Bioinformatics, 38(8), 2022, 2341–2343
https://doi.org/10.1093/bioinformatics/btac087
Advance Access Publication Date: 14 February 2022
Applications Note

Genome analysis

Effectidor: an automated machine-learning-based web server for the prediction of type-III secretion system effectors

Naama Wagner[1], Oren Avram [1], Dafna Gold-Binshtok[1], Ben Zerah[1], Doron Teper[2] and Tal Pupko [1,*]

[1]The Shmunis School of Biomedicine and Cancer Research, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel and [2]Department of Plant Pathology and Weed Research, Institute of Plant Protection Agricultural Research Organization (ARO), Volcani Center, Rishon LeZion 7505101, Israel

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on July 20, 2021; revised on January 31, 2022; editorial decision on February 2, 2022; accepted on February 8, 2022

Abstract

Motivation: Type-III secretion systems are utilized by many Gram-negative bacteria to inject type-3 effectors (T3Es) to eukaryotic cells. These effectors manipulate host processes for the benefit of the bacteria and thus promote disease. They can also function as host-specificity determinants through their recognition as avirulence proteins that elicit immune response. Identifying the full effector repertoire within a set of bacterial genomes is of great importance to develop appropriate treatments against the associated pathogens.

Results: We present Effectidor, a user-friendly web server that harnesses several machine-learning techniques to predict T3Es within bacterial genomes. We compared the performance of Effectidor to other available tools for the same task on three pathogenic bacteria. Effectidor outperformed these tools in terms of classification accuracy (area under the precision–recall curve above 0.98 in all cases).

Availability and implementation: Effectidor is available at: https://effectidor.tau.ac.il, and the source code is available at: https://github.com/naamawagner/Effectidor.

Contact: talp@tauex.tau.ac.il

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

1 Introduction

Many Gram-negative pathogenic bacteria use Type-III secretion systems (T3SSs) to inject type-3 effector (T3E) proteins to eukaryotic hosts and thus promote disease (Pinaud *et al.*, 2018; Wagner *et al.*, 2018). In addition to their contribution to pathogenicity, some T3Es can be recognized as avirulence proteins (Khan *et al.*, 2016). Therefore, T3Es harbor a dual function as host-specificity determinants by contributing to virulence in susceptible hosts and restricting the bacteria in resistant hosts (Bent and Mackey, 2007).

While the T3SSs are conserved across species (Burkinshaw and Strynadka, 2014), the effectors arsenal varies even between different strains of the same species (Jalan *et al.*, 2013; Jiménez-Guerrero *et al.*, 2020). Therefore, traditional homology-based annotation is usually insufficient to unmask the full effectors repertoire of a specific bacterial strain.

Here, we present Effectidor: a computational pipeline designed for the prediction of unknown effectors within bacterial genomes. This work was motivated by our previous experience applying machine-learning (ML) algorithms for the task of predicting effectors (Burstein *et al.*, 2015; Jiménez-Guerrero *et al.*, 2020; Nissan *et al.*, 2018; Ruano-Gallego *et al.*, 2021; Teper *et al.*, 2016). While other methods for effectors prediction exist (Arnold *et al.*, 2009; Dong *et al.*, 2015; Goldberg *et al.*, 2016; Hobbs *et al.*, 2016; Hui *et al.*, 2020), our approach differs in various key aspects. The main aspect is that when training the ML classifiers, our training data are extracted only from the genomic sequence of a strain in question. Thus, separate classifiers are trained for different species. In addition, we perform the analysis on full genomes, rather than on selected proteins, which allows us to include features related to genomic organization and regulatory regions. Lastly, we train several ML classifiers, combining dozens of different features unique to our approach, and select the optimal one using cross validation. Of note, the average running time of Effectidor is <20 min, during which a full genome analysis is performed.

2 Materials and methods

Effectidor is written in Python 3.7, and the ML procedures are performed using the Scikit-Learn package (Pedregosa *et al.*, 2012). Effectidor can be divided to the following steps: (i) establishing the list of positive and negative samples; (ii) extracting features; (iii) training an ML classifier; and (iv) executing the trained classifier for all open reading frames (ORFs) in the analyzed genome.

2.1 Establishing the positive and negative samples

A DNA FASTA file with all ORFs of a given genome must be provided. An additional 'effectors file' listing known effector-coding ORFs in that genome (the positive-labeled samples) may be provided. If such a file is not provided, the list of positive samples is computed based on sequence similarity to previously validated effectors. Negative samples are computed based on sequence similarity to proteins of bacteria that do not encode a T3SS (see Supplementary Data S1 for a detailed description of this step).

2.2 Extracting features

A total of 51 obligatory features are extracted from a mandatory input file, which includes all the DNA ORF sequences of the bacterium in a FASTA format. A ZIP archive with multiple FASTA files is also allowed, e.g. for including plasmid sequences or if the user prefers to upload each contig in a separate file (all contigs may also be uploaded as a single FASTA file). These features include, for example, the GC-content, protein length, relative frequencies of amino acids in the full protein and in the N-terminal region, homology to known T3Es in other bacteria and in the analyzed strain, and existence of SecYEG secretion signal, computed using SignalP 4.1 (Petersen et al., 2011). Additional 19 optional features can be extracted if relevant input information is supplied by the user (Supplementary Data S2 provides a full list of all features and their extraction): (i) if one or more protein FASTA files with protein records of the eukaryotic host are provided, they are used to search for eukaryotic motifs using BLASTp (Altschul et al., 1990); (ii) if a ZIP archive with protein FASTA file(s) containing protein records of closely related bacteria that do not encode a T3SS is provided, it is used to identify ORFs that are unlikely to be T3Es; and (iii) one or more full genome FASTA files and corresponding GFF3 features files can be provided. The GFF3 file contains information regarding the starting and ending point of each ORF and is used to compute genomic-organization features. Together with the full genomic sequence, they are used to search for regulatory motifs in the ORFs' promoters. Specifically, Effectidor can search for the following regulatory motifs: PIP-box (Koebnik et al., 2006), which is relevant for *Xanthomonas*, *Ralstonia* and *Acidovorax*; hrp-box (Zwiesler-Vollick et al., 2002), which is relevant for *Pseudomonas syringae* and plant pathogens of the family of Enterobacteria; mxiE-box (Bongrand et al., 2012; Mavris et al., 2002), which is relevant for *Shigella*; exs-box (Brutinel et al., 2009), which is relevant for *Pseudomonas aeruginosa*; and tts-box (Krause et al., 2002), which is relevant for rhizobia. These files can be provided both for fully assembled genomes and for draft genomes that are composed of several contigs. While for a fully assembled genome these features will be computed in full, for a draft genome these features will be missing for ORFs at the edges of the contigs.

2.3 Training an ML classifier

Several classification algorithms are evaluated, including Linear Discriminant Analysis, Naïve Bayes, K-Nearest Neighbors, Logistic Regression, Support Vector Machine, and Random Forest (RDF). To select the best performing algorithm, 20% of the labeled data (stratified sampled) are kept as a test set. The classifiers are trained on the remaining 80% (including feature selection). This training is done using stratified tenfold cross validation. The trained classifiers are then evaluated on the test data. This procedure minimizes the probability of overfitting the classifiers to the training data. We use the area under the precision–recall curve (AUPRC) as a scoring method since unbalanced labeled data are typically provided. Additional details regarding the ML procedures we use are provided in Supplementary Data S3.

2.4 Output

The main output of Effectidor is a downloadable Excel file with the full predictions for all the ORFs in the genome, sorted in a descending order by their likelihood to encode effectors. The 10 best predictions of unlabeled samples, as well as the scores of the positive samples are displayed in two tables on the screen. The 10 most contributing features, as evaluated using RDF, are displayed as a bar plot and comparisons between effectors and non-effectors for each of these features are displayed in violin plots. The full list of features and their contribution to the classification are available to download as a csv file, and so is the raw features file. Genomic ORFs with significant sequence similarity to T3SS proteins are displayed in a separate table on the screen along with their amino-acid sequence.

3 Results

3.1 Effectidor's predictions evaluation

We first evaluated Effectidor on *Xanthomonas citri* subsp. citri 306 (*X.citri* 306). To demonstrate the utility of Effectidor for identifying effectors in a newly sequenced genome, we did not provide the algorithm with a positive set of known effectors. Instead, we ran the algorithm without a known T3Es file. In such a case, the first step of the algorithm is to query each ORF against a dataset we established of previously published T3Es. This step identified 31 effector homologs in *X.citri* 306. These ORFs were subsequently considered as the positive set and were used as input to the ML algorithms implemented within Effectidor. A literature survey and a manual curation revealed that there are 34 known effectors in *X.citri* 306 (the full list of these effectors is provided in Supplementary Data S4). Following the ML step, Effectidor was able to discover all the *X.citri* effectors. The known effectors scores ranged from 0.998 to 0.534 while all the non-effectors had scores lower than 0.390, i.e. the algorithm could separate with confidence the known effectors from the rest of the ORFs and had neither false-positive predictions (non-effectors erroneously identified as effectors) nor false-negative predictions (effectors erroneously identified as non-effectors). These results were obtained using 0.5 as a threshold for identifying effectors (changing this threshold can affect classification results). The AUPRC score reflects inference precision averaging over all possible cutoffs. The AUPRC of Effectidor on these data was 1.0.

Several running configurations were tested to demonstrate the utility of Effectidor (see Supplementary Data S5). The above results were obtained using a fully assembled genome, which allows computing genomic-organization features as well as features extracted from the regulatory regions such as the existence of a PIP-box. These features may be (partially) missing when only an incomplete genome made of several contigs is available or when only cDNA data are provided. Running Effectidor excluding these features again yielded an AUPRC score of 1.0, albeit, with one effector scoring below 0.5 (the lowest scoring effector had a score of 0.415). Finally, we also excluded features that rely on comparisons to bacteria that do not encode a T3SS. In this case, Effectidor accuracy deteriorated with four false-negative and one false-positive predictions and an AUPRC score of 0.996. The results demonstrate the importance of including features reflecting similarity to bacteria that do not harbor T3SS.

We repeated the above analyses for *Ralstonia solanacearum* GMI1000 (*R.solanacearun* GMI1000) and *Citrobacter rodentium* ICC168 (*C.rodentium* ICC168). For these genomes, as in the former runs on *X.citri* 306, no T3Es input was provided for Effectidor. The results of these runs showed similar patterns to those obtained for *X.citri* 306 (all AUPRC scores above 0.98; see Supplementary Data S6 for full details).

Of note, the full running time of each of these runs was <15 min.

3.2 Comparisons with other web servers

Several web servers aiming to predict T3Es from bacterial genomic sequences have been previously published. We were able to run predictions on six previously published methods: BEAN 2.0 (Dong et al., 2015), pEffect (Goldberg et al., 2016), T3Sepp (Hui et al., 2020), EffectiveT3 (Arnold et al., 2009), Bastion 3 (Wang et al., 2019) and EP3 (which includes EP3_1 and EP3_2) (Li et al., 2021). In Supplementary Data S7, we list additional tools which were not functional during the work on this manuscript. Of note, the input to all these web servers is a single FASTA file of protein sequences, i.e. unlike Effectidor they cannot account for such features as regulatory elements or genomic organization. In addition, some of these web servers can only handle a limited number of input sequences: EP3 can handle up to 100 proteins, BEAN can handle up to 50 proteins, and pEffect, in practice was limited to 25 proteins. Lastly, some of these web servers yield only a binary classification without a confidence score, so the AUPRC score could not be calculated. Instead, we used the Matthews Correlation Coefficient (MCC) score. Effectidor outperformed all these web servers on the *X.citri* data. After Effectidor, the most accurate web server was Bastion 3 (AUPRC score of 0.938). The least accurate web server was EffectiveT3 with an AUPRC score of 0.066. Similar results were obtained for other bacteria. For *R.solanacearum*, the AUPRC of Effectidor was 0.985, while the next best performing web server was Bastion 3 with an AUPRC of 0.932. For *C.rodentium*, Effectidor had an AUPRC of 1.0, while the next best performing web server was T3Sepp with an AUPRC of 0.981 (see Supplementary Data S7, Supplementary Table S7b, for detailed comparisons).

4 Discussion

We hereby present Effectidor, a user-friendly web server that applies several ML algorithms to predict T3Es in bacterial genomes. It is the only web server that trains a different classifier tailored for the analyzed genome in every run. To our knowledge, Effectidor is the only web server that combines features from the analyzed genomic sequence, including both coding and non-coding regions. Moreover, it is the only web server, to our knowledge, which provides the extracted features for further analysis, in addition to the T3Es prediction. Lastly, the performance of Effectidor exceeds its competitors in prediction accuracy.

Funding

This work was supported in part by the Manna Center Program for Food Safety and Security at Tel Aviv University [fellowship to N.W.]; the Edmond J. Safra Center for Bioinformatics at Tel Aviv University [fellowships to N.W. and to O.A.]; and the Dalia and Eli Hurvits foundation [fellowship to O.A.].

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Arnold,R. *et al.* (2009) Sequence-based prediction of type III secreted proteins. *PLoS Pathog.*, **5**, e1000376.

Bent,A.F. and Mackey,D. (2007) Elicitors, effectors, and R genes: the new paradigm and a lifetime supply of questions. *Annu. Rev. Phytopathol.*, **45**, 399–436.

Bongrand,C. *et al.* (2012) Characterization of the promoter, MxiE box and 5′ UTR of genes controlled by the activity of the type III secretion apparatus in *Shigella flexneri*. *PLoS One*, **7**, e32862.

Brutinel,E.D. *et al.* (2009) Functional domains of ExsA, the transcriptional activator of the *Pseudomonas aeruginosa* type III secretion system. *J. Bacteriol.*, **191**, 3811–3821.

Burkinshaw,B.J. and Strynadka,N.C.J. (2014) Assembly and structure of the T3SS. *Biochim. Biophys. Acta Mol. Cell Res.*, **1843**, 1649–1663.

Burstein,D. *et al.* (2015) Novel type III effectors in *Pseudomonas aeruginosa*. *MBio*, **6**, e00161-15.

Dong,X. *et al.* (2015) BEAN 2.0: an integrated web resource for the identification and functional analysis of type III secreted effectors. *Database (Oxford)*, **2015**, bav064.

Goldberg,T. *et al.* (2016) Computational prediction shines light on type III secretion origins. *Sci. Rep.*, **6**, 34516.

Hobbs,C.K. *et al.* (2016) Computational approach to predict species-specific type III secretion system (T3SS) effectors using single and multiple genomes. *BMC Genomics*, **17**, 1048.

Hui,X. *et al.* (2020) T3SEpp: an integrated prediction pipeline for bacterial type III secreted effectors. *mSystems*, **5**, e00288-20.

Jalan,N. *et al.* (2013) Comparative genomic and transcriptome analyses of pathotypes of *Xanthomonas citri* subsp. *citri* provide insights into mechanisms of bacterial virulence and host range. *BMC Genomics*, **14**, 551.

Jiménez-Guerrero,I. *et al.* (2020) Show me your secret(ed) weapons: a multifaceted approach reveals a wide arsenal of type III-secreted effectors in the cucurbit pathogenic bacterium *Acidovorax citrulli* and novel effectors in the *Acidovorax* genus. *Mol. Plant Pathol.*, **21**, 17–37.

Khan,M. *et al.* (2016) Of guards, decoys, baits and traps: pathogen perception in plants by type III effector sensors. *Curr. Opin. Microbiol.*, **29**, 49–55.

Koebnik,R. *et al.* (2006) Specific binding of the *Xanthomonas campestris* pv. vesicatoria AraC-type transcriptional activator HrpX to plant-inducible promoter boxes. *J. Bacteriol.*, **188**, 7652–7660.

Krause,A. *et al.* (2002) Mutational and transcriptional analysis of the type III secretion system of *Bradyrhizobium japonicum*. *Mol. Plant Microbe Interact. MPMI*, **15**, 1228–1235.

Li,J. *et al.* (2021) EP3: an ensemble predictor that accurately identifies type III secreted effectors. *Brief. Bioinf.*, **22**, 1918–1928.

Mavris,M. *et al.* (2002) Identification of the cis-acting site involved in activation of promoters regulated by activity of the type III secretion apparatus in *Shigella flexneri*. *J. Bacteriol.*, **184**, 6751–6759.

Nissan,G. *et al.* (2018) Revealing the inventory of type III effectors in *Pantoea agglomerans* gall-forming pathovars using draft genome sequences and a machine-learning approach. *Mol. Plant Pathol.*, **19**, 381–392.

Pedregosa,F. *et al.* (2012) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Petersen,T.N. *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.

Pinaud,L. *et al.* (2018) Host cell targeting by enteropathogenic bacteria T3SS effectors. *Trends Microbiol.*, **26**, 266–283.

Ruano-Gallego,D. *et al.* (2021) Type III secretion system effectors form robust and flexible intracellular virulence networks. *Science*, **371**, eabc9531.

Teper,D. *et al.* (2016) Identification of novel *Xanthomonas euvesicatoria* type III effector proteins by a machine-learning approach. *Mol. Plant Pathol.*, **17**, 398–411.

Wagner,S. *et al.* (2018) Bacterial type III secretion systems: a complex device for the delivery of bacterial effector proteins into eukaryotic host cells. *FEMS Microbiol. Lett.*, **365**, fny201.

Wang,J. *et al.* (2019) Bastion3: a two-layer ensemble predictor of type III secreted effectors. *Bioinformatics*, **35**, 2017–2028.

Zwiesler-Vollick,J. *et al.* (2002) Identification of novel hrp-regulated genes through functional genomic analysis of the *Pseudomonas syringae* pv. tomato DC3000 genome. *Mol. Microbiol.*, **45**, 1207–1218.