

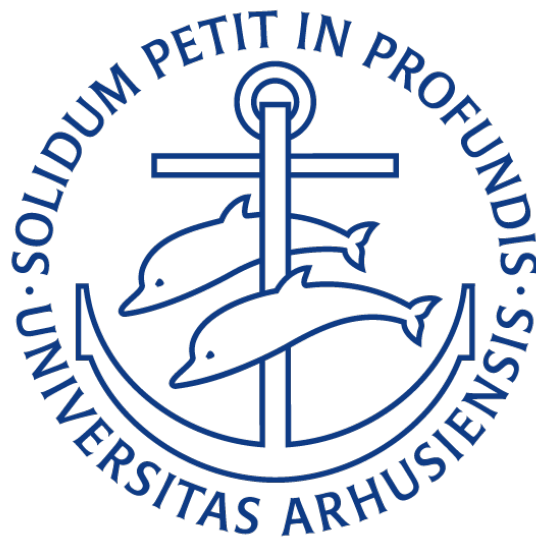
Antibiotic potential of soil bacteria targeting ESKAPE pathogens

Laura Marie Hagen Fur

MSc Student in Bioinformatics

Bioinformatics Research Center, Aarhus University

Spring 2025



Supervisors

Thomas Bataillon

Thomas Tørring

Bioinformatics Research Center Biological and Chemical Engineering

Aarhus University

Abstract

Antibiotic resistance is a growing global concern, with the emergence and spread of antibiotic-resistant bacteria continuing to increase. Antibiotics are small metabolites naturally produced by microorganisms to ensure their survival in competitive environments by inhibiting or killing other competitive microorganisms. Conventional treatment options are getting exhausted, calling for an urgent need to discover novel antibiotics.

Here, I apply an *in silico* approach to discover natural products with potential antibacterial effects against ESKAPE pathogens by utilizing the bioinformatics pipeline, antiSMASH, to mine genomes of bacterial soil isolates for biosynthetic gene clusters (BGCs). In the study, unsupervised and supervised learning methods along with phylogenetic analysis, were integrated to find associations between observed bioactivity and the presence of specific BGCs. A total of 558 BGCs were predicted by antiSMASH, of which several BGCs were flagged in the analyses after correlating these with the results from the bioactivity assays. Some of these BGCs have reported antimicrobial activities with proposed broad- and narrow-spectrum activity, such as epilancin 15x, paenilamicin, and tridecaptin, which highlights the promising potential of combining supervised learning methods with laboratory work to the discovery of novel antibiotics.

Contents

1	Introduction	1
1.1	Bacteria - in prosperity and adversity	1
1.2	The antimicrobial resistance crisis	1
1.3	The pursuit of novel antibiotics	3
1.4	antiSMASH	4
2	Data and methods	6
2.1	Initial hypothesis and testing	8
2.2	GTDB-Tk	9
2.3	Processing of pipeline-generated data	10
2.4	Downstream analysis	11
2.4.1	Phylogenetics	11
2.4.2	Unsupervised learning methods	14
2.4.3	Supervised learning methods	16
2.4.4	Manual approach to find candidate BGCs	19
3	Results and discussion	20
3.1	Analysis of the raw bioactivity data	20
3.1.1	Phylogenetic logistic regression	23
3.1.2	PCA and UMAP	25
3.1.3	Summary	28
3.2	BGC analysis	28
3.2.1	Pairwise comparisons and phylogenetics	35
3.2.2	Supervised learning methods	37
3.2.3	Results from the proportion-based approach	49

3.2.4	A short note on the prophage potential related to the bioactivity from <i>Enterococcus faecalis</i> (<i>E. faecalis</i>)	49
4	Concluding remarks	50
5	Supplementary	61
5.1	Profile hidden Markov models	61
5.2	List of significant BGCs from PLR	63
5.3	List of proposed ESKAPE-specific BGCs from the proportion-based ap- proach	68

Acronyms

<i>A. baumannii</i> <i>Acinetobacter baumannii</i>	6
<i>A. baylyi</i> <i>Acinetobacter baylyi</i>	6
ALS-PDC amyotrophic lateral sclerosis–parkinsonism and dementia complex	1
AMR antimicrobial resistance	1
antiSMASH antibiotics and Secondary Metabolites Analysis SHell	4
BGC biosynthetic gene cluster	4
BUSCO Benchmarking Universal Single-Copy Orthologue	8
CDC Center for Disease Control and Prevention	2
<i>E. cloacae</i> <i>Enterobacter cloacae</i>	6
<i>E. coli</i> <i>Escherichia coli</i>	6
EDA exploratory data analysis	11

<i>E. faecalis</i> <i>Enterococcus faecalis</i>	2
<i>E. faecium</i> <i>Enterococcus faecium</i>	6
<i>E. mori</i> <i>Enterobacter mori</i>	6
<i>E. mundtii</i> <i>Enterococcus mundtii</i>	6
GTDB-Tk the Genome Taxonomy Database Toolkit	8
<i>K. oxytoca</i> <i>Klebsiella oxytoca</i>	6
MAG metagenomics-assembled genome	7
MIBiG Minimum Information about a Biosynthetic Gene cluster	6
<i>P. aeruginosa</i> <i>Pseudomonas aeruginosa</i>	6
<i>P. putida</i> <i>Pseudomonas putida</i>	6
pHMM profile hidden Markov model	5

Prodigal procaryotic gene recognition and translation initiation site identification	5
<i>S. aureus</i> <i>Staphylococcus aureus</i>	6
<i>S. epidermidis</i> <i>Staphylococcus epidermidis</i>	6