Protocol for detection of bacterial proteins involved in efflux mediated antibiotic resistance (ARE) and their sub-families

Deeksha Pandey
Department of Biophysics, University of Delhi South Campus
Bandana Kumari
Department of Biophysics, University of Delhi South Campus
Neelja Singhal
Department of Biophysics, University of Delhi South Campus
Manish Kumar (manish@south.du.ac.in)
Department of Biophysics, University of Delhi South Campus https://orcid.org/0000-0002-7936-9892

Method Article

Keywords: Bacterial Efflux Proteins, Antibiotic Resistance, In-silico Tool, Machine Learning

Posted Date: March 3rd, 2021

DOI: https://doi.org/10.21203/rs.3.pex-1371/v1

Abstract

This protocol describes a method for detection of bacterial proteins involved in efflux mediated antibiotic resistance (ARE) and their sub-families as described in the research paper entitled "**BacEffluxPred: A two-tier system to predict and categorize bacterial efflux mediated antibiotic resistance proteins"** published in Scientific Reports. BacEffluxPred is a support vector machine based two-tier prediction method, that can be used for the detection of efflux proteins responsible for antibiotic resistance in bacteria and to identify the families to which it belongs. The overall prediction cycle includes three important steps:

1) The query protein is presented to the prediction algorithm.

2) If the query protein would be predicted to be a non-ARE protein, the prediction would stop at tier-I.

3) If the query protein would be predicted as an ARE protein at the tier-I, the query protein would be forwarded to tier-II for ARE family prediction.

By using these steps it is possible to generate the models that can be used on proteomic data to predict whether the given data have potential ARE proteins or not if yes it will further classified into their following families. This is the first *in-silico* tool for predicting bacterial ARE proteins and their families and it is freely available as both web-server and standalone versions at ***http://proteininformatics.org/mkumar/baceffluxpred/***

Introduction

Efflux proteins are responsible for transportation of different types of molecules from inside of the cell to the outside environment. Many times efflux proteins are also capable to carry out one or more type of antibiotics. The efflux proteins are responsible for multi-drug resistance in many microbial pathogens [1–5]. In the past several highly accessed and useful antibiotic resistance databases have been established to catalogue the known antibiotic resistance genes at both the whole genome as well as at genes/proteins levels [6–15]. But an *in-silico* tool to predict and annotate efflux proteins responsible for antibiotic resistance (ARE) has not been developed yet.

As per the WHO list of priority pathogens *(https://www.who.int/news/item/27-02-2017-who-publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed)* several pathogens use efflux to attain resistance against antibiotics. Therefore, an efficient tool for predicting the ARE proteins and their family is urgently needed. This protocol describes a machine learning based two-tier *in-silico* tool, named as BacEffluxPred, to predict the bacterial ARE proteins and classify them into their respective families. The prediction cycle of BacEffluxPred completes into two tiers. Tier-I discriminates between ARE and non-ARE however Tier-II predicts the family of classified ARE proteins. Prediction capability of BacEffluxPred has also been evaluated using an independent dataset that showed a very good performance. The details of

the dataset, training methodology and other details of BacEffluxPred can be obtained from Pandey et al[16].

Overview of BacEffluxPred

BacEffluxPred is a web-server to predict efflux proteins that are also involved in antibiotic resistance. It also classifies the predicted efflux proteins into their respective efflux protein families. For prediction protein sequences in FASTA format are required. First the submitted protein sequence will be converted into numerical encoding in the form of Position Specific Scoring Matrix (PSSM). Using PSSM as an input, the BacEffluxPred SVM models predict the ARE protein (tier-I) and classify into their relevant family of efflux proteins at tier-II.

The final outcome of BacEffluxPred depends on the user-selected threshold. Higher thresholds would result in more specific predictions i.e. low false positives, while lower threshold would result in low specificity predictions i.e. more false-positive results. The complete workflow of BacEffluxPred is shown in Figure 1.

Application of the Protocol

Recent advances in DNA technology and the advent of high throughput genomic technologies have led to the identification of a large number of new efflux pump/proteins. Since efflux proteins are one of the factors behind emergence of multi-drug resistance (MDR) in microbial pathogens, hence, BacEffluxPred can be used for annotation of novel efflux based antibiotic resistance genes in a bacterial proteome/genome.

Reagents

Protein sequence(s) for prediction.

Equipment

Hardware

No specific hardware configuration is recommended.

Software

- Linux or Unix operating system
- SVM_{Light} *(http://download.joachims.org/svm_light/current/svm_light.tar.gz)*

 PSI-BLAST *(ftp://ftp.ncbi.nlm.nih.gov/blast/executables/)*

 Non-redundant protein database of NCBI after reducing redundancy to 90%

Procedure

Procedure

Development of the Protocol

With the aim to reduce the requisite time and resources for classification, identification and characterization of ARE proteins we came up with the novel *in-silico* tool BacEffluxPred. It is a two-tier SVM based method that can do quick and highly accurate prediction of bacterial efflux proteins responsible for AR and also predict the efflux protein family to which a predicted ARE may belong.

Implementation of BacEffluxPred Web-server

The BacEffluxPred web-server is hosted on a Linux operating system. Web pages were designed using HTML/CSS scripting languages. The Perl language is used for the back-end pipeline and gcc compiler was used for compilation of all the programs.

Web-server usage guide

Step 1: At the top of the HOME page of **BacEffluxPred** (http://proteininformatics.org/mkumar/baceffluxpred), links to other pages are present. In the middle of the HOME page a brief introduction about the efflux based AR mechanism and link to detailed information of each family can be found. At the end of the page the working schema of BacEffluxPred tool and performance of different prediction models during training is shown in the form of ROC plot.

Step 2: The protein sequences can be submitted on the submission page by either typing/pasting or uploading the sequence file. The input sequence(s) must be in FASTA format and must contain only 20 standard single letter codes for amino acids in either upper or lower case. At present the BacEffluxPred web-server can predict only five sequences at a time. In case more than five sequences are submitted, only the first five sequences will be processed. The result of prediction depends on the SVM threshold that can be changed by the user (the default value is -0.4). Selection of high threshold will result in prediction with high specificity and less sensitivity/false positives while low threshold would result in prediction with low specificity with high sensitivity/false positives.

Installation of standalone version of BacEffluxPred

For Bulk analysis of proteins or for whole proteome analysis we recommend the use of a standalone version of BacEffluxPred, which can be downloaded from ***http://proteininformatics.org/mkumar/baceffluxpred***. In the standalone version of BacEffluxPred, there is

no limit of the number of sequences that can be predicted and annotated. The benchmark data sets that were used to train and test BacEffluxPred predictor can also be downloaded from the download page.

Usage of standalone versions of BacEffluxPred requires a little familiarity with Linux. To help a Linux novice we also provided a detailed help and tutorial page to guide through installation and running the program (*http://proteininformatics.org/mkumar/baceffluxpred/help.html*) is also provided.

Following steps are required to install the standalone version of BacEffluxPred:

1. Download the standalone version of BacEffluxPred.

2. Unpack "baceffluxpred.tar.gz" using command: **tar -xvf baceffluxpred.tar.gz**

3. Download Blast package from *ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release* and install it to run PSI-BLAST into BacEffluxPred installation folder.

4. Define the path for 'blastpgp of Blast package' and the 'database' at the end of the psiblast_pssm.pl (specified as $path_blastpgp and $path_database).

5. Download and install the SVM_{Light} package using the link: *http://download.joachims.org/svm_light/current/svm_light.tar.gz* in directory of BacEffluxPred. Extract the content of tar file (svm_light.tar) and copy the program svm_classify into the BacEffluxPred folder.

Troubleshooting

In case the standalone version of BacEffluxPred does not work, verify (a) the BacEffluxPred folder contained all required programs like SVM_{Light}, BLAST, NR protein database, (b) rechecks whether path of the all programs are mentioned correctly, and (c) the input sequences are in FASTA format. While performing prediction on BacEffluxPred web-server, the reason for an error might be due to (a) sequences are uploaded as Word files, or (b) sequences are not in FASTA format.

Time Taken

A typical prediction cycle for five protein sequences normally takes five minutes to complete on the BacEffluxPred web-server. As per our benchmarking on a computer with Intel(R) Xeon(R) 4 Core E5507 2.27 GHz processor with 6 GB DDR4 RAM, running 64-bit Red Hat Enterprise Linux operating system (Release 6.2), it took one minute per sequence to do the prediction.

Anticipated Results

The prediction result of BacEffluxPred web-server is presented in a tabular format. It displays the result of prediction in two columns (Figure 2). The First column shows the ID of query protein and the second

column displays the prediction result of BacEffluxPred. The prediction results depend on the SVM prediction threshold that the user has selected during the job submission.

Limitations

The input feature of BacEffluxPred is based on PSSM constructed during PSI-BLAST search. Though we have used 90% non-redundant NR protein database of NCBI for the PSI-BLAST search, still it takes slightly longish time to complete a prediction cycle.

Figure legends

Figure-1: Steps involved in development of BacEffluxPred. The development of BacEffluxPred involves six major stages, (a) Data curation and compilation: 'non-antibiotic resistance efflux (Non-ARE), non-efflux antibiotic resistance (Non-EAR), non-efflux prokaryotic and antibiotic resistance efflux (ARE)' protein sequences were retrieved from UniProtKB, Patric, and BacARscan database. These proteins were divided into positive and negative datasets depending upon their capability to efflux out the antibiotics (Tier-I) and the family to which efflux protein belongs (Tier-II); (b) Feature encoding in which the position-specific scoring matrix (PSSM), generated by PSI-BLAST search against 90% redundant NR protein, was used to encode each protein sequence; (c) Using the feature encoded negative and positive datasets support vector machine models were trained using various parameters and kernel features; (d) Performance of each SVM model was evaluated in terms of sensitivity, specificity, accuracy and MCC using leave-one-out cross-validation and the best performing SVM model was also evaluated using independent test dataset; (e) Complete working schema of BacEffluxPred; (f) Classification schema of BacEffluxPred predictions on the basis of actual and prediction state.

Figure-2: Screenshots of the web pages of the BacEffluxPred web-server. (a) The home page that gives brief description about the efflux mechanism and BacEffluxPred web-server; (b) The submission page where the user can submit the query protein for prediction; (c) The help page where user can build an understanding of using the tool; (d) The download page where the user can download different datasets and standalone package of BacEffluxPred.

References

1. Wright, G. D. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat. Rev. Microbiol.* **5**, 175–186 (2007).

2. Marquez, B. Bacterial efflux systems and efflux pumps inhibitors. *Biochimie* **87**, 1137–1147 (2005).

3. Paulsen, I. T., Sliwinski, M. K. & Saier, M. H. Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol.* **277**, 573–592 (1998).

4. Nikaido, H. & Pagès, J.-M. Broad-specificity efflux pumps and their role in multidrug resistance of Gram-negative bacteria. *FEMS Microbiol. Rev.* **36**, 340–363 (2012).

5. Li, X.-Z. & Nikaido, H. Efflux-mediated drug resistance in bacteria: an update. *Drugs* **69**, 1555–1623 (2009).

6. Kumar, R., Srivastava, A., Kumari, B. & Kumar, M. Prediction of β-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. *J. Theor. Biol.* **365**, 96–103 (2015).

7. Srivastava, A., Kumar, R. & Kumar, M. BlaPred: Predicting and classifying β-lactamase using a 3-tier prediction system via Chou's general PseAAC. *J. Theor. Biol.* **457**, 29–36 (2018).

8. Arango-Argoty, G. *et al.* DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 23 (2018).

9. Pesesky, M. W. *et al.* Evaluation of Machine Learning and Rules-Based Approaches for Predicting Antimicrobial Resistance Profiles in Gram-negative Bacilli from Whole Genome Sequence Data. *Front. Microbiol.* **7**, 1887 (2016).

10. Chowdhury, A. S., Call, D. R. & Broschat, S. L. Antimicrobial Resistance Prediction for Gram-Negative Bacteria via Game Theory-Based Feature Evaluation. *Sci. Rep.* **9**, 14487 (2019).

11. Kim, J. *et al.* VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. *PLoS Comput. Biol.* **16**, e1007511 (2020).

12. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).

13. McArthur, A. G. *et al.* The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* **57**, 3348–3357 (2013).

14. Gibson, M. K., Forsberg, K. J. & Dantas, G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2015).

15. Srivastava, A., Singhal, N., Goel, M., Virdi, J. S. & Kumar, M. CBMAR: a comprehensive β-lactamase molecular annotation resource. *Database J. Biol. Databases Curation* **2014**, bau111 (2014).

16. Pandey, D., Kumari, B., Singhal, N. & Kumar, M. BacEffluxPred: A two-tier system to predict and categorize bacterial efflux mediated antibiotic resistance proteins. *Sci. Rep.* **10**, 9287 (2020).

Acknowledgements

DP is supported by the Department of Science and Technology Govt. of India (INSPIRE Program), (DST INSPIRE Fellowship/2016/IF160262 [Grant Number: DST/INSPIRE 03/2015/003022]. BK was a recipient of ICMR-SRF (Grant Number: BIC/11(33)/2014). NS is supported by CSIR Senior Research Associate-ship (Scientist's Pool Scheme) [Grant Number: 13(9089-A)/2019-Pool]. All authors thank University of Delhi South Campus, New Delhi (India) for providing facilities to pursue the research work.

Figures

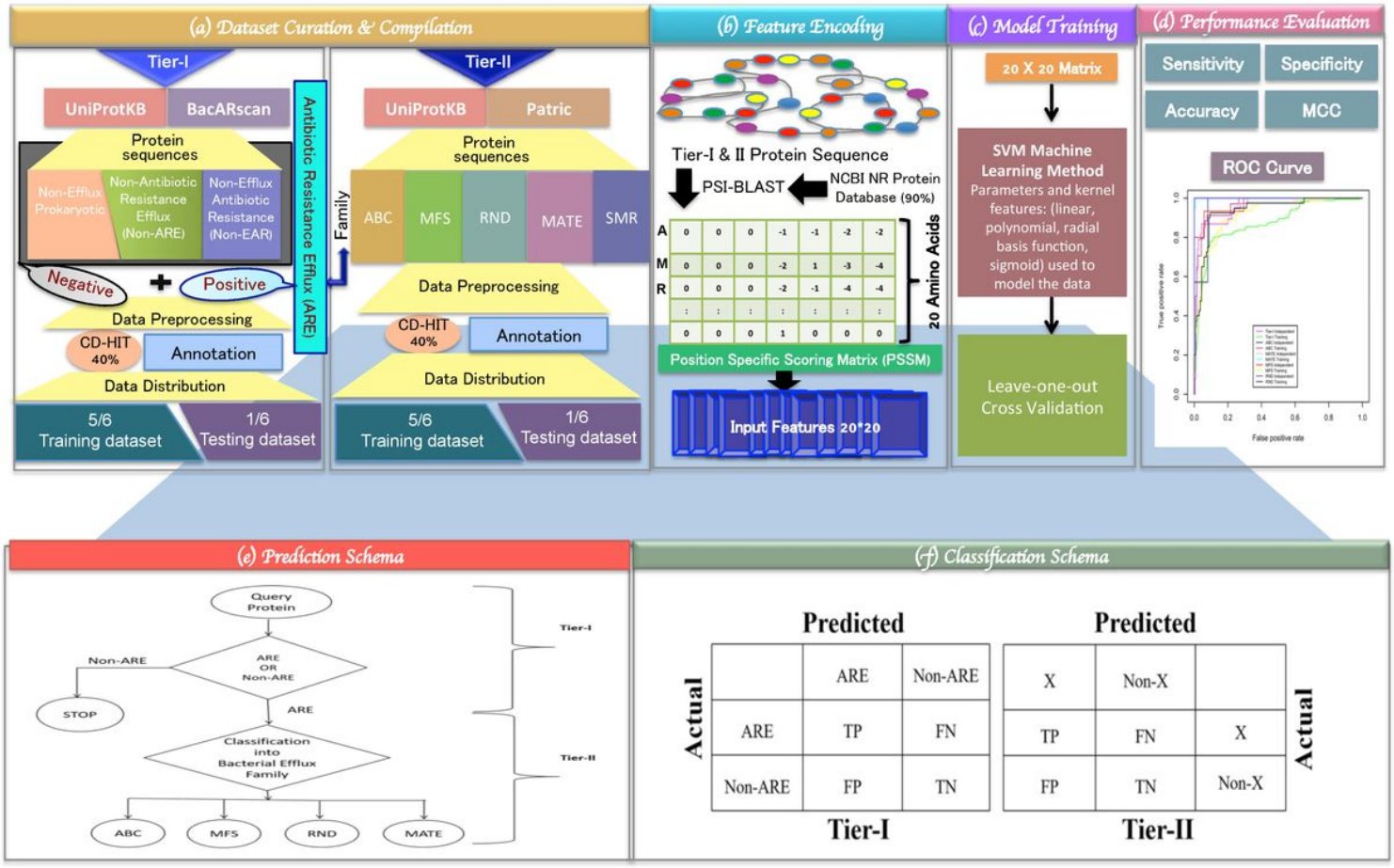

Figure 1

Steps involved in development of BacEffluxPred. The development of BacEffluxPred involved six major steps, (a) Data curation and compilation: 'non-antibiotic resistance efflux (Non-ARE), non-efflux antibiotic resistance (Non-EAR), non-efflux prokaryotic and antibiotic resistance efflux (ARE)' protein sequences were retrieved from UniProtKB, Patric, and BacARscan database. These proteins were divided into positive and negative datasets depending upon their capability to efflux the antibiotics (tier-I) and the family to which efflux protein belongs (tier-II); (b) Feature encoding in which the position-specific scoring matrix (PSSM), generated by PSI-BLAST search against 90% redundant NR protein database, was used to encode each protein sequence; (c) Using the feature encoded negative and positive datasets support

vector machine models were generated using various learning parameters and kernel features; (d) Performance of each SVM model was evaluated in terms of sensitivity, specificity, accuracy and MCC using leave-one-out cross-validation and the best performing SVM model was also evaluated using independent test dataset; (e) Complete working schema of BacEffluxPred; (f) Classification schema of BacEffluxPred predictions on the basis of actual and predicted state.

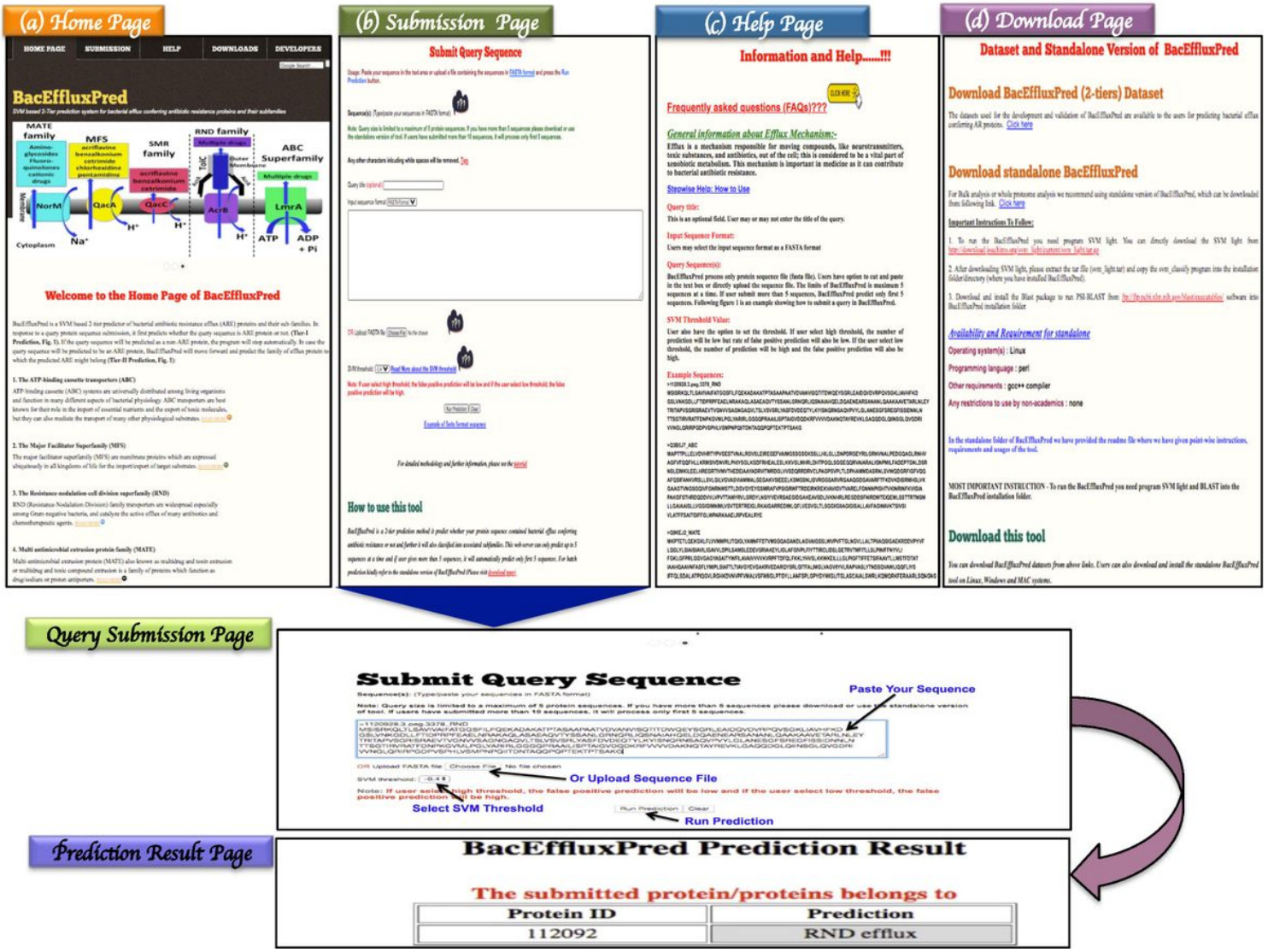

Figure 2

Screenshots of the web pages of the BacEffluxPred web-server. (a) The home page provides brief description about the efflux mechanism and BacEffluxPred web-server; (b) The submission page where the user can submit the query protein for prediction; (c) The help page; (d) The download page where the user can download different datasets and standalone version of BacEffluxPred.