TALEs of virulence and biotechnology

Jan Grau[1], Annett Wolf[1], Maik Reschke[2], Ulla Bonas[2], Stefan Posch[1], and Jens Boch[2]

[1] *Institute of Computer Science, Martin Luther University Halle–Wittenberg*

[2] *Department of Genetics, Institute of Biology, Martin Luther University Halle–Wittenberg*

grau@informatik.uni-halle.de

Abstract: Transcription activator-like effectors (TALEs) are injected into host plant cells by *Xanthomonas* pathogens to function as transcriptional activators. Their DNA-binding domain is composed of conserved amino acid repeats containing repeat-variable diresidues (RVDs) that determine DNA binding specificity.

We present TALgetter, a new approach for predicting TALE target sites based on a statistical model. The predictions of TALgetter indicate a previously unreported positional preference of TALE target sites relative to the transcription start site. In addition, several TALEs are predicted to bind to the TATA-box, which might constitute one general mode of transcriptional activation by TALEs.

1 Introduction

Transcription activator-like effectors (TALEs) are injected into the cells of host plants by plant-pathogenic *Xanthomonas* bacteria, where they act as transcription factors for the benefit of the pathogen [B$^+$09]. The hosts of different *Xanthomonas* strains span a variety of important crop plants including rice, sweet orange, tomato, pepper, and cabbage.

The DNA-binding domain of TALEs is composed of highly conserved tandem repeats, where each repeat usually spans 34 amino acids. These repeats bind to the nucleotides of a DNA target site in a contiguous, non-overlapping fashion. The DNA-specificity of an individual repeat depends on the two amino acids at position 12 and 13, termed repeat-variable diresidues (RVDs) [B$^+$09, MB09].

The computational prediction of virulence targets of natural TALEs is a key step to provide candidates for subsequent experimental validation. However, less than 30 virulence targets have been validated, often only one for an individual TALE. This set is complemented by a few hundred artificial target sites from reporter assays. Hence, novel approaches are required that i) make use of the available data in a holistic manner, and ii) allow for predicting target sites of TALEs without any known target site. In [GWR+13], we propose such an approach for predicting TALE target sites called *TALgetter*.

2 Computational predictions provide insights into the biology of TALE target sites

In [GWR+13], we propose TALgetter, which uses a new statistical model representing *importance* of RVDs and their *binding specificity* independently in a local mixture model. The concept of importance is related to the *efficiency* of RVDs [S+12], but additionally affects the penalty for non-matching nucleotides. In the proposed model, the importance and binding specificity of an RVD are independent of its context in the TALE. For details of the statistical model, we refer to [GWR+13].

In contrast to previous approaches, the parameters of this model are estimated computationally, where different TALEs and their known target sites can be combined in a common training set due to independence assumptions. TALgetter is part of version 2.1 of the open-source Java library Jstacs [G+12].

In [GWR+13], we show that TALgetter yields an improved prediction performance compared to the existing approach, *Target Finder* of the TALE-NT suite [D+12]. Using TALgetter, we predict target sites of *Xanthomonas* TALEs in the important crop plants rice and sweet orange. These predictions elucidate novel putative virulence targets of several TALEs (c.f. Tables 6 to 8 of [GWR+13]).

In addition, we demonstrate that computational approaches are able to gain new insights into the biology of TALE targeting. Specifically, we combine predictions of TALgetter with gene expression data to identify functional TALE target sites. We find that functional target sites are preferentially located in a region from 300 bp upstream to 200 bp downstream of the transcription start (c.f. Figure 7 of [GWR+13]). Our predictions also indicate that many TALEs bind to the TATA-box in the promoters of their target genes. Based on these observations, we propose four biological models (c.f. Figure S7 of [GWR+13]) that may explain the apparent target site preference of TALEs.

The modular architecture of TALEs allows for a rearrangement of repeats to easily generate any desired DNA-specificity. Hence, TALEs have become a preferred biotechnology tool for targeted DNA binding. Although TALgetter has been created for predicting virulence targets of natural TALEs, it is readily applicable to biotechnology problems as, for instance, the off-target prediction of artificial TALE activators.

TALEs are also the basis of TALE nucleases (TALENs), which have been established as a second genome-editing technique besides zinc-finger nucleases [GGB13]. In TALENs, the DNA-binding domain of TALEs is fused with a Fok1 endonuclease domain, where homo- or hetero-dimers of TALENs specifically cut the DNA double strand. Although TALENs cut DNA highly specific, undesired *off-targets* in addition to the targeted genomic region remain an important issue [O$^+$13]. Recently [GBP], we developed a novel tool for the genome-wide prediction of TALEN off-targets, named *TALENoffer*. TALENoffer is based on the same statistical model as TALgetter, and features an optimized runtime to scan complete genomes for TALEN off-target sites within a few minutes.

3 Availability

Web-applications of TALgetter and TALENoffer are available at `http://galaxy.informatik.uni-halle.de`, and can also be installed in a local Galaxy [B$^+$10] server. In addition, we provide command line version of TALgetter and TALENoffer at `http://jstacs.de/index.php/TALgetter` and `http://jstacs.de/index.php/TALENoffer`, respectively. TALgetter also allows users to estimate new model parameters from custom training data. Hence, users can adapt the parameters of the TALgetter model to improved sets of validated TALE target sites, which are to be expected in the near future.

4 Talk outline

We start our talk with an introduction to TALEs and the specific bioinformatics problems that arise in the prediction of TALE target sites. After a brief description of the statistical model of TALgetter, we focus on the biological findings that have been discovered using our computational predictions, namely the previously unreported target site preferences of TALEs and biological models explaining these. We finally succinctly introduce TALENoffer for predicting off-target sites of TALE nucleases.

References

[B$^+$09] Jens Boch et al. Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. *Science*, 326(5959):1509–1512, 2009.

[B$^+$10] Daniel Blankenberg et al. *Galaxy: A Web-Based Genome Analysis Tool for Experimentalists.* John Wiley & Sons, Inc., 2010.

[D$^+$12] Erin L. Doyle et al. TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Research*, 40(W1):W117–W122, 2012.

[G$^+$12] Jan Grau et al. Jstacs: A Java Framework for Statistical Analysis and Classification of Biological Sequences. *Journal of Machine Learning Research*, 13(Jun):1967–1971, 2012.

[GBP] Jan Grau, Jens Boch, and Stefan Posch. TALENoffer: genome-wide TALEN off-target prediction. *Under review.*

[GGB13] Thomas Gaj, Charles A. Gersbach, and Carlos F. Barbas. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in biotechnology, doi: 10.1016/j.tibtech.2013.04.004*, 05 2013.

[GWR$^+$13] Jan Grau, Annett Wolf, Maik Reschke, Ulla Bonas, Stefan Posch, and Jens Boch. Computational Predictions Provide Insights into the Biology of TAL Effector Target Sites. *PLOS Computational Biology*, 9(3):e1002962, 03 2013.

[MB09] Matthew J. Moscou and Adam J. Bogdanove. A Simple Cipher Governs DNA Recognition by TAL Effectors. *Science*, 326(5959):1501, 2009.

[O$^+$13] Mark J Osborn et al. TALEN-based Gene Correction for Epidermolysis Bullosa. *Molecular Therapy, doi:10.1038/mt.2013.56*, 04 2013.

[S$^+$12] Jana Streubel et al. TAL effector RVD specificities and efficiencies. *Nature Biotechnology*, 30(7):593–595, 07 2012.