Published in final edited form as:

Methods. 2014 September ; 69(2): 121–127. doi:10.1016/j.ymeth.2014.06.006.

Tools for TAL effector design and target prediction

Nicholas J. Booher¹ and Adam J. Bogdanove^{*}

Plant Pathology and Plant-Microbe Biology, 334 Plant Science, Cornell University, Ithaca NY 14853 USA

Abstract

TAL effectors are transcription factors injected into plant cells by pathogenic bacteria during infection. They find their specific DNA targets via a string of contiguous, structural repeats that individually recognize single nucleotides (with some degeneracy) by virtue of polymorphisms at residue 13. The number of repeats and sequence of the amino acids at position 13 determine the nucleotide sequence of the DNA target. Due to this modularity, TAL effectors are readily engineered and have been used alone or as molecular fusions for targeted gene activation, gene repression, chromatin modification, chromatin tagging, and most broadly, for genome editing as TAL effector nucleases (TALENs). Several moderate and high-throughput cloning methods are in place for assembling TAL effector-based genetic constructs. Targeting is complicated to an extent by a general requirement for thymine to precede the DNA target, a requirement of TALENs to bind paired opposing sites separated by a defined range of distances, differential contributions of different repeat types to overall affinity, and a polarity to mismatch tolerance. Several computational tools are available online to aid in design and the identification of candidate off-target binding sites, as well as assembly and implementation. These tools vary in their approaches, capabilities, and relative utility for different types of TAL effector applications. Accuracy of off-target prediction is not well characterized yet for any of the tools and will require a better understanding of the qualitative and quantitative variation in the nucleotide preferences of individual repeats.

1. Introduction

TAL effectors are heterogenous transcription factors produced and injected into plant cells by pathogenic bacteria. They bind to DNA in a sequence-specific way by virtue of a modular mechanism that pairs individual, contiguous, and polymorphic structural repeats in the proteins with contiguous nucleotides in their DNA targets, generally in a one-to-one fashion but with some degeneracy [1, 2]. In their native context, TAL effectors activate host disease susceptibility (*S*) genes that promote bacterial multiplication and or disease development, but in some cases they trigger disease resistance genes. As such, TAL

© 2014 Elsevier Inc. All rights reserved.

^{*}For correspondence: 1 (607) 255-7831; ajb7@cornell.edu.

¹njb98@cornell.edu

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

effectors have evolved under opposing selective forces – selection for targeting flexibility to accommodate *S* gene polymorphism that might be encountered across different host genotypes, and selection for targeting stringency to activate *S* genes uniquely without turning on resistance genes [3]. Likely as a result of this, qualitative and quantitative variation in the nucleotide preferences of individual repeats generates diversity in the stringency of specificity of different TAL effectors. And, the modular DNA recognition mechanism itself allows for rapid, recombinational evolution of new specificities.

The nucleotide specificity of a TAL effector repeat is determined by the amino acid at position 13 through direct interaction with the plus strand of the DNA. This residue and the amino acid at position 12 together constitute the repeat-variable diresidue (RVD). The RVD is the unit originally identified as the specificity determinant [1, 2] but structural elucidation later demonstrated that the residue at position 12 stabilizes the repeat and does not directly interact with the DNA [4, 5]. Common RVDs (using the single letter amino acid code) include HD, specifying cytosine, NG, specifying thymine, NN, specifying guanine or adenine, and NI, specifying adenine. Next most common are NS, and N* (in which the asterisk denotes a missing 13th residue), which have generally lax specificity. The rare RVD NH specifies G uniquely [6, 7]. Less common RVDs that have yet to be characterized can generally be assumed to have the specificity of a more common one if it shares the same 13th amino acid.

As discussed in other articles in this issue, TAL effectors are readily engineered and have proven to be a broadly applicable DNA targeting platform (see also [8] for a review). Several cloning methods have been developed that enable individual researchers to rapidly assemble custom repeat arrays into various structural contexts, including the native TAL effector backbone and derivatives fused to other protein domains [9–13]. High-throughput, automatable methods based on solid-state assembly have also been developed [14–16]. Reagents for many of these methods are available as kits from the non-profit plasmid repository Addgene (www.addgene.org). Custom TAL effector constructs can also be purchased from commercial sources. TAL effectors with customized specificities have been used not only for targeted gene activation, but also gene repression, chromatin modification, chromatin tagging for visualization and capture [17–19], and most broadly, for genome editing, as TAL effector nucleases (TALENs). TALENs are paired fusions with the catalytic domain of the *FokI* endonuclease, which functions as a dimer.

The modular DNA recognition mechanism of TAL effectors makes it relatively straightforward to design an array of repeat modules to target a desired nucleotide sequence. A general requirement for thymine to precede the DNA target sequence [1, 2] and the requirement of TALENs to target paired sites within a defined range of distances apart on opposing strands [20, 21], as well as differential contributions of different repeat types to overall activity [6], and a polarity to mismatch tolerance [22] add complexity however. Several computational tools have been developed to aid in design, as well as assembly and off-target prediction. This article reviews these tools, organized by application and with attention to relative merits, to provide a guide for potential users and to highlight challenges that remain.

2. TAL Effector Design

2.1. TAL Effector Targeter

TAL Effector Targeter designs repeat arrays to target particular DNA sequences. It is available both as an open source downloadable program and as a web-based tool as part of the TAL Effector Nucleotide Targeter (TALE-NT) suite [23] (<http://tale-nt.cac.cornell.edu/>). Target sequences must be provided in FASTA format and can be input either by pasting the sequences into a text box on the page or uploading an already prepared file. Users can specify a minimum and maximum length for designed arrays, choose whether to allow a thymine (default), cytosine (observed in at least one native target [24]), or either on the 5' end, and select if they want to use the RVD NH to target guanine for greater specificity, or NN for better affinity. By default TAL Effector Targeter only outputs RVD sequences for target sites that conform to several base composition rules thought to increase TAL effector affinity [6], but these restrictions can be disabled through checkboxes on the page.

A useful feature of TAL Effector Targeter is its ability to assess the specificity of designed TAL effectors by predicting binding sites in the intended target sequence (pre-loaded genomes / promoter sets / NCBI ID accepted). This uses the Target Finder tool with a 3x score cutoff (Target Finder is described under section 3.1, below). Optimizations enabled by batch processing allow TAL Effector Targeter to count binding sites for several TAL effectors quicker than if Target Finder were run for each effector. The results are summarized in the output as the number of sites found for each TAL effector so that the most specific are readily apparent.

Output from TAL Effector Targeter is a tab-delimited text file suitable for import into spreadsheet programs that is also displayed as a table on the website. Each row of the table is a TAL effector with columns for the name of the target sequence, the start position of the TAL effector in the target sequence, the length of the repeat array, a space-separated list of the RVDs, which strand of the target the array is designed to bind, and the plus strand sequence of the target, including the 5' T/C. The 'Resources' page of the TALE-NT site includes a spreadsheet with sequences of plasmids used in the Golden Gate assembly method published by the authors [9], as well as links to the reagents and protocols for that method.

2.2. TAL Plasmids Sequence Assembly Tool

The TAL Plasmids Sequence Assembly Tool (<http://bit.ly/assembleTALsequences>) is a web-based utility developed by the Bao lab at Georgia Institute of Technology that generates the plasmid DNA sequences needed to make TAL effector constructs using any of several different assembly methods. Input is accepted either as a target site in FASTA format, or an RVD sequence in a FASTA-like format. Users can select from the Golden Gate, FLASH, and ICA assembly methods [9, 15, 16, 25], and for the Golden Gate assembly method, which destination vector they would like to use (links to AddGene pages for each vector are at the bottom of the page) and whether to use the RVD NK, NN, or NH to target guanine. Output for plasmid constructs is provided in both FASTA format and as an annotated

GenBank file. The site also provides an additional, handy tool that can perform an alignment of sequencing reads to your destination vector to confirm that cloning was successful.

The TAL Plasmids Sequence Assembly Tool requires a particular target site or TAL effector to have been previously identified and thus provides no tools for binding site selection or target prediction. It fills a particular niche that makes it a valuable complement to all other tools reviewed in this section.

3. TAL Effector Target Prediction

3.1. Target Finder

Target Finder is another tool in the TALE-NT suite that identifies potential binding sites for a TAL effector with a user-specified RVD sequence, using a position weight matrix based on RVD-nucleotide association frequencies observed in natural TAL effectors. The web-based version of the tool has genomes and promoter sequences for several model organisms preloaded for searching, and also accepts IDs for genome assemblies on the NCBI website, or sequences uploaded or pasted into a text box on the site in FASTA format. Options are available to change the score threshold for output sites (explained in the next paragraph), find sites preceded by a T (default), C, or both, whether to search the reverse complement of the DNA sequence, and how many results to display in the table on the website. Output from Target Finder is available for download in both tab-delimited text and gff3 formats. The top binding sites found are displayed in a table on the results page with columns for sequence name and binding site position (for searches on pre-loaded genomes this column links to a genome browser with a track loaded showing the binding site), the strand the hit is on, the score of the binding site, the target sequence, and the plus strand sequence of the target.

Target sites are ranked and sorted by a score calculated as the sum of the negative log probability of each RVD in the TAL effector pairing with the corresponding nucleotide in the target sequence over the whole target sequence (lower scores are better). The probability of each RVD-DNA base interaction is based on the frequency of that association observed among known, naturally occurring TAL effector-target combinations, and is calculated as:

$$\left(\frac{9}{10}\right)\left(\frac{p}{n}\right) + \left(\frac{1}{10}\right)\left(\frac{1}{4}\right)$$

where p is the number of times the RVD is seen paired with the nucleotide in known, naturally occurring TAL effector-target combinations and n is the total number of observations for this RVD in those TAL-effector-target combinations [1, 23]. The small 2.5% bonus given to each interaction ensures a nonzero probability for that pairing. This allows the tool to find binding sites with pairings not already observed in nature. The RVD-DNA frequency counts currently used can be found in the supplemental material of [23], though these will change as new verified targets are identified. By default, results are filtered to those with a score less than three times that of the best score that results from each RVD pairing with its preferred nucleotide, based on the observation that most verified targets of natural TAL effectors have scores lower than this threshold.

3.2. Talvez

Talvez (<http://bioinfo-prod.mpl.ird.fr/cgi-bin/talvez/talvez.cgi>) scans for possible bindings sites using a position weight matrix (PWM) similar to the one used by Target Finder [26]. It differs from Target Finder by assigning to some rare RVDs the same observation counts as the common RVD with the same 13th position residue, which, as noted in the Introduction (Section 1), is the sole specificity determinant. Talvez also makes use of the authors' observation that RVDs closer to the C terminus in natural TAL effectors more often pair with non-preferred nucleotides than those in the N-terminal end, in agreement with experimental evidence for this polarity [22]. Accordingly, users can specify a position in the RVD array after which an alternative PWM more forgiving of mismatches is used for scoring. On the web-based version of Talvez this position defaults to zero, although the authors indicate that using 19 improves prediction performance.

Primary input to Talvez consists of a set of RVD sequences and a set of FASTA formatted nucleotide sequences to search within for potential targets of those RVD strings. Several plant and fungal promoter sequences are pre-loaded. Optional parameters include the number of top-ranked targets to return for each RVD string, the minimum score of the reported sites, the pseudo-counts (bonus) to add to each entry in the PWM, the number of RVDs after which to apply the alternative PWM, and counts for both the primary and alternative PWM. Talvez accounts for the T/C preceding the TALE target by prepending a pseudo-RVD called "OO" to each TALE sequence that in the default scoring matrix has 30 observations for C and 70 observations for T. When pseudo-counts are added to the matrix this allows finding binding sites not preceded by a T/C.

Output from Talvez is a tab-delimited file. If a small number of results are returned for each TAL effector this file will be displayed as a table on the website. Columns in the output are TAL effector name, RVD sequence, target sequence ID, binding site score, distance of the site from the end of the target sequence, start position and end of the site, binding site sequence, DNA sequence of the site, and the rank of the site. If the user selected a preloaded set of promoter sequences to search, the output will also contain the chromosome, strand, name of the gene the match occurred in, and start and end position of that gene on the chromosome. Binding site scores are the sum of the log of the probability of each RVD-nucleotide matchup divided by the frequency of that nucleotide over all search sequences [27]. Higher scores are better.

Although Talvez runs significantly slower than the other tools described in this section, it is the only tool reviewed that allows users to set an explicit position after which mismatches are penalized less, and, unlike Target Finder, Talvez allows input of new observation counts for the position weight matrix without modifying the software's source code.

3.3. TALgetter

TALgetter (<http://www.jstacs.de/index.php/TALgetter>) is part of the Jstacs framework [28] and identifies potential binding sites using a local mixture model [29]. In this model, in addition to a probability for each RVD pairing with each nucleotide, each RVD is considered to have an "importance", indicating the probability that it interacts with the DNA

at all. The base preceding the target site is modeled separately from the rest, and thus the probability score of TAL effector binding is the probability of the preceding base multiplied by the probability of each base in the binding site given its corresponding RVD and the previous two bases in the site. The latter probability is calculated as:

$$wx+yz$$

where w is the probability of the RVD interacting with DNA, x is the probability of the RVD pairing with the corresponding base in the binding site, y is the probability of the RVD not interacting with DNA, and z is the probability of this base in the binding site given the previous two bases in the site. The rationale for considering the previous two bases in the site is not explained, but is presumably informed by the prior work of the authors using higher-order Markov models to predict transcription factor binding sites [30]. The parameters of the model are estimated with independent Dirichlet priors using the Bayesian maximum a-posteriori learning principle over a set of verified binding sites for natural TAL effectors and some artificially designed TAL effectors. Artificial TAL effectors in the training set were weighted so that all from the same experiment together count the same as one natural TAL effector.

TALgetter accepts as input a FASTA file of DNA sequences to scan for binding sites, and the RVD sequence of a single TAL effector to query the DNA sequences with. Genomes for several model animals, rice, and *Arabidopsis thaliana* are pre-loaded for searching, as well as promoter sequences for rice. Users can select if they want a model that assigns the same probabilities to RVDs ending with the same residue, and whether to use the default model parameters or re-estimate them from a custom training set. In addition, options are available to limit returned target sites by rank or maximum p -value.

TALgetter produces output in tab-delimited text and FASTA format. The FASTA file contains the binding site sequence with the ID and other columns of the table in the FASTA header. Columns in the table include the sequence ID, distance of the binding site from the start and end of the sequence, binding site sequence, a string representing matches/mismatches in this site compared to the perfect site, the site score, and the empirical p -value and empirical Evalue for the site. When searching rice promoters, the table displayed on the website links locus IDs to the gene information page on the Rice Genome Annotation Project website (<http://rice.plantbiology.msu.edu/>) [31]. The empirical p -value is the percentile of the binding site score in the distribution of binding site scores obtained by scoring the RVD sequence against sequences of the same length drawn from a randomly generated DNA sequence at least as long as the combined length of all of the DNA sequences being searched. A lower p -value is better. The website output page also shows sequence logos for theoretical target sites and the sequences of target sites actually found.

4. TALEN Design and Off-Target Prediction

4.1.1. TALEN Targeter / Paired Target Finder—TALEN Targeter, also part of the TALE-NT tool suite, designs TALENs that target user-specified sequences. Users can input a minimum and maximum length for the two RVD arrays and a length range for the spacer

or choose from presets for four common architectures, and choose to allow T, C, or both on the 5' end of the binding site for each monomer. By default the tool restricts output to the TALEN with the longest monomers and the shortest spacer that targets each position in the sequence, but can be set to output all TALENs targeting a specific position, or all possible TALENs targeting the sequence. Additionally, users can restrict output to TALENs for which the binding site for each monomer is at least 25% C+G with no runs of 5 or more A+T, rules suggested by Streubel and colleagues for creating TALEs with good binding efficiency [6].

A key feature of TALEN Targeter is its ability to assess the specificity of designed TALENs by predicting all potential target sites in the input sequence (pre-loaded genomes / promoter sets / NCBI ID accepted). It uses a modified version of Target Finder to find predicted sites for the monomers of each TALEN with a score under 3x their best score that are separated by distances in the spacer size range, including both homodimers and heterodimers. The results of the search are summarized in the output as the number of sites found for each TALEN so users can identify at a glance the TALENs predicted to be most specific. If a user wants to know the location and scores of predicted binding sites for a TALEN, another tool called Paired Target Finder is available for this. Though TALEN Targeter does not output these data, like TAL Effector Targeter, optimizations enabled by batch processing allow it to count binding sites for several TALENs quicker than if Paired Target Finder were run for each TALEN.

TALEN Targeter produces a tab-delimited text file. For each TALEN there is a column for sequence name, cut site position (estimated as the middle or immediately left of middle position in the spacer), start position of each monomer, spacer range, RVD sequence of each monomer, the plus strand sequence of the target site, and % strong RVDs (HD, NN/NH) over both monomers. TALEN Targeter also prints restriction enzyme sites (New England Biolabs) that occur in the spacer region and nowhere else in the 250bp region on either side that can be used as a diagnostic for TALEN-mediated mutations in the spacer. If the user selects off-target counting, the output file shows these for different heterodimer and homodimer combinations of the TALEN monomers in the order RVDSEQ1/RVDSEQ1, RVDSEQ1/RVDSEQ2, RVDSEQ2/RVDSEQ1, RVDSEQ2/RVDSEQ2, TOTAL. The table displayed on the webpage shows only the total.

4.1.2. Mojo Hand—Mojo Hand (<http://talendesign.org/>) designs TALENs to target genes and sequences from the NCBI database [32]. Primary input consists of an email address and a gene or nucleotide sequence ID. Nucleotide sequences in FASTA format are also accepted either through a textbox on the site or by uploading a file. For gene inputs users can choose to target exons, introns, or both, specify which mRNA feature to use for exons by index in the GenBank record or accession number, or choose to target a misc. RNA feature or coding sequence by index in the GenBank record. Additional options allow changing the 5' base, setting the minimum and maximum repeat array and spacer lengths, extending the binding site search outside exon boundaries by providing a “short flank length”, and highlighting CpG islands that may interfere with binding. Mojo Hand can also identify restriction enzyme sites in the binding site spacer. Users can specify the minimum distance between

multiple restriction sites for the same enzyme, and how far out from the TALEN target to look for restriction sites (“long flank length”).

Output from Mojo Hand consists of a FASTA file containing the sequences analyzed, a FASTA file with the analyzed sequences and the flanking sequence that was searched for restriction sites, and a web report of possible TALENs. The TALEN web report displays all possible TALENs targeting the input sequences for inputs less than 1000 bp, those with a first monomer binding site starting every 10 bp for inputs between 1000 bp and 5000 bp, and those every 20 bp for inputs greater than 5000 bp. For all listed TALENs the report contains the plus strand sequence of the first monomer site, spacer, and second monomer site, the binding (reverse) strand sequence of the second monomer site, and the RVD sequence for each site, with NN used to specify guanine. Potentially useful restriction enzyme sites are also shown along with suppliers and activity levels in common PCR conditions. To get information about a particular TALEN site users can click the “Download as CSV file” button to download the site in CSV format or click the “Save as candidate” link to send it to their email address.

4.1.3. E-TALEN—E-TALEN (<http://www.e-talen.org/>) is a web-exclusive tool that streamlines the process of designing TALENs for particular locations within genes for different purposes (although it can also design TALENs for generic sequences) [33]. Users select from several model genomes, provide ENSEMBL accession numbers for up to 50 genes they want to target, choose whether they want to do a gene knockout, 5′ sequence replacement for N-terminal tagging, or a 3′ one for C-terminal tagging, and which assembly kit they would like to use, and the tool automatically designs and screens possible TALEN designs based on best location within the gene for the chosen task and uniqueness to the intended target. Optional parameters are available to customize nearly every part of the process, and restriction enzyme site analysis is also available. One useful feature of E-TALEN missing from other tools is the ability to check for off-targets simultaneously in the target genome and any repair, replacement, or insertion sequences to be introduced as part of the experiment.

To assess possible off-target effects E-TALEN treats prospective binding sites as paired-end short sequencing reads and uses Bowtie2 [34] to search for matches. This approach has the benefit of being extremely fast, but by treating all mismatches as equal it discounts variation in the tolerance of individual RVDs for different nucleotides. In particular, four of the five assembly kits pre-loaded into the website can use the RVD NN, which has dual specificity for adenine and guanine. This, combined with the default setting to only identify off-targets with zero mismatches from the intended binding site, may mislead novice users into believing their TALENs are more specific than they really are. In spite of these shortcomings, E-TALEN remains a useful to quickly screen sequences and design potentially useful TALENs.

E-TALEN provides results as a web report and as downloadable tab-delimited text and GFF3 files. For each TALEN, the strand, binding site nucleotide sequence, exons hit, RVD sequences, and a score based on length and compositional variance (higher is better) are given. The web report also shows statistics regarding how many possible TALEN designs

were filtered out due to each design criterion, and a summary graphic showing the targeted gene and the location of target sites for the output TALENs.

4.1.4. SAPTA—SAPTA (http://baolab.bme.gatech.edu/Research/BioinformaticTools/TAL_targeter.html) designs TALENs that are predicted to have good activity, specifically considering TALENs in which the RVD NK is used to more specifically bind guanine than the often-used NN, at the cost of reduced activity [35]. Potential TALEN-monomer binding sites are scored using a model that successfully related binding site length and base composition to activity in a single-strand annealing assay with a test set of 130 NK-containing TALENs. Scores of appropriately spaced monomer sites are added to form a composite score, and TALENs are ranked based on this score; larger scores are better, and TALENs with a score over 30 are predicted to have good activity.

To use the SAPTA tool, users provide a DNA sequence in plain-text format and enter a minimum and maximum binding site length and spacer length. To target a specific base in the sequence, that base can be surrounded in square brackets. By default, sites are required to be preceded by T on the 5' end, redundant sites with identical sequences are hidden, and only the top 100 sites are shown, but there are options for each. Output consists of a table sorted by score with columns for the starting position of the TALEN, the DNA sequence bound by the left monomer, the DNA sequence bound by the right monomer, the size of each monomer and the spacer between them, the scores for each monomer and the composite score, and names and sequences of restriction enzyme sites in the spacer. Further, a scrollable textbox is displayed listing all positions within the DNA sequence, and the score for the best TALEN that targets that position, or a zero if no TALENs target the position.

Additional tools on the SAPTA webpage can identify and rank individual TALEN-monomer sites within a DNA sequence, or calculate scores for given TALEN or TALEN-monomer sites.

4.1.5. TALENoffer—TALENoffer (<http://www.jstacs.de/index.php/TALENoffer>) is an extension of the TALgetter model that predicts target sites for TALENs [36]. The tool takes as input FASTA formatted query sequences, an optional annotation file, and RVD sequences for each TALEN monomer. Options are available to use different spacer ranges appropriate for several common architectures or to provide custom values, to search for heterodimers only (by default both heterodimers and homodimers are considered), and to select whether to find targets assuming the endonuclease is fused to the standard C-terminal (default) or the non-standard N-terminal end. The model training settings for TALgetter are also available.

TALENoffer ranks possible binding sites by assigning each a score that is the sum of the scores for the two monomers of the TALEN, calculated for each as

$$\left(\frac{1}{x+1}\right)^y$$

where x is the length of the monomer and y is the log-probability of the monomer binding this site under the TALgetter model. Returned results are also filtered based on a q -value (defaults to 0.4), to include those sites for which the average likelihood over both monomer-DNA binding sites is equal to or greater than $q * 100\%$ the likelihood of the best pairing. Additionally, each monomer-DNA must have a likelihood at least $0.9 * q$ as good as that of its best pairing.

Results from TALENoffer are downloadable in both tab-delimited text and GFF2/GFF3 formats. For each predicted binding site the target sequence, position of each monomer start, sequence of each monomer binding site, and a string representing matches/mismatches in each monomer binding site compared to its perfect site are listed. Spacer length, architecture (heterodimer/homodimer), score, and full binding site sequence are also given. If an annotation was provided, the names of any features in the target site are listed. The web interface results page also shows a sequence logo for each monomer to represent its target site as a motif.

4.1.6. PROGNOS—PROGNOS (<http://baolab.bme.gatech.edu/Research/BioinformaticTools/prognos.html>) predicts off-targets of TALENs and provides primer sequences for amplicon sequencing across the target following TALEN treatment [37]. It is hosted on the same website as SAPTA but is not integrated with SAPTA. The only required input is the RVD sequences or intended target sites of a TALEN and a choice of genome to search from several pre-loaded ones. Options for spacer size, ideal primer melting temperature, number of mismatches tolerated in candidate targets sites, and an email address to notify when the job completes are also available.

PROGNOS ranks candidate off-target sites using three different algorithms and provides separate output files for each. The first is a simple homology score based entirely on the number of mismatches in the site; a lower score is better. The second is a weighted sum of the score ratios for the two monomers using the TALE-NT Target Finder scoring scheme; again a lower score is better. The third, dubbed “TALEN v2.0”, is similar in structure to the second, but uses RVD-DNA association frequencies from verified TALEN off-targets rather than those from verified TAL effector targets. Additionally, TALEN v2.0 has trained parameters that let the strong binding RVDs NN and HD [6] compensate for mismatches in adjacent RVD-nucleotide pairs, and penalize mismatches after the 14th RVD [22]. Scores for this third algorithm are out of 100, and higher is better.

The online and downloadable versions of PROGNOS both provide output in HTML and CSV formats. For the top sites identified by each algorithm the score and sequence are given, along with the genomic context (exon, intron, or intergenic), and the name of the nearest gene. A link to the region in a genome browser is also shown. Primer sequences and melting temperatures are provided for amplification and sequencing after mutagenesis, as well as the expected product length and a link to expected PCR results. The HTML report also contains a form populated with all primers to assist in ordering.

5. Going Forward

The existing online resources for TAL effector and TALEN design and target prediction complement each other and provide a good working toolkit for DNA targeting with these reagents (Tables 1, 2). Yet no software is fully capable of exploiting the inherent variation in targeting specificity of TAL effectors for design, or accounting for it in target and off-target prediction. Nor can any of the current tools accurately estimate the affinities of different TAL effectors for different target sequences, or even the relative affinities. Our understanding of the variables that influence effector binding specificity and affinity is still too incomplete. Several of the tools attempt to compensate by including speculative parameters that fit observed patterns in the small set of validated TAL effector and TALEN target pairs. Preferable however is to incorporate values based on eventual comprehensive biochemical quantification of the affinity contributions of different RVD-nucleotide combinations, accounting for effects of position and context in the binding site. In their publication of PROGNOS [37], Fine and colleagues present and discuss refinement of off-target prediction through iterative experimental validation. Accumulation of sufficient data of that sort as well as identification of a large collection of bona fide, naturally occurring TAL effector-target pairs may enable a machine learning approach that bypasses biochemical analysis and instead generates powerful statistical models for fine-tuned design and more accurate target prediction in the next generation of tools.

Acknowledgments

Research on TAL effectors in our laboratory is supported by grants from the National Science Foundation (IOS 1238189) and the National Institutes of Health (R01 GM098861).

References

1. Moscou MJ, Bogdanove AJ. *Science*. 2009; 326:1501. [PubMed: 19933106]
2. Boch J, Scholze H, Schornack S, Landgraf A, Hahn S, Kay S, Lahaye T, Nickstadt A, Bonas U. *Science*. 2009; 326:1509–1512. [PubMed: 19933107]
3. Bogdanove AJ, Schornack S, Lahaye T. *Current Opinion in Plant Biology*. 2010; 13:394–401. [PubMed: 20570209]
4. Mak AN, Bradley P, Cernadas RA, Bogdanove AJ, Stoddard BL. *Science*. 2012; 335:716–719. [PubMed: 22223736]
5. Deng D, Yan C, Pan X, Mahfouz M, Wang J, Zhu JK, Shi Y, Yan N. *Science*. 2012; 335:720–723. [PubMed: 22223738]
6. Streubel J, Blucher C, Landgraf A, Boch J. *Nat Biotechnol*. 2012; 30:593–595. [PubMed: 22781676]
7. Cong L, Zhou R, Kuo Y-c, Cunniff M, Zhang F. *Nature Communications*. 2012; 3:968.
8. Doyle EL, Stoddard BL, Voytas DF, Bogdanove AJ. *Trends Cell Biol*. 2013; 23:390–398. [PubMed: 23707478]
9. Cermak T, Doyle EL, Christian M, Wang L, Zhang Y, Schmidt C, Baller JA, Somia NV, Bogdanove AJ, Voytas DF. *Nucleic Acids Research*. 2011; 39:e82. [PubMed: 21493687]
10. Zhang F, Cong L, Lodato S, Kosuri S, Church GM, Arlotta P. *Nat Biotechnol*. 2011; 29:149–153. [PubMed: 21248753]
11. Morbitzer R, Elsaesser J, Hausner J, Lahaye T. *Nucleic Acids Research*. 2011; 39:5790–5799. [PubMed: 21421566]
12. Weber E, Gruetzner R, Werner S, Engler C, Marillonnet S. *PLoS ONE*. 2011; 6:e19722. [PubMed: 21625552]

13. Li L, Atef A, Piatek A, Ali Z, Piatek M, Aouida M, Sharakou A, Mahjoub A, Wang G, Khan S, Fedoroff NV, Zhu J-K, Mahfouz M. *Molecular Plant*. 2013
14. Wang Z, Li J, Huang H, Wang GC, Jiang MJ, Yin SY, Sun CH, Zhang HS, Zhuang FF, Xi JJ. *Angewandte Chemie-International Edition*. 2012; 51:8505–8508.
15. Reyon D, Tsai SQ, Khayter C, Foden JA, Sander JD, Joung JK. *Nat Biotechnol*. 2012; 30:460–465. [PubMed: 22484455]
16. Briggs AW, Rios X, Chari R, Yang L, Zhang F, Mali P, Church GM. *Nucleic Acids Research*. 2012; 40:e117. [PubMed: 22740649]
17. Miyanari Y, Ziegler-Birling C, Torres-Padilla ME. *Nat Struct Mol Biol*. 2013; 20:1321–1324. [PubMed: 24096363]
18. Byrum SD, Taverna SD, Tackett AJ. *Nucleic acids research*. 2013; 41:e195. [PubMed: 24030711]
19. Fujita T, Asano Y, Ohtsuka J, Takada Y, Saito K, Ohki R, Fujii H. *Sci Rep*. 2013; 3:3171. [PubMed: 24201379]
20. Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ, Voytas DF. *Genetics*. 2010; 186:757–761. [PubMed: 20660643]
21. Miller JC, Tan S, Qiao G, Barlow KA, Wang J, Xia DF, Meng X, Paschon DE, Leung E, Hinkley SJ, Dulay GP, Hua KL, Ankoudinova I, Cost GJ, Urnov FD, Zhang HS, Holmes MC, Zhang L, Gregory PD, Rebar EJ. *Nat Biotechnol*. 2011; 29:143–148. [PubMed: 21179091]
22. Meckler JF, Bhakta MS, Kim MS, Ovadia R, Habrian CH, Zykovich A, Yu A, Lockwood SH, Morbitzer R, Elsaesser J, Lahaye T, Segal DJ, Baldwin EP. *Nucleic acids research*. 2013; 41:4118–4128. [PubMed: 23408851]
23. Doyle EL, Booher NJ, Standage DS, Voytas DF, Brendel VP, Vandyk JK, Bogdanove AJ. *Nucleic acids research*. 2012; 40:W117–122. [PubMed: 22693217]
24. Yu Y, Streubel J, Balzergue S, Champion A, Boch J, Koebnik R, Feng J, Verdier V, Szurek B. *Mol Plant Microbe Interact*. 2011; 24:1102–1113. [PubMed: 21679014]
25. Sanjana NE, Cong L, Zhou Y, Cunniff MM, Feng G, Zhang F. *Nat Protocols*. 2012; 7:171–192.
26. Pérez-Quintero AL, Rodríguez-R LM, Dereeper A, López C, Koebnik R, Szurek B, Cunnac S. *PLoS ONE*. 2013; 8:e68464. [PubMed: 23869221]
27. Megraw, M.; Hatzigeorgiou, A. *Plant MicroRNAs*. Meyers, BC.; Green, PJ., editors. Humana Press; 2010. p. 149-161.
28. Grau J, Keilwagen J, Gohr A, Haldemann B, Posch S, Grosse I. *J Mach Learn Res*. 2012; 13:1967–1971.
29. Grau J, Wolf A, Reschke M, Bonas U, Posch S, Boch J. *PLoS Comput Biol*. 2013; 9:e1002962. [PubMed: 23526890]
30. Grau, J.; Keilwagen, J.; Kel, AE.; Grosse, I.; Posch, S. In: Falter, C.; Schliep, A.; Selbig, J.; Vingron, M.; Walther, D., editors. *German Conference on Bioinformatics, GI2007*; p. 123-134.
31. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, Childs KL, Davidson RM, Lin H, Quesada-Ocampo L, Vaillancourt B, Sakai H, Lee SS, Kim J, Numa H, Itoh T, Buell CR, Matsumoto T. *Rice*. 2013; 6:4. [PubMed: 24280374]
32. Neff K, Argue D, Ma A, Lee H, Clark K, Ekker S. *BMC Bioinformatics*. 2013; 14:1. [PubMed: 23323762]
33. Heigwer F, Kerr G, Walther N, Glaeser K, Pelz O, Breinig M, Boutros M. *Nucleic Acids Research*. 2013; 41:e190. [PubMed: 24003033]
34. Langmead B, Salzberg SL. *Nat Meth*. 2012; 9:357–359.
35. Lin YN, Fine EJ, Zheng ZL, Antico CJ, Voit RA, Porteus MH, Cradick TJ, Bao G. *Nucleic Acids Research*. 2014; 42
36. Grau J, Boch J, Posch S. *Bioinformatics*. 2013
37. Fine EJ, Cradick TJ, Zhao CL, Lin Y, Bao G. *Nucleic Acids Research*. 2013

TAL effector tools

Table 1

	TAL-NT TAL Effector Targeter ¹	TAL Plasmids Sequence Assembly Tool ²	TALE-NT Target Finder ¹	TALVEZ ³	TALgetter ⁴
Version or Date Tested	2.0	1/14/2014	2.0	3.1	1.0
Purpose	TAL effector design, target prediction ⁵	TAL effector plasmid design	Target prediction	Target prediction	Target prediction
Novelty	Outputs best TAL effector sequences to target DNA sequence based on RVD composition and number of predicted off-targets	Outputs plasmids sequences or aligns sequencing reads for TAL effector targeting pre- determined sequence	Can load sequences directly from NCBI genome assembly and nucleotide sequence databases	Algorithm attempts to correct for polarity effects; web interface accepts custom values for PWM	Can estimate model parameters for a new training set through web interface
Off-target prediction sequence sources	NCBI, pre-loaded genomes	n/a	NCBI, FASTA, pre-loaded genomes	FASTA, pre-loaded genomes	FASTA, pre-loaded genomes
Off-target prediction method	Modular PWM	n/a	Modular PWM	Modular PWM	Local mixture model
Off-target prediction output	Number of predicted off- targets	n/a	Sequence, score, location, genome browser link	Sequence, score, location, annotation	Sequence, score, location, annotation, genome browser link
Web server	Yes	Yes	Yes	Yes	Yes
Open Source / License	Yes, ISC	Yes, Unknown	Yes, ISC	Yes, GPL	Yes, GPL
Programming Language	Python, CUDA	JavaScript	C	Perl	Java
Cross-Platform Executable Provided	Yes ⁶	Yes ⁷	No	Yes	Yes

¹ <http://tale-nt.cac.cornell.edu>

² <http://bit.ly/assembleTALsequences>

³ <http://bioinfo-prod.mpl.ird.fr/cgi-bin/talvez/talvez.cgi>

⁴ <http://www.jstacs.de/index.php/TALgetter>

⁵ TAL Effector Targeter only reports the number of predicted targets. Target Finder reports target sequences and locations.

⁶ Off-target counting is supported on Linux only and requires an NVIDIA GPU and separate compilation. TAL effector design functionality works on all platforms.

⁷ The TAL Plasmids Sequence Assembly tool is implemented entirely in the browser. Saving a copy of the web page would enable using it offline.

TAL effector nuclease tools

Table 2

	TALE-NT ¹	E-TALEN ²	Mojo Hands ³	SAPTA ⁴	TALENoffer ⁵	PROGNOS ⁶
Version or Date Tested	2.0	2.5	1.5	6/7/2014	1.0	1/14/2014
Purpose	TALEN design, target prediction	TALEN design, target prediction	TALEN design	TALEN design	Target prediction	Target prediction
TALEN target sequence source(s)	FASTA	ENSEMBL, FASTA	NCBI, FASTA	FASTA	n/a	n/a
TALEN cutsite design considerations	RVD composition, restriction enzymes, predicted off-targets	Genomic context, restriction enzymes, experiment type, predicted off-targets	Genomic context, restriction enzymes	Predicted activity based on RVD composition and site length, restriction enzymes	n/a	n/a
Off-target prediction sequence sources	NCBI, FASTA, pre-loaded genomes	FASTA, pre-loaded genomes	n/a	n/a	FASTA, pre-loaded genomes	FASTA, pre-loaded genomes
Off-target prediction method	Modular PWM	# mismatches from perfect target	n/a	n/a	Local mixture model	Modular PWM
Off-target prediction output	Sequence, score, location, genome browser link	n/a ⁷	n/a	n/a	Sequence, score, location, annotation, genome browser link	Sequence, score, location, annotation, genome browser link, primers for validation
Web server	Yes	Yes	Yes	Yes	Yes	Yes
Open Source / License	Yes, ISC	No	Yes, custom non-commercial	Yes, Unknown	Yes, GPL	Yes, Unknown
Programming Language	Python, CUDA	Unknown	Perl	JavaScript	Java	Perl
Cross-Platform Executable Provided	Yes ⁸	No	On request	Yes ⁹	Yes	Yes

¹ <http://tale-nt.cac.comell.edu>

² <http://www.e-talen.org/>

³ <http://talendesign.org/>

⁴ http://baolab.bme.gatech.edu/Research/BioinformaticTools/TAL_targeter.html

⁵ <http://www.jsiacs.de/index.php/TALENoffer>

⁶ <http://baolab.bme.gatech.edu/Research/BioinformaticTools/prognos.html>

- ⁷ E-TALEN only counts off-targets as an exclusionary criterion for TALEN design. It does not output the location of matches.
- ⁸ Off-target counting is supported on Linux only and requires an NVIDIA GPU and separate compilation. TALEN design functionality works on all platforms.
- ⁹ SAPTA is implemented entirely in the browser. Saving a copy of the web page would enable using it offline.