The Effects of Evolutionary Distance on the Phylogenetic Placement of Metagenomic Data

Elisa Loza
Computational and Systems Biology
Rothamsted Research
Harpenden, AL5 2JQ, UK

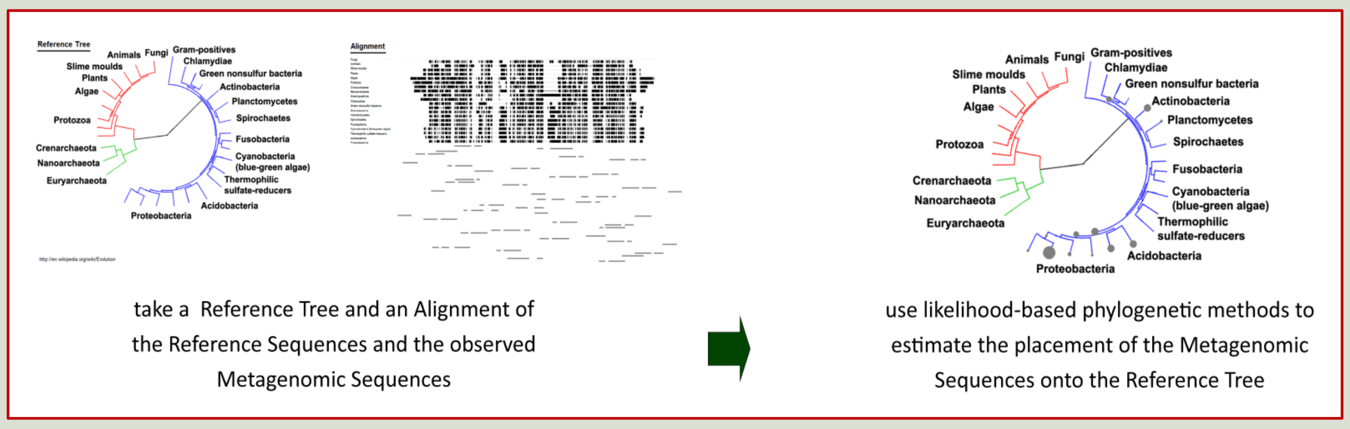
Nick Goldman
EMBL-European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton Cambridge, CB10 1SD, UK

The Problem

Metagenomic data are DNA sequences that are directly sampled from a habitat (e.g. agricultural soil, ocean water, human gut) and, typically, sequenced using high-throughput technologies.

The observed metagenomic sequences (or reads) come without information on organismal origin.

One approach to identify the organismal origin of metagenomic reads is to use **phylogenetic placement** methods (e.g. [1]), as follows:



take a Reference Tree and an Alignment of the Reference Sequences and the observed Metagenomic Sequences

use likelihood-based phylogenetic methods to estimate the placement of the Metagenomic Sequences onto the Reference Tree

We aim to answer: '*what are the effects of increasingly distant metagenomic data on the quality of phylogenetic placement?*'

Method

1. A reference tree, *T*, with two main clades – one fully balanced and one fully unbalanced – was designed. The edge length from the most recent common ancestor (MRCA) of a main clade to a tip within that clade was set to 1 (Fig.1).
2. A set of reference sequences was generated on *T* using evolutionary-model estimates from the 'human vaginal microbiome' (16S rRNA gene) distributed as part of the `pplacer` software package [1, 2].
3. A total of 1,450 metagenomic reads, with an average length of 500 bp, were generated using `MetaSim` [3]. The reads were sampled from hypothetical taxa diverging an **evolutionary distance** (or edge length), λ, from the midpoint (i.e. location $\chi = 0.5$) of the edges in *T*.
4. The metagenomic reads were placed onto *T* using `pplacer`. *Phylogenetic placement quality* was measured by the number of reads that were placed onto the correct edge.
5. The generation of the reference sequences and metagenomic reads, together with the placement exercise, was replicated five times.
6. Placement quality was further studied for $\chi = 0, 0.25, 0.75$, where $\chi = 0$ denotes the most exterior point of an edge and $\chi = 0.75$ denotes ¾ of edge length, from the most exterior point of the edge.
7. Placement quality results were statistically analysed using ANOVA methods [e.g. 4].

Results

Figure 1 shows the reference tree and the pattern of placement quality against evolutionary distance, λ, for eleven illustrative edges. In this figure $\chi = 0.5$.

There is a significant decrease in placement quality as λ increases ($F_{10,125}$=25.9; *P*<0.001), which is stronger for short than long edges ($F_{85,125}$=2.8; *P*<0.001).

Placement quality is significantly better on exterior than on interior edges ($F_{1,27}$=7.4; *P*=0.011).

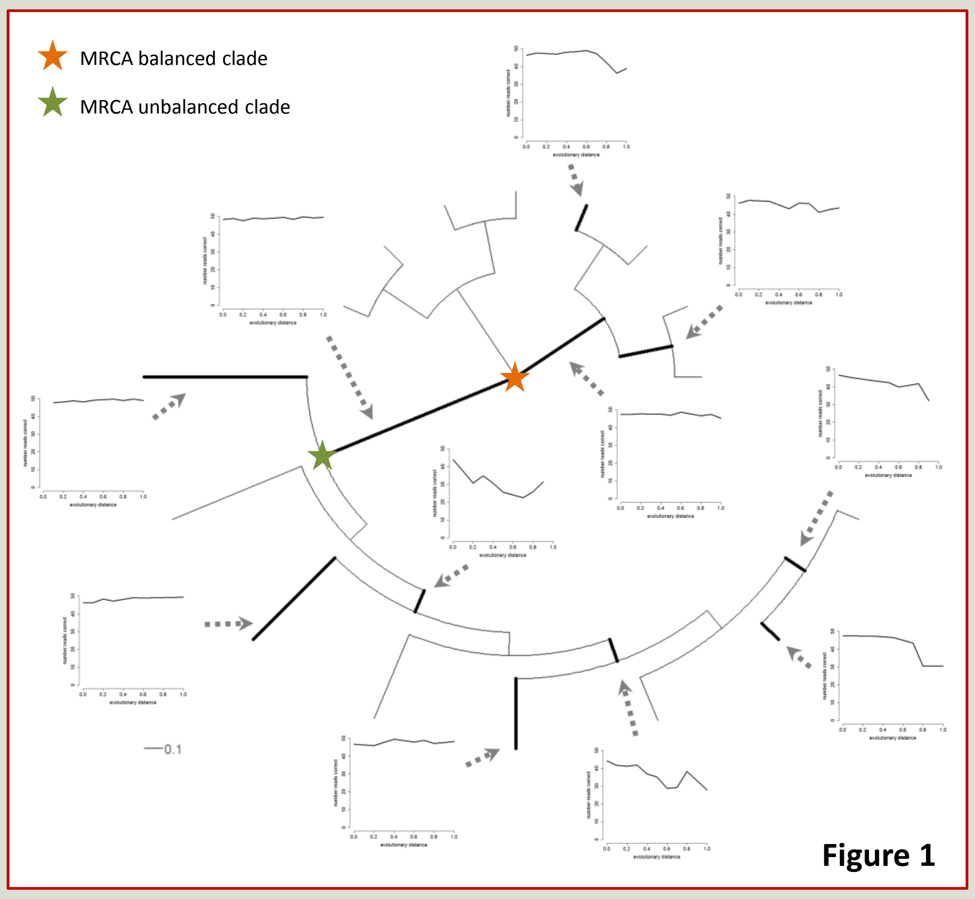
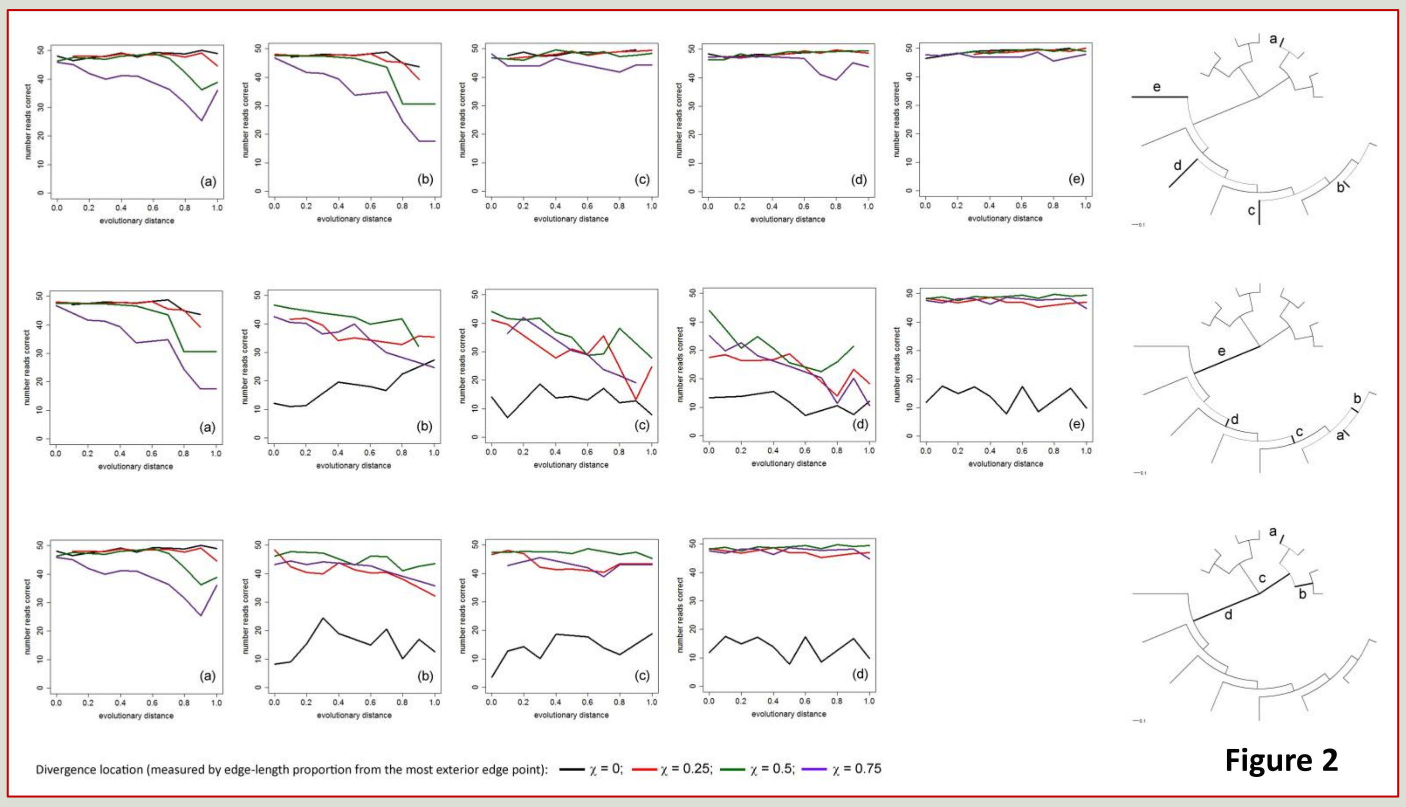


Figure 1

Figure 2 shows the pattern of placement quality against evolutionary distance, at divergence points along an edge $\chi = 0, 0.25, 0.5, 0.75$.

Overall placement quality does not differ significantly between $\chi = 0, 0.75$ nor between $\chi = 0.25, 0.5$, only across pairs of χ's ($F_{3,881}$=87.6; *P*<0.001).



Divergence location (measured by edge-length proportion from the most exterior edge point): — χ = 0; — χ = 0.25; — χ = 0.5; — χ = 0.75

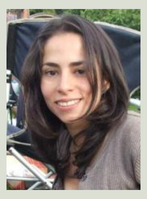
Figure 2

Conclusions and future work

- Use reference trees with as many taxa as possible that are relevant to the habitat of study.
- Our results coincide with [1, 5] but go beyond the aspects investigated therein.
- Currently conducting longer simulations and investigating 'distance' quality scores.
- Our ultimate goal is to produce guidelines for the design of reference trees in Metagenomic studies.

For further information

Please contact me at
elisa.loza@rothamsted.ac.uk



With thanks to the BBSRC and
the Software Sustainability Institute.

References

[1] Matsen *et al. BMC Bioinformatics* 2010, 11(1):538.
[2] http://microbiome.fhcrc.org/apps/refpkg [Accessed 11 June 2012].
[3] Richter *et al. PLoS ONE* 2008, 3(10): e3373.
[4] Montgomery, *Design and Analysis of Experiments*, 2nd ed., John Wiley & Sons, 1984.
[5] Stark *et al. BMC Genomics* 2010, 11:461.