Bioinformatics, 36(3), 2020, 942–944
doi: 10.1093/bioinformatics/btz677
Advance Access Publication Date: 3 September 2019
Applications Note

Sequence analysis

DDAP: docking domain affinity and biosynthetic pathway prediction tool for type I polyketide synthases

Tingyang Li [1], Ashootosh Tripathi[2,3], Fengan Yu[2], David H. Sherman[2,3,4] and Arvind Rao[1,5,6,*]

[1]Department of Computational Medicine and Bioinformatics, [2]Natural Products Discovery Core, Life Sciences Institute, [3]Department of Medicinal Chemistry, [4]Department of Chemistry, Department of Microbiology and Immunology, [5]Department of Radiation Oncology, Michigan Institute for Data Science and Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, USA and [6]Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 2, 2019; revised on June 25, 2019; editorial decision on August 23, 2019; accepted on August 30, 2019

Abstract

Summary: DDAP is a tool for predicting the biosynthetic pathways of the products of type I modular polyketide synthase (PKS) with the focus on providing a more accurate prediction of the ordering of proteins and substrates in the pathway. In this study, the module docking domain (DD) affinity prediction performance on a hold-out testing dataset reached 0.88 as measured by the area under the receiver operating characteristic (ROC) curve (AUC); the Mean Reciprocal Ranking (MRR) of pathway prediction reached 0.67. DDAP has advantages compared to previous informatics tools in several aspects: (i) it does not rely on large databases, making it a high efficiency tool, (ii) the predicted DD affinity is represented by a probability (0–1), which is more intuitive than raw scores, (iii) its performance is competitive compared to the current popular rule-based algorithm. DDAP is so far the first machine learning based algorithm for type I PKS DD affinity and pathway prediction. We also established the first database of type I modular PKSs, featuring a comprehensive annotation of available docking domains information in bacterial biosynthetic pathways.

Availability and implementation: The DDAP database is available at https://tylii.github.io/ddap. The prediction algorithm DDAP is freely available on GitHub (https://github.com/tylii/ddap) and released under the MIT license.

Contact: ukarvind@umich.edu

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

1 Introduction

Natural products (NPs) such as penicillin, erythromycin, artemisinin, taxol and tetrodotoxin are small chemical compounds produced by bacteria, fungi, plants and animals. These small molecules are found to display a broad range of biological activities (Katz and Baltz, 2016). About 40% of new drugs approved in the past 30 years were either unaltered NPs or derived from NPs (Newman and Cragg, 2016). Interest is growing in the search for novel NPs in both industrial and academic fields. Although new mass spectrometry (MS), and nuclear magnetic resonance (NMR) technologies have improved efficiency, complete characterization of isolated NPs is still an intellectual challenge, which renders structure elucidation of new NPs time consuming. Developing computational tools to predict NP structures based on DNA/protein sequences [e.g. antiSMASH (Blin *et al.*, 2019), NP.searcher (Li *et al.*, 2009), PRISM (Skinnider *et al.*, 2017) and SeMPI (Zierep *et al.*, 2017)] is of high interest to facilitate and streamline structure characterization, and offer new automated approaches.

Polyketide synthases (PKSs) are one of the most important classes of biosynthetic enzymes. Type I modular PKS (T1PKS) consists of a series of genes encoding multifunctional proteins, including a loading module and multiple extension modules (Dutta *et al.*, 2014). Each extension module is responsible for adding one acyl-monomer to the polyketide chain. The assembly order of polyketide substrates is not always coincident with gene cluster architecture in the bacterial genome. Therefore, finding the correct order of modules and substrates in the polyketide biosynthetic pathway is a crucial step in structure prediction. Previous research has demonstrated that the substrate assembly order is determined by cognate docking domain (DD) pairs at the N-/C-terminus of PKS proteins (Gokhale and Khosla, 2000; Gokhale *et al.*, 1999). In 2009, Yadav *et al.* published a rule-based DD affinity prediction algorithm, based on a general assumption about the 6-deoxyerythronolide B synthase (DEBS) DD structure (Yadav *et al.*, 2009). This method is used by many well-known NP discovery tools including antiSMASH and NP.searcher, despite its several defects (see Supplementary

Materials): (i) The assumed DD structure does not apply to all DDs (Moss *et al.*, 2019; Whicher *et al.*, 2013). (ii) This algorithm typically yields a large number of tied ranks in pathway prediction. (iii) The upper/lower limits of the pathway likelihood scores vary with the total number of genes in the PKS, which makes the scores difficult to interpret. The Yadav *et al.* study is the only available algorithm that specifically predicts DD affinity in T1PKS.

In this study, we collected pathway information of T1PKS from manuscripts published over the past 24 years (1995–2018) and developed a machine learning based docking domain affinity prediction tool, DDAP. DDAP uses protein sequences to predict the assembly order of the compounds produced by T1PKS. The DDAP database includes 172 T1PKS and 764 docking domains. This is the first and most comprehensive database of DDs in bacterial type I modular PKSs.

2 Materials and methods

2.1 The DDAP database

The DDAP database contains 172 records of type I modular PKS proteins, among which 92 are annotated with published pathway information, 80 are annotated with pathways predicted by DDAP. Pathway information includes the docking domain sequences (764 DDs) and the order of genes in the biosynthetic pathway. Users are able to download all above-mentioned datasets from the database and browse the pathway data through an interactive table on the web page.

2.2 Docking domain affinity prediction

DDAP has two main functions: (i) predicting the likelihood of interaction given the amino acid (AA) sequences of a C-terminal DD (Head) and a N-terminal DD (Tail); (ii) predicting the most likely pathways based on the predicted DD affinity. DDAP uses machine learning models to predict DD affinity (Fig. 1A–C). DDAP takes DD sequences in FASTA format as input and returns the predicted affinity of each DD pair. The affinity will always be a number between 0 and 1, where 1 indicates high affinity (Fig. 1D).

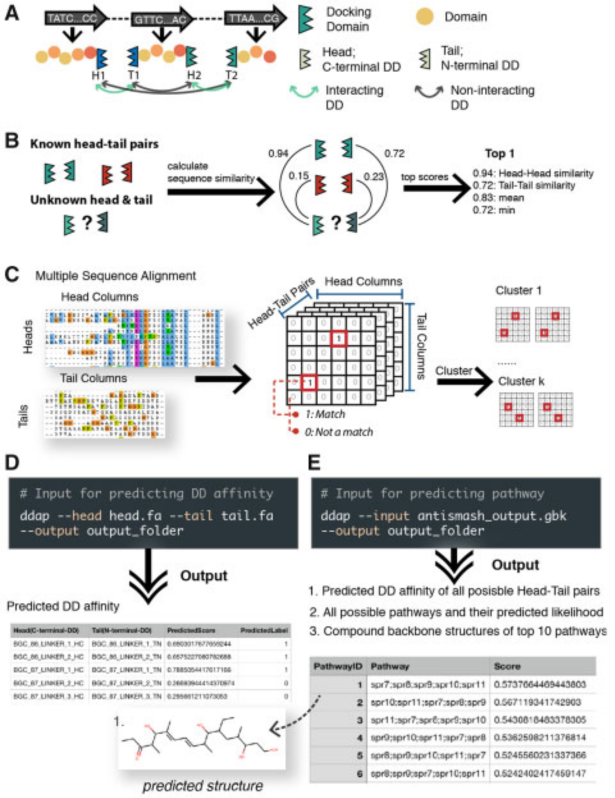
2.3 Pathway prediction

antiSMASH 5 is a state-of-the-art tool for natural product gene cluster discovery (Blin *et al.*, 2019). Following antiSMASH identification of a biosynthetic gene cluster, DDAP can read the PKS annotation and its docking domain affinity prediction algorithm predicts the order of modules/substrates in the biosynthetic pathway (see Supplementary Materials). Alternatively, users can also provide the AA sequences of PKS proteins in FASTA or CSV format. In the output, DDAP returns an exhaustive list of all possible pathways, each associated with a probability score (0–1), and a SMILES string representing the backbone structure of the proposed biosynthetic product. DDAP also provides a plot of the compound structure for the top ten most likely pathways (Fig. 1E).

3 Results

According to the five-fold cross-validation results of the best performing model, the area under the receiver operating characteristic (ROC) curve (AUC) for DD affinity prediction was 0.80 (95% CI: 0.78–0.81). The Mean Reciprocal Ranking (MRR) of the true pathways was 0.63 (95% CI: 0.59–0.67). Approximately 71% of the time, the true pathway ranked among the top three. We further tested the best performing model on the hold-out testing set. The final model achieved AUC = 0.88 (95% CI: 0.77–0.98) for DD affinity prediction. The MRR for pathway prediction was 0.67 (95% CI: 0.27–1.00). The true order received the highest likelihood score in 4 of 7 testing pathways (see Supplementary Materials).

Finally, we compared our method with the most widely used method, which was originally developed by Yadav *et al.*, and later adopted by antiSMASH and NP.searcher. We used antiSMASH 4.2.0 (Blin *et al.*, 2017) to test the performance of Yadav's method.



```
# Input for predicting DD affinity
ddap --head head.fa --tail tail.fa
--output output_folder
```

Predicted DD affinity

Head(C-terminal-DD)	Tail(N-terminal-DD)	PredictedScore	PredictedLabel
BGC_86_LINKER_1_HC	BGC_86_LINKER_1_TN	0.6903017677659244	1
BGC_86_LINKER_2_HC	BGC_86_LINKER_2_TN	0.6575227080762688	1
BGC_87_LINKER_1_HC	BGC_87_LINKER_1_TN	0.7885354417617166	1
BGC_87_LINKER_2_HC	BGC_87_LINKER_2_TN	0.26683944414370974	0
BGC_87_LINKER_3_HC	BGC_87_LINKER_3_TN	0.2959861211073053	0

```
# Input for predicting pathway
ddap --input antismash_output.gbk
--output output_folder
```

1. Predicted DD affinity of all posisble Head-Tail pairs
2. All possible pathways and their predicted likelihood
3. Compound backbone structures of top 10 pathways

PathwayID	Pathway	Score
1	spr7;spr8;spr9;spr10;spr11	0.5737664469443803
2	spr10;spr11;spr7;spr8;spr9	0.567119341742903
3	spr11;spr7;spr8;spr9;spr10	0.5430818483378305
4	spr9;spr10;spr11;spr7;spr8	0.5362598211376814
5	spr8;spr9;spr10;spr11;spr7	0.5245560231337366
6	spr8;spr9;spr7;spr10;spr11	0.5242402417459147

Fig. 1. Overview of DDAP algorithm. (**A**) A schematic view of polyketide synthase modules and docking domains. (**B**) The pipeline of the feature selection method in Model 1. The similarity measure utilized in this study is the bit-score which is not in the range of (0–1) as depicted in the plot. (**C**) The pipeline of the feature selection method in Model 2. (**D**) An example of the input and output of DDAP for docking domain affinity prediction. (**E**) An example of the input and output of DDAP for pathway prediction

Seventy pathways were used to compare the performance (see Supplementary Materials). For these 70 pathways, antiSMASH achieved MRR = 0.48 (95% CI: 0.38–0.57); DDAP achieved MRR = 0.62 (95% CI: 0.57–0.66).

4 Conclusion

In this study, we established a database for pathways and docking domains of type I modular polyketide synthases. We also built a machine learning based algorithm that predicts T1PKS pathways. The DDAP algorithm is shown to outperform the state-of-the-art without relying on large databases of proteins/compounds. The prediction tool can be readily incorporated into natural product discovery pipelines and used as a complementary tool along with genome mining software to provide accurate predictions of bacterial type I modular PKS pathways and backbone structures of the secondary metabolites.

Funding

Partial funding for this research was provided by National Institutes of Health (NIH) grant R35 GM118101, and the H. W. Vahlteich Professorship (to D.H.S.). A.R. and T.L. were supported by Cancer Center Support Grant Bioinformatics Shared Resource 2P30CA046592, startup funds from the University of Michigan, NCIR37CA214955-01A1, and a Research Scholar Grant from the American Cancer Society (RSG-16-005-01). We are also grateful to the University of Michigan Biosciences Initiative and UM Natural Products Discovery Core (NPDC) for support.

Conflict of Interest: A.R. is a member of Voxel Analytics, LLC.

References

Blin,K. *et al*. (2017) antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res*., **45**, W36–W41.

Blin,K. *et al*. (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res*., **47**, W81–W87.

Dutta,S. *et al*. (2014) Structure of a modular polyketide synthase. *Nature*, **510**, 512–517.

Gokhale,R.S. *et al*. (1999) Dissecting and exploiting intermodular communication in polyketide synthases. *Science*, **284**, 482–485.

Gokhale,R.S. and Khosla,C. (2000) Role of linkers in communication between protein modules. *Curr. Opin. Chem. Biol*., **4**, 22–27.

Katz,L. and Baltz,R.H. (2016) Natural product discovery: past, present, and future. *J. Ind. Microbiol. Biotechnol*., **43**, 155–176.

Li,M.H. *et al*. (2009) Automated genome mining for natural products. *BMC Bioinformatics*, **10**, 185.

Moss,N.A. *et al*. (2019) Nature's combinatorial biosynthesis produces vatiamides A-F. *Angew. Chem. Int. Ed*., **58**, 9027–9031.

Newman,D.J. and Cragg,G.M. (2016) Natural products as sources of new drugs from 1981 to 2014. *J. Nat. Prod*., **79**, 629–661.

Skinnider,M.A. *et al*. (2017) PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res*., **45**, W49–W54.

Whicher,J.R. *et al*. (2013) Cyanobacterial polyketide synthase docking domains: a tool for engineering natural product biosynthesis. *Chem. Biol*., **20**, 1340–1351.

Yadav,G. *et al*. (2009) Towards prediction of metabolic products of polyketide synthases: an in silico analysis. *PLoS Comput. Biol*., **5**, e1000351.

Zierep,P.F. *et al*. (2017) SeMPI: a genome-based secondary metabolite prediction and identification web server. *Nucleic Acids Res*., **45**, W64–W71.