Prediction of Translation Initiation Sites in Prokaryotic Genomes with TICO

Maike Tech

Abt. Bioinformatik
Institut für Mikrobiologie und Genetik (IMG)
Universität Göttingen

September 24, 2005

Project TICO

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

Maike Tech, Nico Pfeifer, Burkhard Morgenstern und Peter Meinicke,
Bioinformatics, 2005

Problems in prokaroptic gene prediction

- Annotation of translation starts is difficult
- Lage number of False Positives
- Insufficient results for *heterogeneous* genomes

Algorithm of TICO

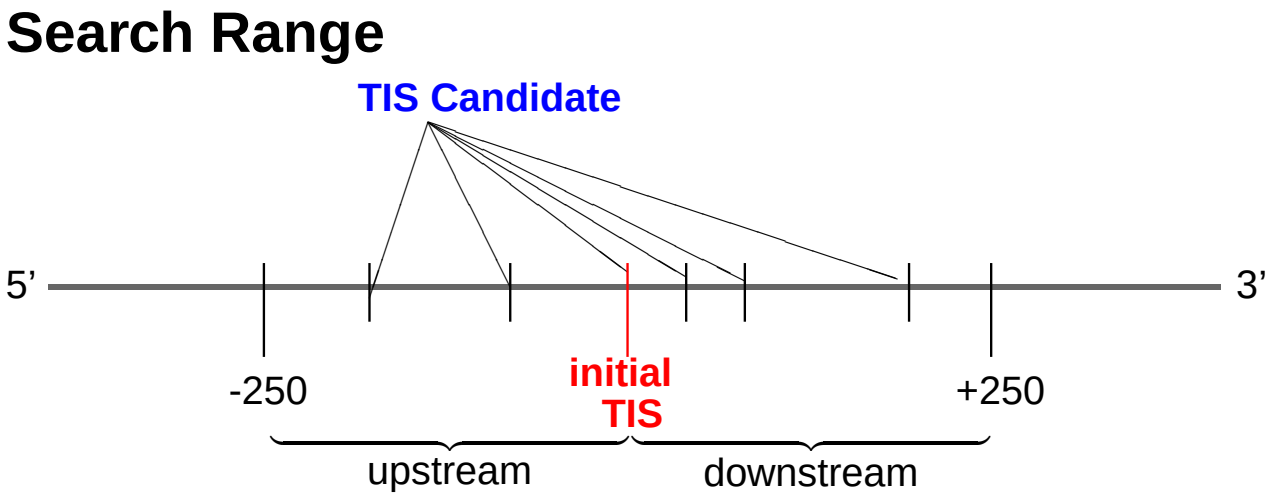
Initialization: Search for TIS candidates with the initial annotation.

Iterative Optimization:

1. Estimation of the smoothed trinucleotide probabilities for all candidates
2. Calculation of a Position Weight Matrix (PWM)
3. Calculation of a score for each candidate
4. Reallocation of the labels

Break condition: No alteration of the labels during classification.

TIS candidates



Algorithm of TICO

Initialization: Search for TIS candidates with the initial annotation.

Iterative Optimization:

1. Estimation of the smoothed trinucleotide probabilities for all candidates
2. Calculation of a Position Weight Matrix (PWM)
3. Calculation of a score for each candidate
4. Reallocation of the labels

Break condition: No alteration of the labels during classification.

Calculation of the PWM

The smoothing is realized as simple matrix product with a smoothing matix **S**.

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

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.

Calculation of the smoothing matrix

The smoothing matrix $\mathbf{S} \in \mathbb{R}^{L \times L}$ is calculated as

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

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

Algorithm of TICO

Initialization: Search for TIS candidates with the initial annotation.

Iterative Optimization:

1. Estimation of the smoothed trinucleotide probabilities for all candidates
2. Calculation of a Position Weight Matrix (PWM)
3. Calculation of a score for each candidate
4. Reallocation of the labels

Break condition: No alteration of the labels during classification.

Algorithm of TICO

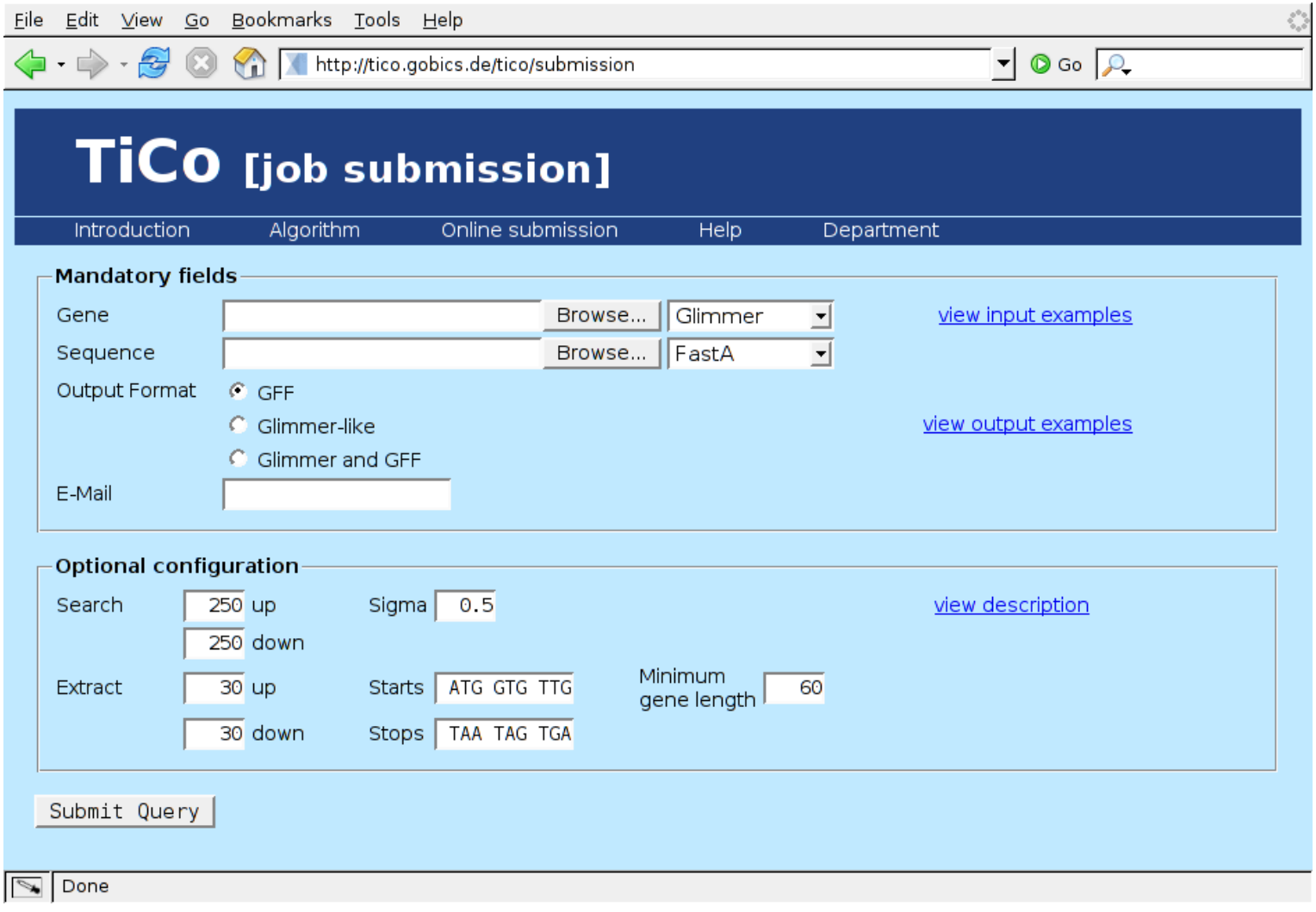
Initialization: Search for TIS candidates with the initial annotation.

Iterative Optimization:

1. Estimation of the smoothed trinucleotide probabilities for all candidates
2. Calculation of a Position Weight Matrix (PWM)
3. Calculation of a score for each candidate
4. Reallocation of the labels

Break condition: No alteration of the labels during classification.

Web-Interface: http://tico.gobics.de/



Comparison of TICO's performance with other post processors

Data		GLIMMER	Post processors			
			GS-f.	MED	RBSf.	TICO
EcoGene	854	63.2	90.3	92.0	81.9	**94.3**
Bsub	1248	61.3	87.9	89.2	78.5	**89.4**
PseudoCAP	3281	57.8	83.6	3.6	67.7	**84.7**

Notation: % correct predicted TIS

Outlook TICO

- Evaluation of TIS predicition with TICO on *heterogeneous* genomes
- Reduce False Positives in predictions through the score calucated by TICO
- Add our own ORF-finder based on a clustering algorithm

Acknowledgements

Göttingen Genomics Laboratory

Elzbieta Brzuszkiewicz, Florian Fricke
Frank Hoster, Heiko Liesegang,
Axel Strittmatter

Universität Regensburg

Inst. für Biophysik und
physikalische Biochemie
Rainer Merkl

Universität Göttingen

Institut für Informatik
Stephan Waack, Carsten Damm, Oliver Keller
Thomas Brodag, Katharina Surovcik

Inst. Mikrobiologie und Genetik
Burkhard Morgenstern,
Peter Meinicke, Nico Pfeifer