XLIII Congreso de la Sociedad Española de Genética
SEG2023 Valencia, del 21 al 23 de junio

Calculating genomic signature distances between phages and their bacterial host for distinguishing lytic and lysogenic phages

Vicente Arnau[1,2,3] , Wladimiro Diaz-Villanueva[1,2,3], Jorge Mifsut[1], Paula Villasante[4], Pablo Román[1], Mária Džunková[1]

1.Institute for Integrative Systems Biology(I2SysBio), Universitat de València(UV) and Consejo Superior de Investigaciones Científicas (CSIC), Valencia, Spain
2.Genomic and Health Area, Foundation for the Promotion of Sanitary and Biomedical Research of the Valencia Region(FISABIO), Valencia, Spain
3.Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública(CIBEResp), Madrid, Spain
4.Universitat Oberta de Catalunya i Universitat de Barcelona

INTRODUCTION

Environmental impact of uncultured phages is shaped by their preferred life cycle (lytic or lysogenic), however, our ability to predict it is very limited. In recent years several studies have shown that Homology-free methods (genomic signature) can be useful for the classification of viral genomes (1) and for characterizing bacteriophages by comparing their genomic signature with that of their hosts to obtain host-phage relationships and determine their lifestyle (2).

METHODS

We present two approaches to discriminate lysogenic and lytic phages based on the comparison of the similarity of their genomic signatures with those of their hosts which may reflect their co-evolution.

A) The Euclidean distance between the relative frequencies of short length k-mers, in our case k = 4 (k4freq) and

Word in Seq_1	Frequency	Relative Frequency
AAAA	185588	0.219
AAAC	47630	0.056
AAAG	57613	0.068
AAAT	137216	0.162
AACA	39934	0.047
. . .		
TTTG	48929	0.058
TTTT	184609	0.218

Words in Seq_2	Frequency	Relative Frequency
AAAA	18	0.175
AAAC	5	0.049
AAAG	6	0.058
AAAT	15	0.146
AACA	4	0.039
. . .		
TTTG	5	0.049
TTTT	19	0.184

For a given k-mer w, its occurrence in a contig X is defined as X_w and the relative frequency of this k-mer is defined as:

$$f_w^X = \frac{X_w}{\sum_w X_w}$$

Following the guidelines of Vinga & Almeida (2003) [44], we calculated the Euclidean distance (k4freq) between the pairs of genomes:

$$Eu(X,Y) = \sqrt{\sum_{w \in S^k} |f_w^X - f_w^Y|^2}$$

B) Alignment-free comparison based on exact k=14 oligonucleotide matches (k14exact). We proposed a new distance of similarity for high values of k (k > 14) [3], where the value of 4k is two orders of magnitude larger than the size of the largest genome.

$$SX = \sum_w X_w \qquad SY = \sum_w Y_w$$

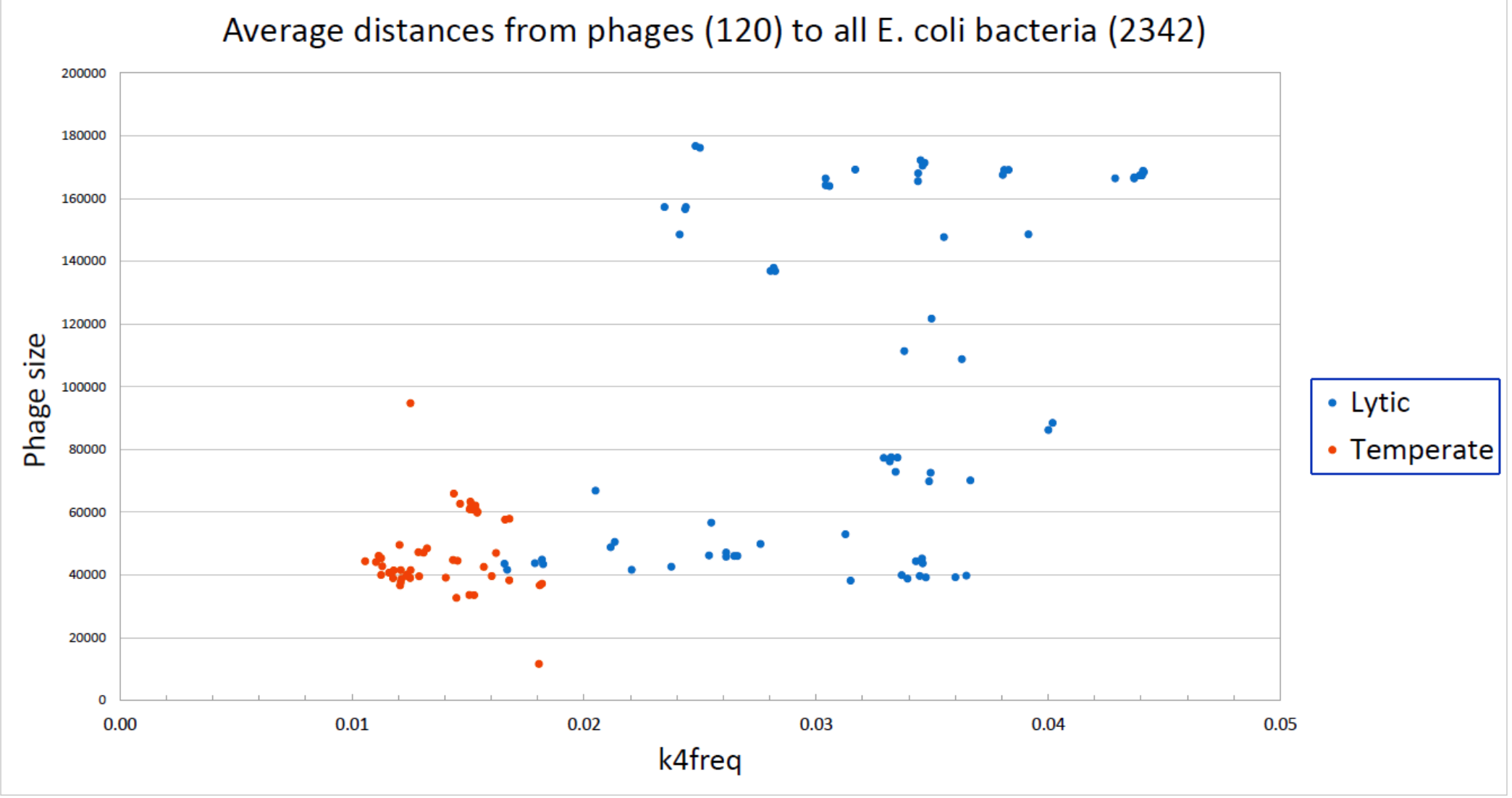
If SX < SY, we define the similarity function SIM between two sequences as:

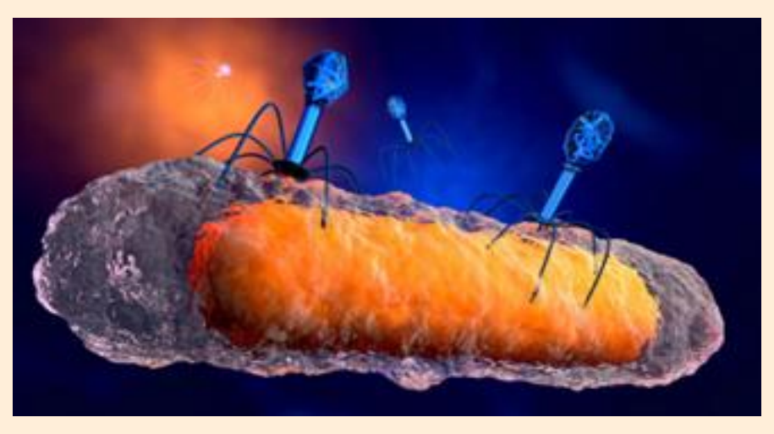
$$SIM(X,Y) = \frac{\sum_w X_w}{SX} \qquad \forall (Y_w > 0)$$

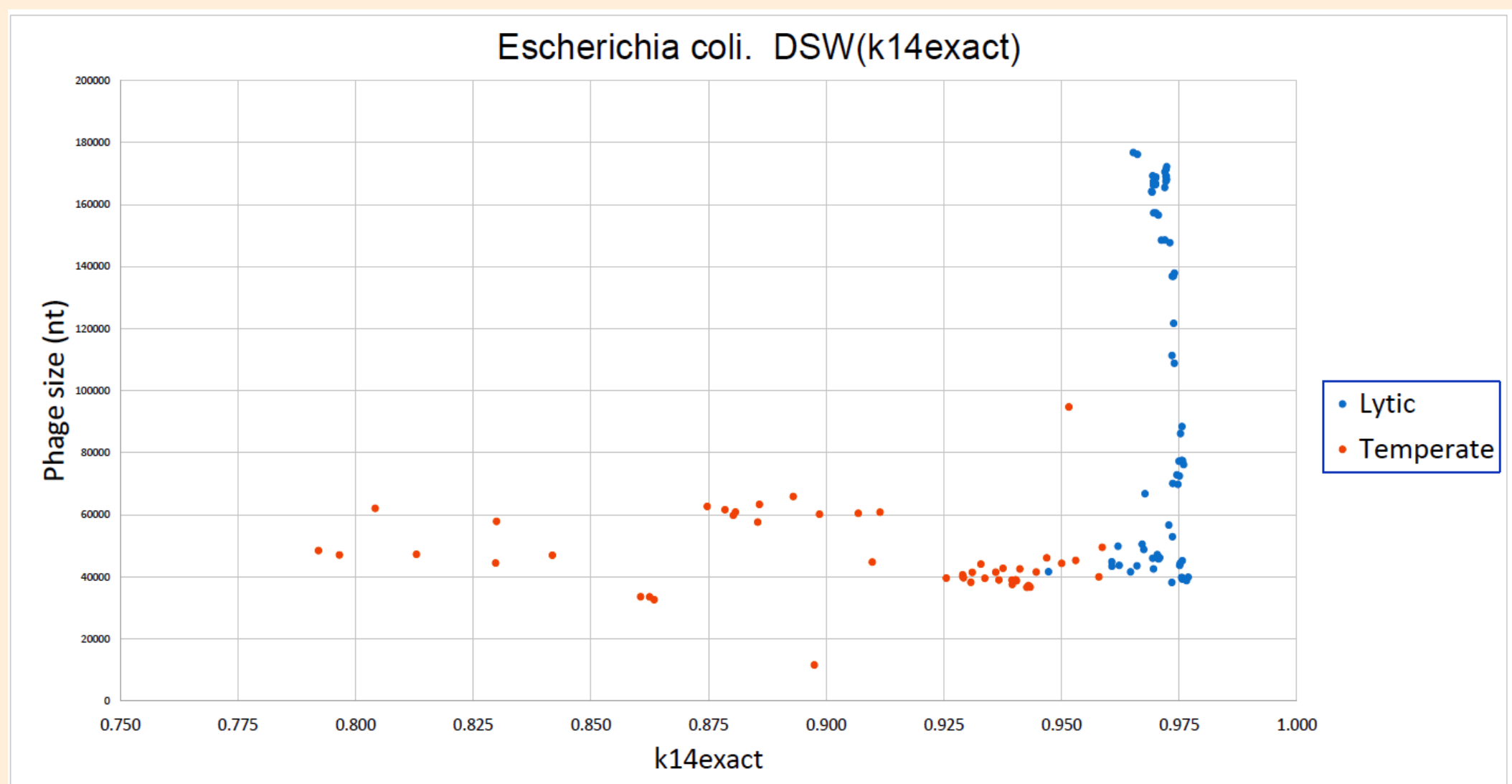
Finally, we define the distance measure DSW(k14exact) as the inverse of the Similarity function, as follows:

$$DSW(X,Y) = (1 - SIM(X,Y))$$

RESULTS







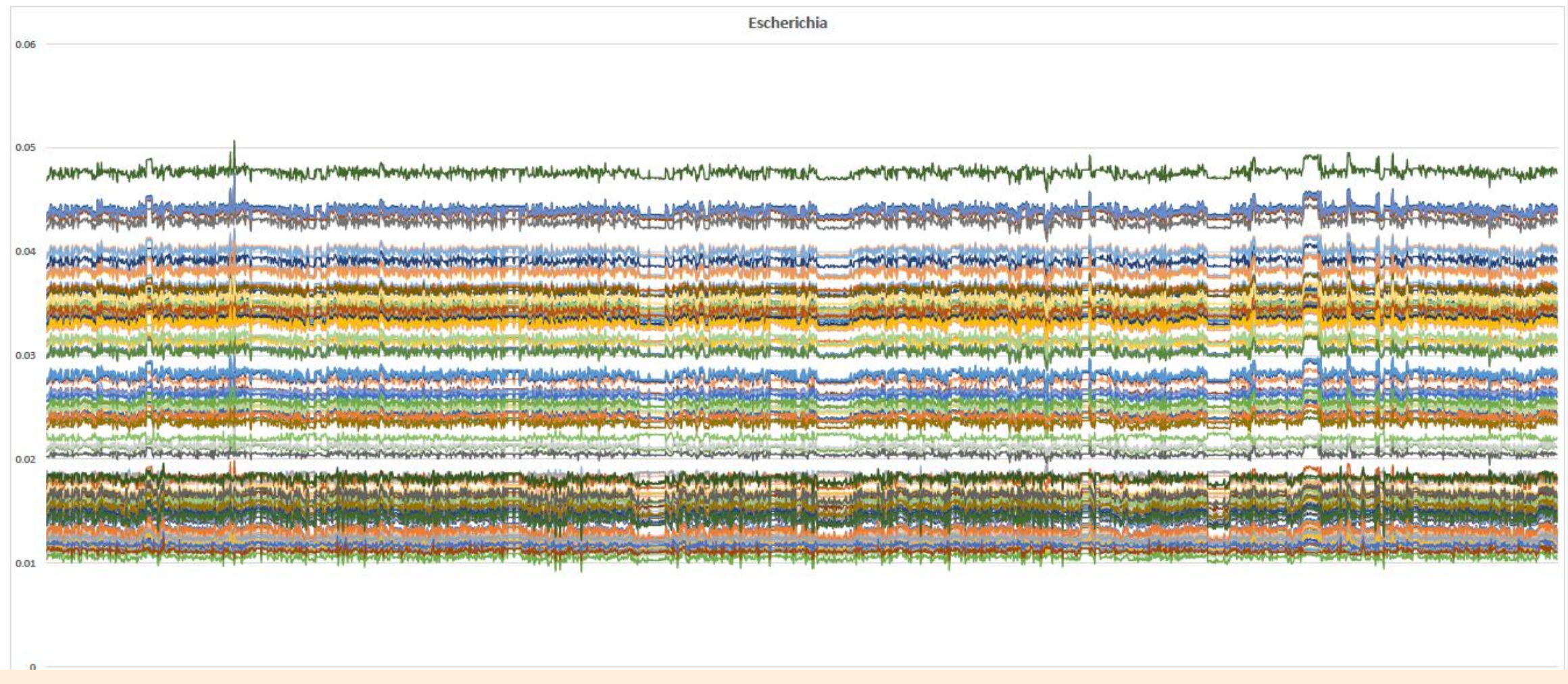
We explored 5126 reference bacterial host strains and 284 associated phages from NCBI RefSeq.

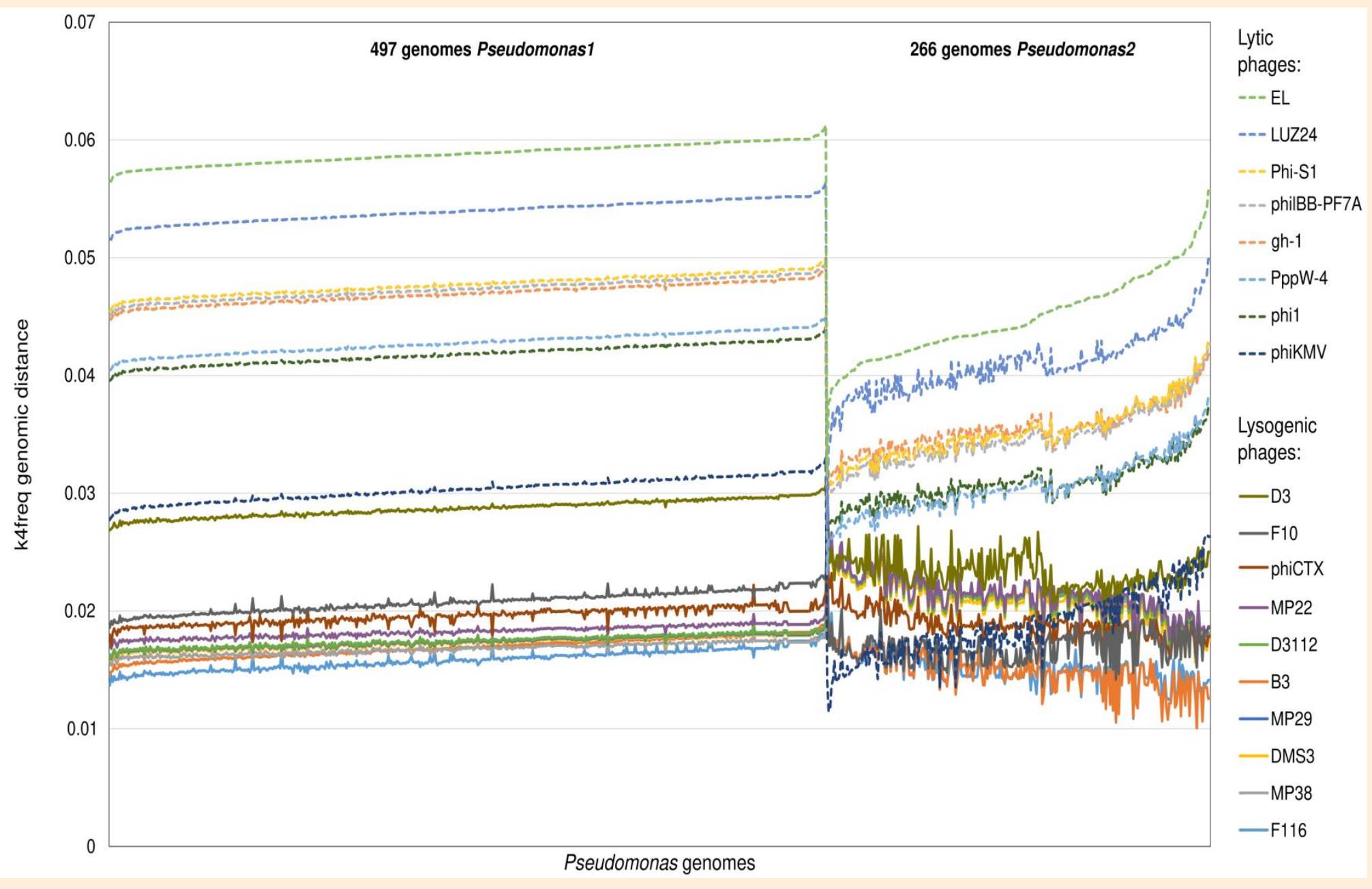
The thresholds for distinguishing lysogenic and lytic phages using the **k4freq** method was 0.026, and 0.955 using the **k14exact** method.

The k14exact performed better than k4freq. The example shows E. coli phages.

Most of the phages are associated to their host on the level of genus or species. Nevertheless, the different strains of the same species can have very different morphological and physiological characteristics and different reactivity with the phages. Therefore, we assess genomic distances of lytic and lysogenic phages to all bacterial strains available in NCBI. We clustered the strains by their hexamers frequencies to obtain strain groups with similar genomic content.

All 2342 Escherichia strains formed a single hexamer-based group and they had similar distances to the set of 120 Escherichia phages.





In contrast, the 727 Pseudomonas strains were split into two hexamer-based groups which had different distances to the set of 18 Pseudomonas phages. Some discrepancies were observed: e.g. lytic phage phiKMV appeared among the lysogenic phages when compared with Pseudomonas strains group 2.

For both bacterial genera, clear differenciation between lysogenic and lytic phages was observed.

CONCLUSIONS & NEAR FUTURE

- The oligonucleotide-based genome analysis methods can be used for predictions of life cycles of phages
- In the near future, we plan to study uncultured environmental phages by applying this method to large metagenomic and single-cell genomics data sets

CONTACT

Vicente.Arnau@uv.es or Maria.Dzunkova@uv.es

@Vicent_Arnau @MDzunkova

ACKNOWLEDGMENTS

The computations were performed on the HPC cluster *Garnatxa* at the Institute for Integrative Systems Biology (I2SysBio). I2SysBio is a mixed research center formed by the University of Valencia (UV) and Spanish National Research Council (CSIC).

REFERENCES

1. Pride DT et al. (2006) Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 7, 8.
2. Deschavanne P et al. (2010) The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. Virol J 7, 163.
3. Swain, M.T. et al. Interpreting alignment-free sequence comparison: What makes a score a good score? NAR Genom. Bioinform. 2022, 4, lqac062.