BIOINFORMATICS APPLICATIONS NOTE
Vol. 29 no. 22 2013, pages 2931–2932
doi:10.1093/bioinformatics/btt501

Sequence analysis
Advance Access publication August 30, 2013

TALENoffer: genome-wide TALEN off-target prediction

Jan Grau[1,*], Jens Boch[2] and Stefan Posch[1]

[1]Institute of Computer Science and [2]Department of Genetics, Institute of Biology, Martin Luther University Halle–Wittenberg, D-06099 Halle (Saale), Germany

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Transcription activator-like effector nucleases (TALENs) have become an accepted tool for targeted mutagenesis, but undesired *off-targets* remain an important issue. We present TALENoffer, a novel tool for the genome-wide prediction of TALEN off-targets. We show that TALENoffer successfully predicts known off-targets of engineered TALENs and yields a competitive runtime, scanning complete mammalian genomes within a few minutes.

Availability: TALENoffer is available as a command line program from http://www.jstacs.de/index.php/TALENoffer and as a Galaxy server at http://galaxy.informatik.uni-halle.de.

Contact: grau@informatik.uni-halle.de

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on May 31, 2013; revised on July 27, 2013; accepted on August 21, 2013

1 INTRODUCTION

The DNA binding domain of transcription activator-like (TAL) effectors is composed of highly conserved tandem repeats, where amino acids 12 and 13 of a repeat [repeat-variable diresidue, (RVD)] determine DNA binding specificity. Each repeat binds to 1 bp of the DNA in a contiguous non-overlapping fashion (Boch *et al*., 2009; Moscou and Bogdanove, 2009). Recently, we developed TALgetter, a tool for predicting putative targets of natural TAL effectors, and we showed that this tool achieves an improved prediction accuracy compared with previous approaches (Grau *et al*., 2013).

The DNA binding domain of TAL effectors can be fused with a Fok1 endonuclease domain to yield TAL effector nucleases (TALENs), where homo- or hetero-dimers of TALENs need to bind to opposite strands of the DNA in 5′–3′ orientation and in a restricted distance range to specifically cut the DNA double strand. TALENs have been established as a second genome-editing technique besides zinc-finger nucleases (Gaj *et al*., 2013; Miller *et al*., 2011). Although the binding of TALENs is highly specific, undesired *off-targets* in addition to the targeted genomic region remain an important issue (Hockemeyer *et al*., 2011; Mussolino *et al*., 2011; Osborn *et al*., 2013; Tesson *et al*., 2011) that may cause severe side effects. Hence, tools for the computational prediction of TALEN off-targets have been developed, namely, idTALE (http://idtale.kaust.edu.sa) and Paired Target Finder (PTF) [https://tale-nt.cac.cornell.edu, Doyle *et al.* (2012)]. Here, we only consider PTF because the ‘Search for TALEN target’ application of idTALE is not applicable to custom input data.

In this article, we present TALENoffer, an alternative tool for predicting TALEN off-targets. TALENoffer applies the statistical model of TALgetter to the more complex problem of TALEN off-target prediction. This requires novel methods for ranking off-targets and accelerated scanning approaches to achieve acceptable runtimes, which are explained in the following section.

2 METHODS

2.1 Statistical model

The statistical model of TALENoffer assumes that the probability of a nucleotide of a target site depends on the RVD of the corresponding repeat. In addition, it reflects that different RVDs contribute differently to the activity of TAL effector constructs (Streubel *et al*., 2012) (details in Supplementary Methods). Given RVD sequence $\boldsymbol{y} = y_1, \ldots, y_L$ and model parameters $\boldsymbol{\lambda}$, this model assigns each putative monomer target site $\boldsymbol{x} = x_0, \ldots, x_L$ a likelihood $P(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\lambda})$. Based on the likelihood, we define a relative score $s(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\lambda}) := \frac{1}{L+1} \log P(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\lambda})$

2.2 Ranking and filtering off-targets

Given two TALEN monomers with RVD sequences $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ of length L_1 and L_2, respectively, a distance d between TALEN monomers and a putative off-target site $\boldsymbol{x} = x_0, \ldots, x_{L_1+d+L_2+1}$, we first determine the relative scores $s_1 = s(x_0, \ldots, x_{L_1}|\boldsymbol{y}_1, \boldsymbol{\lambda})$ and $s_2 = s(x^{\mathrm{c}}_{L_1+d+L_2+1}, \ldots, x^{\mathrm{c}}_{L_1+d+1}|\boldsymbol{y}_2, \boldsymbol{\lambda})$ of the two monomer target sites, where x^{c} denotes the complement of nucleotide x. We define the score s of the complete off-target site $\boldsymbol{x}$ as the sum $s = s_1 + s_2$ of the two relative scores. This scoring scheme allows for ranking off-target sites, given TALEN monomers of different lengths, although typically $L_1 = L_2$.

We report off-targets yielding a score s that exceeds a threshold $t = s^* + 2\log(q)$, where s^* denotes the score of the best-matching theoretical off-target site for the current pair of TALEN monomers. Parameter q specifies that the average likelihood over all positions of the off-target site shall be at least $q \cdot 100$ % of the average likelihood over all positions of the best-matching site. We additionally require each monomer score s_i to exceed a threshold $t_i = s_i^* + \log(q \cdot 0.9)$, where s_i^* denotes the best monomer score of $\boldsymbol{y}_i$. This allows for only mild compensation between the two monomers of a common off-target site. In the TALENoffer application, users may select pre-defined or enter custom values of q (see also Supplementary Fig. S2). In addition, we limit the ranked sites that are reported to a user-specified number. We consider homo- and hetero-dimers of the TALEN monomers and both DNA strands because all can lead to off-target effects.

2.3 Runtime optimization

Using a naive scanning approach for predicting TALEN off-targets, we presumably shift a sliding window of width $L_1 + 1$ along the DNA sequence and compute the score s_1, given the first TALEN monomer $\boldsymbol{y}_1$

*To whom correspondence should be addressed.

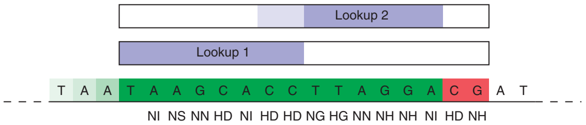
Lookup 2
Lookup 1
T A A T A A G C A C C T T A G G A C G A T
NI NS NN HD NI HD HD NG HG NN NH NH NI HD NH

Fig. 1. Speed-up strategy of TALENoffer. Two lookup tables of partial scores are represented by boxes, where the light shaded part of lookup table 2 serves as condition for computing the partial likelihood of the dark shaded part. The rightmost part of the putative off-target site only needs to be considered if both lookup tables indicate a score above the threshold

within this window. Whenever we find a hit, i.e. $s_1 \geq t_1$, we also scan the reverse complement of the downstream sequence for sufficiently good hits, given the second monomer $\boldsymbol{y}_2$ within the user-specified distance range.

The scanning approach of TALENoffer is based on this naive approach but uses a speed-up strategy, which is illustrated in Figure 1 for one TALEN monomer. Given the TALEN monomer, we compute a partial score for each possible 8mer prefix and store it in lookup table 1, together with the information whether a target site with this prefix may yield a sufficiently large total score s_i (details in Supplementary Methods). For lookup table 2, we proceed in complete analogy using partially overlapping 8mer infixes. Both lookup tables can be accessed efficiently, given the prefix and infix of a putative monomer target site. Scanning input sequences for off-target sites, we test whether both lookup tables indicate that the putative target site under the sliding window might exceed threshold t_1. If this is the case, we only need to compute the remaining score for the nucleotides beyond lookup Table 2 to yield the total score s_1. If s_1 exceeds the threshold, we apply the same strategy for putative target sites of the second monomer on the opposite strand within the user-specified distance range.

In addition to this speed-up strategy, TALENoffer is multithreaded to allow for simultaneously loading, parsing and scanning input data.

3 RESULTS

3.1 Finding known off-targets

For evaluating predictions, we use TALENs and reported off-targets from several recent studies (details in Supplementary Methods, Supplementary Table S2, Supplementary Fig. S4). For all of these datasets, the intended TALEN target is reported on rank 1 by TALENoffer and PTF. However, we observe differences between both tools for the predicted off-targets.

Tesson *et al.* (2011) designed a TALEN pair for targeting *IgM* in *Rattus norvegicus*. The off-target reported by Tesson *et al.* (2011) is predicted by TALENoffer on rank 2 and by PTF on rank 6.

Mussolino *et al.* (2011) targeted *CCR5* in human. The off-target *CCR2* is reported only by TALENoffer (rank 2) because of an atypical A at position 0 of one monomer target site not allowed by PTF.

Hockemeyer *et al.* (2011) targeted *PPP1R12C* in human and reported two off-targets, which are both predicted by TALENoffer (ranks 47 and 195), whereas PTF predicts only one of these off-targets (rank 159).

Osborn *et al.* (2013) targeted human *COL7A1* and report three off-targets. Off-target GGT1 is reported by TALENoffer on rank 106 and by PTF on rank 72. PRMT2 is reported by TALENoffer (rank 3) but not by PTF. The third off-target is reported by neither approach due to the large number of 11 mismatch positions.

3.2 Runtime comparison

We compare the runtime of TALENoffer with runtime optimization with that of PTF in Table 1 for example datasets of different sizes. We find that PTF requires 3.0–5.8 times the runtime of TALENoffer on the same input datasets. Considering memory consumption, PTF consistently allocates less memory than TALENoffer. However, for all input datasets, TALENoffer requires at most 2 GB of memory, which allows for execution on current standard computers (details in Supplementary Table S1).

Table 1. Runtime of TALENoffer compared with PTF on example datasets of different sizes

Dataset (size)	PTF	TALENoffer
Arabidopsis thaliana 1 kb up (34 Mb)	12 s	4 s
Homo sapiens exome (170 Mb)	1 min 5 s	13 s
Oryza sativa genome (373 Mb)	2 min 8 s	24 s
Homo sapiens genome (2.8 Gb)	17 min 57 s	3 min 5 s

Note: As an example TALEN, we use two TALEN monomers with 15 repeats (details in Supplementary Methods) with a distance of 12–24 bp between monomer target sites and with multithreading enabled. All values are measured on a standard laptop (Intel Core i7, ULV, dual core 2 GHz).

4 CONCLUSION

We present TALENoffer, a novel tool for the genome-wide prediction of TALEN targets and off-targets, which successfully predicts known off-targets of engineered TALENs and yields a competitive runtime. TALENoffer is implemented using the open-source Java library Jstacs (Grau *et al.*, 2012) and is available as a command line program and as a Galaxy (Blankenberg *et al.*, 2010) web application, which can also be installed to a local Galaxy server.

Conflict of interest: JB is part owner of a patent application regarding the use of TAL effectors.

REFERENCES

Blankenberg,D. *et al.* (2010) Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. *Curr. Protoc. Mol. Biol.*, **89**, 19.10.1–19.10.21.

Boch,J. *et al.* (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509–1512.

Doyle,E.L. *et al.* (2012) TAL effector-nucleotide targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Res.*, **40**, W117–W122.

Gaj,T. *et al.* (2013) ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.*, **31**, 397–405.

Grau,J. *et al.* (2013) Computational predictions provide insights into the biology of TAL effector target sites. *PLoS Comput. Biol.*, **9**, e1002962.

Grau,J. *et al.* (2012) Jstacs: a Java framework for statistical analysis and classification of biological sequences. *J. Mach. Learn. Res.*, **13**, 1967–1971.

Hockemeyer,D. *et al.* (2011) Genetic engineering of human pluripotent cells using TALE nucleases. *Nat. Biotechnol.*, **29**, 731–734.

Miller,J.C. *et al.* (2011) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.*, **29**, 143–148.

Moscou,M.J. and Bogdanove,A.J. (2009) A simple cipher governs DNA recognition by TAL effectors. *Science*, **326**, 1501.

Mussolino,C. *et al.* (2011) A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res.*, **39**, 9283–9293.

Osborn,M.J. *et al.* (2013) TALEN-based gene correction for epidermolysis bullosa. *Mol. Ther.*, **21**, 1151–1159. doi:10.1038/mt.2013.56.

Streubel,J. *et al.* (2012) TAL effector RVD specificities and efficiencies. *Nat. Biotechnol.*, **30**, 593–595.

Tesson,L. *et al.* (2011) Knockout rats generated by embryo microinjection of TALENs. *Nat. Biotechnol.*, **29**, 695–696.