Al-Farabi Kazakh National University

TYGS: a novel high-throughput platform for state-of-the-art genome-based taxonomy

Online Lecture 8

by

Mashzhan Akzhigit

2020

aj.akzhigit@gmail.com

Outline

- What is the TYGS?
- What are the main advantages of using GGDC?
- VICTOR tool
- What are the main advantages of using this phylogeny pipeline?

DSMZ server

- TYGS (Type (Strain) Genome Server) https://tygs.dsmz.de/
- GGDC (Genome to Genome Distance Calculator) http://ggdc.dsmz.de/ggdc.php#
- VICTOR (Virus Classification and Tree Building Online Resource) http://ggdc.dsmz.de/victor.php#
- Single-Gene Trees http://ggdc.dsmz.de/phylogeny-service.php#

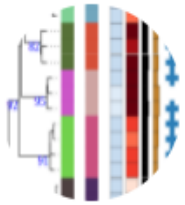
Main page

Genome-to-Genome Distance Calculator | Home | News | GGDC | TYGS | VICTOR | Single-Gene Trees | FAQ | Leibniz Institute DSMZ

TYGS: a novel high-throughput platform for state-of-the-art genome-based taxonomy.

The Type (Strain) Genome Server, TYGS, allows for genome-based prokaryote taxonomy using a comprehensive database of genomic and taxonomic data.

Try TYGS

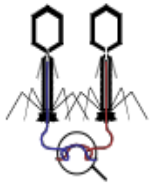


TYGS

TYGS, the Type (Strain) Genome Server, is a user-friendly high-throughput web server for genome-based prokaryote taxonomy. Microbial taxonomy needs genome-based analyses but these can be complex and require expert knowledge. TYGS fills this gap. TYGS is as reliable as the

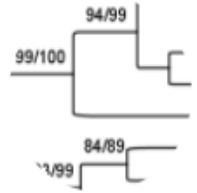
GGDC

The pragmatic species concept for *Bacteria* and *Archaea* is ultimately based on DNA-DNA hybridization (DDH), a method known to be tedious. The GGDC is a state-of-the-art in silico method for genome-to-genome comparison, thus reliably mimicking conventional DDH, except for its



VICTOR

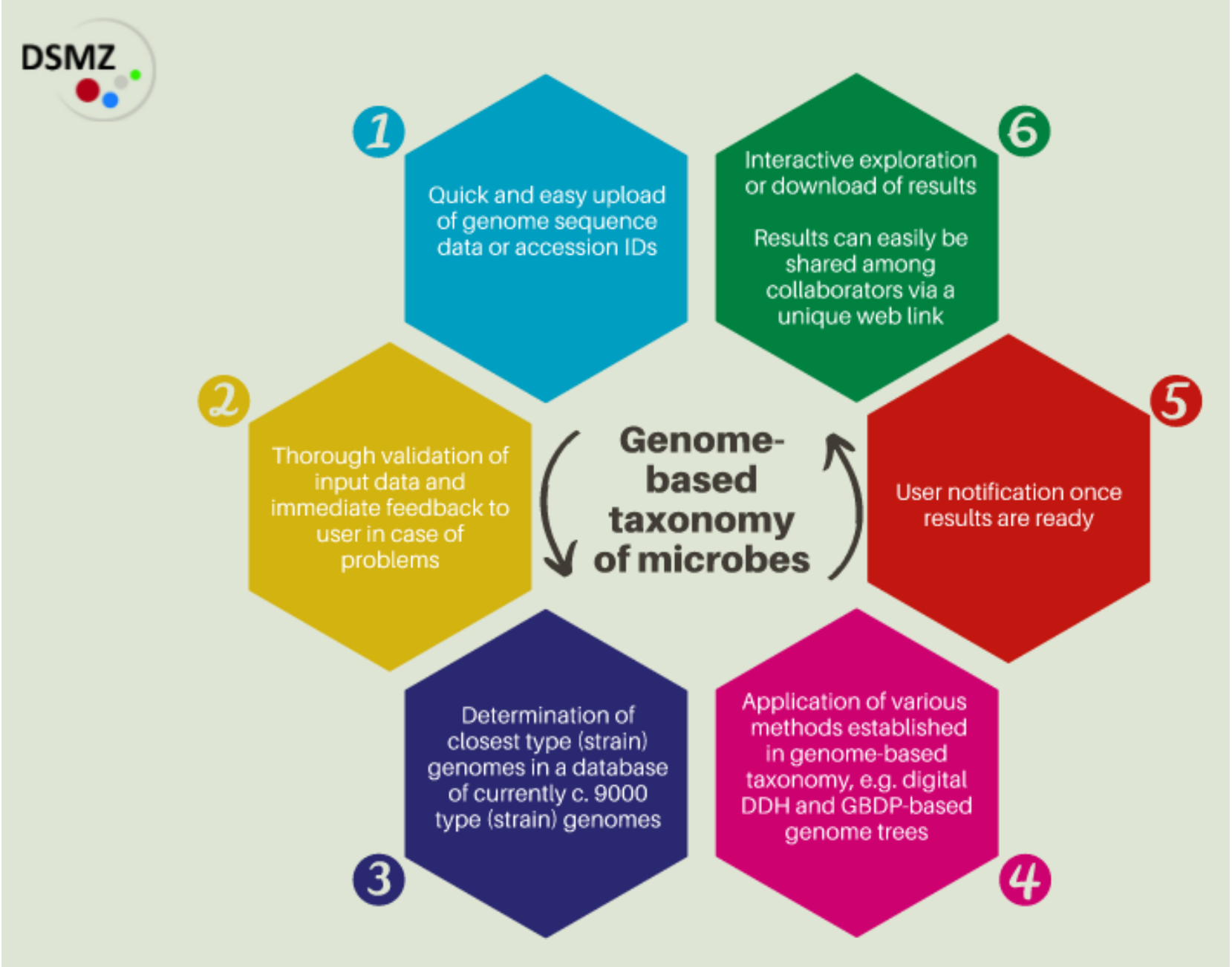
Virus taxonomy should be strongly informed by sequence-based methods. The web service VICTOR compares phages using their genome or proteome sequences. Results include phylogenomic trees with branch support and suggestions for the classification at the species, genus



Single-Gene Trees

A service which (i) infers maximum-likelihood and maximum-parsimony **phylogenies** using state-of-the-art methods and (ii) calculates exact **pairwise similarities** between gene sequences under settings qualifying for the application of **phylum-specific** 16S similarity thresholds.

Scientific Background



How is my privacy respected?

- All uploaded sequence data are deleted on the server within 24 hours after the calculations have been completed. Both the sequence data and the e-mail addresses of the users are stored only on the server and only during that time period. Both sequence data and the e-mail addresses are not made available to third parties.

What kind of format can I upload data?

- FASTA
- GenBang or GenBank accession number

What is the TYGS?

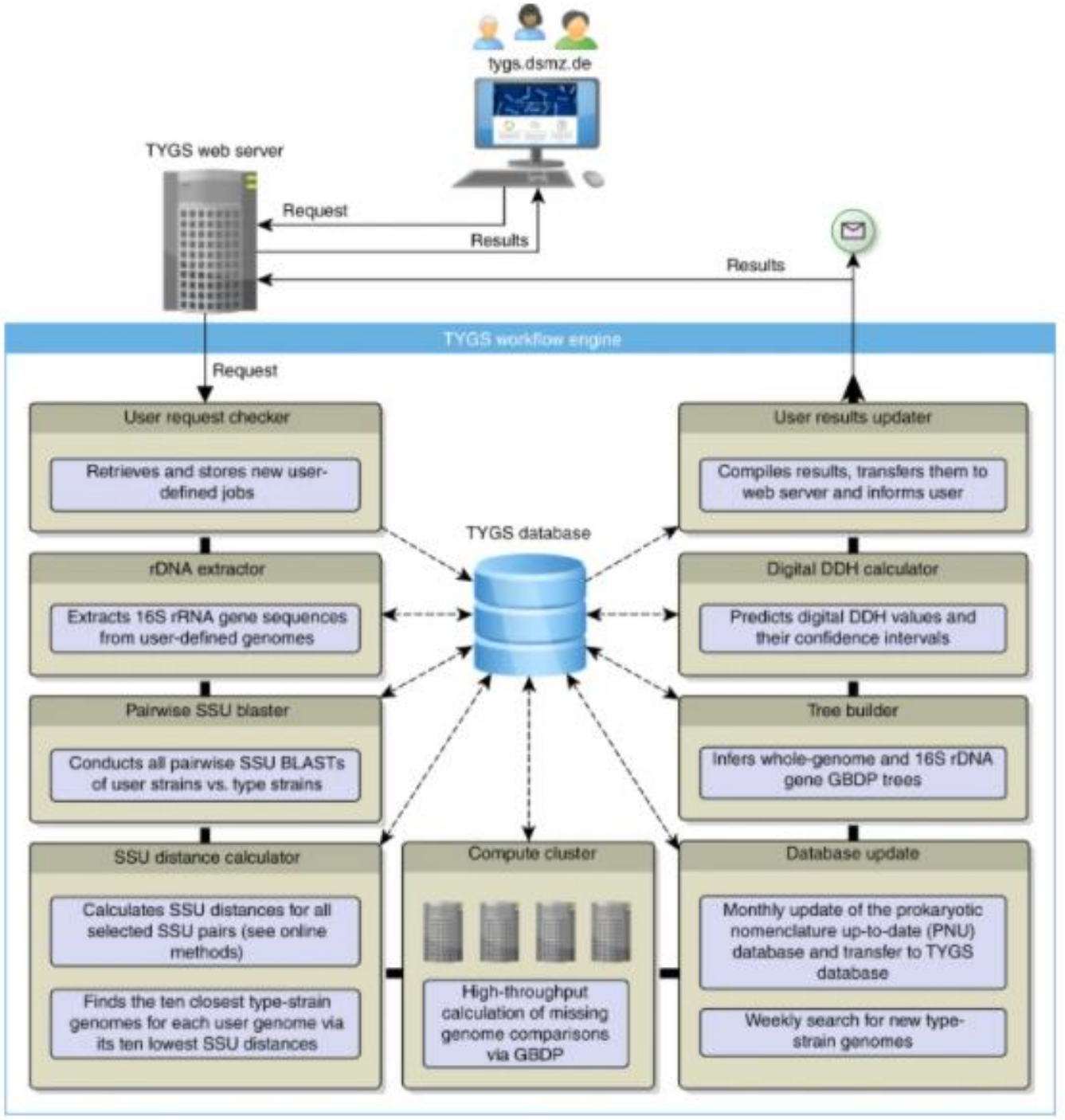
- YGS, the Type (Strain) Genome Server, a user-friendly high-throughput web server for genome-based prokaryote taxonomy, connected to a large, continuously growing database of genomic, taxonomic and nomenclatural information. It infers genome-scale phylogenies and state-of-the-art estimates for species and subspecies boundaries from user-defined and automatically determined closest type genome sequences. TYGS also provides comprehensive access to nomenclature, synonymy and associated taxonomic literature.

More information here:

Meier-Kolthoff, J.P., Göker, M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat Commun 10, 2182 (2019). https://doi.org/10.1038/s41467-019-10210-3

TYGS workflow.

First, a request is prepared and submitted by a user to the TYGS web server.
The TYGS workflow engine periodically checks for new user requests and imports these data into the central TYGS database.
The data are processed by several independent services.
Once the results arrive at the web server the user gets informed via e-mail and can conveniently access the results via the web browser of choice. Solid lines indicate the logical program flow, whereas data flow from and to the database is indicated by dashed lines



Results ec388433-020e-47c2-97b0-0d6430982f72

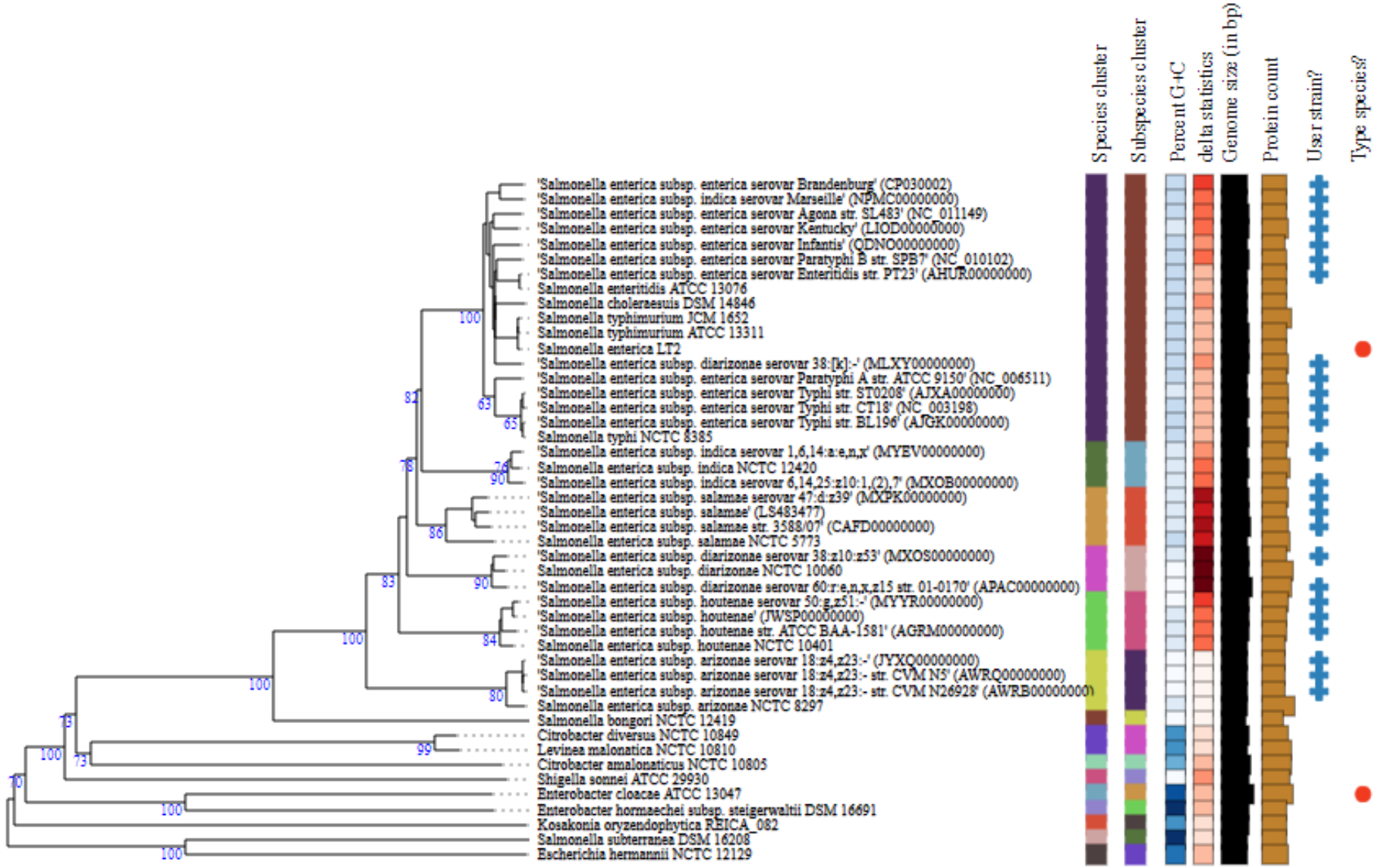
Job runtime 281 m | Example

Download PDF Report | Identification | Trees | Pairwise comparisons | Strains in your dataset | Methods

Table 1: Phylogenies

Figure	View tree	No. strains	Average branch support	δ statistics	Download tree
Figure 1: GBDP tree (whole-genome sequence-based)		46	54.1 %	0.16	Newick
Figure 2: GBDP tree (16S rDNA gene sequence-based)		46	49.1 %	0.342	Newick

Whole-genome sequence-based tree



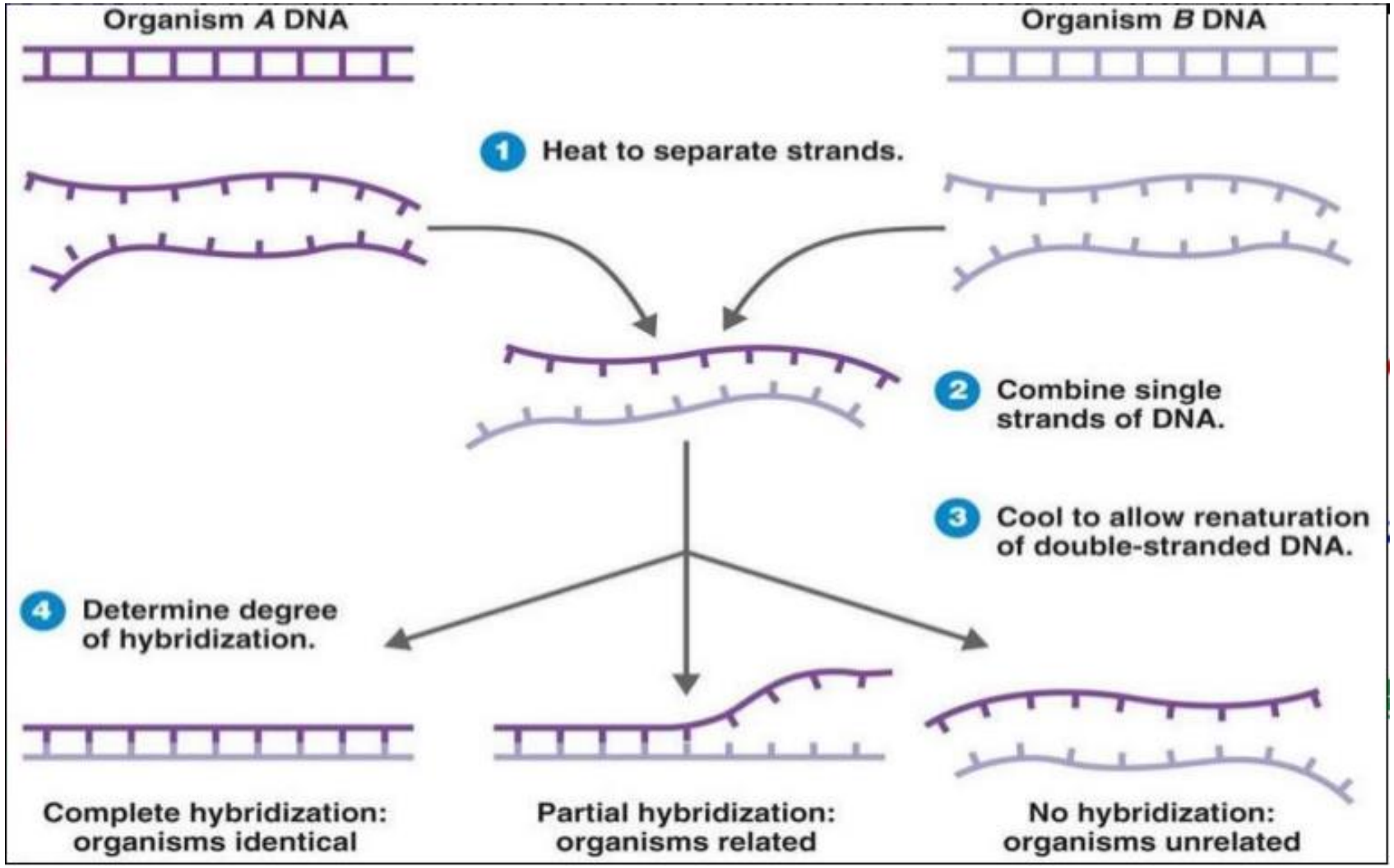
GGDC (Genome to Genome Distance Calculator)

- The pragmatic species concept for Bacteria and Archaea is ultimately based on DNA-DNA hybridization (DDH). While enabling the taxonomist, in principle, to obtain an estimate of the overall similarity between the genomes of two strains, this technique is tedious and error-prone and cannot be used to incrementally build up a comparative database. Recent technological progress in the area of genome sequencing calls for bioinformatics methods to replace the wet-lab DDH by **in-silico** genome-to-genome comparison.

- More information here:

Auch, A.F., von Jan, M., Klenk, H. et al. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand in Genomic Sci 2, 117–134 (2010).

Main principle of DNA-DNA hybridization



Workflow

=-=

NG-25130_4BAk_lib397215_6820_1_1_(paired)_assembly.1.fa (query) vs. subterraneus_ASM18354v3_genomic (reference):

=-=

Formula: 1 (HSP length / total length)
Distance: 0.1522
DDH estimate (GLM-based): 76.10% [72.1 - 79.6%]
Probability that DDH > 70% (i.e., same species): 86.18% (via logistic regression)
Probability that DDH > 79% (i.e., same subspecies): 49.19% (via logistic regression)

- -

Formula: 2 (identities / HSP length) (RECOMMENDED)
Distance: 0.0574
DDH estimate (GLM-based): 56.80% [54.1 - 59.6%]
Probability that DDH > 70% (i.e., same species): 41.42% (via logistic regression)
Probability that DDH > 79% (i.e., same subspecies): 9.37% (via logistic regression)

- -

Formula: 3 (identities / total length)
Distance: 0.2008
DDH estimate (GLM-based): 74.50% [71 - 77.7%]
Probability that DDH > 70% (i.e., same species): 87.45% (via logistic regression)
Probability that DDH > 79% (i.e., same subspecies): 36.97% (via logistic regression)

- -

Difference in % G+C: 0.31 (interpretation: either distinct or same species)

How do I interpret the results?

- The GGDC compares a query genome with a reference genome and calculates an intergenomic distance under three different distance formulae. The formulae support your decision about the relatedness of your novel strain to known (type) strains.

- Formula 1: length of all HSPs (high-scoring segment pairs) divided by total genome length
- Formula 2: sum of all identities found in HSPs divided by overall HSP length
- Formula 3: sum of all identities found in HSPs divided by total genome length
- Note

- Formula 2 is independent of genome length and is thus robust against the use of incomplete draft genomes.
- For other reasons for preferring formula 2, see this FAQ entry . If there are any significant differences between the three formulae, please base your decision on the recommended formula 2.

Can I use GGDC with incompletely sequenced genomes?

- Yes, if one uses formula 2 (which is the recommended formula anyway), one needs only about 20% of the genome to get the same result as with the full genome. The other two formulas will be severely affected by genome incompleteness.

- Note, however, that this does not mean that arbitrarily short genome fragments can successfully be analyses with the GGDC. For instance, sometimes users unwisely upload 16S rRNA gene sequences to the GGDC. This is not expected to work.

VICTOR
(Virus Classification and Tree Building Online Resource)

- This web service compares bacterial and archaeal viruses using their genome or proteome sequences. The results include phylogenomic trees inferred using the Genome-BLAST Distance Phylogeny method (GBDP), with branch support, as well as suggestions for the classification at the species, genus, subfamily and family level.

- More information here:

Jan P Meier-Kolthoff, Markus Göker, VICTOR: genome-based phylogeny and classification of prokaryotic viruses, Bioinformatics, Volume 33, Issue 21, 01 November 2017, Pages 3396–3404

Which distance formula should be preferred?

- The VICTOR study indicates that formula d6 should be preferred when amino-acid sequences of prokaryotic viruses are analysed — unless incomplete proteome sequences are contained in the data set. In that case d4 is the formula of choice.
- All distances are calculated from matches (local alignments) between two genome or two proteome sequences. in BLAST jargon, these matches are known as HSPs (high-scoring segment pairs). The meaning of the three distance formulas is as follows:

- d0: length of all HSPs divided by total genome length
- d4: sum of all identities found in HSPs divided by overall HSP length
- d6: sum of all identities found in HSPs divided by total genome length
- The branch lengths of the resulting VICTOR trees are scaled in terms of these distance formulas.

- The GGDC uses the same formulas but for historical reasons it applies a different terminology. GGDC formula 1 is VICTOR d0, GGDC formula 2 is VICTOR d4, and GGDC formula 3 is VICTOR d6.

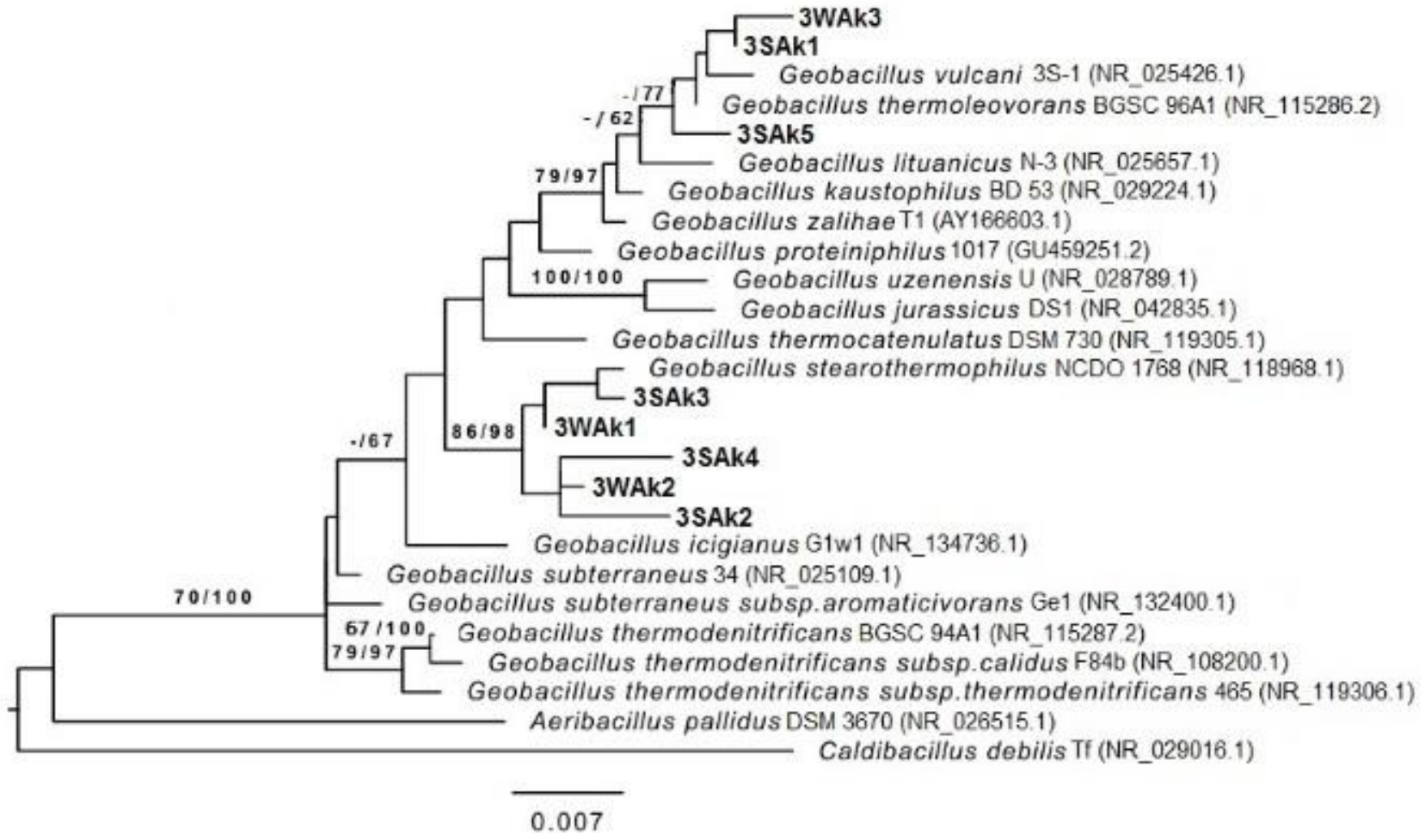
Single-Gene Trees

- For phylogeny reconstruction, this service combines state-of-the-art software for multiple sequence alignment, maximum likelihood (ML) and maximum parsimony (MP) analysis. Nucleotide data are optionally downloaded from GenBank and always automatically checked for reverse-complement sequences and duplicated labels.
- In the case of amino-acid data, the optimal model for ML is automatically determined. The pipeline is thus ideally suited for moderately sized single-gene data sets as used, e.g., in the description of new bacterial or other species.

What are the main advantages of using this phylogeny pipeline?

- Uploaded RNA sequences are automatically converted to DNA sequences.
- When long sequences such as genome sequences are encountered, an attempt is made to extract 16S rRNA gene sequences from each genome sequence. Thus the user needs not normally care about this step.
- Moreover, optionally pairwise nucleotide similarities are calculated.
- Finally, the results e-mails already include publication-ready text describing all methods used the pipeline, the results, and the according literature references.

An example


3WAk3
3SAk1
Geobacillus vulcani 3S-1 (NR_025426.1)
Geobacillus thermoleovorans BGSC 96A1 (NR_115286.2)
3SAk5
Geobacillus lituanicus N-3 (NR_025657.1)
Geobacillus kaustophilus BD 53 (NR_029224.1)
Geobacillus zalihae T1 (AY166603.1)
Geobacillus proteiniphilus 1017 (GU459251.2)
Geobacillus uzenensis U (NR_028789.1)
Geobacillus jurassicus DS1 (NR_042835.1)
Geobacillus thermocatenulatus DSM 730 (NR_119305.1)
Geobacillus stearothermophilus NCDO 1768 (NR_118968.1)
3SAk3
3WAk1
3SAk4
3WAk2
3SAk2
Geobacillus icigianus G1w1 (NR_134736.1)
Geobacillus subterraneus 34 (NR_025109.1)
Geobacillus subterraneus subsp.aromaticivorans Ge1 (NR_132400.1)
Geobacillus thermodenitrificans BGSC 94A1 (NR_115287.2)
Geobacillus thermodenitrificans subsp.calidus F84b (NR_108200.1)
Geobacillus thermodenitrificans subsp.thermodenitrificans 465 (NR_119306.1)
Aeribacillus pallidus DSM 3670 (NR_026515.1)
Caldibacillus debilis Tf (NR_029016.1)
-/77
-/62
79/97
100/100
86/98
-/67
70/100
67/100
79/97
0.007

Questionnaire

- How long is your data can be kept on the **DSMZ** server?
- What kind of format can you upload to the server?
- How do you explain DNA-DNA hybridization?
- Are there any differences between DNA-DNA hybridization and GGDC?
- What is a server?
- Which formula is recommended in GGTC?

References

- Meier-Kolthoff, J.P., Göker, M. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nat Commun 10, 2182 (2019). https://doi.org/10.1038/s41467-019-10210-3
- Auch, A.F., von Jan, M., Klenk, H. et al. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand in Genomic Sci 2, 117–134 (2010).
- Jan P Meier-Kolthoff, Markus Göker, VICTOR: genome-based phylogeny and classification of prokaryotic viruses, Bioinformatics, Volume 33, Issue 21, 01 November 2017, Pages 3396–3404
- https://tygs.dsmz.de/
- http://ggdc.dsmz.de/ggdc.php#
- http://ggdc.dsmz.de/phylogeny-service.php#