TSSAR: Transcription Start Site Annotation Regime for dRNA-seq data

Fabian Amman[1], Michael T. Wolfinger[2,3,4], Ivo L. Hofacker[2,5,9], Peter F. Stadler[1,2,5,6,7,8] and Sven Findeiß[2,9]

[1] Bioinformatics Group, Department of Computer Science and the Interdisciplinary Center for Bioinformatic, University of Leipzig, Härtelstr. 16-18, 04107 Leipzig, Germany. [2] Institute for Theoretical Chemistry, University of Vienna, Währingerstr. 17, A-1090 Vienna, Austria. [3] Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories , University of Vienna & Faculty of Computer Science, University of Vienna, Dr. Bohr-Gasse 9, A-1030 Vienna, Austria. [4] Department of Biochemistry and Molecular Cell Biology, Max F. Perutz Laboratories, University of Vienna, Dr. Bohr-Gasse 9, A-1030 Vienna, Austria. [5] Center for RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, Frederiksberg C, Denmark. [6] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany. [7] Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany. [8] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501. [9] Research group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Währingerstr. 29, A-1090 Vienna, Austria.

Introduction

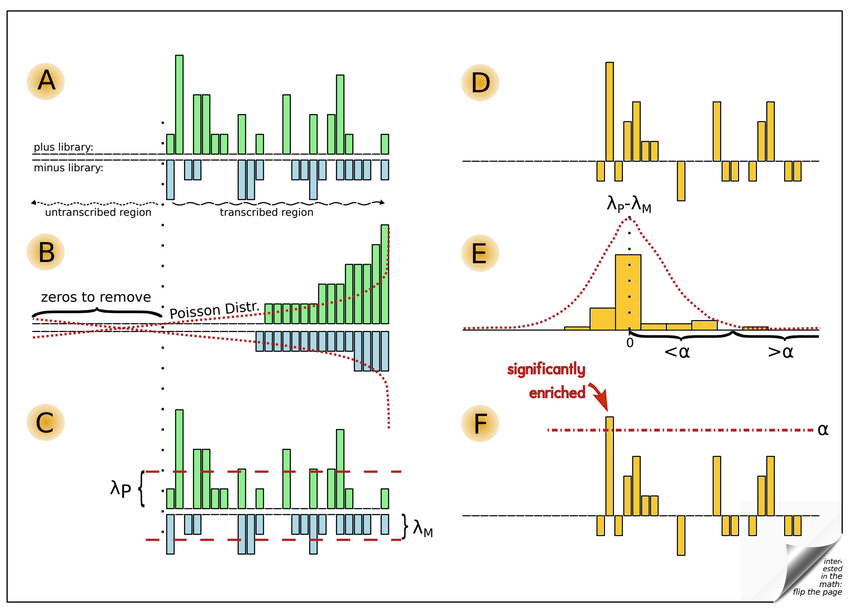
To fully comprehend a bacterial cell the when and where of transcription initiation is of pivotal interest. The first informs what mRNAs exist at a given time, thus are potentially translated into effector proteins. The later describes how the substrate of translation and post transcriptional gene regulation looks like. There is a variety of different techniques to find the exact position of a transcription start site (TSS). But only a few are able to screen whole genomes for TSS in a high-throughput manner. One of these methods is dRNA-seq [1], which works by enriching primary transcription starts in a TEX treated library compared to an untreated library. TEX specifically degrades RNA fragments which are not protected by a triphosphate at its 5' end, a characteristic of RNA fragments originating from primary transcription starts. Since the depletion is not infallible, not every signal represents an original TSS. Hence, a statistical analysis of the read counts in the treated versus the untreated library has to be performed [2]. Therefore we developed TSSAR, a Transcription Start Site Annotation Regime, with the intention to set the interpretation of dRNA-seq data on a sound statistical basis combined with a user friendly interface.

Method

To account for the different transcription dynamics in the genome, each site is evaluated in the context of its local surrounding by a sliding window approach.

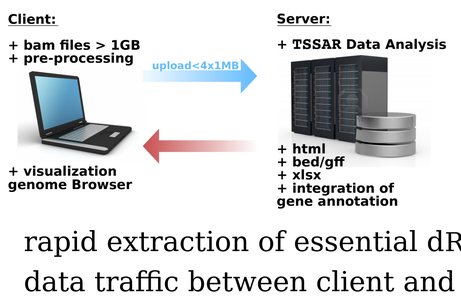
Background Modeling

An arbitrary region in the genome might be a mixture of transcribed and not transcribed sections. For the first, read start counts can be described by a Poisson distributed random variable, the later is expected to be uniformly zero distributed A. To estimate the parameters which describe only the Poisson part, TSSAR applies a zero-inflated Poisson model regression [3]. All excess zeros are believed to originate from untranslated regions and are removed from the sample B. Finally, the mean value λ of the remaining sample is calculated, describing the background distribution of the transcribed part of the considered window C.



TSS Annotation

TSSAR aims for finding positions with a significantly enriched signal in the TEX treated library, considering the expected variability from the background model. Thereto, the read start count difference between treated and untreated library for each position is considered D. The derived sample of differences follows a Skellam distribution [4]. The distribution's shape and position is characterized by the prior deduced λ parameters. Regarding the whole sample, each value can be evaluated how well it fits the model E. Given a p-value cutoff, a minimal difference α can be deduced above which all positions are annotated as TSS F.



Architecture

TSSAR is available both in a stand alone version and as a RESTful Web Service. Client-side pre-processing by means of a platform-independent client application allows for rapid extraction of essential dRNA-seq input (mapped reads) and avoids huge data traffic between client and server. The statistical TSSAR model is then applied to the data on the server. Predicted TSS are available for screening in modern Web browsers and can be downloaded in various file formats.

Analysis and post-processing

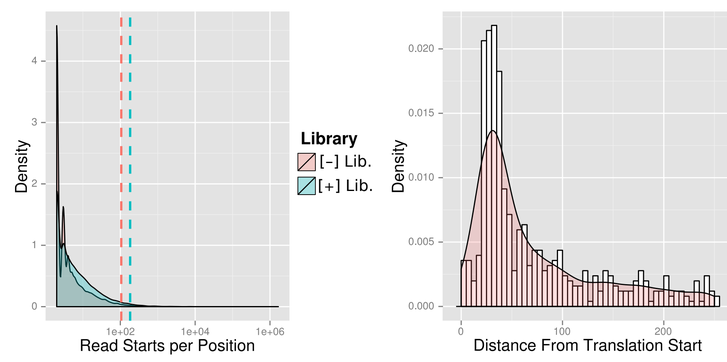
chrom	start	stop	name	score	strand
chr	42064	42065	TSS_000097	34	-
chr	27532	27533	TSS_000025	30	+
chr	54524	54525	TSS_000038	10	+
chr	97138	97139	TSS_000150	13	-
chr	43183	43184	TSS_000036	28	+

annotated TSS in BED format

position	strand	id	score	class	comment
42065	-	TSS_000097	34	Ad	antisense to gene HP0043 (3nt downstream)
27533	+	TSS_000025	30	P	24nt upstream of gene HP0027
54525	+	TSS_000038	10	Ai	antisense to gene HP0054
97139	-	TSS_000150	13	O	-
43184	+	TSS_000036	28	IP	within gene HP0044; 59nt upstream of HP0045

TSS classification related to gene annotation

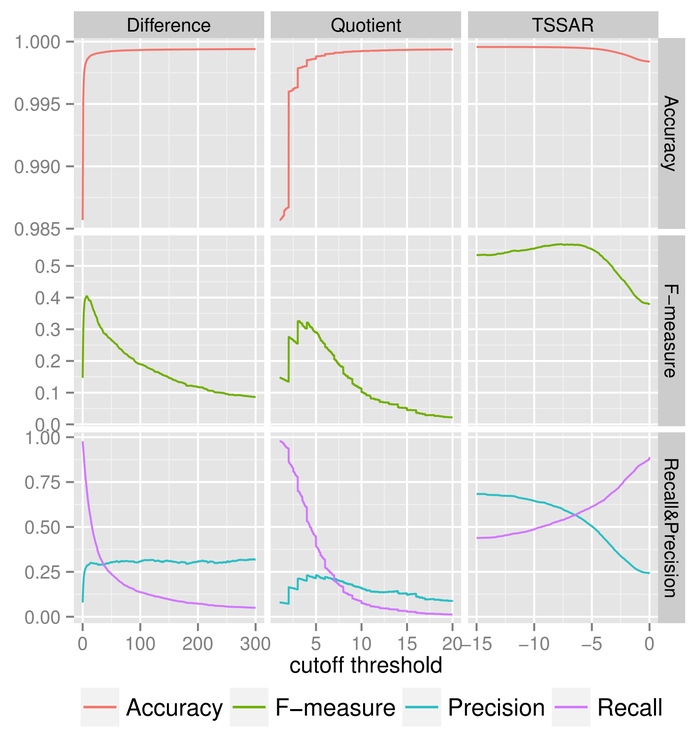
TSSAR's main output lists significantly enriched positions. In addition, consecutive TSS are clustered together to the most prominent signal. If the reference genome's annotation is provided, TSSAR uses this information to classify each annotated TSS according to its genomic context.



Quality control of dRNA-seq data *Analysis of annotated TSS in their genomic context*

Evaluation

To assess the performance of TSSAR we used a published *H. pylori* dRNA-seq data set [1]. Our approach was compared to two basic approaches. There, the TSS annotation was done based on the simple classifier '*Difference*' and '*Quotient*' between read start counts in the treated and untreated library. For all methods the results were compared to the manual annotation from [1]. To quantify the performance, recall, precision, accuracy and F_1-measure were calculated. TSSAR shows a higher precision and simultaneously a less sharp drop of the recall rate. Hence, in terms of the F_1-measure, it excels the basic approaches.



Discussion and Conclusion

TSSAR provides several advantages over previous dRNA-seq interpretations. Among others, bias from prior notion is eliminated, the analysis is automated (or semi-automated, since manually inspection is still highly advised) which reduces time and effort, and enables to shift resources from technical issues to focus more on biological questions.

TSSAR is available online at **http://rna.tbi.univie.ac.at/TSSAR**.

References

[1] Sharma et al. (*Nature*; 2010)
[2] Schmidtke et al. (*Nucleic acids research*; 2012)
[3] Yee (*Journal of Statistical Software*; 2010)
[4] Skellam (*Journal of the Royal Statistical Society*; 1946)

This work was partly funded by:
Bundesministerium für Bildung und Forschung
ebio:RNAsys an Initiative by the BMBF
RNA REG
FWF Der Wissenschaftsfonds.
Gedruckt im Universitätsrechenzentrum Leipzig
UNIVERSITÄT LEIPZIG
universität wien