Medium and large scale taxonomic identification of prokaryotic sequences

Guy Perrière

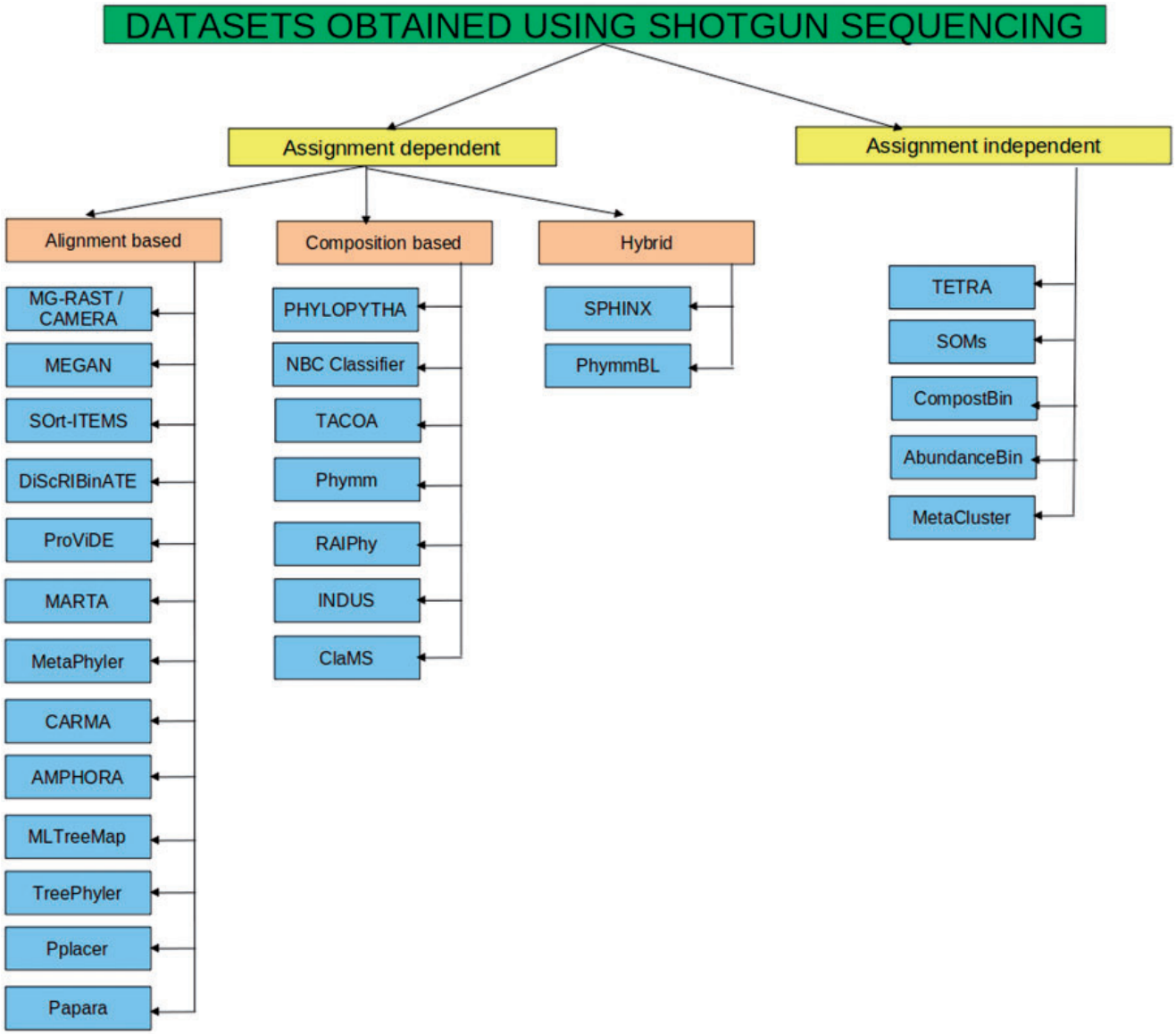
Pôle Rhône-Alpes de Bioinformatique
Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS 5558

September 25th 2014

Why taxonomic identification?

- Required in various domains such as:
  - Identification of pathogens (human, animals and plants).
  - Detection of contaminations (*e.g.*, food industry).
  - Ecology and environmental studies (*e.g.*, bioremediation, biodiversity estimation).
- The special case of public health:
  - For a long time use of phenotypic approaches (*e.g.*, the well-known Api® kits).
  - Use of proteomics-based identification such as mass spectrometry.
- Other fields:
  - Now mainly genomic sequences coupled with bioinformatics approaches.

Bioinformatics approaches

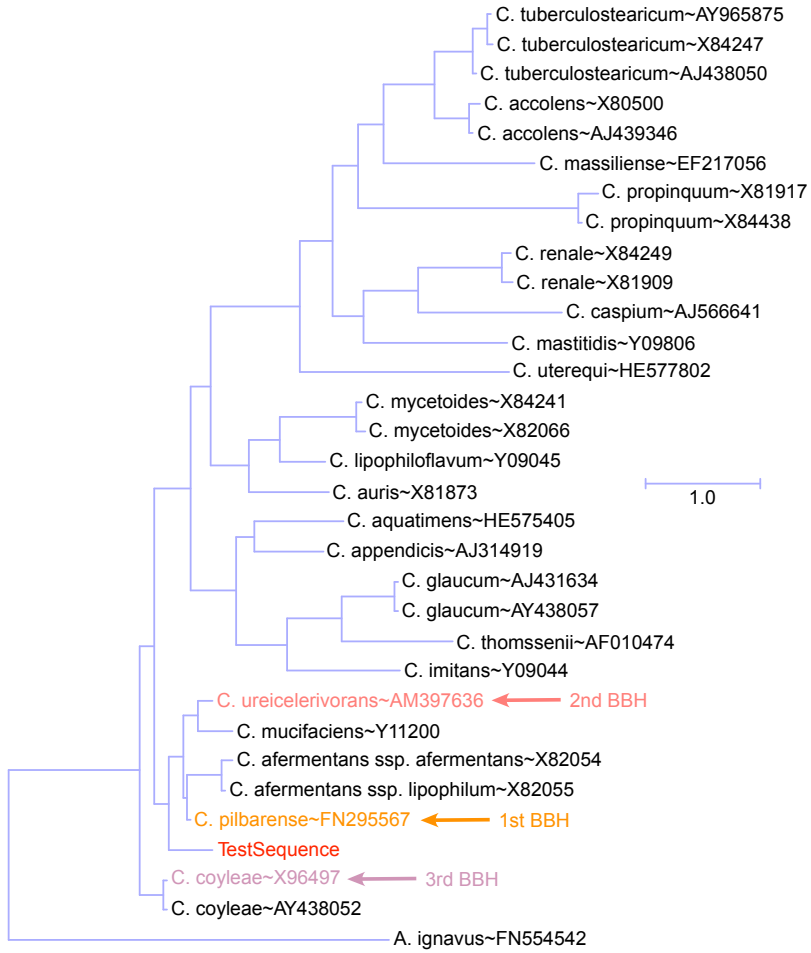


Alignment-based approaches

- Alignment step:
  - Similarity search, usually with `BLAST`.
  - Selection of a set of candidates on the basis of a similarity score (like E-value).
  - Identification using Best BLAST Hit (BBH) or Best Reciprocal BLAST Hit (BRH).
- Optional phylogenetic step:
  - Multiple alignment on a set of selected similar sequences.
  - Phylogenetic tree building.
  - Identification by phylogenetic proximity.

Why bothering with phylogeny?

- The BBH is frequently not the nearest neighbor on a phylogenetic sense:
  - Up to 40% of erroneous assignations.
  - Need to look at:
    - Topological distance.
    - Patristic distance.
  - Most of the microbial species are still not sequenced:
    - $>$ 99% of micro-organisms cannot be grown *in vitro*.

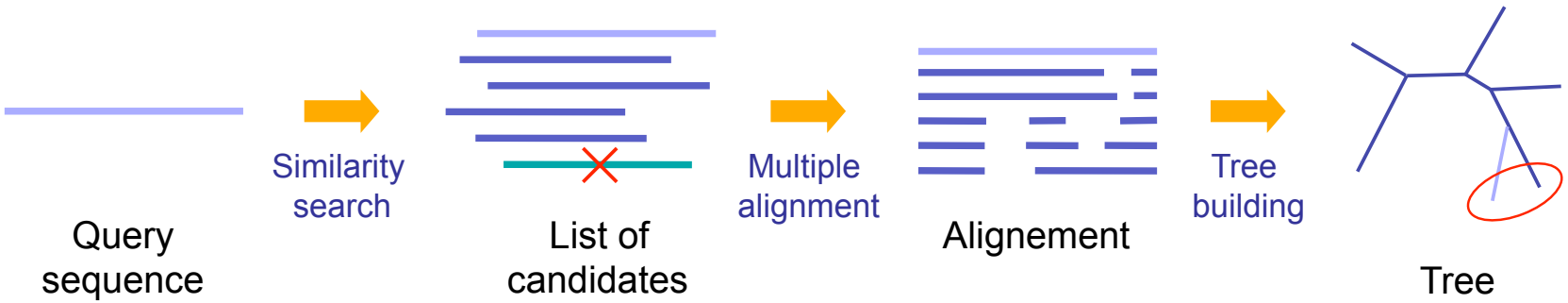


Markers commonly used

- The only truly universal marker for taxonomic identification is SSU rRNA (16S/18S):
  - 5 892 778 sequences in `GenBank` (18 September 2014).
  - Dedicated databases associated with some identification systems:
    - `Greengenes` with a set of basic tools such as `BLAST`.
    - `SILVA` with `ARB`.
    - `RDP` with a "naive Bayesian classifier".
- More specialized markers:
  - *groEL*, *gyrB*, *dnaJ/dnaK*, *recA*, *rpoB*, *sodA*, etc.

Simplified workflow

- Detection of homologs through a similarity search procedure.
- Multiple alignment computed with the query sequence and a selection of homologs.
- Phylogenetic tree computed with this alignment.
- Taxonomic assignation by phylogenetic proximity.



The precursor

JOURNAL OF CLINICAL MICROBIOLOGY, Apr. 2003, p. 1785–1787 Vol. 41, No. 4
0095-1137/03/$08.00+0 DOI: 10.1128/JCM.41.4.1785–1787.2003

BIBI, a Bioinformatics Bacterial Identification Tool

G. Devulder,[1]* G. Perrière,[2] F. Baty,[1] and J. P. Flandrois[1]

UMR CNRS 5558, Laboratoire de Bactériologie, Faculté de Médecine Lyon-Sud, 69921 Oullins Cedex,[1] and UMR CNRS 5558, Université Claude Bernard-Lyon 1, 69622 Villeurbanne Cedex,[2] France

Received 23 September 2002/Returned for modification 27 November 2002/Accepted 20 January 2003

BIBI was designed to automate DNA sequence analysis for bacterial identification in the clinical field. BIBI relies on the use of BLAST and CLUSTAL W programs applied to different subsets of sequences extracted from GenBank. These sequences are filtered and stored in a new database, which is adapted to bacterial identification.

Use since original publication:

- 150 000 identifications/year on average since 2003.
- 80 citations in indexed journals.

BIBI^{PQP}

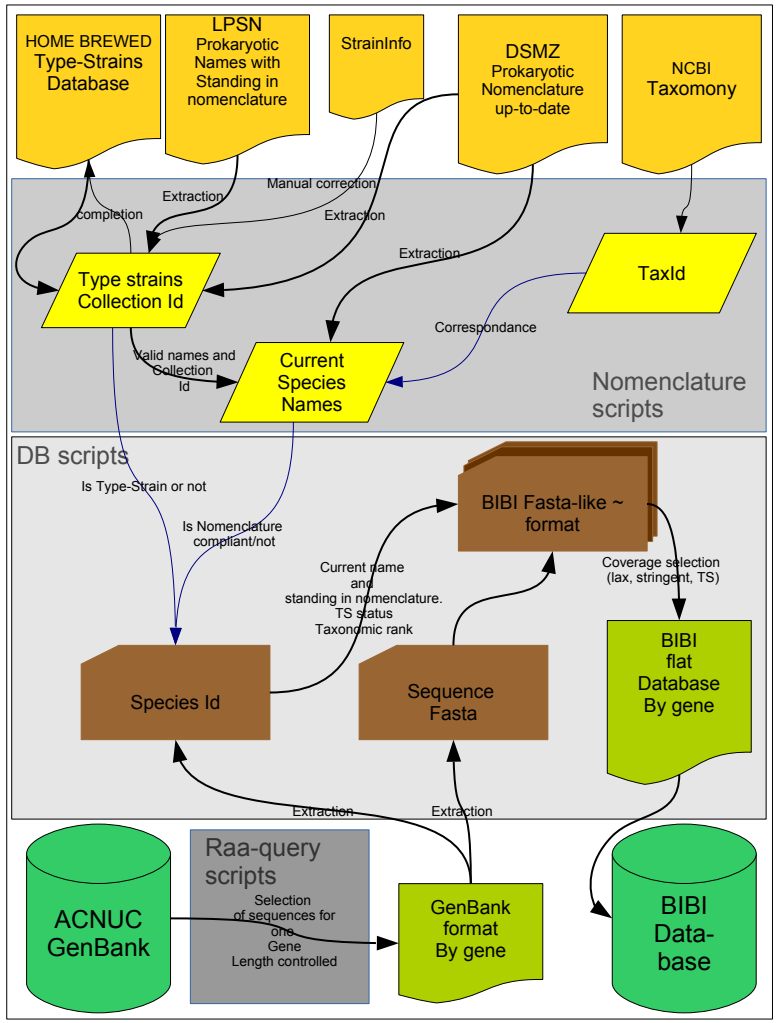
- Latest version of `BIBI` (Flandrois *et al.*, in prep.)
- Stands for Bioinformatics Prokaryotes Quick Phylogeny.
- Choice of a broad range of databases:
  - SSU rRNA.
  - Protein genes (*dnaJ/dnaK*, *fusA*, *glyA*, *groEL/hsp*60, *groEL2/hsp*65, *groES/cpn*10, *gyrB*, *recA*, *sodA*, *tuf*).
- Emphasis on speed without sacrificing the accuracy:
  - Can handle batches containing up to 3 000 sequences.

On-line access:

- `https://umr5558-bibiserv.univ-lyon1.fr/lebibi/lebibi.cgi`

Databases building

- All sequences taken from `GenBank`:
  - Database implementation with the `ACNUC` system.
- Poor quality of taxonomic annotations:
  - Complementation with LPSN and DSMZ.
- Automated scripts with a few manual expertises:
  - Regular updates (three releases/year).



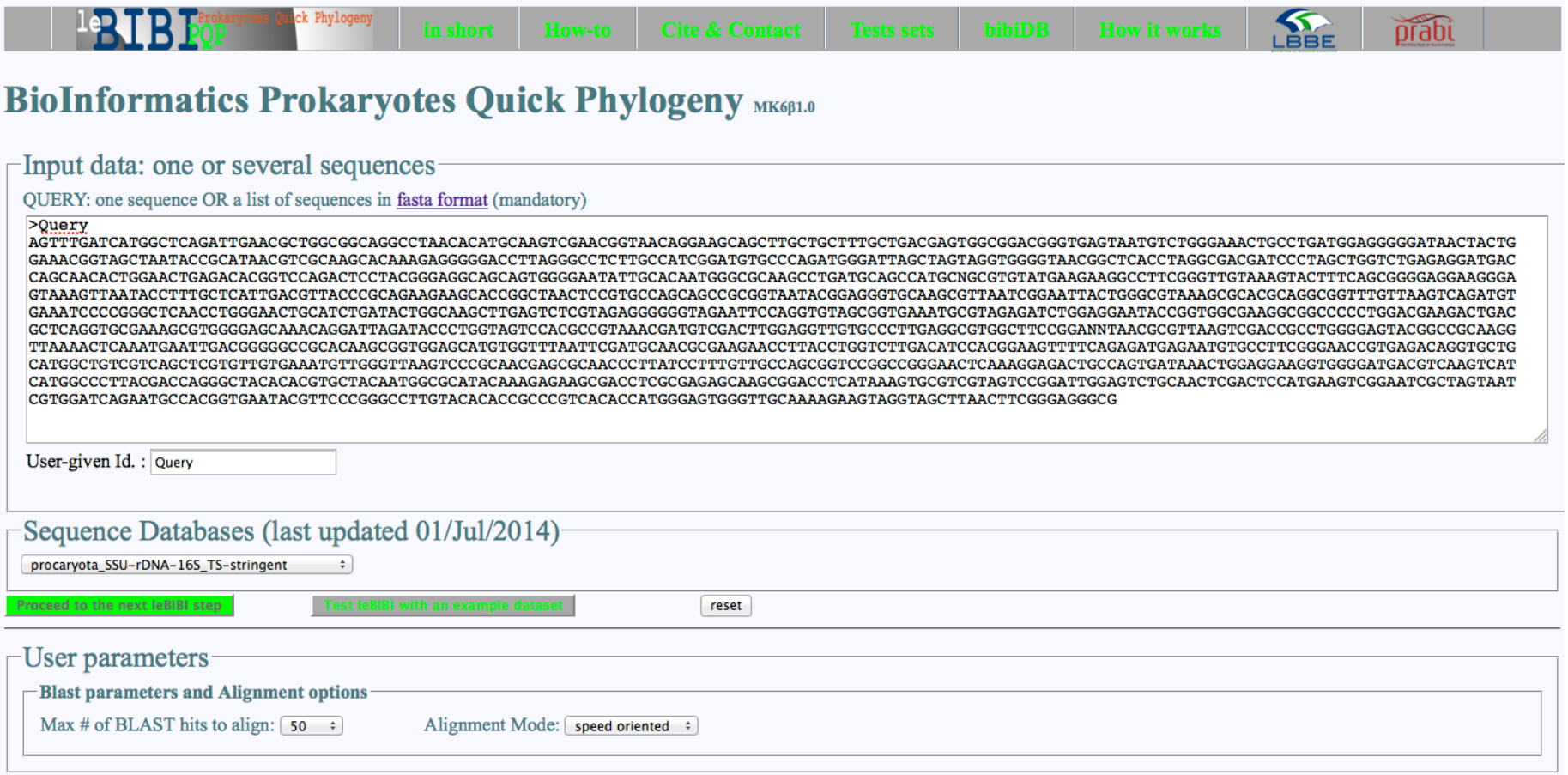
Available rRNA databases (1)

- `Lax`:
  - All SSU rRNA sequences from Bacteria and Archaea of length > 300 bp.
  - Large amount of unidentified or not fully identified sequences.
  - Lot of redundancies and erroneous species identifications.
- `Stringent`:
  - Subset of `Lax` retaining only sequences of validly denominated species and sequences corresponding to a type strain.
  - Lot of redundancies and erroneous species identifications.
- `TS stringent`:
  - Subset of `Stringent`, retaining only sequences from type strains.
  - Newly described species or non validly published species may be missing.
  - Less susceptible to contain sequences with erroneous species identification.

Available rRNA databases (2)

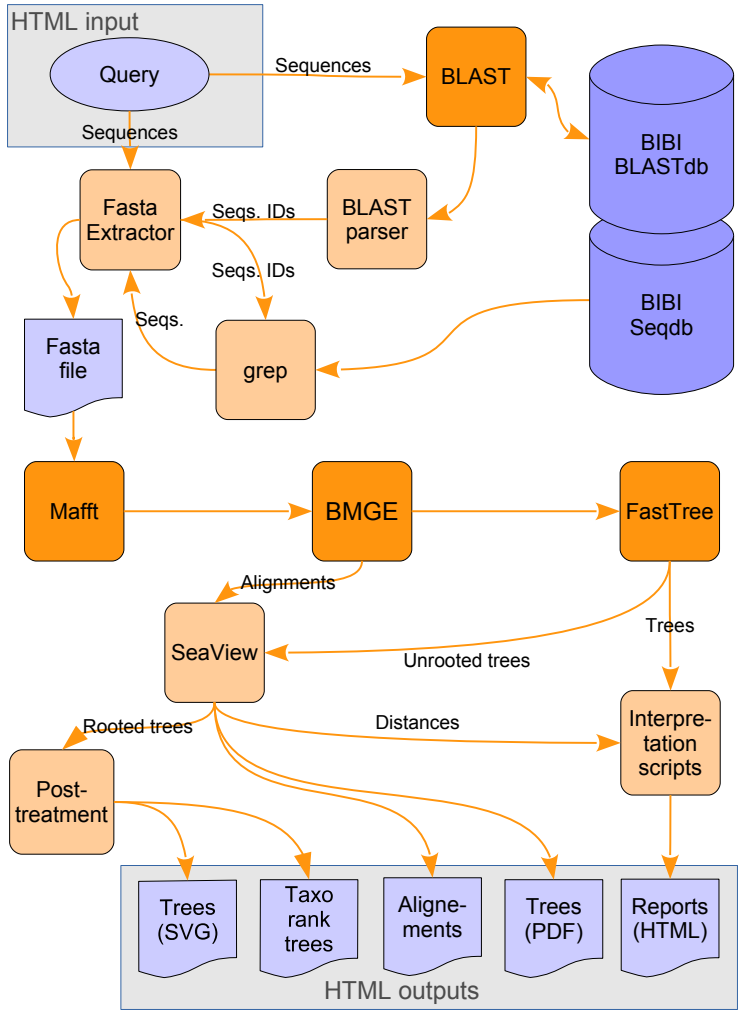
- `Superstringent`:
  - Subset of `TS stringent` containing only one type strain sequence per species.
  - Sequences are those cited in LPSN (List of Prokaryotic names with Standing in Nomenclature).
- `Genus level`:
  - Subset of `Superstringent` containing only one type strain (from the type species) sequence per genus.
- `Undetermined`:
  - Sequences of unidentified/uncultivable Bacteria and Archaea (typically environmental samples).
  - Warning: very special database, do not use without understanding its specificity.

Form for data input



Query workflow

- Input and output with basic HTML forms.
- Core programs:
  - Similarity searches with `BLAST`.
  - Multiple alignments with `Mafft`.
  - Trimming with `BMGE`.
  - Phylogenies with `FastTree`.
- `SeaView` for graphical outputs (SVG and PDF).

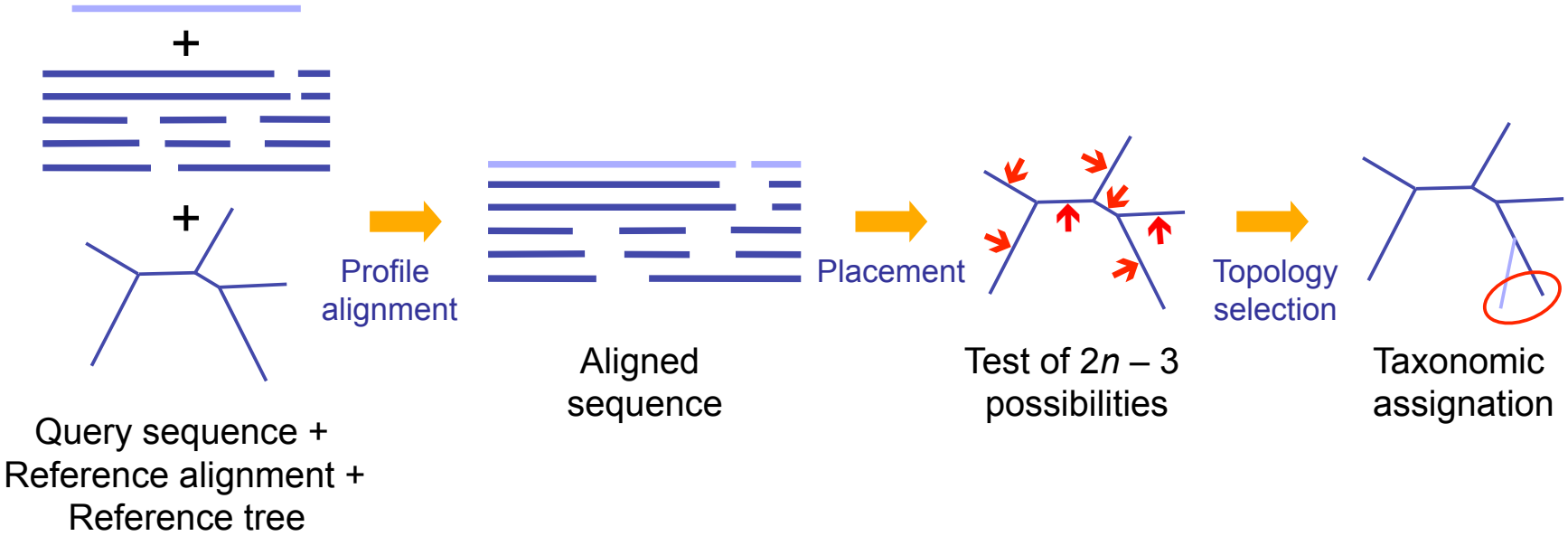


Advantages and limitations

- No need to install locally a complex set of databases and software.
- Much faster than all available on-line phylogeny services, including `phylogeny.fr`:
  - Full processing of a single SSU rRNA sequence takes < 7 seconds when using all default parameters.
- The broad range of available databases.
- Not suited for large-scale analyses involving millions of sequences.
- Performs poorly when using short reads that located in regions containing a low amount of phylogenetic signal.

Simplified workflow

- Profile alignment of the query sequence on a reference alignment.
- Placement of the query sequence on each of the $2n - 3$ possible positions on a reference tree containing n leaves.
- Selection of the tree showing the highest score.



Implementation at PRABI

- Initially developed as a "classical" pipeline.
- Soon to be implemented as a `Galaxy` workflow.
- Prerequisites:
  - Reference alignments built with software package `Infernal` (HMM profiles).
  - Corresponding phylogenetic trees built with `FastTree`.
- Programs flow:
  - Filtering of SSU rRNA sequences with `SortMeRNA`.
  - Aligment of the query sequence on a reference alignment with `HMMalign`.
  - Phylogenetic placement with `pplacer`.

pplacer

Matsen *et al. BMC Bioinformatics* 2010, **11**:538
http://www.biomedcentral.com/1471-2105/11/538

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree

Frederick A Matsen[1*], Robin B Kodner[2,3], E Virginia Armbrust[2]

Abstract

Background: Likelihood-based phylogenetic inference is generally considered to be the most reliable classification method for unknown sequences. However, traditional likelihood-based phylogenetic methods cannot be applied to large volumes of short reads from next-generation sequencing due to computational complexity issues and lack of phylogenetic signal. "Phylogenetic placement," where a reference tree is fixed and the unknown query sequences are placed onto the tree via a reference alignment, is a way to bring the inferential power offered by likelihood-based approaches to large data sets.

Results: This paper introduces `pplacer`, a software package for phylogenetic placement and subsequent visualization. The algorithm can place twenty thousand short reads on a reference tree of one thousand taxa per hour per processor, has essentially linear time and memory complexity in the number of reference taxa, and is easy to run in parallel. `Pplacer` features calculation of the posterior probability of a placement on an edge, which is a statistically rigorous way of quantifying uncertainty on an edge-by-edge basis. It also can inform the user of the positional uncertainty for query sequences by calculating expected distance between placement locations, which is crucial in the estimation of uncertainty with a well-sampled reference tree. The software provides visualizations using branch thickness and color to represent number of placements and their uncertainty. A simulation study using reads generated from 631 COG alignments shows a high level of accuracy for phylogenetic placement over a wide range of alignment diversity, and the power of edge uncertainty estimates to measure placement confidence.

Conclusions: `Pplacer` enables efficient phylogenetic placement and subsequent visualization, making likelihood-based phylogenetics methodology practical for large collections of reads; it is freely available as source code, binaries, and a web service.

Advantages and limitations

- Theoretically suited for large-scale analyses involving hundreds of millions of reads.
- Necessity to have reference alignments that can be very large:
  - Poor quality of multiple alignments containing thousands of sequences:
    - Poor quality of the inferred trees.
  - Use of a reduced set of taxa:
    - Accuracy of the assignation?
- Not accurate when using short sequences due to the lack of phylogenetic information:
  - Assignment to the Lowest Common Ancestor (LCA):
    - Many assignations as “Bacteria” or “Cellular organism”!
  - Clearly not suited for Illumina paired end sequencing.

Take home messages

- Reasoning in terms of similarity scores is often misleading (especially with metagenomic data):
  - Makes you think that the organism bearing the sequence with the highest score is indeed the closest one.
- Medium-scale taxonomic identification with phylogeny is now reasonably quick AND efficient:
  - High-quality databases adapted to the biological question asked are mandatory!
- Large-scale taxonomic identification with phylogenetic placement is still subject to important drawbacks:
  - A solution could be the use of improved mapping strategies, such as the one implemented in `EMIRGE`.

Acknowledgements

- Laboratoire de Biométrie et Biologie Évolutive:
  - Jean-Pierre Flandrois (PU-PH Université Lyon 1).
  - Manolo Gouy (DR CNRS).
- Pôle Rhône-Alpes de Bioinformatique:
  - François Bartolo (IE France Génomique).
  - Dominique Guyot (IE Université Lyon 1).
  - Clément Lionnet (M1 Université de Rouen)
  - Christine Oger (IR Université Lyon 1).