

A web server for the ancestral genome reconstruction

Abdoulate Baniré Diallo¹, Vladimir Makarenkov², Mathieu Blanchette³

1 Introduction

The recent interest for ancestral genome reconstruction provides clues on several aspect of evolutionary changes [1]. Ancestral genome reconstruction attempts to predict the DNA sequences of all ancestral species in a given phylogeny according to a multiple sequence alignment. Accurate ancestral genome reconstruction can contribute to the study of adaptation, behavioral changes and functional divergence [1, 2]. Two of the most important steps in ancestral genome reconstruction procedure are the prediction of substitution and insertion and deletion (indel) that may have produced a given set of aligned regions [1]. Although the inference of indel evolutionary scenarios is useful in several problems, it has received relatively little attention. We have recently proposed a statistical framework that enables one to infer the most likely indel scenario and to estimate uncertainties of predictions based on fixed indel rate parameters and a given multiple sequence alignment. The developed framework is adequate for small-scale genomic regions with insertions, deletions and substitutions [2]. Substitutions are predicted using an adaptation of the Felsenstein algorithm [3, 4]. The maximum likelihood indel scenario is predicted by an exact algorithm and a fast heuristic described in [2]. These methods also permit the computation of the uncertainty associated to ancestral sequence reconstruction. To facilitate the computation of ancestral sequence reconstruction, we made available the *Ancestors 1.0* web server. This server performs the computation of multiple sequence alignments using several widely used algorithms, the inference of exact or heuristic based indel scenarios and the prediction of substitution scenarios. All the steps have been combined in a single web interface. The results are presented as colored output indicating the level of confidence of each prediction. *Ancestors 1.0* is available at the following URL address: <<http://ancestors.bioinfo.uqam.ca/ancestorWeb/>>.

2 User inputs and *Ancestors 1.0* outputs

The *Ancestors 1.0* web form is divided in three parts (see Figure 1). The first part allows users to supply a set of orthologous sequences in Fasta format. The sequences could either be aligned or not. Users can choose a multiple sequence alignment method between the following: Clustal-W, Dialign, Mavid, Mafft and T-Coffee. Those methods have been chosen for one or more of the following reason(s): they are widely used in comparative genomics, they have been shown to be accurate on genomic data, and/or they can handle a reasonable large datasets. The second part allows users to supply a rooted phylogenetic tree related to the multiple sequence alignment data that can be used to guide the ancestral sequence reconstruction. The tree format is the widely used Newick format. The third part gives the choice of indel parameters. Users can either ask to report the most likely indel scenario or the posterior decoding for predicting the presence or absence of characters at each position of the ancestral sequences. The posterior decoding is used to compute the confidence levels of the predictions as described in [2]. Future versions of the program will allow adding sequence using genbank accession number, the reconstruction of the phylogenetic tree by supplying several phylogenetic tree reconstruction methods. Moreover, it will propose practical tools for analyzing the obtained results.

The *Ancestors 1.0* results are made available using different plain text and HTML format files. The summary of those files is given as default output (see Figure 2). The results present each input file, followed by the result of the alignment method. The results of the indel predictions are presented in three files (the indel scenario, the tree-HMM state created, the posterior probability for each position of each sequence). The results of the ancestral nucleotides predicted contain the characters in plain text and HTML file. The HTML file presents a colored output according to the level of prediction confidence, as shown in Figure 2. Those confidence levels are given in separate file together with the obtained posterior probabilities. For specific help request, users can use the bug report form. Finally, it is worth noting that this web server can easily compute ancestral sequences for more than 30 species. For instance the benchmark CFTR regions (1Mb) of 28 mammals have been reconstructed with less than three minutes using the implemented heuristics for indel reconstruction.

¹Departement of computer science, Université du Québec à Montréal, Qc, Canada. E-mail: diallo.abdoulaye@uqam.ca

²Departement of computer science, Université du Québec à Montréal, Qc, Canada. E-mail: makarenkov.vladimir@uqam.ca

³McGill Centre for bioinformatics, McGill University, Qc, Canada. E-mail: blanchem@mcb.mcgill.ca

The Ancestors program
web This is the online version of the ancestor program. Source code is available for [download](#).

This program reconstruct the ancestral sequences in two steps. The inference of the maximum likelihood indel scenario and the inference of ancestral characters.

Enter your sequence in the FASTA format (Sample)

```
>D29
GGGCGCCAGTGGAGCGGTGAG---GTGACATGAAAGCCATCACCAAGGGCCACATGACCTC
>L5
GGGCGCCAGTGGAGCGGTGAG---GTGACATGAAAGCCATCACCAAGGGCCACATGACCTC
>Bx2
CTGGCTCTGATGGAGCGG---GCGAGCCGAGCCAGCCATCACCAAGGGCCACATGACCTC
>Bb1
CTGGCTCTGATGGAGCGG---GCGAGCCGAGCCAGCCATCACCAAGGGCCACATGACCTC
>T4
CGGTGAGTCTGGAGCT-GACC---CGAAGCCGCTGAGCCAGCCAGGGTGGCGCCAA
>A118
CGGTGAGTCTGGAGCT-GACC---CGAAGCCGCTGAGCCAGCCAGGGTGGCGCCAA
```

Pasted Sequence File

Alignment Method:

Enter the phylogenetic tree in the NEWICK format (Tree sample)

```
((A118:0.23397,ph111:0.23474):0.0954,((Bb1:0.12134,(D29:0.08015,
L5:0.07621):0.03223):0.04305,Bx2:0.13289):0.08952,tm4:0.27211):0.05
```

Pasted Tree File

The best exact scenario
 The best heuristic scenario
 Absolute Proportional

The exact posterior decoding
 The heuristic posterior decoding
 Absolute Proportional

Tree Hidden Markov Model probabilities:
 0.01 Insertion Start
 0.1 Insertion Extension
 0.01 Deletion Start
 0.1 Deletion Extension

DNA substitution probabilities: HKY

Tools:
[Ancestors](#)
[Consensus](#)
[Missing data](#)
[Rooting or unrooting trees](#)
[T-Rex Online](#)

Useful links:
[NCBI](#)
[UCSC genome browser](#)
[Ensembl genome browser](#)
[Phylogeny programs](#)
[Sanger bioinformatics tools](#)
[McGill University](#)
[Université du Québec à Montréal](#)
[Tourism in Montreal](#)

My source of information
[Radio Canada](#)
[BBC](#)
[Le monde](#)
[Guinee news](#)

Figure 1: The *Ancestors 1.0* user input form. The interface contains the section for alignment, the phylogenetic tree and the indel parameters. On the right of the display, links are given to commonly used tools. The alignment and the tree presented here is a set of phage genes.

The Ancestors program
web This is the online version of the ancestor program. Source code is available for [download](#).

Input Files
[Input Sequence File](#)
[Input Phylogenetic Tree](#)

Alignment files
 No Alignment method has been chosen! The given sequences are already aligned
[Alignment File](#)

Indels Results
[Indels by characters](#)
[Indels state scenarios](#)
[States created by the tree-HMM](#)

Substitutions Results
[Ancestral characters](#)
[Confidences](#)
[Posterior probabilities](#)
[Colored output](#)
[back](#)
[Report bugs](#)

```
> A118+PH11+
CGGTGAGTCTGAGCT-GACC---CGAAGCCGCTGAGCCAGCCAGGGTGGCGCCAA
> D29+L5+
GGGCGCCAGTGGAGCGGTGAG---GTGACATGAAAGCCATCACCAAGGGCCACATGACCTC
> Bx2+D29+L5+
CTGGCTCTGATGGAGCGG---GCGAGCCGAGCCAGCCATCACCAAGGGCCACATGACCTC
> Bb1+D29+L5+BX22+TM4+
CTGGCTCTGATGGAGCGG---GCGAGCCGAGCCAGCCATCACCAAGGGCCACATGACCTC
> A118+PH11+BX21+D29+L5+BX22+TM4+
CGGTGAGTCTGAGCT-GACC---CGAAGCCGCTGAGCCAGCCAGGGTGGCGCCAA
```

Confidence levels: 0  100

Figure 2: The *Ancestors 1.0* main output gives links to all obtained result files as well as input files. On the right, the ancestral sequence predictions and the confidence percentage for each nucleotide is presented.

References

- [1] Blanchette, M., Green, E.D., Miller, W. and David Haussler. 2004. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Research*, 14(12). pp. 2412-2423.
- [2] Diallo, A.B., Makarenkov, V. and Blanchette M. 2007. Exact and heuristics methods to indel maximum likelihood problem. *Journal of Computational Biology*, 14(4). pp. 446-461.
- [3] Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17. pp. 368-376.
- [4] Felsenstein, J. and Churchill, G. 1996. A hidden markov model approach to variation among sites in rate of evolution. *Molecular Biology Evolution*, 13. pp. 93-104.