Open resource of bacterial biofilm for the prediction of biofilm-associated proteins

Zhiyuan Zhang[1, 3, *], Wajid Hussain[2], Erguang Li[3,*], Yuanyuan Pan[1], Qingqing Li[1]

[1]Department of Medical Information Engineering, School of Medical Information, Wannan Medical College, Wuhu 241000, China

[2]Advanced Biomaterials and Tissue Engineering Center, College of Life Sciences and Technology, Huazhong University of Science and Technology, Wuhan 430074, China.

[3]Medical School of Nanjing University, Nanjing, Jiangsu, 210093, China.

*Corresponding authors: Tel/Fax: 0553-3932370; Email: zhangzhiyuan@wnmc.edu.cn, erguang@nju.edu.cn

Abstract

Biofilm-associated proteins plays an important role in biofilm formation. However, systematical mining and comprehensive arrangement of biofilm-associated proteins in big data is still lacking, which requires new data analysis strategy and high-quality datasets under different conditions. Here, we described a new data analysis strategy to search for potential biofilm-associated proteins and provided a comprehensive open resource of biofilm-associated proteins for bacteria, which could be an infrastructure for the biofilm research community.

Introduction

Biofilm are adhesion structures formed by single or multiple bacteria and their metabolites. The National Institutes of Health (NIH) has revealed that among all microbial and chronic infections, 65% and 80%, respectively[1], are associated with biofilm formation. In clinical practice, biofilms can greatly improve the ability of pathogenic bacteria to resist antibiotics, thus increasing the risk of infection, which is difficult to cure after infection[2]. Biofilm-associated proteins are a type of protein molecules closely related to the formation of bacterial biofilms, which include constitutive proteins located downstream of the biofilm regulatory network and upstream transcriptional regulators. During the development and formation of biofilms, the transcription profile of bacteria changes, and some genes with variable expression levels often play an important role in biofilm formation. Therefore, it is necessary for researchers to analyze gene expression profile information in the background of bacterial biofilms to identify potential biofilm-associated proteins. The discovery of biofilm-associated genes and proteins can help us to better understand the molecular mechanisms of bacterial biofilm formation. At the same time, by studying the structure of the active site of the protein, we can design repressors that bind to it, which also provides a direction for the development of new antibiotics[3]. Only a few databases provide biofilm data, such as Quorumpeps[4] (http://quorumpeps.ugent.be) for QS-derived signaling peptides, BiofOmics[5] (http://biofomics.org) for biofilm experimental information, BaAMPs[6] (http://www.baamps.it), aBiofilm[7] (http://bioinfo.imtech.res.in/manojk/abiofilm/), dpABB for antibiofilm Agents, BSD[8] (http://bsd.cme.msu.edu) for structural information. However, systematical mining and comprehensive arrangement of biofilm-associated proteins in big data is still lacking, which requires new data analysis strategy and high-quality datasets under different conditions. A comprehensive open resource of biofilm-associated proteins for bacteria is not available and it will be helpful for the research community.

To systematically explore biofilm-associated proteins in bacteria, we combined differentially expressed genes (DEGs) obtained through gene differential expression analysis with a constructed biofilm-associated protein predictive model to search for potential biofilm-associated proteins.

Results

Firstly, we performed text-mining analysis using a customized Python pipeline by searching for biofilm-associated proteins (genes) in the titles or abstracts of literature from the NCBI PubMed database (https://www.ncbi.nlm.nih.gov/pubmed). We focused on experimentally validated biofilm-associated proteins in bacteria, which will provide important features and information regarding biofilm-associated proteins. We found 1,321 entries of supported evidence from 942 articles, which proved 405 experimentally validated biofilm-associated proteins from different bacteria whose sequences will be used for data input in the subsequent construction of a predictive model for biofilm-associated proteins. We proposed a feature representation learning scheme that integrates different types of sequence-based feature descriptors[9,10]. We utilized different feature extraction methods for feature representation and obtained 46 feature descriptors in total. A list of the 46 descriptors is provided in **Supplementals**. We compared all 46 feature descriptors using RF classifiers, and the top 3 best feature descriptors (**CKSAAP**, **TPC,** and **AAC_DPC**) with the RF classifier obtained 0.7 0.72, 0.75 AUC scores, respectively. Each of the 3 feature descriptors were trained and evaluated using the RF classifier on the datasets. Finally, the predicted class label for each trained RF model is regarded as an attribute that forms a new feature vector, named combined_feature. The combined_feature using the RF classifier obtained an AUC score 0.80 (Fig. 1B), which were generally higher than those obtained using other classifiers. For specific details, please refer to **Supplementals**. We curated transcriptome datasets from biofilm research, including 105 RNA-seqs / 1205 samples from 16 bacteria, from the Gene Expression Omnibus of the national center for Biotechnology Information (NCBI GEO, https://www.ncbi.nlm.nih.gov/gds/) and Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra/) databases. All transcriptome data were processed as described in **Methods**.

After removed the batch effect for combined RNA-seq expression data, we classified them into two groups: (1) Human microbial pathogens, in total 15: *E.coli*, *Burkholderia cenocepacia*, *Streptococcus pneumoniae*, *Streptococcus mutans*, *Staphylococcus epidermidis*, *Staphylococcus aureus*, *Salmonella enterica*, *Pseudomonas aeruginosa*, *Porphyromonas gingivalis*, *Listeria monocytogenes*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Mycobacterium Tuberculosis*, *Vibrio cholerae*, and *Listeria monocytogenes*; (2) others: *Bacillus subtilis*. For human pathogenic bacteria, we analyzed the gene expression of different bacterial strains, and found that in the background of biofilm, the number of DEGs of different bacteria is quite different, which may be due to the number of genes in the bacterial genome itself is different. The proportion of DEGs to the number of genes in the bacterial genome range from 0.3% to 53% (Fig. 1C). We also analyzed gene expression changes in individual bacteria under different conditions. According to the differential expression analysis, the shared genes and specific genes of bacteria were obtained, which were differentially expressed under different experimental condition and only under single one specific conditions.

In this study, we used *E. coli* as an example to represent how information on of potential biofilm-associated proteins can be obtained using transcriptome data and constructed predictor.

We calculated DEGs of 3 RNA-seq experiments for *E. coli* (**Experiment_1**: aerobic biofilm cultures vs. aerobic planktonic cultures, **Experiment_2**: anaerobic biofilm cultures vs. anaerobic planktonic cultures, and **Experiment_3**: biofilm vs. planktonic cultures), in which the expression abundances were normalized as TPM values[11], using a cutoff of |log2 FC| > 1.5 (FC, fold change) and p-value <0.05 to define DEGs between experiments[12]. We found that the number of DEGs in Experiments1, 2, and 3 were 614, 1507, and 2252, respectively. Among these experiments, the number of shared genes of Experiment_1 with Experiment_2, Experiment_2 with Experiment_3, and Experiment_1 with Experiment_3 were 194, 1360, and 284, respectively (Fig. 1D). In total, 173 shared genes were obtained from the 3 RNA-seq experiments. Meanwhile, we constructed protein-protein interaction (PPI) networks for 173 shared genes using STRING[13] (Fig. 1E), a current collection of known and predicted direct physical binding and indirect functionally related interactions between proteins / genes. The results indicated that these genes were involved in propanoate, butanoate and pyruvate metabolism.

Next, when exploring the top expression genes for the three experiments, we found that they shared 23 genes in the top 500 DEGs, including one commonly downregulated gene *garP* (Fig. 1F), which was associated with the plasma membrane, and 22 commonly upregulated genes (*b1551, infA, tyrU, cspI, groS, csrA, exbD, argQ, greA, sraF, metU, b4140, yifE, rplU, yheL, secG, zapA, wzzB, ydfK, cspB, cspG,* and *yjcB*) in three experiments (Fig. 1G), which are mainly were involved in RNA-binding activity. Subsequently, we used the constructed classifier to predict 23 DEGs, in total, 15 genes (*garP, yjcB, yifE, yheL, tyrU, sraF, secG, metU, b1551, argQ, wzzB, csrA, zapA, ydfK, b4140*) were predicted to be biofilm-associated proteins (Fig. 1H), which may play a key role in biofilm formation and can provided references for experimental biologists.

To make it more efficient and pain-free for researchers to obtain transcriptome data in biofilm research, query information of experimental validated biofilm-associated protein, and use constructed predictor for protein prediction, we constructed BBSdb, an state-of-the-art open resource of bacterial biofilm data for prediction of biofilm-associated protein (BBSdb, http://124.222.145.44/#!/) (Fig. 2).

In conclusion, we systematically presented a new strategy using *E. coli* as an example to combine the DEGs obtained with a constructed biofilm-associated protein predictive model to search for potential biofilm-associated proteins. In addition, we provide a comprehensive database for biofilm research, which could be an infrastructure for the biofilm research community. In the future, several questions remain to be addressed, including the long-term maintenance of BBSdb and improvement of predictive ability upon big data accumulation.

Methods

Data processing

For the processing of RNA-seq raw sequencing reads, we used FastQC to evaluate the overall quality of the downloaded data, followed by the Trim_galore to remove sequencing vectors and low-quality bases. Salmon[14] is the latest computational algorithm for transcript quantification from RNA-seq data that could make the expression data compared across experiments and projects. In this study, we performed transcript quantification using Salmon, which adopted TPM (Transcript Per Million) for normalization.

Text-mining analysis

In this study, text-mining analysis is used to mine experimentally validated biofilm-associated proteins in the literature. The specific steps include: 1) obtaining literature, which must focus on bacterial biofilm at the gene or protein level; 2) obtaining certain sentence as supported evidence, which must describe gene or protein and contain the following keyword: biofilm; 3) extracting information of biofilm-associated gene or protein..The retrieved information was further verified by three rounds of manual inspection.

Feature Descriptors for Protein sequences

A comprehensive review of existing sequence-based feature-encoding algorithms for protein feature representation was conducted (**Supplementals**).

Machine learning algorithm

In this study, we employed the RF, SVM, KNN, XGBoost and LR algorithms to build models and make predictions. According to the result, RF, which has been widely used in bioinformatics and computational biology[15,16], had stronger power than other four classifiers.

Receiver operating characteristic curve

The receiver operating characteristic (ROC) curve is used to evaluate the overall performance of predictive model. It is plotted by calculating sensitivity (True Positive Rate; TPR) against 1-specificity (False Positive Rate; FPR) under different thresholds. The area under the ROC curve, termed as AUC, is often used as a metric to access the overall performance. The AUC score ranges from 0.5 to 1. The larger score a predictor achieves, the better performance it has.

Database design and implementation

BBSdb, which integrated data and a predictor, was designed as a relational database. All data were loaded into a MySQL database, and the prediction function could be used in BBSb website. The frontend of the website was coded using JavaScript and HTML, while the backend was coded using Python with a Flask framework to support queries to the MySQL database and provide representational state transfer (REST) application programming interfaces (APIs) for programmable access to our data. The AngularJS framework was used to bride the front- and back-ends. Echarts.js and plotly.js used for visualizations at the front end.

Data availability

Publicly available datasets were analyzed in this study. This data can be found here: **http://124.222.145.44/#!/terms**.The source code for the developed platform is available at the BBSdb GitHub repository (**https://github.com/895548727/Bacterial_biofilm**).

Acknowledgements

The work was supported by grants from the NSFC (81871636 to EL).

References

1. Jamal, M. et al. Bacterial biofilm and associated infections. *J Chin Med Assoc.* **81**, 7-11 (2018).

2. Schwarzer, S. et al. The efficacy of topical agents used in wounds for managing chronic biofilm infections: A systematic review. *J Infect.* **80**, 261-270 (2020).

3. Shukla, SK. et al. Targeting hydrophobicity in biofilm-associated protein (Bap) as a novel antibiofilm strategy against Staphylococcus aureus biofilm. *Biophys. Chem.* **289**, 106860 (2022).

4. Wynendaele, Evelien. et al. Quorumpeps database: chemical space, microbial origin and functionality of quorum sensing peptides. *Nucleic acids research* **41**, D655-9 (2013).

5. Lourenço, Anália. et al.BiofOmics: a Web platform for the systematic and standardized

collection of high-throughput biofilm data. *PloS one* **7**, e39960 (2012).

6. Di, Luca, Mariagrazia. et al. BaAMPs: the database of biofilm-active antimicrobial peptides. *Biofouling* **31**, 193-9 (2015).

7. Rajput, Akanksha. et al. aBiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. *Nucleic acids research* **46**, D894-D900 (2018).

8. Magalhães, Rita, P. et al. BSD: the Biodegradative Strain Database.*Trends in biotechnology* **38**, 937-940 (2020).

9. Rajput, A. et al. Prediction and analysis of quorum sensing peptides based on sequence features. *PLoS One* **10**, e0120066. (2015).

10. Wei, L. et al. Prediction and analysis of quorum sensing peptides based on sequence features. *Brief Bioinform.* **21**,106-119 (2020).

11. Mortazavi, A. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods.* **5**, 6218 (2008).

12. Zhang, Z. et al. Mr.Vc v2: An updated version of database with increased data of transcriptome and experimental validated interactions. *Front Microbiol.***13**,1047259 (2022).

13. Szklarczyk, D. et al. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*. **51**, D638-D646 (2023).

14. Patro, R et al. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods.* **14**, 417-419 (2017).

15. Xu, Y. et al. Alternative splicing links histone modifications to stem cell fate decision. *Genome Biol.* **19**, 133 (2018).

Figure legends

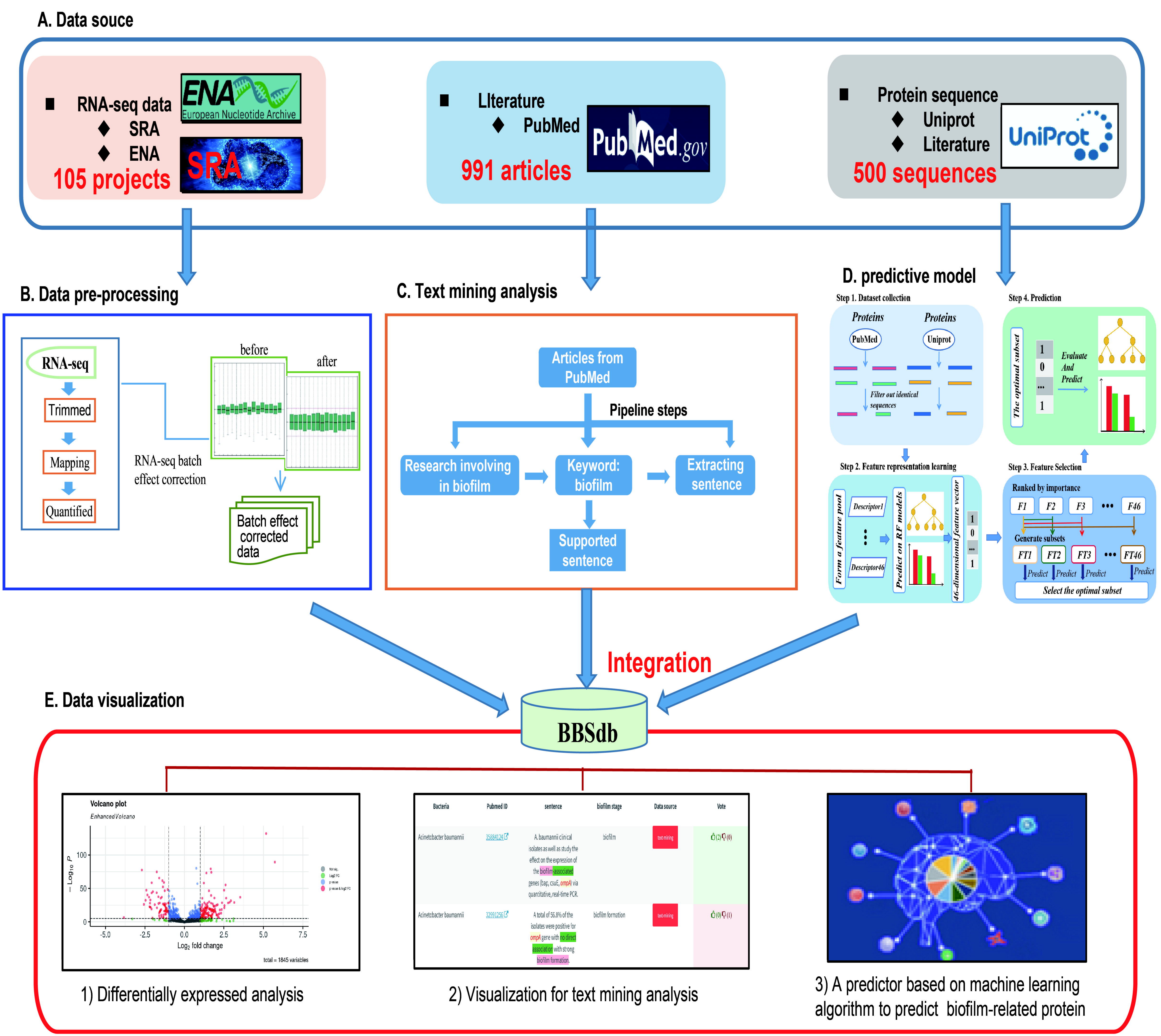
Fig. 1 Introduction and analysis

(A) Pipeline of the feature representation learning scheme. Firstly, a feature pool with 46 feature descriptors is constructed by nine feature-encoding algorithms . Afterwards, each descriptor is trained and evaluated using the RF classifier on the biofilm-associated protein sequence dataset. Finally, the predicted class label for each trained RF model is regarded as an attribute to form a new feature vector. **(B)** ROC curves of the best-performing feature descriptors using RF classifier. **(C)** The number of DEGs of 15 Human microbial pathogens, and the proportion of DEGs to the number of genes in the bacterial genome. **(D)** The shared and specific DEGs of 3 RNA-seq experiments for *E. coli.* **(E)** The analysis of protein-protein interaction (PPI) interaction networks for 173 shared genes obtained from 3 RNA-seq experiments. **(F)** Top 500 DEGs were obtained by differentially expressed analysis for each RNA-seq experiment, then, up-regulated genes were retained to screen out overlapped up-regulated genes of three experiments. **(G)** Same as above, the down-regulated genes were retained to screen out overlapped down-regulated genes of three experiments. **(H)** Combination of DEGs with constructed prediction model for potential biofilm-associated protein mining.

Fig. 2 Overview of BBSdb

(A) The data source. All data including the raw sequencing data, information of genomic annotation, literature, and experimental validated biofilm-associated proteins were collected from public databases, such as NCBI GEO, ENA, Uniprot, PubMed and so on. **(B)** The data pre-processing. The RNA-seq raw data perform quality control, mapping, transcript quantification and normalization to obtain high-quality data for further analysis. **(C)** Text mining analysis. A customized Python pipeline was used by searching for biofilm-associated proteins (genes) in the titles or abstracts of literature from the NCBI PubMed database. **(D)** Construction of predictive model. Firstly, the targeted protein sequences are subjected to the feature representation learning scheme, and as a result, a 46-dimensional feature vector will be generated. Secondly, the feature vector generated at the previous step is optimized to a 3-dimensional optimal feature vector. Ultimately, the proteins are predicted and scored by the well-trained RF model.**(E)** The data integration. BBSdb integrated polytype data with a biofilm-associated proteins predictior, which placed in “Bacteria”, “Genes(Proteins)”, “Diff. Expr.” and “Prediction” pages.

-


A. Data souce
RNA-seq data
SRA
ENA
105 projects
ENA European Nucleotide Archive
SRA
LIterature
PubMed
991 articles
PubMed.gov
Protein sequence
Uniprot
Literature
500 sequences
UniProt
B. Data pre-processing
RNA-seq
Trimmed
Mapping
Quantified
RNA-seq batch effect correction
before
after
Batch effect corrected data
C. Text mining analysis
Articles from PubMed
Pipeline steps
Research involving in biofilm
Keyword: biofilm
Extracting sentence
Supported sentence
D. predictive model
Step 1. Dataset collection
Proteins
Proteins
PubMed
Uniprot
Filter out identical sequences
Step 2. Feature representation learning
Form a feature pool
Descriptor1
Descriptor46
Predict on RF models
46-dimensional feature vector
1
0
...
1
Step 3. Feature Selection
Ranked by importance
F1
F2
F3
F46
Generate subsets
FT1
FT2
FT3
FT46
Predict
Predict
Predict
Predict
Select the optimal subset
Step 4. Prediction
The optimal subset
1
0
...
1
Evaluate And Predict
Integration
BBSdb
E. Data visualization
Volcano plot
EnhancedVolcano
− Log10 P
Log2 fold change
total = 1845 variables
Bacteria
Pubmed ID
sentence
biofilm stage
Data source
Vote
Acinetobacter baumannii
35884124
biofilm
text mining
Acinetobacter baumannii
32591256
biofilm formation
text mining
1) Differentially expressed analysis
2) Visualization for text mining analysis
3) A predictor based on machine learning algorithm to predict biofilm-related protein

(A)
Protein sequences on the biofilmSeq dataset
Feature representation
Step 1. Forming a feature pool
AAC
CTD
AAindex
CKSAAP
CTriad
Geary
188B
WAAC
DDE
Parameter varying approach
Descriptor1
Descriptor2
Descriptor3
Descriptor44
Descriptor45
Descriptor46
Fed
Step 2. Training learning model
RF model
Predict
Label (0/1)
Concatenate
Step 3. Forming a new feature pool
46-dimensional 0/1 feature vector
(B)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Combined feature (AUC = 0.80)
Chance_level (AUC = 0.5)
(C)
The ratio of DEGs to total genes in genome
Total genes in genome
DEGs
0
2000
4000
6000
8000
3,286 / 5,669 (58.0%)
2,837 / 5,957 (47.6%)
913 / 6,632 (13.8%)
1,283 / 4,611 (27.8%)
583 / 4,967 (11.7%)
1,547 / 3,809 (40.6%)
220 / 4,611 (4.8%)
709 / 4,007 (17.7%)
1,284 / 3,425 (37.5%)
1,517 / 3,038 (50.0%)
957 / 2,689 (35.6%)
583 / 2,974 (19.6%)
10 / 2,934 (0.3%)
290 / 2,510 (11.6%)
239 / 2,224 (10.7%)
P.aeruginosa
K.pneumoniae
B.cenocepacia
E.coli
S.enterica
A.baumannii
M.tuberculosis
V.cholerae
E.faecalis
S.aureus
S.epidermidis
P.gingivalis
L.monocytogenes
S.mutans
S.pneumoniae
(D)
Experiment 1
Experiment 2
Experiment 3
309
21
126
173
111
1187
781
Size of DEGs s in each experimental condition
0
252.5
505
614
1507
2252
Experiment 1
Experiment 2
Experiment 3
(E)
preA
yqeB
hyuA
ygeW
ssnA
xdhB
gloA
tdcB
rhaD
aldA
fadB
tdcE
sdhD
glpC
pta
atoB
glpB
atoE
frdC
fumB
fadE
atoA
yebO
slyB
yneM
mgrB
phoP
(F)
Experiment 1
Experiment 2
Experiment 3
209
3
64
22
14
142
258
Overlapped up-regulated genes in top 500 DEGs
(G)
Experiment 1
Experiment 2
Experiment 3
239
11
209
1
1
48
14
Overlapped down-regulated genes in top 500 DEGs
(H)
8
15
8
Overlapped DEGs
Genes using predictor