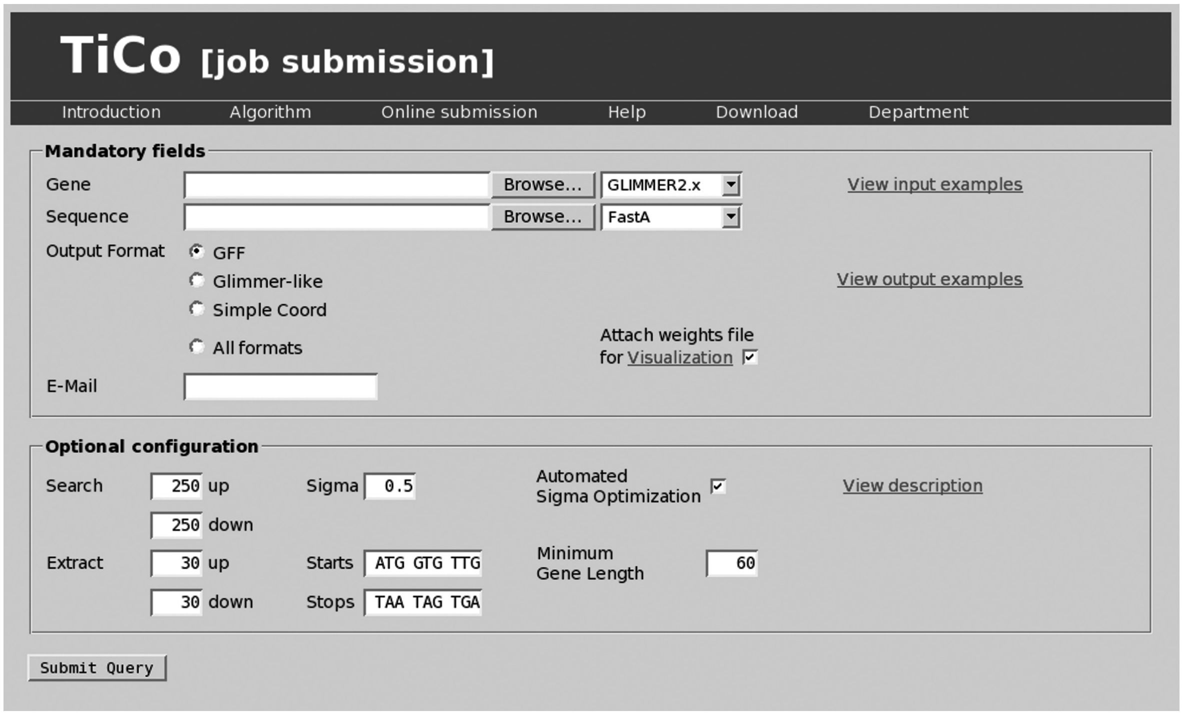
The submission page of the web interface of TICO is shown in Figure 1. The user should upload the sequence file and the initial annotation. In addition, one or several output formats have to be selected and an email address has to be provided to receive the results. The commandline version is applicable to annotation pipelines and it is executable on a common desktop PC. The user interface has been implemented in Java and can be configured by means of a properties file or via commandline parameters.

Since the first version of TICO was published (9) a number of extensions have been added. The most important extensions include an automatic adaptation of the smoothing parameter *sigma* (7) and the visualization of the PWM features, which is detailed in the next section.

VISUALIZATION OF LEARNED CHARACTERISTICS

The latest version of TICO provides a companion tool for visualization of the discriminative features learned for the classification of TIS candidates. The basis for the visualization is provided by the PWM which results from the clustering algorithm. The corresponding weights specify how trinucleotide occurrences within the sequence window contribute to the score used for labelling of TIS candidates. If a certain trinucleotide occurrence within that window is associated with a high positive weight, the occurrence provides evidence for a strong TIS. Similarly, a high negative weight provides evidence for a weak TIS.

In the current implementation, the visualization of learned characteristics is realized by transforming the PWM into a colour image (Figure 2). In the resulting PWM image, colour represents the weight value associated with a certain



Mandatory fields

Gene GLIMMER2.x

Sequence FastA

Output Format
● GFF
○ Glimmer-like
○ Simple Coord
○ All formats

Attach weights file for Visualization ☑

E-Mail

Optional configuration

Search 250 up
250 down

Sigma 0.5

Automated Sigma Optimization ☑

Extract 30 up
30 down

Starts ATG GTG TTG

Stops TAA TAG TGA

Minimum Gene Length 60

Figure 1. The submission page of the TICO web interface. The sequence and an annotation have to be uploaded. The user may choose one or more output formats and should give an email address to receive the results. In addition, some optional parameters can be adjusted, including the range of extracted sequences around the start candidates.

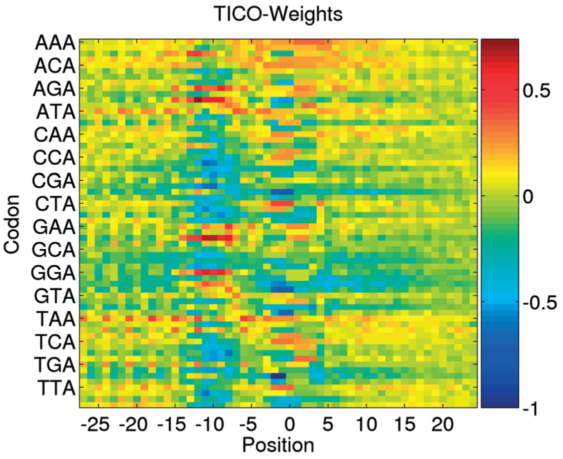


Figure 2. Visualization of the PWM calculated for *E.coli* using the default settings. On the horizontal axis the position within the sequence window is denoted; on the vertical axis trinucleotides are denoted in alphabetical order. The potential start codon is located at position 0. The upstream and downstream regions are represented by negative and positive position indices, respectively. A colour scale indicates the numerical values of the weights.

trinucleotide (row) at a specific position (column). High positive and high negative weights result in red and deep blue spots, respectively. Intermediate weights produce orange, yellow or green areas in the image. A colour scale at the right of the PWM image displays the colours associated with certain weight values. To obtain more precise information about the weights, the numerical value can be accessed by clicking on the corresponding image location. In that context, a zoom function may be used for close inspection of the weights.

By means of the visualization of the PWM, the resulting TIS model can be interpreted by the user. In combination with biological background knowledge, the position-dependent trimer weights can provide information about relevant TIS signals. Figure 2, for example, shows a visualization of the weights calculated for *Escherichia coli*. The Shine–Dalgarno sequence of *E.coli* has been proposed to contain the pattern AGGAG (6) ∼12 to 4 nt upstream of the translation start (10). In accordance with these findings, the image shows high positive weights for the trimers AGG, GAG, AGA and GGA at positions −12 to −7. Also, a slight positive maximum for the trimer AAA immediately following the start codon is observable. Such a triple-A downstream box has been proposed to provide another TIS-related signal (11), which has been confirmed in (12).

We should point out that the visualization of PWMs is only useful if strong TIS-related signals are actually present in the data. If statistically significant signals cannot be found in the sequence, no characteristic features will be observable. However, weak signals with bad visibility in the PWM image do not automatically imply a bad prediction of TIS locations. Although for several high-G+C genomes the resulting PWM images show a bad signal-to-noise ratio, in many cases the prediction of TIS can nevertheless be improved considerably using our algorithm (7).

PERFORMANCE

The performance of TICO has been tested on the genomes of *E.coli* and *Bacillus subtilis*, as well as on the high-G+C genomes of *Pseudomonas aeruginosa*, *Burkholderia pseudomallei* and *Ralstonia solanacearum*. Using these genomes is reasonable because for the corresponding organisms a reliable gene annotation is available. The results in comparison with other recent tools for improvement of TIS prediction can be found at http://tico.gobics.de/results.jsp. A detailed comparison of our tool with the tools RBSfinder (4), MED-Start (2) and GS-Finder (1) has been published in (7).

ACKNOWLEDGEMENTS

The work was partially supported by the Bundesministerium für Buildung und Forschung (BMBF) project MediGrid (01AK803G). Funding to pay the Open Access publication charges for this article was provided by the annual budget of the authors' department.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Zhu,H.-Q., Hu,G.-Q., Ouyang,Z.-Q., Wang,J. and She,Z.-S. (2004) Accuracy improvement for identifying translation initiation sites in microbial genomes. *Bioinformatics*, **20**, 3308–3317.
3. Tech,M. and Merkl,R. (2003) YACOP: enhanced gene prediction obtained by a combination of existing methods. *In Silico Biol*., **3**, 441–451.
4. 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.
5. 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 Acides Res*., **31**, 1780–1789.
6. Shine,J. and Dalgarno,L. (1974) The 3′ terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementary to nonsense triplets and ribosome binding sites. *Proc. Natl Acad. Sci. USA*, **71**, 1342–1346.
7. Tech,M. and Meinicke,P. (2006) An unsupervised classification scheme for improving predictions of prokaryotic TIS. *BMC Bioinformatics*, **7**, 121.
8. Delcher,L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res*., **27**, 4636–4641.
9. Tech,M., Pfeifer,N., Morgenstern,B. and Meinicke,P. (2005) TICO: a tool for improving predictions of prokaryotic translation initiation sites. *Bioinformatics*, **21**, 3568–3569.
10. Draper,D.E. (1996) *Escherichia coli and Salmonella, Volume I: Translational Initiation, 2nd edn*. ASM Press, Washington, DC, pp. 902–908.
11. Sato,T., Terabe,M., Watanabe,H., Gojobori,T., Hori-Takemoto,C. and Miura,K. (2001) Codon and base biases after the initiation codon of the open reading frames in the *Escherichia coli* genome and their influence on the translation efficiency. *J. Biochem*., **129**, 851–860.
12. Meinicke,P., Tech,M., Morgenstern,B. and Merkl,R. (2004) Oligo kernels for datamining on biological sequences: a case study on prokaryotic translation initiation sites. *BMC Bioinformatics*, **5**, 169.