BMC Bioinformatics

RESEARCH

Open Access

DePolymerase Predictor (DePP): a machine learning tool for the targeted identification of phage depolymerases

Damian J. Magill[1] and Timofey A. Skvortsov[2*]

*Correspondence:
t.skvortsov@qub.ac.uk

[1] Saint-Avertin, France

[2] School of Pharmacy, Queen's University Belfast, 97 Lisburn Road, Belfast BT9 7BL, Northern Ireland, UK

Abstract

Biofilm production plays a clinically significant role in the pathogenicity of many bacteria, limiting our ability to apply antimicrobial agents and contributing in particular to the pathogenesis of chronic infections. Bacteriophage depolymerases, leveraged by these viruses to circumvent biofilm mediated resistance, represent a potentially powerful weapon in the fight against antibiotic resistant bacteria. Such enzymes are able to degrade the extracellular matrix that is integral to the formation of all biofilms and as such would allow complementary therapies or disinfection procedures to be successfully applied. In this manuscript, we describe the development and application of a machine learning based approach towards the identification of phage depolymerases. We demonstrate that on the basis of a relatively limited number of experimentally proven enzymes and using an amino acid derived feature vector that the development of a powerful model with an accuracy on the order of 90% is possible, showing the value of such approaches in protein functional annotation and the discovery of novel therapeutic agents.

Keywords: Bacteriophage, Depolymerase, Machine-learning

Background

Biofilms are the most common form of bacterial lifestyle in nature [1]. Biofilm formation by pathogenic bacteria allows for the establishment of a multicellular consortium of clinical significance due to the role such communities play in the persistence of bacterial infection and their resistance to various modes of treatment and disinfection. Indeed, such assemblages confer antimicrobial resistance on multiple levels including limiting the penetrability of antimicrobial compounds, the presence of metabolically inactive persister cells exhibiting intrinsic resistance, and the internal structure of such communities providing an optimal environment facilitating horizontal gene transfer (HGT) of resistance determinants [2]. A critical component for the establishment of biofilms and a significant contributor to the resistant phenotype they exhibit is the production of a matrix embedding the biofilm cells consisting of various polymeric compounds, including proteins, extracellular DNA, and polysaccharides. The latter can be broadly

categorised as lipopolysaccharides (LPS), which are integral components of cell walls of Gram-negative bacteria, capsular polysaccharides (CPS), loosely associated with bacterial surface, and exopolysaccharides (EPS), released by bacteria into the surrounding environment [3]. The ability to remove such polymeric barriers in order to expose the underlying community of cells is a desirable one from the practical point of view, be it for the purposes of surface disinfection, de-fouling, or to improve the biocidal effects of antibiotic treatment.

Barrier properties of bacterial biofilms also pose a problem for bacterial viruses (bacteriophages) whose diffusion and ability to infect host cells is reduced within biofilms [4]. Targeted degradation of biofilm polysaccharides is a feature of many bacteriophages (phages) which increases the probability of successful infection; this is the result of enzymatic activity of a class of phage-encoded enzymes called depolymerases (DP). The majority of DPs are phage-associated enzymes and belong to lyase and hydrolase classes, with the former constituting a large majority of the well characterised and experimentally validated DPs [5–7]. Given the global antibiotic crisis we now face, there is a resurgence of interest in both phages and phage-derived therapeutic agents as alternatives. Several recently published reviews describe the structural and functional characteristics of phage DPs and outline their potential applications as biotechnological tools and therapeutic agents [8–11].

The therapeutic potential of phage DPs was recognised more than 60 years ago [12]. Phage DPs are of particular interest due to their potential use in combinatorial therapies with antibiotics or other antimicrobial agents and in the removal of biofilms from medical devices most notably catheters [13, 14]. Moreover, as the depolymerases do not kill bacteria, it is posited that they could be employed on their own as anti-virulence agents, decreasing bacterial fitness and facilitating the clearance of the bacteria by the human immune system [10]. Therefore, any approach that enhances our ability to identify novel DPs is of great value, especially since it is not always trivial to attribute depolymerase activity to a specific gene. As the polysaccharides produced by even closely related bacterial species may have subtle but significant structural differences, phage DPs acting on them also demonstrate high variability, to the point that the depolymerase domains will sometimes be among the only genomic DNA fragments showing no conservation between phages of the same genus or species [14]. Although the majority of known DPs are elements of phage receptor-binding proteins (RBPs) such as tail spikes and thus have conserved N-terminal domains responsible for virion attachment, some depolymerases can be encoded as truncated RBPs (presumably acting as diffusible DPs), further complicating their reliable prediction [15].

Machine learning based approaches are proving to be an extremely valuable avenue in all realms of science and this is no less true of phage biology whereby success has been demonstrated through the application of such techniques towards the identification of phage structural proteins [16], host-phage pairs [17], RBPs [18] and lifecycle [19] amongst others [20]. Recently published papers expand this list to include endolysins [21] and depolymerases [22]. Nevertheless, the ultimate success or failure of machine learning algorithms depends on many factors, including but not limited to the size and composition of training sets, the algorithm used for the problem at hand, and the careful construction of a vector capturing adequately discriminant features [23]. Therefore,

more ML solutions are needed to expand the computational phage characterisation toolkit and allow for a series of complementary approaches to be available.

In this manuscript we describe the development and application of a machine learning approach towards the identification of phage DPs, highlighting that such models should form an integral part of our toolkit enabling the discovery of novel therapeutics. We demonstrate that even a relatively small training set of experimentally verified sequences is sufficient to produce a machine learning model capable of accurately predicting DPs in a multitude of phages infecting different bacterial species. Indeed, an accuracy of 90% was attained on the test data set and a similar result for genome context predictions that detected the DP within the top 10 predictions.

Methods

Data set preparation

In order to establish a database of DP sequences that would ultimately fuel our model, we focussed our attention on publications within which depolymerase activity had been experimentally demonstrated. A comprehensive literature search was conducted, and a database established consisting of 50 depolymerase sequences. Additional file 1: Table S1 presents an overview of this sequence database including the phages and references from which they were found. 28 of the sequences exclusively state CPS as an enzymatic target, 20 target EPS, and the remaining two target LPS and a combination of targets. The majority of sequences were *Podoviridae* derived and the database concerned phages infecting Gram-negative bacteria. The size range of sequences varied from 150 amino acids to 1267 amino acids in length.

To complete this dataset, we required 50 sequences that would serve as the negative non-depolymerase set and thus provide a 1:1 positive to negative sequence set. To do this, we randomly extracted 50 sequences from a soil metagenome (SRR15048733) that were sampled across the size distribution of sequences so as to avoid the introduction of sequence size biases. BLAST searches were conducted with these sequences against the positive depolymerases to ensure the absence of homology followed by HHPred analysis to confirm the absence of domains known to be associated with depolymerase activity [24].

To highlight the dissimilarity in the dataset, we calculated pairwise similarity scores across the entire dataset and represented this as a heatmap (Fig. 1).

Feature extraction and selection

A diverse range of features were generated which were derived solely from the amino acid sequences. Eleven of these features were directly calculated using the ProteinAnalysis feature from the BioPython (version 1.73) ProtParam module [25]. These were the molecular weight (MW), aromaticity, predicted instability and isoelectric point, GRAVY score, predicted secondary structure (sequence proportion engaged in helices, strands, and turns), extinction coefficients (ox/red), and a combined flexibility score. Beyond this the relative abundance of each amino acid and the total sequence length were also taken into account. As a final set of features, we considered dipeptides and tripeptides as a function of conserved physicochemical properties. Seven groups were established consisting of amino acids with a hydrocarbon R group, those with an uncharged aromatic

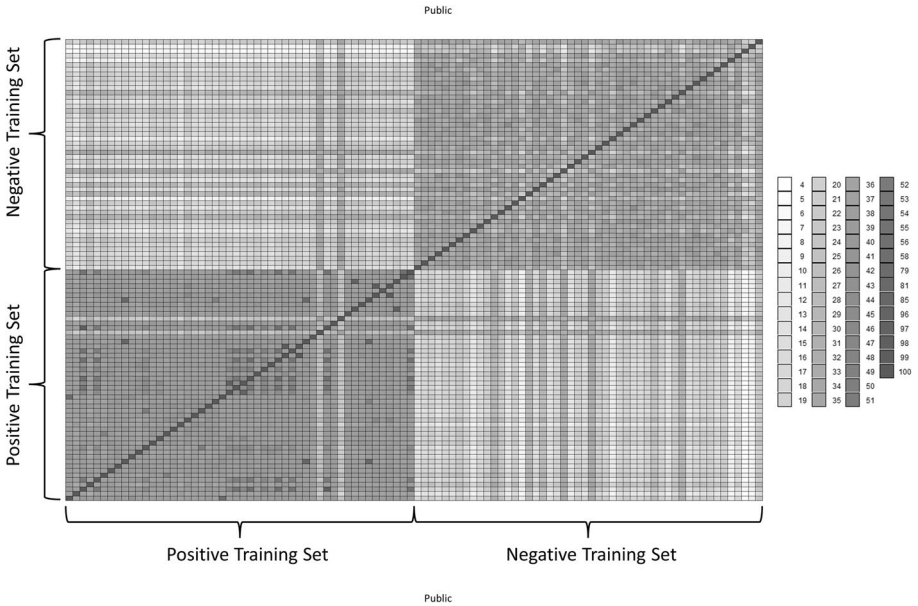

Fig. 1 Heatmap of pairwise similarity scores calculated for the training dataset. Grayscale colours correspond to percentage identity as provided in the associated legend. The negative and positive components of the dataset are highlighted with braces and associated labels. As highlighted by the scale of the legend, the global identities of the matrix are rather low, showing a high level of dissimilarity between the sequences

side chain, sulphur containing, positively charged, negatively charged, polar uncharged, and proline. According to this schema, the dipeptides AE and LD were considered as both belonging to group 15. Whilst allowing us to incorporate dipeptide and tripeptide properties into the model, this also reduced the overall feature set compared to using all possible combinations of amino acids. Python scripts used for calculation of di- and tri-peptides have been incorporated into the depolymerase prediction tool we describe below.

Model selection, training, and evaluation

With respect to the appropriate choice of machine learning algorithm, we decided to test both support vector machine (SVM) and random forest (RF) approaches [26, 27]. This was due to the fact that our data set constituted a small number of samples exhibiting a high feature space. In both cases, we leveraged a grid search with a nested validation scheme in order to assess the hyperparameter space and find the best model configuration for both algorithms. We conducted 20 independent trials with inner and outer cross-validation loops defined with 4 splits on a randomly shuffled dataset using the trial number as the random state. This was conducted using the scikit-learn library (version 0.23.2) [28].

To evaluate final model performance, we particularly focussed on the overall accuracy and recall on the defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

where TP FP, TN, and FN correspond to true positive, false positive, true negative, and false negative respectively regarding the classification performed on the test data. All scores reported are the average obtained following the cross-validation.

With respect to the hyperparameters tuned, for SVM both linear and RBF kernels were evaluated along with cost and gamma functions when applicable. For RF, differing numbers of estimators were evaluated using a step size of 100 along with total tree depth, and the minimum samples supporting a branch and split of the tree. In addition to this, we also integrated a two-degree polynomial feature transformation, min/max scaling, and applied entropy-based impunity.

Software package DePolymerase Predictor

Once optimal parameters were determined for the model following evaluation, the final RF-based version of the DePolymerase Predictor (DePP) tool was created incorporating the entirety of the training set. Both the source code of the tool and a standalone ready-to-use version of the application are available as detailed in the "Availability of data and materials" section. A simple user-friendly GUI has been developed through which users can step-by-step upload their sequences, generate the feature vector, and carry out predictions and view the output. Finally, our tool is also available as a web application. We continue actively working in the field of phage depolymerase research and will be regularly updating, expanding and improving our tool as new experimentally verified depolymerases become available.

Results and discussion

Feature generation

The application of the feature generation script was carried out on the 100 amino acid sequence input data set. This resulted in the construction of a feature vector with 424 descriptors for each of the sequences. An additional column was added to distinguish the depolymerases from the negative cases. This entire training set is presented in Additional file 2: Table S2 and can be used directly in the reproduction of our analysis with the parameters outlined below.

Model evaluation and final selection

The SVM approach was initially applied to the dataset with no hyperparameter tuning, with the application of a linear kernel. This resulted in a model exhibiting an overall accuracy score of 0.70 across all folds as presented in the normalised confusion matrix in Fig. 2. This model performed extremely poorly with respect to true and false positives but handled negative cases well. Indeed, hyperparameter tuning did little to resolve this problem. The overall accuracy remained unchanged, but the model improved in its ability to correctly identify non-depolymerase sequences with 100% success rate. This was at the cost of decreased performance on positive cases with only 45% of true depolymerases being correctly identified as such.

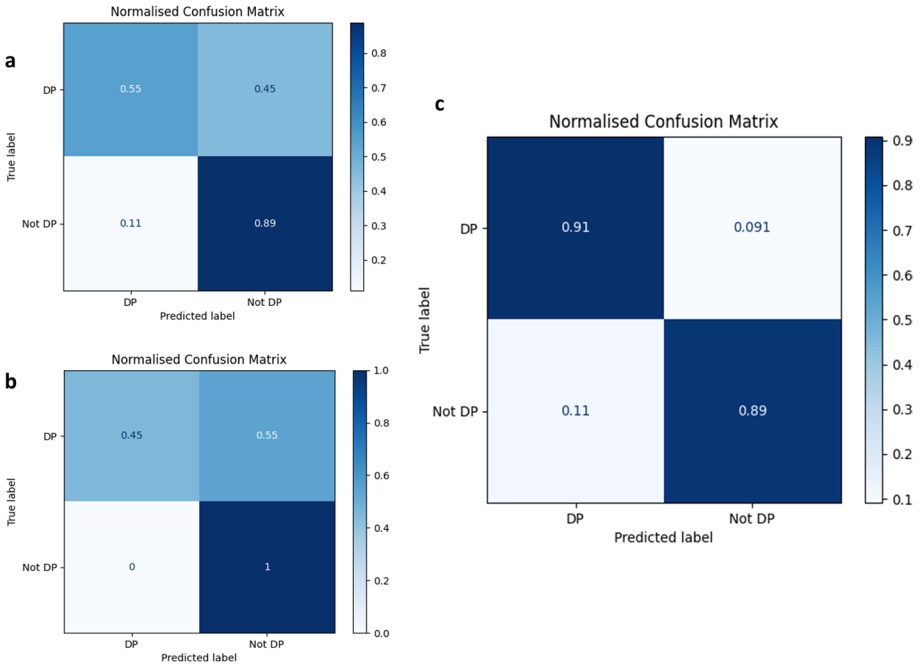

Fig. 2 Normalised confusion matrices summarising model performance on test data. Matrices give the proportion of depolymerase (DP) and non-depolymerase (Not DP) that are correctly identified by the model, corresponding thus to the true/false positive and true/false negative proportions. Matrices are shown for non-optimised (before hyperparameter tuning) SVM (**a**), optimised SVM (after hyperparameter tuning) (**b**), and optimised RF (**c**) models

Subsequent application of an RF approach yielded more promising results. This is an ensemble machine learning method that leverages multiple decision trees in order to reduce variance and provide better model generalization. It performs especially well with small sample sets and large feature spaces and so it was expected it may be the best approach to this problem. Application of a tuned RF model indeed showed a much higher level of performance (Fig. 2). An overall accuracy score of 0.90 was obtained with similar performance observed with respect to the correct classification of positive and negative cases. It was found that for this case, the following parameters provided optimal performance of the model: use of 1500 estimators with automatic definition of maximum features to be used by each tree. A maximum depth of 30 was applied with a minimum sample support of 3 required for each leaf. Each tree split was evaluated using the entropy-based criterion. The pipeline also integrated a two-degree polynomial feature transformation along with application of min/max scaling. The RF approach was implemented in the final version of our tool (DePP).

Application of model towards depolymerase identification

In order to further assess the performance of our model, we first tested it within the context of whole phage genomes to see whether it could correctly identify depolymerases amongst the other genes. Due to the fact that our model leverages experimentally active depolymerases and we have thus exhausted this option, we were limited to performing this test on computationally predicted enzymes. The first such case was *Pseudomonas* phage pf16; a phage previously characterised by our group [29].

Depolymerase activity was previously observed in this phage and extensive computational analysis identified gp215 as the likely candidate with probable pectate lyase activity. We proceeded to analyse pf16 gene products using our model and ranked the probabilities accordingly. These results are presented in Fig. 3. We immediately observed that the predicted depolymerase was ranked 4th by our model in the context of the whole genome. This in itself is a reasonable result; however, further analysis of the other higher ranked candidates revealed that they possess domains not unrelated to what is observed in depolymerases including endosialidase and VrlC-like domains, the latter speculated to have sialidase activity [30]. Although it is considered somewhat uncommon, it has been postulated that certain phages (e.g. Camphawk) could have multiple depolymerases associated with different proteins, so the presence of multiple depolymerase genes predicted by our model in some phage genomes should be investigated further [11].

We next tested the performance of our model by analysing the phage genomes containing computationally predicted depolymerases from a review by Pires et al. [11]. This provided a good opportunity to test the generalisability of our model as the sequences described in this paper exhibit significant diversity in terms of the domains present and nature of the hosts infected by the phages. We downloaded the genomes of the associated phages, removing some for which the records no longer exist. This resulted in 155 genomes on which we applied our model. Predictions were performed, the probabilities ranked, and the position of the putative depolymerase identified. Additional file 3: Table S3 presents all of the genomes, the depolymerases and the associated ranking provided by our model. Across all sequences the depolymerase featured as the first prediction 40.6% of the time. This increased to 69.7% and 78.1% for top 3 and top 5 predictions respectively. When considering top 10 and top 20 this grows to a large majority with 87.1% and 94.8%. Most poorly predicted sequences were those containing domains that did not feature in our model, especially DUF867. When we look closer at the distribution of these results we observe a good level of model generalisability in a number of aspects (Fig. 4). Despite being

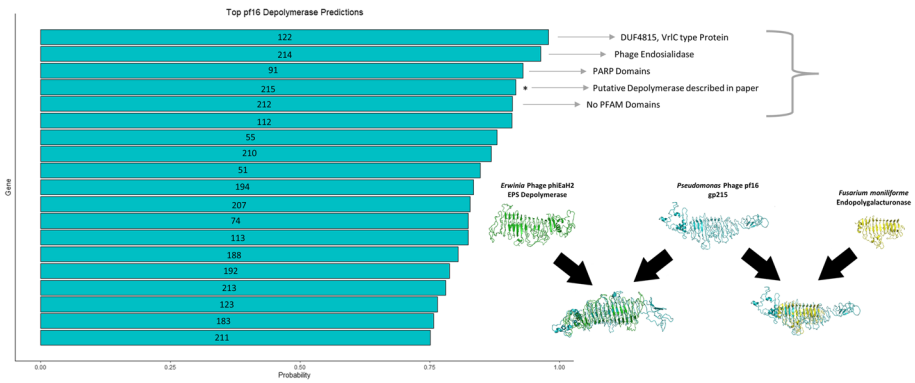

Fig. 3 Top Predictions of *Pseudomonas* phage pf16 depolymerases. The graph highlights that probability reported by the model of the gene product being a depolymerase. Gene products are labelled accordingly. The putative depolymerase previously reported is highlighted on the graph and the modelling of this protein shown with respect to a known EPS depolymerase and endopolygalacturonase as reported in Magill et al. [29]

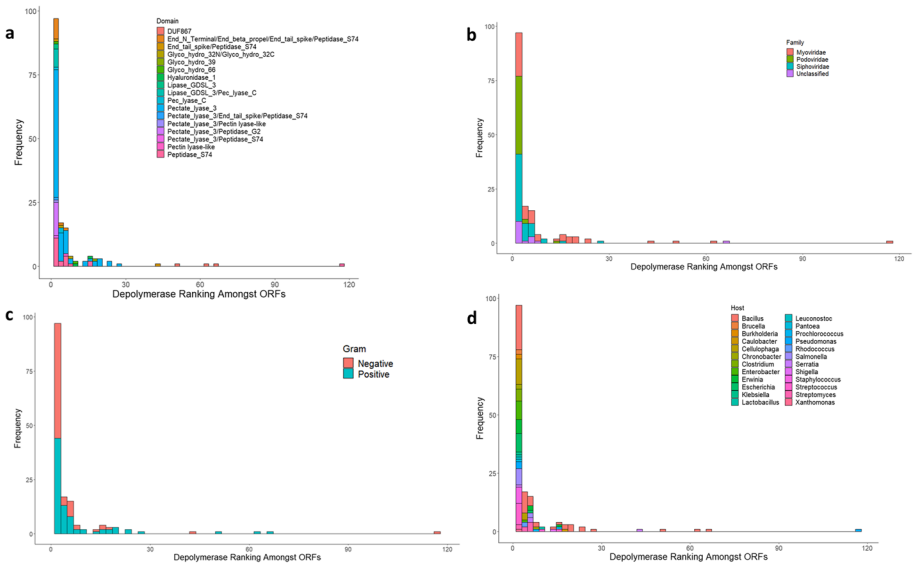

Fig. 4 Graphs showing ranking of depolymerases predicted by the model. Rankings performed on depolymerase predictions from genomes described by Pires et al. [11]. Rankings are coloured by depolymerase domains (**a**), family of the phage described (**b**), whether the host is Gram-positive or negative (**c**), and by the host genus (**d**)

based on depolymerases in phages infecting Gram-negative bacteria, the model performs equally well for phages infecting both types. This fact also holds when considering the family of phage and the genus of the host. This implies that the model is leveraging features that are common to a large majority of known depolymerase enzymes.

Benchmarking DePP against other approaches

Phage research, like other domains, is starting to reap the benefits of widespread AI application. This is no less true of depolymerase prediction and it is therefore necessary to benchmark our tool against existing approaches, notably PhageDPO [22]. In order to do this, we generated PhageDPO predictions across the same genome set from Pires et al. [11] using both SVM and ANN approaches and noting the result which ranked the depolymerase highest amongst the list of gene products for each phage (Additional file 3: Table S3). Globally across the 155 genomes, PhageDPO outperformed DePP on 36% of the dataset, was equivalent to our approach on 34% of the sequences, and was outperformed by our model on the remaining 30% of the sequences. Much of the superior performance of PhageDPO is attributable to better predictions on phages infecting Gram-positive bacteria, notably *Bacillus*. This is not wholly surprising considering the fact that the training dataset used by DePP consists of only of sequences of experimentally verified depolymerases and thus is limited in size and coverage.The lack of depolymerases of *Bacillus* phages in our dataset clearly represents an opportunity for further work involving isolation, purification and biochemical characterisation ofthese putative depolymerases. Indeed, this knowledge gap has been already recognised and being addressed; for instance, two novel *Bacillus pumilus* phages and their experimentally characterised depolymerases were described in a recently published study [31].

Despite the better performance of PhageDPO on phages of Gram-positive bacteria, our model displayed a noteworthy advantage in its ability to predict depolymerases possessing infrequently encountered domains such as Glyco_hydro_32N/Glyco_hydro_32C and Peptidase_S74 for which PhageDPO showed greater difficulty. This may be due to a large overrepresentation of Pectate_lyase_3 domain containing depolymerases in the training set of the latter. Given this performance, there is a clear advantage in using both DePP and PhageDPO in a complementary manner alongside other approaches such as HHPred to provide additional confirmation of predictions [24].

Novelty, limitations, and future directions

Our approach to training the model deviates from the most common one in the area of protein function prediction; indeed, most other groups have opted to use as many suitable potential template sequences for training ML models as possible to improve their sensitivity and performance [16, 17]. In contrast, by using only experimentally verified phage depolymerase sequences we have deliberately restricted the size of the training set to only 50 sequences primarily belonging to phages of Gram-negative bacteria. We made this decision for several reasons.

Firstly, we aimed to minimise the number of false positive hits. Admittedly, while this has noticeably reduced the number of false positive predictions made by DePP (Fig. 2), this has also somewhat limited the predictive power of the resulting model. Nevertheless, as we envision that our tool will be used by experimentalists, we sought to increase the reliability of the positive predictions even if it would result in discarding some of the true positive predictions. Our approach, being focussed on experimental sequences, is therefore intentionally orthogonal and complementary to the one used in PhageDPO.

Secondly, we were interested to examine how a smaller training set including only highly-reliable (experimentally verified) sequences would perform in comparison to a larger set consisting of true and probable depolymerases. As we demonstrate in this paper, not only the performance of DePP is overall comparable to that of PhageDPO, but we have also been able to identify certain classes of depolymerases which our model was able to predict better. This should not be interpreted as PhageDPO being inferior, but should rather serve as a reminder that the performance of ML/AI models ultimately depends on multiple parameters, with the composition and quality of the training set being among the most critical factors.

Finally, training the model exclusively on experimentally validated enzymes has enabled us to identify the underrepresented depolymerase types that warrant further investigation to verify their predicted enzymatic activities. We are currently performing experimental validation of some of the predicted enzymes.

Conclusion

Bacteriophage depolymerases offer a host of promising clinical and biotechnological applications, including the synergistic treatment of infections via biofilm removal. There is, however, a need for rapid and accurate identification of such enzymes. In this work we have described the development and application of a machine learning approach based on a small training set of experimentally verified sequences that allows for depolymerase

prediction with an overall accuracy of 90% using a sequence-derived feature vector. We demonstrated that this model was extendable to depolymerases from a variety of phages, robustly predicting them in the context of the genomes across several hosts and enzyme classes. This highlights the power that such approaches can offer in the identification of industrially and/or clinically useful enzymes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05341-w.

Additional file 1: Table S1. Overview of the phage depolymerase database obtained following literature search for experimentally demonstrated enzymes. The table outlines the phages and their hosts along with the target of the depolymerase. The article referencing the phage is given in full at the bottom of the table.

Additional file 2: Table S2. Training set used to fuel the machine learning model. The table contains the sequences of the positive and negative cases along with the entire feature set generated for the downstream modelling.

Additional file 3: Table S3. Rankings of depolymerase predictions in the context of whole phage genomes. Sequences were obtained for phages containing computationally predicted depolymerases described in Pires et al.. The ranking is given relative to the total number of ORFs predicted for each of the phages presented in the table.

Acknowledgements
Not applicable

Author contributions
DM and TS conceived the study; DM developed the software; DM and TS prepared and processed the data and analysed the results; DM and TS wrote the manuscript and approved the final version along with figures. Both authors read and approved the final manuscript.

Funding
The authors declare that they received no specific funding for this work.

Availability of data and materials
The source code for the application can be found via the following URL: https://github.com/DamianJM/Depolymerase-Predictor. We include both a version of the DePP package with a graphical user interface and a command-line only version that can be easily integrated into bioinformatics pipelines. In addition, standalone versions of the application for Windows and MacOS are available with all dependencies and training set compiled within at the following address: https://sourceforge.net/projects/depolymerase-predict. Finally, we provide a user-friendly online version of DePP (Web-DePP) that allows the users to quickly familiarise themselves with our tool and can be used for quick and convenient analysis of small datasets: https://timskvortsov.github.io/WebDePP/. The training dataset has been provided as part of the supplementary data which allows for our work to be reproduced. Depolymerases used in the development of the model are detailed in Additional file 1: Table S1, with all accession numbers and associated literature references provided.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
Not applicable.

Received: 9 September 2022 Accepted: 16 May 2023
Published online: 19 May 2023

References

1. Flemming HC, Wuertz S. Bacteria and archaea on Earth and their abundance in biofilms. Nat Rev Microbiol. 2019;17(4):247–60.
2. Uruén C, Chopo-Escuin G, Tommassen J, Mainar-Jaime RC, Arenas J. Biofilms as promoters of bacterial antibiotic resistance and tolerance. Antibiotics. 2020;10(1):3.
3. Mostowy RJ, Holt KE. Diversity-generating machines: genetics of bacterial sugar-coating. Trends Microbiol. 2018;26(12):1008–21.
4. Simmons M, Drescher K, Nadell CD, Bucci V. Phage mobility is a core determinant of phage–bacteria coexistence in biofilms. ISME J. 2018;12(2):531–43.

5. Majkowska-Skrobek G, Łątka A, Berisio R, Maciejewska B, Squeglia F, Romano M, Lavigne R, Struve C, Drulis-Kawa Z. Capsule-targeting depolymerase, derived from Klebsiella KP36 phage, as a tool for the development of anti-virulent strategy. Viruses. 2016;8(12):324.
6. Olszak T, Shneider MM, Latka A, Maciejewska B, Browning C, Sycheva LV, et al. The O-specific polysaccharide lyase from the phage LKA1 tailspike reduces Pseudomonas virulence. Sci Rep. 2017;7:16302.
7. Thompson JE, Pourhossein M, Waterhouse A, Hudson T, Goldrick M, Derrick JP, Roberts IS. The K5 lyase KflA combines a viral tail spike structure with a bacterial polysaccharide lyase mechanism. J Biol Chem. 2010;285(31):23963–9.
8. Knecht LE, Veljkovic M, Fieseler L. Diversity and function of phage encoded depolymerases. Front Microbiol. 2020;10:2949.
9. Latka A, Maciejewska B, Majkowska-Skrobek G, Briers Y, Drulis-Kawa Z. Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers during the infection process. Appl Microbiol Biotechnol. 2017;101(8):3103–19.
10. Oliveira H, Drulis-Kawa Z, Azeredo J. Exploiting phage-derived carbohydrate depolymerases for combating infectious diseases. Trends Microbiol. 2022;30(8):707–9.
11. Pires DP, Oliveira H, Melo LD, Sillankorva S, Azeredo J. Bacteriophage-encoded depolymerases: their diversity and biotechnological applications. Appl Microbiol Biotechnol. 2016;100(5):2141–51.
12. Adams MH, Park BH. An enzyme produced by a phage-host cell system: II. The properties of the polysaccharide depolymerase. Virology. 1956;2(6):719–36.
13. Shahed-Al-Mahmud M, Roy R, Sugiokto FG, Islam MN, Lin MD, Lin LC, Lin NT. Phage φAB6-borne depolymerase combats Acinetobacter baumannii biofilm formation and infection. Antibiotics. 2021;10(3):279.
14. Rice CJ, Kelly SA, O'Brien SC, Melaugh EM, Ganacias JC, Chai ZH, Gilmore BF, Skvortsov T. Novel phage-derived depolymerase with activity against Proteus mirabilis biofilms. Microorganisms. 2021;9(10):2172.
15. Latka A, Leiman PG, Drulis-Kawa Z, Briers Y. Modeling the architecture of depolymerase-containing receptor binding proteins in Klebsiella phages. Front Microbiol. 2019;10:2649.
16. Cantu VA, Salamon P, Seguritan V, Redfield J, Salamon D, Edwards RA, Segall AM. PhANNs, a fast and accurate tool and web server to classify phage structural proteins. PLoS Comput Biol. 2020;16(11):e1007845.
17. Boeckaerts D, Stock M, Criel B, Gerstmans H, De Baets B, Briers Y. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. Sci Rep. 2021;11(1):1–14.
18. Boeckaerts D, Stock M, De Baets B, Briers Y. Identification of Phage receptor-binding protein sequences with hidden Markov models and an extreme gradient boosting classifier. Viruses. 2022;14(6):1329.
19. Hockenberry AJ, Wilke CO. BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. PeerJ. 2021;9:e11396.
20. Nami Y, Imeni N, Panahi B. Application of machine learning in bacteriophage research. BMC Microbiol. 2021;21(1):1–8.
21. Criel B, Taelman S, Van Criekinge W, Stock M, Briers Y. PhaLP: a database for the study of phage lytic proteins and their evolution. Viruses. 2021;13(7):1240.
22. Vieira MF, Duarte J, Domingues R, Oliveira H, Dias O. PhageDPO: phage depolymerase finder; 2023. bioRxiv, p. 2023–02.
23. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23(1):40–55.
24. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucl Acids Res. 2005;33:W244–8.
25. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, De Hoon MJ. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–3.
26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
27. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
28. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory; 1992, p. 144–52.
29. Magill DJ, Krylov VN, Shaburova OV, McGrath JW, Allen CC, Quinn JP, Kulakov LA. Pf16 and phiPMW: expanding the realm of pseudomonas putida bacteriophages. PLoS ONE. 2017;12(9):e0184307.
30. Billington SJ, et al. Complete nucleotide sequence of the 27-kilobase virulence related locus (vrl) of Dichelobacter nodosus: evidence for extrachromosomal origin. Infect Immun. 1999;67:1277–86.
31. Skorynina AV, Koposova ON, Kazantseva OA, Piligrimova EG, Ryabova NA, Shadrin AM. Isolation and characterization of two novel siphoviruses novomoskovsk and bolokhovo, encoding polysaccharide depolymerases active against bacillus pumilus. Int J Mol Sci. 2022;23:12988.

Publisher's Note