EffectiveDB--updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems

Eichinger, Valerie; Nussbaumer, Thomas; Platzer, Alexander; Jehl, Marc-André; Arnold, Roland; Rattei, Thomas

DOI:
10.1093/nar/gkv1269

Published online 20 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D669–D674***
doi: 10.1093/nar/gkv1269

EffectiveDB—updates and novel features for a better annotation of bacterial secreted proteins and Type III, IV, VI secretion systems

Valerie Eichinger[1,†], Thomas Nussbaumer[1,†], Alexander Platzer[1,†], Marc-André Jehl[1], Roland Arnold[2] and Thomas Rattei[1,*]

[1]Division of Computational System Biology, Department of Microbiology and Ecosystem Science, University of Vienna, 1090 Vienna, Austria and [2]Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario M5G 1X8, Canada

Received September 16, 2015; Revised October 23, 2015; Accepted November 3, 2015

ABSTRACT

Protein secretion systems play a key role in the interaction of bacteria and hosts. EffectiveDB (http://effectivedb.org) contains pre-calculated predictions of bacterial secreted proteins and of intact secretion systems. Here we describe a major update of the database, which was previously featured in the NAR Database Issue. EffectiveDB bundles various tools to recognize Type III secretion signals, conserved binding sites of Type III chaperones, Type IV secretion peptides, eukaryotic-like domains and subcellular targeting signals in the host. Beyond the analysis of arbitrary protein sequence collections, the new release of EffectiveDB also provides a 'genome-mode', in which protein sequences from nearly complete genomes or metagenomic bins can be screened for the presence of three important secretion systems (Type III, IV, VI). EffectiveDB contains pre-calculated predictions for currently 1677 bacterial genomes from the EggNOG 4.0 database and for additional bacterial genomes from NCBI RefSeq. The new, user-friendly and informative web portal offers a submission tool for running the EffectiveDB prediction tools on user-provided data.

INTRODUCTION

Interactions between bacteria and eukaryotes are widespread in all ecosystems on earth and often lead to symbiotic relationships. The most prominent themes in current research are different types of human-microbe interactions, such as the interplay of human microbiomes with their host or human infections by bacterial pathogens. Understanding of bacterial interactions with other hosts, such as livestock animals and crop plants, are becoming crucial for sustaining nutrition and gaining renewable energy. Despite the major and fundamental progress in these fields due to novel molecular and computational methods, predictive modeling of complex host-microbe interactions is still limited (1). Among other challenges a better understanding of molecular mechanisms underlying microbe host interactions is needed.

Protein secretion is one of the major mechanisms for direct molecular interaction between bacteria and hosts and thus of fundamental importance. Several databases and web applications have been developed to search for proteins involved in bacterial protein secretion and for genomes encoding these. T346Hunter (2), SecReT4 (3) and SecReT6 (4) are specialized in the detection of bacterial secretion systems that are able to inject effector proteins directly into eukaryotic cells. T346Hunter (2) identifies core members of the Type III, IV and VI secretion systems by sequence similarity to conservation models. The tool provides the percentage of detected core components as utility for interpretation and estimating the potential functionality of secretion systems. SecReT4 (3) and SecReT6 (4) also focus on the known components of the secretion systems. The identified proteins can be obtained as a physical map. None of these three tools is able to make a binary decision whether a secretion system is intact or not. This decision, non-trivial due to the genetic flexibility and the limited knowledge about bacterial secretion systems (5), has to be made by the user. Considering the crucial role of presence and functionality of particular proteins for virulence mediated by protein secretion systems, as e.g. shown in (6), the role of computational predictions by generic models is to suggest and rank suitable candidates for further, more specific computational and experimental analysis.

Secreted protein sequences are mainly determined by their secretion signals. Among other tools, T3SEdb (7) and

*To whom correspondence should be addressed: Tel: +43 1 4277 76680; Fax: +43 1 4277 8 76680; Email: thomas.rattei@univie.ac.at
†These authors contributed equally to the paper as first authors.

BEAN 2.0 (8) combine various approaches to predict Type III secreted proteins, such as machine learning approaches, domain annotation or by exploiting information of the conserved genomic context with secretion system core genes. Also additional features are included, such as chaperone-binding sites that facilitate the specific binding of chaperones to one or several effectors (9). Type IV secreted proteins can be recognized by their C-terminal signal sequence (10). Four types of distinctive features, amino acid composition, dipeptide composition, position-specific scoring matrix composition and auto covariance transformation of position-specific scoring matrix, were used to develop the classifier T4EffPred (11). T4SEpre (12) contains multiple models representing C-terminal sequential and position-specific amino acid compositions, possible motifs and structural features.

Predictions of intact protein secretion systems and secreted bacterial proteins would ideally be included in genome annotations, as available from primary DNA sequence archives (13). However, no standards for secretion-related annotations have been established so far. Also the microbial genome re-annotation initiative of the NCBI RefSeq team does not include a specialized method for protein secretion prediction (14). Therefore specialized secondary databases are needed to collect, structure and present secretion-related data in the context of a rapidly growing number of published microbial sequence records. Among other resources, PATRIC (15) is a sophisticated database covering diverse fields of pathogenicity. It offers, for example, predictions of antibiotic resistance and of virulence-related proteins. Nevertheless, comprehensive predictions of bacterial secreted proteins and of intact protein secretion systems, based on specialized and continuously updated prediction methods, were so far not available from a single publicly available resource.

We have therefore updated and expanded EffectiveDB (16), a database that provides pre-calculated predictions of bacterial secreted proteins as well as online tools for detecting effectors by their Sec-dependent and Type III secretion signals, and by predicting eukaryotic-like domains (ELD) which are likely to interact with host proteins. In order to create a comprehensive 'one-stop-shop' for analyzing genomes of host interacting microbes, we have improved the secretion prediction methodology of EffectiveDB and expanded its scope towards predictions of intact Type III, IV and VI secretion systems. Owing to the burst of sequenced genomes and metagenomes, and the constant increase in available experimental data, EffectiveDB assists biologists in the systematic analysis of microbial genomes and in short-listing putative genes for experimental studies of host microbe interactions.

RESULTS

EffectiveDB is based on a suite of prediction programs for bacterial protein secretion. Pre-calculated predictions of secreted proteins and secretion systems resemble the core of EffectiveDB. While the previous release provided pre-calculations for only 1160 bacterial genomes, the updated version includes all 1677 bacterial genomes from EggNOG 4.0 (17) and for additional bacterial genomes from NCBI RefSeq (14). The EffectiveDB programs are also made available to the user for predictions of submitted protein sequences. According to the type of user input, the predictions are performed in 'protein mode', predicting secreted proteins for any collection of protein sequences, or in 'genome mode', also enabling the prediction of secretion systems and the discovery of novel ELD for proteins from nearly complete genomes (Figure 1). The EffectiveDB website also hosts documentation, software and data download, and supplementary information for all methods that have been specifically implemented for EffectiveDB.

EffectiveS346 facilitates the prediction of intact secretion systems

The 'EffectiveS346' software has been implemented for EffectiveDB in order to predict secretion systems encoded in bacterial genomes and to provide a clear 'yes/no' prediction on whether they are sufficiently complete or not. This method predicts Type III, IV and VI secretion systems, which are all able to inject eukaryotic cells and to directly transfer proteins into the host cell cytosol. We used the support vector machine (SVM) approach from the recently extended PICA framework (18,19) to create a classification model for each secretion system. Genotypes are represented by COG/NOG presence profiles from the EggNOG 4.0 database (17). The prediction model was trained with COGs and NOGs of bacterial strains encoding the intact secretion system as positive samples (Supplementary Tables S1–S3). Due to the lack of available experimental data indicating non-intact secretion systems we have randomly sampled the negative dataset from all remaining relevant genomes in EggNOG 4.0, excluding all genera that contain positive samples (Supplementary Tables S1–S3). To maximize the predictive power of the secretion system models, we only considered genomes with clear indications for the intact nature of their secretion system, such as contained in SEED (20), SecReT (3,4) AtlasT4SS (21) and own searches in publications. We considered only complete genomes by requiring the existence of at least 39 of 40 phylogenetic marker genes (22). The SVM uses the default values suggested by PICA: type:C = 5, Kernel = linear, gamma = 0 (19). The whole workflow of EffectiveS346 is shown in Supplementary Figure S1. Technical documentation about the PICA models in EffectiveS346, lists of COGs and NOGs ranked by their relevance for the secretion systems, as well as lists of positive and negative training data are provided in the EffectiveDB web portal.

The input data for EffectiveS346 are genotype files listing COGs and NOGs according to EggNOG 4.0, which are present in a particular genome. Genotype files may optionally provide a mapping of the COGs/NOGs to protein names. To our knowledge no public web service is available for the assignment of EggNOG orthologous groups to proteins in user-provided genomes. Therefore, in order to facilitate the genotype prediction for user-provided genomes, the EffectiveDB web portal contains software modules for this purpose. They require protein sequences from nearly complete genomes as input. The initial gene prediction is intentionally left to the user, as the correct prediction of translation initiation sites is crucial for the detection of N-terminal

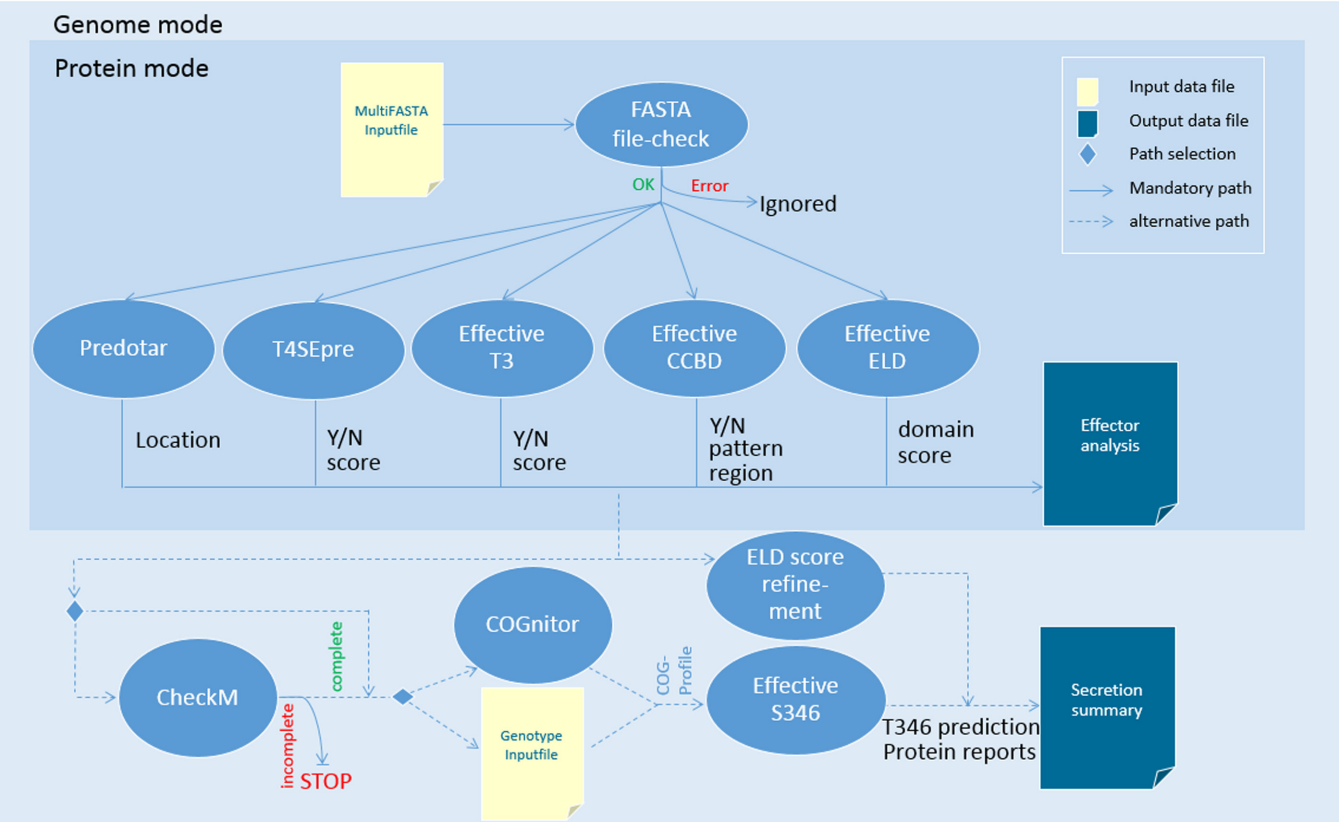

Figure 1. Integration of methods for the prediction of bacterial secreted proteins and intact secretion systems in EffectiveDB. The workflow depicts the protein and genome modes of EffectiveDB. In the protein mode any set of proteins can be analyzed. The genome mode extends the protein mode by enabling EffectiveS346. For proteins from (almost) complete genomes the orthologous groups are calculated or provided by the user. These are the input data for the prediction of intact Type III, IV and VI secretion system.

signal peptides in other EffectiveDB methods. For any user-submitted FASTA file an optional check of genome completeness by CheckM (23) is provided. According to the prediction accuracy of PICA models for incomplete genomes we recommend that at a major fraction (default = 85%) of the marker genes should be present in order to obtain reliable secretion system predictions. The COGnitor program (24) predicts orthologous groups which serve as input for EffectiveS346. Alternatively, the user may submit own lists of EggNOG 4.0 COGs and NOGs present in a genome. In this case, the time-consuming homology search by COGnitor is omitted.

With the genotype list as input, the three SVM models calculate binary classifications whether the input species contains intact Type III, IV and VI secretion systems. These predictions feature a mean balanced accuracy of over 90% and a standard deviation of below 4.5% (Supplementary Tables S4–S7). Nevertheless, we showed that at least for the Types III and IV, in only 85% complete genomes more than 83% of the intact secretion systems could still be recognized. EffectiveS346 additionally provides lists of the 100 most important COGs in regard to each classification and lists of those COGs that are contained in the KEGG maps of the three secretion systems. From proteins of COGs, which are associated with the secretion systems in KEGG, EffectiveDB estimates the copy number of predicted secretion systems. These copy numbers are shown in the result summaries as well as in the detailed output files, combined with lists of the respective protein names. If protein names represent locus tags, these lists are useful to infer putatively intact (complete) clusters of secretion system genes and to their respective genomic regions. Among the 1677 pre-calculated bacterial genomes from EggNOG 4.0 we have predicted 164 intact Type III, 266 intact Type IV and 247 intact Type VI secretion systems. The Supplementary Figures S2–S7, created with Krona (25), visualize the taxonomic distribution of genomes with and without the three secretion systems. Genome contents and prediction overlaps with T346Hunter (2), SecReT4 (3) and SecReT6 (4) are given in Supplementary Figure S8 and Table S8, respectively.

EffectiveT3, EffectiveCCBD and T4SEpre predict Type III and Type IV secreted proteins

The prediction of Type III secreted proteins in EffectiveDB was improved and extended. An updated version of EffectiveT3 (26) facilitates the recognition of N-terminal signal peptides. For the update we have assembled new training datasets, combining 504 verified secreted proteins from T3SEdb (7) along with our original training data (26). The new model is also a Naive Bayesian Classifier, trained with more data. Sequence similarity based elimination of redundancy, creation of features, selection of the most discrim-

inating features, learning and testing procedure were performed as described initially (26). Training and classifying was performed with the Weka package (27). When performing a leave-one-out cross validation test, this yielded in an accuracy of 0.87 that is comparable to our previous report (26). In addition, a leave-one-taxon-out test was applied to prove that the model is still based on ubiquitous features of the signal and can thereby be applied to any taxon. In this test all proteins from one taxon are kept out from the training and are then exclusively used as test data. Overall, an average area under the curve (AUC) of 0.80 was obtained. The new model is now embedded into EffectiveDB and also available as EffectiveT3 module for download. A comparison between the performance of the old and new model, calculated as receiver operating curves, is shown in Supplementary Figure S9. All training data are provided on the website and in Supplementary Table S9. We run also BEAN 2.0 (8) on our new data. The results are similar and shown in Supplementary Figure S9 and Table S10. The default minimal score from the Naive Bayesian Classifier for the class 'secreted' is 0.9999 in the new model. This default value is called 'selective' at the webpage, whereas 0.95 is called 'sensitive'. The threshold can also be freely chosen.

Complementing the signal peptide based prediction of Type III secreted proteins we have integrated a novel method for class IB chaperone prediction. Those chaperones facilitate the correct selection and unfolding of Type III dependent effector proteins (28). A commonly shared sequence within the region of the 70 N-terminal residues of Type III secreted proteins was shown to serve as binding site of the chaperones. This 'conserved chaperone-binding domain' (CCBD) follows the explicit pattern: $(LMIF)_1XXX(IV)_5XX(IV)_8X(N)_{10}$ (29). We implemented EffectiveCCBD, allowing to compare any given collection of protein sequences against this motif by using Biopython (30). The output informs the user whether the pattern was found within the expected region (26–70 amino acids from the N-terminus) or in the surrounding regions (1–25 and 71–150 amino acids from the N-terminus). Across the 1677 pre-calculated bacterial genomes we have observed fewer positive predictions by EffectiveCCBD (49 543) than by EffectiveT3 (361 189). For 8651 proteins both programs agree in their positive prediction, whereas 2 165 946 proteins are consistently predicted as not secreted.

For the prediction of Type IV secreted proteins we have integrated the program T4SEpre (12), which only requires amino acid sequences as input and is thus compatible with the other methods for secreted protein prediction in EffectiveDB. Due to the very high computational costs of the T4SEpre model based on protein secondary structure (Sse), we have only used the sequence-based models T4SEpre_psAac and T4SEpre_bpbAac. We used the published databases to calculate the amino acid properties and bi-residue properties with therein provided methods.

EffectiveELD predicts secreted proteins based on eukaryotic-like domains, independently from the mode of transport

The implementation of EffectiveELD, which predicts secreted proteins based on ELD, was not changed since the initial version of EffectiveDB. However, besides the update of the genome repository and the protein domain database we have changed the presentation of ELD in the EffectiveDB web portal. Mean and standard deviation of the domain frequency in not host-associated genomes are now shown and can be exported into different file formats. This is mainly relevant for the analysis of proteins from metagenomic samples. Metagenome assembly artifacts may artificially increase the copy number of typically single-copy non-effector genes, such as as house-keeping genes. In these cases, the reported Z-score would indicate significant enrichment of such genes, which are certainly not effectors. This type of false positive matches can now be easily detected and excluded from further analysis.

In the 'protein mode' of EffectiveDB, analyzing arbitrary collections of protein sequences, only ELD with significant enrichment in at least one host-associated genome from the EffectiveDB genome repository are reported. In the new 'genome mode' of EffectiveDB, the Z-scores for the enrichment of ELD are automatically calculated *de novo* for all protein domains occurring in eukaryotic genomes. This allows the prediction of novel ELD that have not yet been observed in any of the host-associated genomes from the EffectiveDB genome repository.

Predotar predicts subcellular targeting of effectors in the host cell

Multiple evidences suggest organelles as targets of bacterial secreted proteins (reviewed e.g. in 31 with regard to mitochondria). The program Predotar (32) is a tool that allows to rapidly screen N-terminal targeting sequences and to predict their subcellular localization in eukaryotic host cells. We have included this tool into EffectiveDB in order to facilitate downstream analysis of putative secreted proteins due to their predicted cellular location in the host. Predotar is used for both, the protein and the genome mode of EffectiveDB.

EffectiveDB integrates predictions and Supplementary Data in an improved website

A new website has been implemented for EffectiveDB. It provides interactive access to the pre-calculated predictions of secreted proteins and secretion systems. A submission form is provided to specify input data and parameters for EffectiveDB calculations of user-submitted data. No data from these interactive calculations are stored in EffectiveDB. Results are only provided to the owner of a submission and are deleted after one month. All predictions are presented to the user in an integrative manner: equivalent predictions, such as all predictions based on an individual protein sequence, are grouped (Figure 2). For the genome-based predictions by EffectiveS346 the web site provides a summary as well as the detailed lists of the most relevant proteins for each secretion system.

OUTLOOK

We will continue to update EffectiveDB on a yearly basis. In consideration of promising current research in other teams we expect that further signal peptide based predictions of type IV secreted proteins can be added in the next

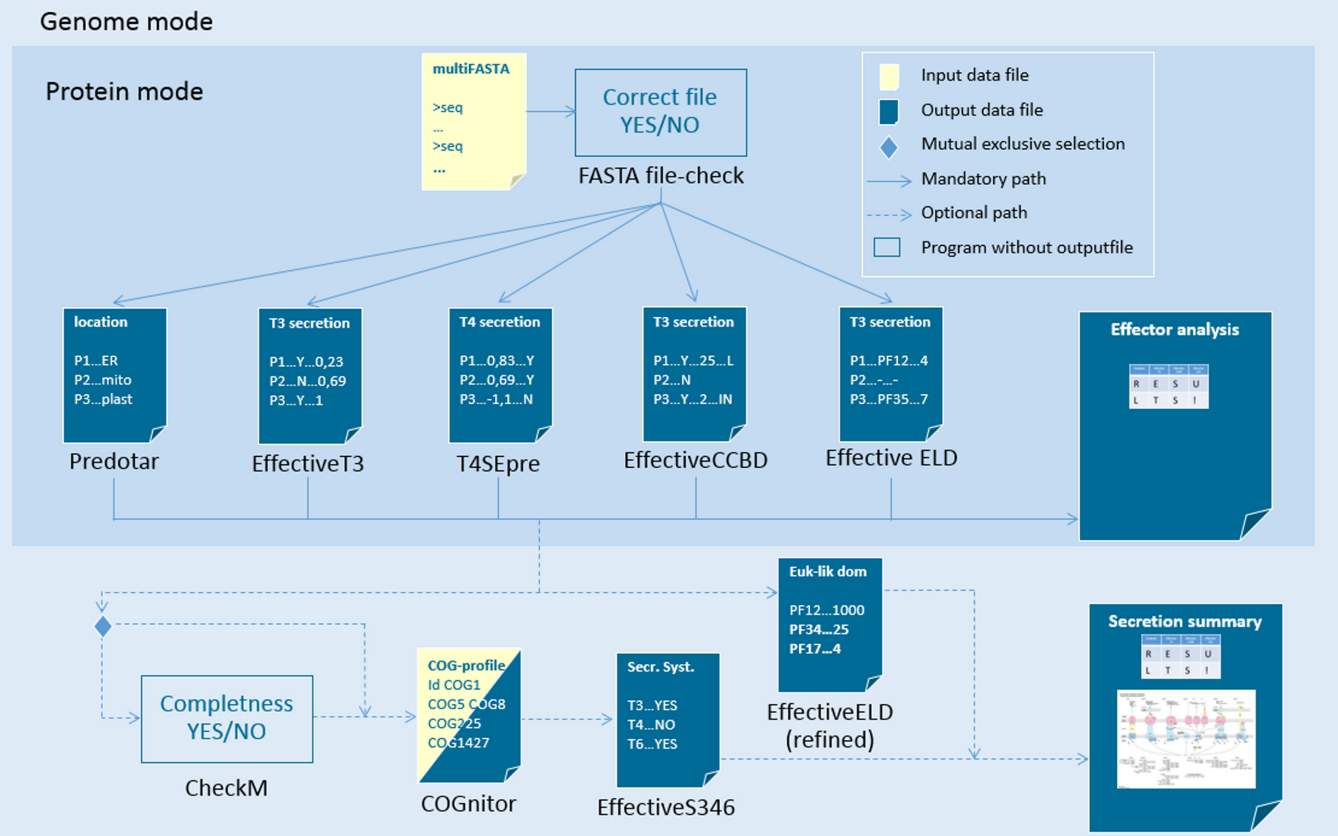

Figure 2. Schematic output of EffectiveDB. The EffectiveDB output consists of predictions for individual proteins and for the entirety of proteins in a complete genome. Protein-based predictions (Type III secretion peptides, Type III chaperone binding sites, Type IV secretion peptides, eukaryotic-like domains and subcellular targeting) are grouped. Genome-based results are provided as summary and as lists of most relevant orthologous groups for each secretion system.

release. The pre-calculation of secretion systems in complete genomes is linked to the EggNOG database, which provides the genotype lists. Any future update of EggNOG will therefore also trigger an update and significant extension of EffectiveDB. We will continue to incorporate additional bacterial genomes from NCBI RefSeq, which are not yet contained in EggNOG.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We thank Professor Ian Small from The University of Western Australia and Dr Claire Lurin from the Unité de Recherche en Genomique Végétale (URGV) for providing and helping us with the integration of Predotar into our web server. We are grateful to three anonymous referees for their constructive suggestions.

FUNDING

Austrian Science Fund (FWF) [I 2191 to T.R.]. Funding for open access charge: Austrian Science Fund (FWF) [I 2191 to T.R.].

Conflict of interest statement. None declared.

REFERENCES

1. Ji,B. and Nielsen,J. (2015) From next-generation sequencing to systematic modeling of the gut microbiome. *Front. Genet.*, **6**, 219.
2. Martinez-Garcia,P.M., Ramos,C. and Rodriguez-Palenzuela,P. (2015) T346Hunter: a novel web-based tool for the prediction of type III, type IV and type VI secretion systems in bacterial genomes. *PLoS One*, **10**, e0119317.
3. Bi,D., Liu,L., Tai,C., Deng,Z., Rajakumar,K. and Ou,H.Y. (2013) SecReT4: a web-based bacterial type IV secretion system resource. *Nucleic Acids Res.*, **41**, D660–D665.
4. Li,J., Yao,Y., Xu,H.H., Hao,L., Deng,Z., Rajakumar,K. and Ou,H.Y. (2015) SecReT6: a web-based resource for type VI secretion systems found in bacteria. *Environ. Microbiol.*, **17**, 2196–2202.
5. Costa,T.R., Felisberto-Rodrigues,C., Meir,A., Prevost,M.S., Redzej,A., Trokter,M. and Waksman,G. (2015) Secretion systems in Gram-negative bacteria: structural and mechanistic insights. *Nat. Rev. Microbiol.*, **13**, 343–359.
6. Pukatzki,S., Ma,A.T., Sturtevant,D., Krastins,B., Sarracino,D., Nelson,W.C., Heidelberg,J.F. and Mekalanos,J.J. (2006) Identification of a conserved bacterial protein secretion system in Vibrio cholerae using the Dictyostelium host model system. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 1528–1533.
7. Tay,D.M., Govindarajan,K.R., Khan,A.M., Ong,T.Y., Samad,H.M., Soh,W.W., Tong,M., Zhang,F. and Tan,T.W. (2010) T3SEdb: data warehousing of virulence effectors secreted by the bacterial Type III Secretion System. *BMC Bioinformatics*, **11**(Suppl. 7), S4.

8. Dong,X., Lu,X. and Zhang,Z. (2015) BEAN 2.0: an integrated web resource for the identification and functional analysis of type III secreted effectors. *Database (Oxford)*, **2015**, bav064.
9. Birtalan,S.C., Phillips,R.M. and Ghosh,P. (2002) Three-dimensional secretion signals in chaperone-effector complexes of bacterial pathogens. *Mol. Cell*, **9**, 971–980.
10. McDermott,J.E., Corrigan,A., Peterson,E., Oehmen,C., Niemann,G., Cambronne,E.D., Sharp,D., Adkins,J.N., Samudrala,R. and Heffron,F. (2011) Computational prediction of type III and IV secreted effectors in gram-negative bacteria. *Infect. Immun.*, **79**, 23–32.
11. Zou,L., Nan,C. and Hu,F. (2013) Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics*, **29**, 3135–3142.
12. Wang,Y., Wei,X., Bao,H. and Liu,S.L. (2014) Prediction of bacterial type IV secreted effectors by C-terminal features. *BMC Genomics*, **15**, 50.
13. Nakamura,Y., Cochrane,G., Karsch-Mizrachi,I. and International Nucleotide Sequence Database, C. (2013) The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.*, **41**, D21–D24.
14. Tatusova,T., Ciufo,S., Fedorov,B., O'Neill,K. and Tolstoy,I. (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, **42**, D553–D559.
15. Wattam,A.R., Abraham,D., Dalay,O., Disz,T.L., Driscoll,T., Gabbard,J.L., Gillespie,J.J., Gough,R., Hix,D., Kenyon,R. *et al.* (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, **42**, D581–D591.
16. Jehl,M.A., Arnold,R. and Rattei,T. (2011) Effective–a database of predicted secreted bacterial proteins. *Nucleic Acids Res.*, **39**, D591–D595.
17. Powell,S., Forslund,K., Szklarczyk,D., Trachana,K., Roth,A., Huerta-Cepas,J., Gabaldon,T., Rattei,T., Creevey,C., Kuhn,M. *et al.* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.*, **42**, D231–D239.
18. MacDonald,N.J. and Beiko,R.G. (2010) Efficient learning of microbial genotype-phenotype association rules. *Bioinformatics*, **26**, 1834–1840.
19. Feldbauer,R., Schulz,F., Horn,M. and Rattei,T. (2015) Prediction of microbial phenotypes based on comparative genomics. *BMC Bioinformatics*, **16**(Suppl. 14), S1.
20. Overbeek,R., Begley,T., Butler,R.M., Choudhuri,J.V., Chuang,H.Y., Cohoon,M., de Crecy-Lagard,V., Diaz,N., Disz,T., Edwards,R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**, 5691–5702.
21. Souza,R.C., del Rosario Quispe Saji,G., Costa,M.O., Netto,D.S., Lima,N.C., Klein,C.C., Vasconcelos,A.T. and Nicolas,M.F. (2012) AtlasT4SS: a curated database for type IV secretion systems. *BMC Microbiol.*, **12**, 172.
22. Mende,D.R., Sunagawa,S., Zeller,G. and Bork,P. (2013) Accurate and universal delineation of prokaryotic species. *Nat. Methods*, **10**, 881–884.
23. Parks,D.H., Imelfort,M., Skennerton,C.T., Hugenholtz,P. and Tyson,G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.
24. Kristensen,D.M., Kannan,L., Coleman,M.K., Wolf,Y.I., Sorokin,A., Koonin,E.V. and Mushegian,A. (2010) A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics*, **26**, 1481–1487.
25. Ondov,B.D., Bergman,N.H. and Phillippy,A.M. (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics*, **12**, 385.
26. Arnold,R., Brandmaier,S., Kleine,F., Tischler,P., Heinz,E., Behrens,S., Niinikoski,A., Mewes,H.W., Horn,M. and Rattei,T. (2009) Sequence-based prediction of type III secreted proteins. *PLoS Pathog.*, **5**, e1000376.
27. Hall,M., Frank,E., Holmes,G., Pfahringer,B., Reutemann,P. and Witten,I.H. (2009) The WEKA Data Mining Software: an Update. *SIGKDD Explorations*, **11**, 10–18.
28. Thomas,N.A., Ma,I., Prasad,M.E. and Rafuse,C. (2012) Expanded roles for multicargo and class 1B effector chaperones in type III secretion. *J. Bacteriol.*, **194**, 3767–3773.
29. Costa,S.C., Schmitz,A.M., Jahufar,F.F., Boyd,J.D., Cho,M.Y., Glicksman,M.A. and Lesser,C.F. (2012) A new means to identify type 3 secreted effectors: functionally interchangeable class IB chaperones recognize a conserved sequence. *MBio*, **3**, doi:10.1128/mBio.00243-11.
30. Cock,P.J., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
31. Lobet,E., Letesson,J.J. and Arnould,T. (2015) Mitochondria: a target for bacteria. *Biochem. Pharmacol.*, **94**, 173–185.
32. Small,I., Peeters,N., Legeai,F. and Lurin,C. (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.