Bioinformatic approach to analysis of plasmid pool in metagenomes from polluted soils

Paweł Krawczyk[1], Adam Sobczak[12], Leszek Lipiński[1], Andrzej Dziembowski[12]

[1]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland
[2] Institute of Genetics and Biotechnology, University of Warsaw, Poland
Contact: p.krawczyk@ibb.waw.pl

INTRODUCTION

Plasmids are bacterial mobile genetic elements that facilitates rapid evolution and adaptation of their hosts to changing environmental conditions. Genes coded on plasmids has a big impact on their bacterial hosts, their importance for soil properties and fertility cannot be disregarded. This is especially important in agricultural soils, which are often treated with toxic chemical compounds, like pesticides. Soils contaminated with pesticides are often enriched in bacterial or fungi species capable to degrade deadly compounds. Moreover genes located on mobile elements are known to play important role in resistance of microorganisms to chemical pollution. In presented work bioinformatic approach to plasmid diversity in pesticide contaminated soils was described..

SAMPLES

Soil samples (coming from 7 sites) were collected in 2010 and 2011 during ellimination of underground infrastructure where obsolete pesticides where stored from 1960's.

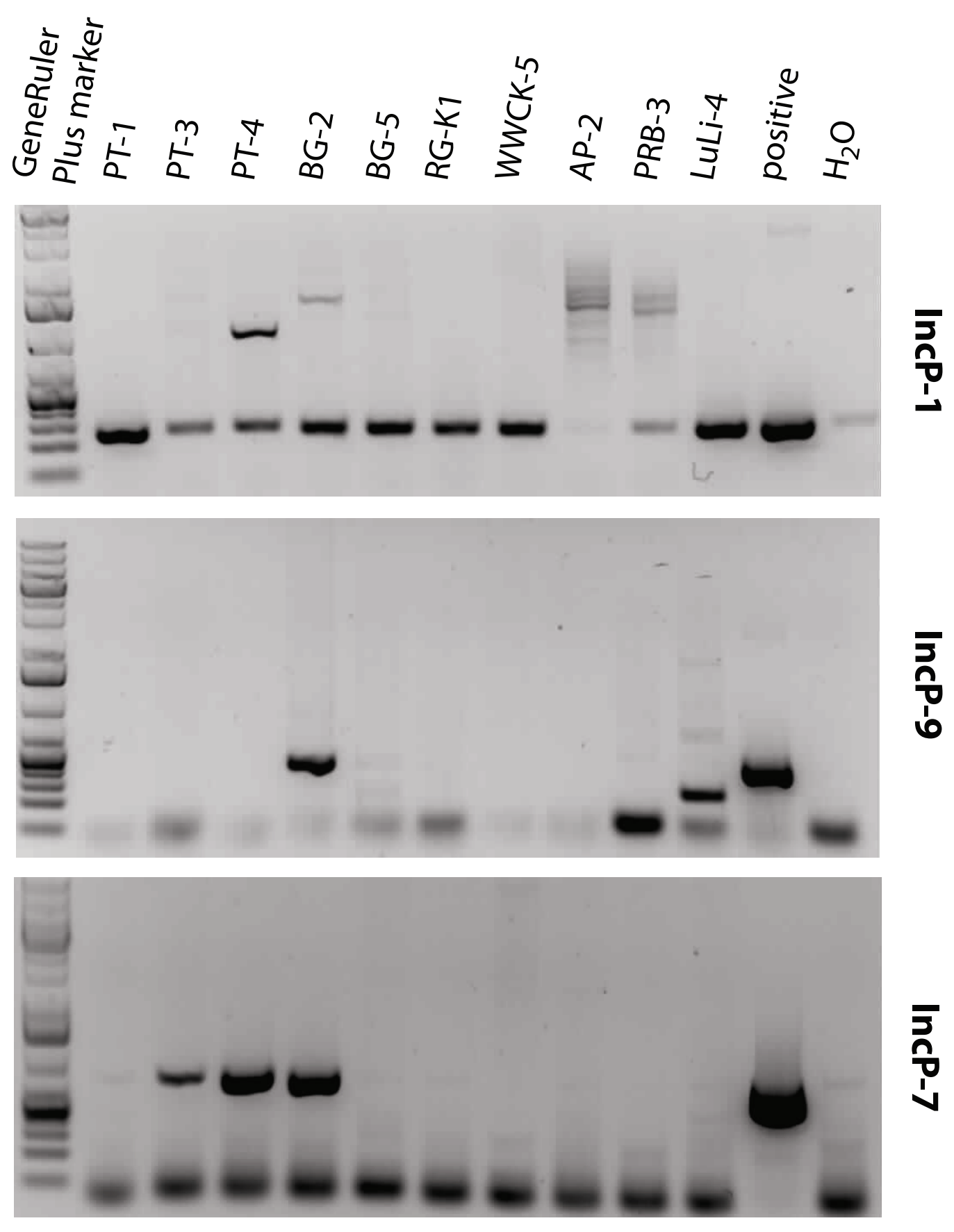
DNA was isolated with modified method of Zhou and collegues (1996). 16S rRNA genes fragments (357-786) were amplified via PCR and sequenced on 454 GS FLX Titanium machine.

Physicochemical properties of collected samples were assesed with standard methods. Pesticides were detected and quantified by GC-MS and HPLC-MS (Tab. 1).

sample	species (16S based)	contaminants concentration [ng/ul]
PT-1	36	73382.33
PT-3	21	26848.33
PT-4	135	221637.72
BG-2	183	2781.32
BG-5	419	28288.39
RG-K1	84	426.37
WWCK-5	360	78.36
AP-2	794	46.00
PRB-3	280	2.0
LuLi-4	535	27216.00

Tab. 1. Basic characteristics of analyzed samples

IncP PLASMIDS DIVERSITY IN CONTAMINATED SOILS



Replicons of selected IncP groups were detected with standard PCR replicon typing method, using primers trfa21 /trfa22 (IncP1, targeting fragment of trfA2 gene), tolRepF/tolRepR (IncP9, targeting fragment of rep gene) and RepRmsF/RepRmsR (IncP7, targeting fragment of rep gene).

Analysis revealed presence of IncP1 plasmids in all analysed samples. IncP9 and IncP7 was restricted to more polluted samples (Fig. 1).

Fig. 1 PCR replicon typing of metagenomic samples coming from pesticide-contaminated soils.

HIGHLIGHTS

* PCR replicon typing revealed presence of IncP-1,IncP-7 and IncP-9 plasmids in organochlorine-poluted soils

* Sequence signatures can be used to predict plasmid sequences in metagenome sequencing

* Proteobacteria plasmids are most abundant in organochlorine polluted soils and carry genes involved in aromatic compounds degradation

MACHINE LEARNING OF PLASMID SEQUENCE SIGNATURES

Kmer profiles (for 3,4 and 5-mers) of 2288 genomic and 1656 plasmid sequences downloaded from NCBI were calculated using Jellyfish. Obtained profiles were used for supervised training of Self Organizing Map (SOM) using kohonen library in R.

Training was performed both for plasmid/chromosome classification as well as for phylogenetic origin. Obtained model got 90% accuracy, validated using publicly available plasmidome data (Brown Kav et al, 2012).

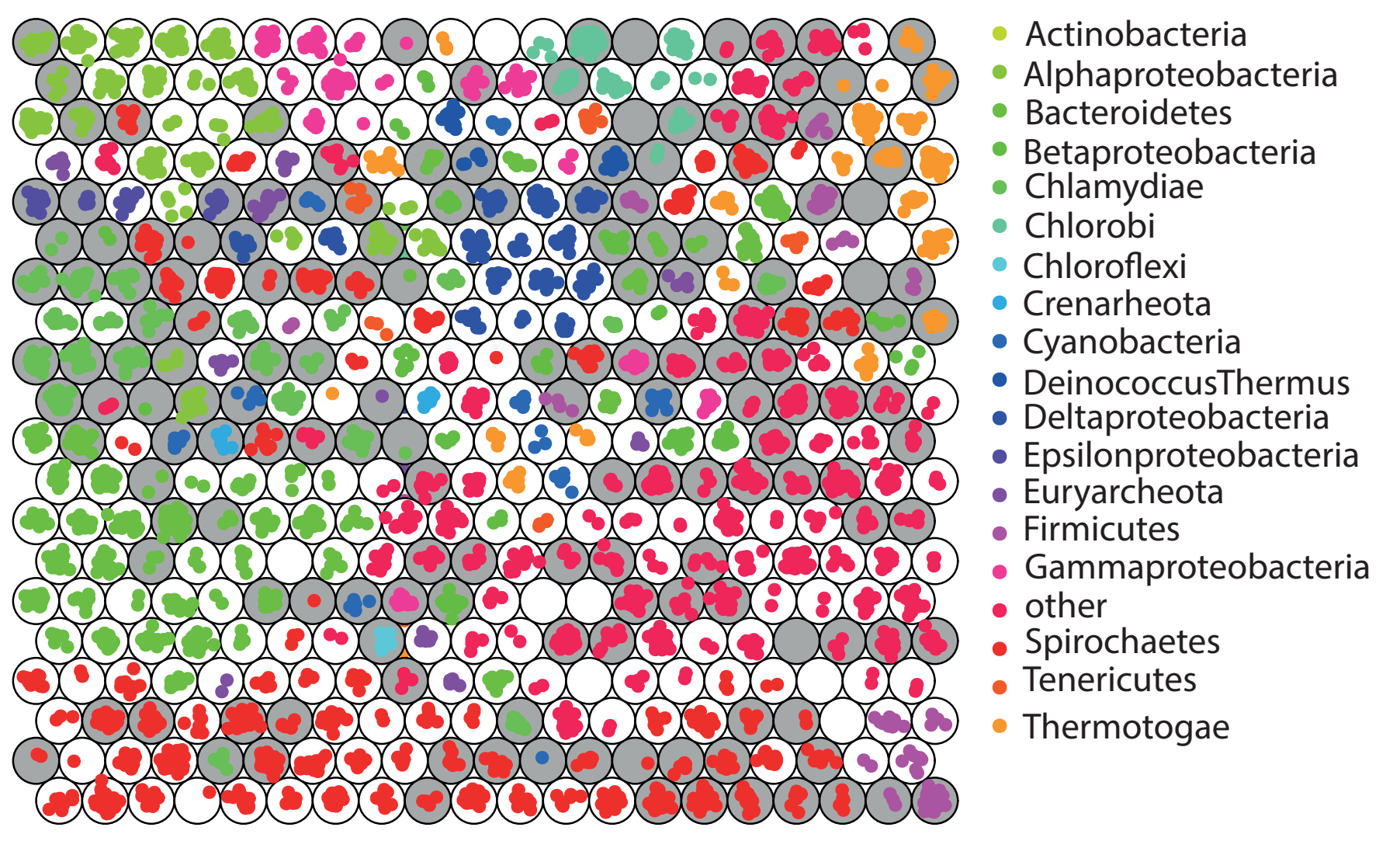


Fig. 2. Phylogenetic clustering of obtained map. Grey - plasmid sequences

APPLICATION OF OBTAINED MODEL TO REAL METAGENOMIC DATA

Samples with the highest pollution levels (PT-4, PT-1, BG-5 and RG-K1) were sequenced on Illumina platform and assembled using MetaVelvet or CLC Genomics Workbench. Obtained contigs were filtered out of sequences shorter than 1000 nt. For remaining sequences kmer frequencies were obtained using the same method as applied to machine learning process and used for prediction of contig origin.

Analysis revealed that most (~60%) of contigs (mostly shorter ones) were identified to be of plasmid origin. Predicted phylogenetic distribution was mostly consistent with annotation data (Fig.3)

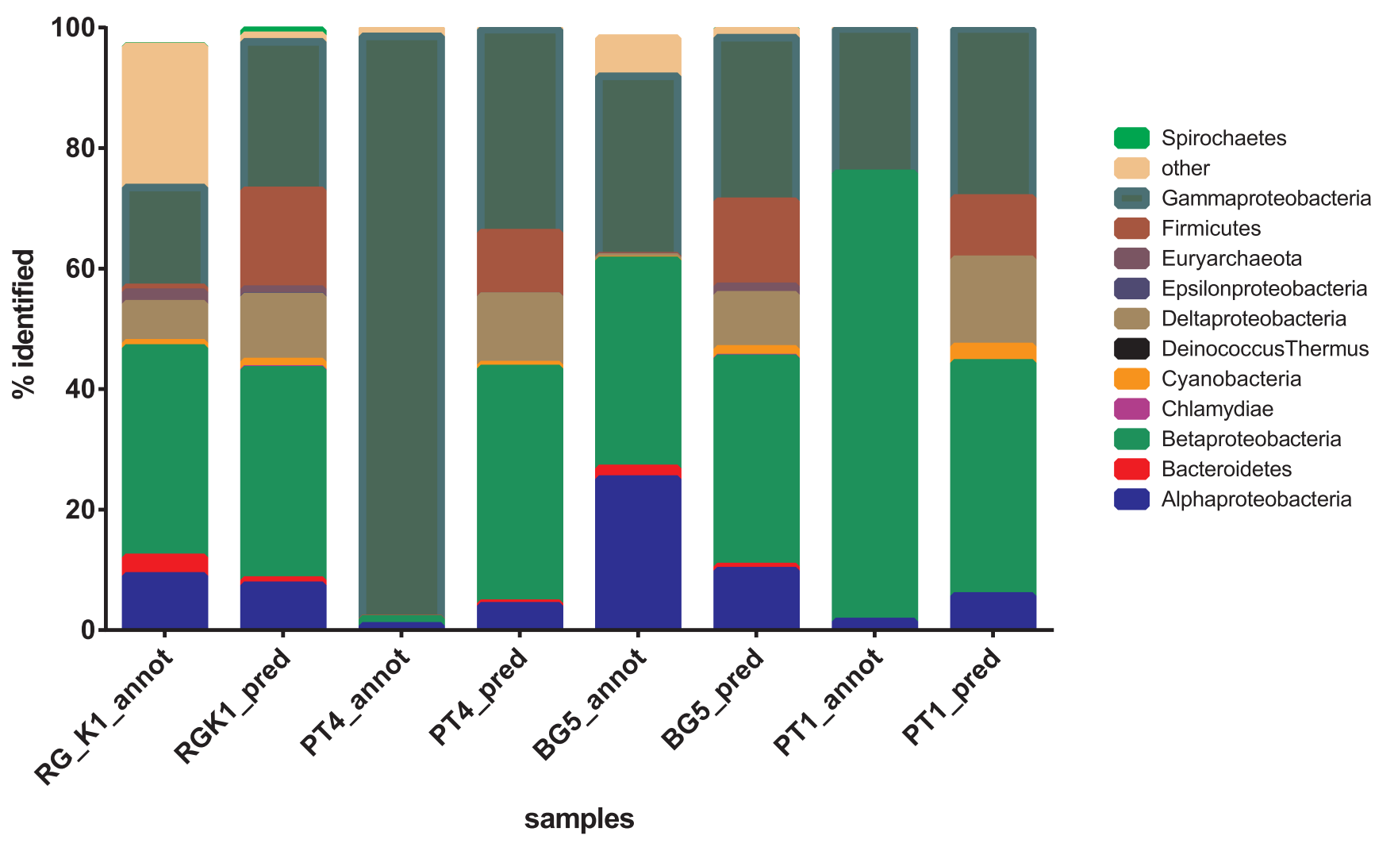


Fig. 3. Phylogenetic distribution of contigs classified to plasmids. Pred - model prediction, annot - MEGAN annotation

ANNOTATION OF PLASMID SEQUENCES

Prodigal was used for identification of coding sequences, then blastp against nr database and MEGAN were used for functional annotation. 40% of orfs were assigned to any functional category.

As expected, many orfs were annotated to SEED categories of virulence or transposable elements (containing plasmid structural genes). Metabolism of Aromatic Compounds constituted from 0.3% in RG_K1 sample, to 1.4% in PT1 sample (Fig. 4).

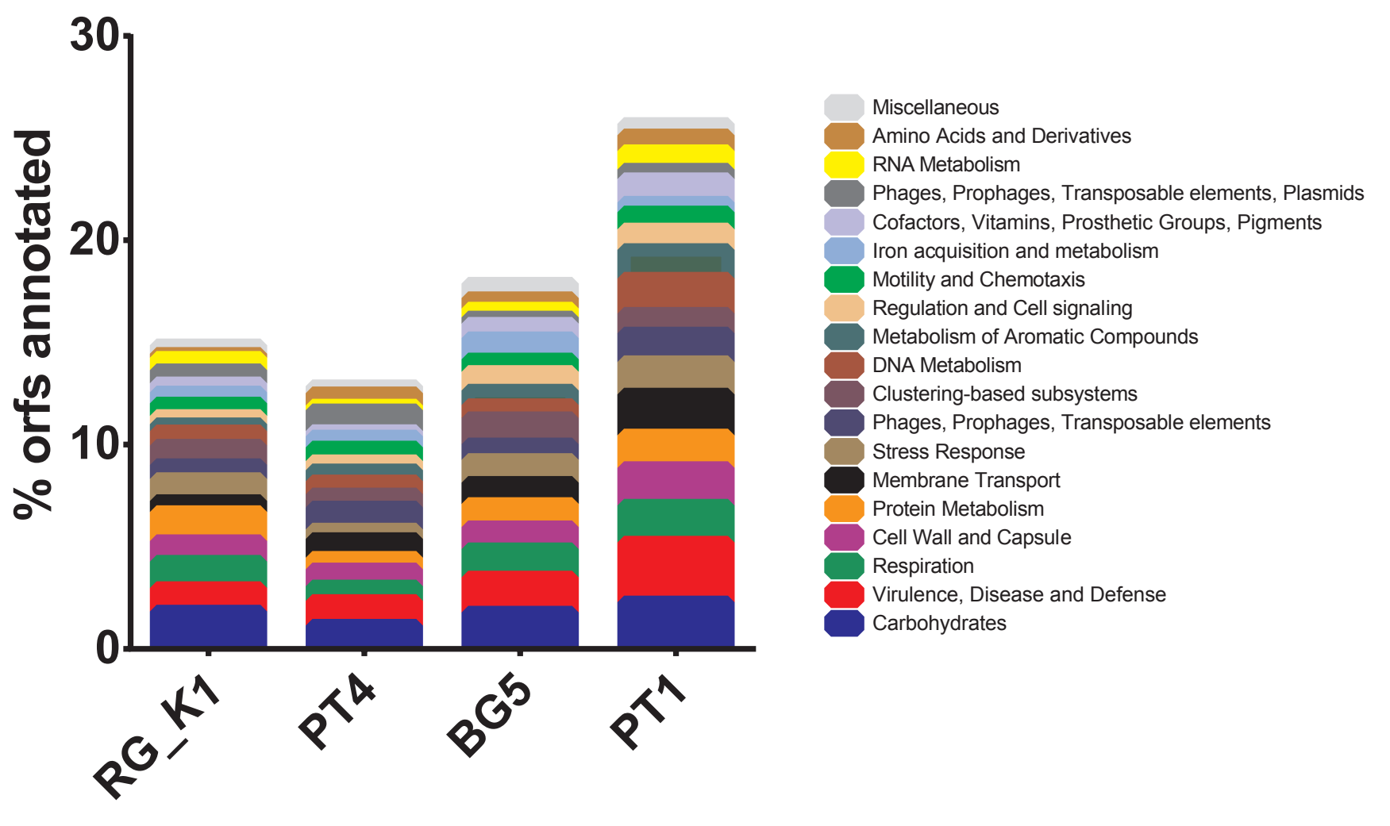


Fig. 4. Most common SEED categories in annotated plasmid contigs

ACKNOWLEDGEMENTS

This work was supported by the Polish National Science Centre grant DEC-2012/05/N/NZ9/01393

INNOVATIVE ECONOMY
NATIONAL COHESION STRATEGY

This investigation was cofinanced by the European Union from the resources of the European Regional Development Fund under the Innovative Economy Programme

EUROPEAN UNION
EUROPEAN REGIONAL
DEVELOPMENT FUND

HUMAN CAPITAL
NATIONAL COHESION STRATEGY

This work has been supported with a scholarship from the European Social Fund, Human Capital Operational Programme for the execution of the Programme "Support for bio med tech scientists in technology transfer
(UDA-POKL.08.02.01-14-041/09)

EUROPEAN UNION
EUROPEAN
SOCIAL FUND