Computing Phylogenies with Phylonium

Fabian Klötzl

kloetzl@evolbio.mpg.de

2020-11-03

Biology textbooks usually contain a two part workflow to reconstruct the phylogenetic history of biological sequences: First, align the sequences and then use a maximum likelihood approach to find the best tree. This method worked well for many years but has become less useful as the number of sequenced genomes grew rapidly and made existing workflows computationally unfeasible. To mitigate this problem so-called *alignment-free* methods have been devised that quickly estimate phylogenies from unaligned sequences at reasonable accuracy. This document explains how to use one of these tools, phylonium, for phylogeny reconstruction.

1 Installation from Source

phylonium requires libdivsufsort for fast suffix array computation. Install it via your package manager, for instance apt.

```
% sudo apt install libdivsufsort
```

At the moment phylonium has to be build directly from the source code. This requires further dependencies.

```
% sudo apt install libdivsufsort-dev build-essential automake
```

Get a local copy of phylonium by cloning the repository:

```
% git clone https://github.com/evolbioinf/phylonium
```

Change into the newly created directory and prepare the configuration of the package.

```
% cd phylonium
% autoreconf -fi
```

Now the directory contains a new file called `configure`. This script will scan your system and configure the build procedure for phylonium accordingly.

```
% ./configure
```

> **Note:** phylonium uses x86 specific instructions to speed up the comparison of sequences. If you have a different CPU (e. g. arm), you should use `./configure --disable-x86simd`. On old x86 systems with old compilers you may have to disable the optimizations for AVX512: `./configure --disable-avx512`.

The configuration step fails if libdivsufsort or the GSL are not found. After successful configuration the program has to be compiled.

```
% make
```

This step compiles the source code into the executable `phylonium` located in the `src` subdirectory. To make it available to all users do:

```
% sudo make install
```

2 Basic Usage

After phylonium has been built, we can check whether it has been correctly installed. For that simply call phylonium without any arguments.

```
% phylonium
Usage: phylonium [OPTIONS] FILES...
  FILES... can be any sequence of FASTA files, each file
       representing one genome.

Options:
  -2, --2pass     Enable two-pass algorithm
  -b, --bootstrap=N Print additional bootstrap matrices
  --complete-deletion Delete the whole aligned column in case of
       gaps
  -r FILE         Set the reference genome
  -t, --threads=N The number of threads to be used; by default,
       all available processors are used
  -v, --verbose   Print additional information
  -h, --help      Display this help and exit
     --version    Output version information and acknowledgments
```

phylonium now prints the available command line arguments. The same can be achieved via `phylonium -h`.

In simplistic terms, phylonium reads in assembled DNA sequences in FASTA format and produces a matrix of their evolutionary distances. It prints an error message if the input is malformed. All contigs belonging to the same genome are in the same input file. phylonium uses the file name as identifier for the sequence. The final distance matrix is in PHYLIP format.

To see how phylonium works, we apply it to a set of 29 *Escherichia coli* genomes. First, download and unpack them.

```
% wget https://github.com/EvolBioInf/life2015/raw/gh-pages/eco29.
    fasta.gz
% gunzip eco29.fasta.gz
% cat eco29.fasta | awk -v RS='>' 'NR>1{print ">" $0 > $1 ".fa"
    }'
```

Now your working directory contains 29 individual genomic files with the extension `.fa`. We begin by comparing only two of them.

```
% phylonium FM180568.fa BA000007.fa
2
BA000007 0.0000e+00 2.4833e-02
FM180568 2.4833e-02 0.0000e+00
```

On the first line of the output is the number of genomes compared followed by a two-by-two matrix. Each value represents an estimated substitution rate between two sequences. As each genome is equal to itself the main diagonal consists only of zeros. So the evolutionary distance between sequences `BA000007` (first row) and `FM180568` (second column) is $2.4833 \cdot 10^{-2}$ substitutions per site. Note that the matrix is symmetric. Next we execute phylonium on the whole dataset.

```
% phylonium *.fa > eco29.out
```

Now the file `eco29.out` contains a distance matrix of size 29 by 29, which can be summarized as a phylogeny using the neighbor-joining algorithm implemented in, for example, my mat tools[1].

```
% mat nj eco29.out > eco29.tree
% cat eco29.tree
((((((((AE005174:-0.000019,BA000007:1.8692e-04)100:0.000738,
    CP001846:7.8206e-04)100:0.005264,CP000034:7.1481e-03)
    100:0.002179,((((((AE014075:0.003806,CU928162:4.0810e-03)
    30:0.000106,((CP000243:0.000287,CP000468:2.8884e-04)
    92:0.000463,CU928161:6.3318e-04)100:3.3526e-03)84:0.000290,
    CP000247:4.3564e-03)100:0.001367,FM180568:5.7900e-03)
    100:0.006160,(CU928164:0.006474,CP000970:6.6612e-03)100:2.9209
    e-03)100:0.001376,CU928163:9.6094e-03)100:3.7308e-03)
    100:0.001825,((((AP009048:0.000021,U00096:-1.2523e-05)
    100:0.000054,CP001396:5.3529e-05)42:0.000005,CP000948:-1.0393e
    -05)100:0.004144,(CP000802:0.002662,CP000946:3.0385e-03)
    100:8.4870e-04)100:1.8957e-03)60:0.000689,((AE005674:0.000240,
    AE014073:7.9602e-05)100:0.000401,CP000266:5.1495e-04)
```

[1] https://github.com/kloetzl/mattools

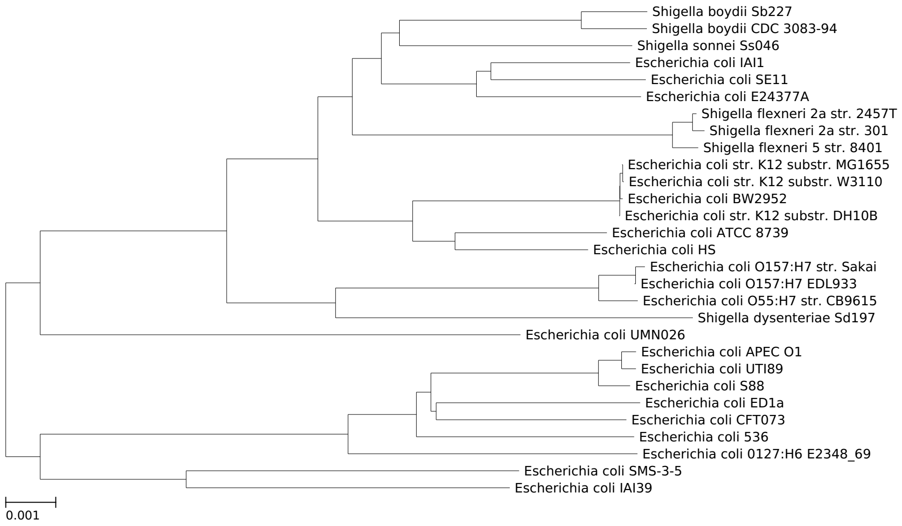
Shigella boydii Sb227
Shigella boydii CDC 3083-94
Shigella sonnei Ss046
Escherichia coli IAI1
Escherichia coli SE11
Escherichia coli E24377A
Shigella flexneri 2a str. 2457T
Shigella flexneri 2a str. 301
Shigella flexneri 5 str. 8401
Escherichia coli str. K12 substr. MG1655
Escherichia coli str. K12 substr. W3110
Escherichia coli BW2952
Escherichia coli str. K12 substr. DH10B
Escherichia coli ATCC 8739
Escherichia coli HS
Escherichia coli O157:H7 str. Sakai
Escherichia coli O157:H7 EDL933
Escherichia coli O55:H7 str. CB9615
Shigella dysenteriae Sd197
Escherichia coli UMN026
Escherichia coli APEC O1
Escherichia coli UTI89
Escherichia coli S88
Escherichia coli ED1a
Escherichia coli CFT073
Escherichia coli 536
Escherichia coli 0127:H6 E2348_69
Escherichia coli SMS-3-5
Escherichia coli IAI39
0.001

Figure 1: A phylogeny of 29 *Escherichia coli* and *Shigella* genomes; This phylogeny was computed using phylonium default arguments, neighbor joining and then visualized with ETE.

```
100:6.4055e-03)78:0.000581,((AP009240:0.003100,CU928160:2.7861
e-03)100:0.000288,CP000800:3.2921e-03)100:1.9029e-03)
89:0.000362,CP000038:4.6574e-03,(CP001063:0.001312,CP000036
:1.3202e-03)100:3.6383e-03);
```

This phylogeny in Newick format, which I am here visualizing using the ETE Toolkit[2].

3 Picking a Reference

Internally, phylonium *aligns* all sequences to a reference. This simplified alignment is then used to estimate the rate of substitutions on homologous regions. The reference can impact the accuracy of the phylogeny. With the `--verbose` switch phylonium prints the reference it has picked automatically.

```
% phylonium *.fa --verbose
chosen reference: AP009240
ref: AP009240
Mapping 29 sequences: 100.0% (29/29), done.
Comparing the sequences: 100.0% (406/406), done.
```

[2] http://etetoolkit.org/

```
...
avg coverage: 0.738057
alignment: 110170145 137065486 0.803777
```

It may be beneficial to explicitly set the reference, for instance if the data set contains an assembly of the type strain. This can is done with the `-r` argument.

```
% phylonium *.fa --verbose -r FM180568.fa
ref: FM180568
...
```

For as yet unexplored data sets it may be hard to choose a good reference. For these cases phylonium comes with the `--2pass` option. If enabled, after a first run, phylonium picks a central sequence as a reference, which is a reasonable choice for most data sets.

```
% phylonium *.fa --verbose --2pass
chosen reference: AP009240
ref: AP009240
ref: CP000948
...
```

4 Dealing with Gaps

phylonium estimated the evolutionary distance as the number of substitutions per site. However, substitutions are only one type of mutations. Indels frequently disturb the homology of sequences. There exist two standard methods to handle the corresponding gaps: pairwise and complete deletion. Consider the following alignment.

```
Seq1  AACTT
Seq2  AGGTT
Seq3  AG-TC
```

Under pairwise deletion (the default in phylonium) `Seq1` is separated by two mismatches both from `Seq2` and `Seq3`; the difference being that there are five homologous nucleotides between `Seq1` and `Seq2`, whereas `Seq1` and `Seq3` share only four nucleotides. So the pairwise deletion model only considers gaps in the two sequences under consideration, hence the name. With complete deletion, this changes. Whenever there is a gap in the data, the whole column is masked. Thus, the above dataset becomes transformed into the following.

```
Seq1  AA-TT
Seq2  AG-TT
Seq3  AG-TC
```

The distance between `Seq1` and `Seq3` remains unchanged; But the distance from `Seq` to `Seq2` changes from $\frac{2}{5}$ to $\frac{1}{4}$.

Activating the `--complete-deletion` option can be advantageous when analyzing data sets with divergent sequences. However, complete deletion biases the estimation towards conserved genetic regions. The resulting distances are often too small.

On the other hand, complete deletion can result in the loss of a large part of the input data, for example, if a single assembly, out of potentially hundreds, failed. So use complete deletion with caution.