GBDP for genome-based phylogenetic analysis and classification of viruses

Jan P. Meier-Kolthoff, Alexander Auch, Johannes Wittmann, Hans-Peter Klenk and Markus Göker*

Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures

* Correspondence: Markus Göker <markus.goeker@dsmz.de>

Challenge

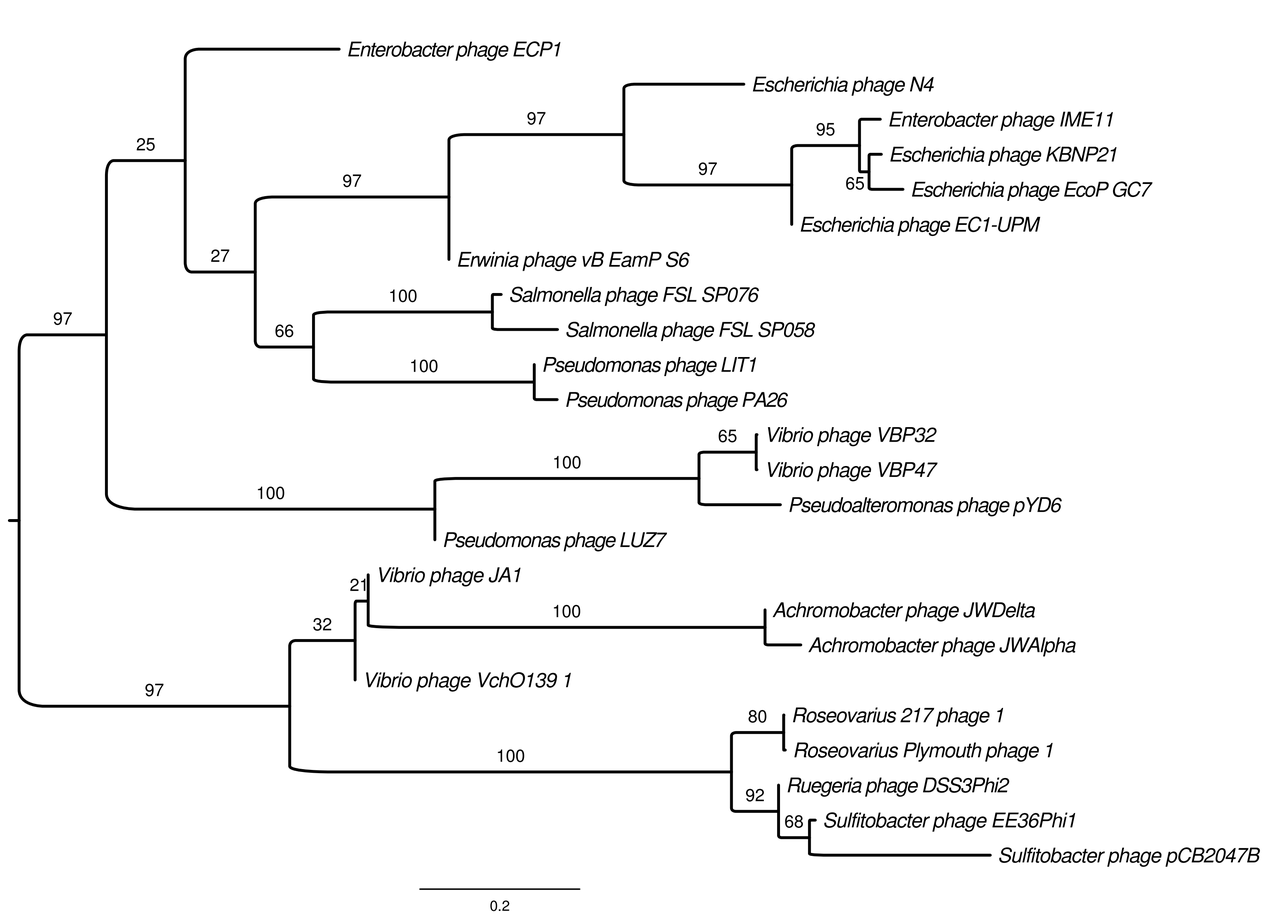
Taxonomy entails the classification, nomenclature and identification of organisms. High-throughput sequencing and high-throughput culturing provide for an increased availability of genome sequences and their subsequent use in phylogenetic analyses and genome-based classification. Challenges particularly related to virus genomes are their vastly deviating compositions and their low degree of sequence similarity. Thus the question arises whether virus phylogeny really is best be tackled by inference from multiple sequence alignments.

Method

The Genome-Blast Distance Phylogeny approach (GBDP), which uses only local pairwise sequence alignments, was originally introduced for the phylogenetic inference from microbial whole genomes [1]. It had to deal with largely differing genome sizes, repetitive sequences and paralogy. Strategies implemented into GBDP also allowed for phylogenetic reconstruction from plastid [2] and fungal genomes [3]. A recent addition is branch support via pseudo-bootstrapping [4]. GBDP is phylogenetically reliable [5] and was successfully applied to species delineation in prokaryotes [6].

Application

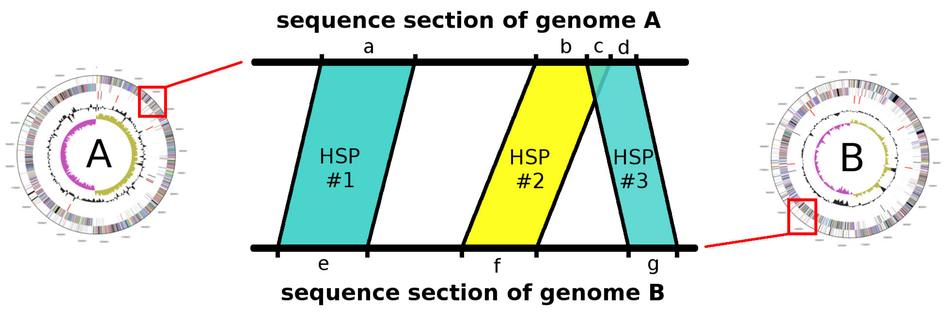
In our study on *Achromobacter* phages [7], GBDP was used to phylogenetically place the new phages within the N4 family.



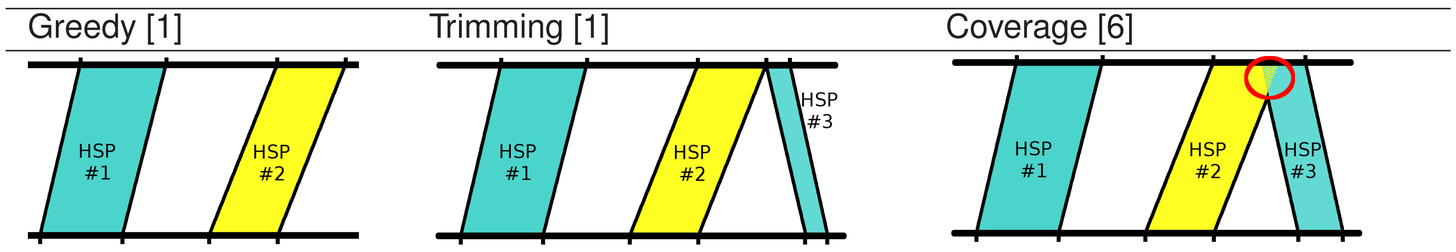
Resolution as indicated by pseudo-bootstrap support was accordingly higher than in single-gene phylogenies. Among the data input, HSP filtering approaches and distance formulas available in GDBP, the combination of protein sequences, "Trimming", relaxed e-value filtering and formula d_5 (logarithmised d_4) worked best for this data set.

GBDP principle

1 BLAST run between two (draft) genomes A and B resulting in the usual set of matches called "**high-scoring segment pairs**" (HSPs). An example for such homologous regions is:



2 Can this information now directly be used to calculate an **intergenomic distance** between A and B? Almost! Beforehand we need to correct for **overlapping HSPs** (segment "c" in our example), most likely caused by **paralogous genes**, which could **bias** the distance between A and B. The **HSP-filtering** approaches are:



Drop smaller HSP
Pro: faster
Con: information loss

Remove overlapping parts
Pro: preserves information
Con: computation complex

Merge overlapping HSPs
Pro: preserves information
Con: implementation difficult

3 Final calculation of a distance between A and B using the filtered HSP set. Ten different **distance formulas** d_0 - d_9 are available [6], examples are:

$$d_0(A,B) = 1 - \frac{\text{sum of HSP lengths}}{\text{sum of genome lengths}}$$

$$d_4(A,B) = 1 - \frac{\text{\# identical base pairs in HSPs}}{\text{total length of all HSPs}}$$

$$d_6(A,B) = 1 - \frac{\text{\# identical base pairs in HSPs}}{\text{sum of genome lengths}}$$

Conclusion

GBDP makes efficient use of the information available in whole virus genomes. Like inference from concatenated multiple sequence alignments, GBDP delivers branch support, but with a lower computational cost. As some settings also allow for incompletely sequenced genomes, GBDP is a universal tool for virus phylogeny and classification.

References

[1] Henz S, Huson D, Auch AF, Nieselt-Struwe K, Schuster S (2005) Whole-genome prokaryotic phylogeny. Bioinformatics 21: 2329–2335.

[2] Auch AF, Henz SR, Holland BR, Göker M (2006) Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences. BMC Bioinformatics 7: 350.

[3] Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M (2014) GBDP on the grid. In: Schulz J, Hermann S, editors, Hochleistungsrechnen in Baden-Württemberg, KIT Scientific Publishing. pp. 83–102.

[4] Meier-Kolthoff JP, Auch AF, Klenk HP, Göker M (2014) Highly parallelized inference of large genome-based phylogenies. Concurrency Computat Pract Exper 26: 1715–1729.

[5] Patil KR, McHardy AC (2013) Alignment-free genome tree inference by learning group-specific distance metrics. Genome Biol Evol 5: 1470–84.

[6] Meier-Kolthoff J, Auch A, Klenk HP, Göker M (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14: 60.

[7] Wittmann J, Dreiseikelmann B, Rohde M, Meier-Kolthoff JP, Bunk B, et al. (2014) First genome sequences of *Achromobacter* phages reveal new members of the N4 family. Virol J 11: 14.