Improving Prediction of Translation Initiation Sites in Prokaryotic Genomes with TICO

gobics.de
Göttingen Bioinformatics

Maike Tech, Nico Pfeifer, Burkhard Morgenstern and Peter Meinicke

Department of Bioinformatics, Institute for Microbiology and Genetics (IMG), University of Göttingen

Introduction

Gene prediction in prokaryotes is still not completely solved:

- Annotation of translation starts is difficult, in most cases ORFs provide many candidates for a potential start site
- Large number of false positives implies much work for the annotators
- Insufficient results for *heterogeneous* genomes (e. g. genomes with pathogenic islands)

```
    GAGAAGCCACAAAAAATGAATGTTAATTACCTGATGCA
GGACTGGATATGCTGATTCTTATTTCACCTGAATGCGCTTAT
ACGGAGACGTTTAGATGGGTAAAATAATTGGTCGACCTGGGT
CGATATGATTTAAGTCGTGCCGATGAATTACTCGATAACTGG
TCACCTGAAAGAGAAAGTAAAGGAAGTGAAATCTGGA
```

Clustering Algorithm

Initialization:
Search for TIS candidates with the initial annotation. Classification: The initial TIS is labelled *strong*, all other candidates from that ORF are labelled *weak*.

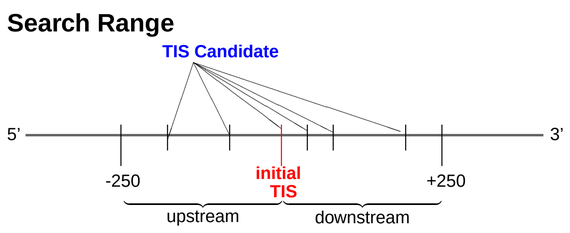
Iterative Optimization:

1. Estimation of the smoothed trinucleotide probabilities for strong and weak TIS models according to the current classification.
2. Calculation of a Position Weight Matrix (PWM)
3. Calculation of a PWM score for each candidate
4. (Re-)Classification: Updating the labels

Break condition:
Labels do not change in classification.

Extraction of ORF-specific TIS candidates

All TIS candidates of an ORF have to share same reading frame and no stop codon has to occur between a candidate and the annotated stop.



Calculation of the PWM

Smoothing is realized by (matrix) multiplication of trinucleotide probability matrix $\mathbf{P}$ with the smoothing matrix $\mathbf{S}$.

$$\widetilde{\mathbf{P}} = \mathbf{P} \cdot \mathbf{S}$$
$$\mathbf{W} = \log \widetilde{\mathbf{P}}_{strong} - \log \widetilde{\mathbf{P}}_{weak}$$

The Position Weight Matrix $\mathbf{W} \in \mathbb{R}^{4^K \times L}$ is calculated from the difference of the elementwise logarithms of the probabilities of *strong* candidates and *weak* candidates.

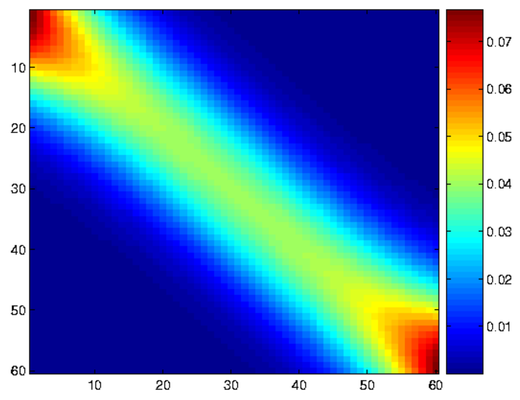
Smoothing

The smoothing matrix $\mathbf{S} \in \mathbb{R}^{L \times L}$ is calculated with a normalized gaussian function:

$$s_{ij} = \frac{e^{\left(-\frac{1}{2\sigma^2}(i-j)^2\right)}}{\sum_k e^{\left(-\frac{1}{2\sigma^2}(k-j)^2\right)}},$$

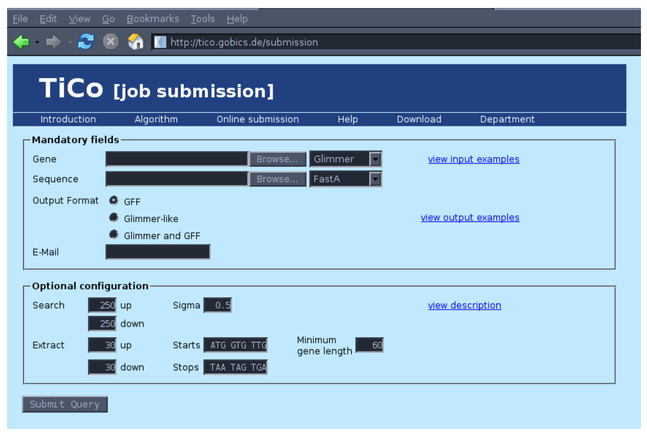
with
$i, j, k \in \{1, \ldots L\}$

Visualization of the smoothing matrix



Implementation

TICO can be accessed via the web at http://tico.gobics.de/ or can be downloaded as a standalone tool for Linux and Windows.



Results

Comparison of TICO's performance with other post processors. The initial prediction was obtained from GLIMMER 2.02 (DELCHER *et al.*). Despite of the generality of our approach, TICO achieves the best results even on the high G+C-genome of *P. aeruginosa*.

		Post processors			
Data	GLIMMER	GS-f.	MED	RBSf.	TICO
EcoGene	63.2	90.3	92.0	81.9	**94.3**
Bsub	61.3	87.9	89.2	78.5	**89.4**
PseudoCAP	57.8	83.6	3.6	67.7	**84.7**

Notation:
% correctly predicted TIS as compared to the reference dataset.

GS-f.	GS-finder (OU *et al.*)
MED	MED-Start (ZHU *et al.*)
RBSf	RBSfinder (SUZEK *et al.*)
EcoGene	854 genes from *E. coli* with verified N-termini (RUDD)
Bsub	1248 *non-y* genes from the genome of *B. subtilis*
PseudoCAP	3281 genes with an annotated function from *P. aeruginosa*

Outlook

- Evaluation of TIS prediction with TICO on *heterogeneous* genomes
- Reduce false positives in predictions through the PWM-score calucated by TICO
- Add our own ORF-finder which obviates the dependence on an initial GLIMMER prediction

References

DELCHER *et al.*, Improving microbial gene identification with GLIMMER. *Nucleic Acids Res.* 1999, 27(23):4636–4641.

MEINICKE *et al.*, Oligo Kernels for datamining on biological sequences: A case study on prokaryotic translation initiation sites. *BMC Bioinformatics*, 2004, 5(169).

OU *et al.*, GS-finder: A program to find bacterial gene start sites with a self-training method. *The International Journal of Biochemistry & Cell Biology*, 2004, 36(3):535–544.

PseudoCAP: Pseudomonas aeruginosa Community annotation project, http://pseudomonas.com/

RUDD, EcoGene: A genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, 2000, 28:60–64.

SUZEK *et al.*, A probabilistic method for identifying start codons in bacterial genomes. *Bioinformatics*, 2001, 17(12):1123–1130.

TECH *et al.*, TICO: A tool for improving predictions of prokaryotic translation initiation sites. *Bioinformatics*, 2005, 21(17):3568–3569.

ZHU *et al.*, Accuracy improvement for identifying translation initiation sites. *Bioinformatics*, 2004, 20(18):3308–3317.