Phylogenetics

Ancestors 1.0: a web server for ancestral sequence reconstruction

Abdoulaye Banire Diallo1,2,* , Vladimir Makarenkov1 and Mathieu Blanchette2
1Department of Computer Science, Université du Québec a Montréal, PO. Box 8888 Downtown Station, Montreal Qc, H3C 3P8 and 2School of Computer Science, McGill University, 3630 University street #3107, Montreal, Qc, H3A 2B2, Canada

ABSTRACT
Summary: The computational inference of ancestral genomes consists of five difficult steps: identifying syntenic regions, inferring ancestral arrangement of syntenic regions, aligning multiple sequences, reconstructing the insertion and deletion history and finally inferring substitutions. Each of these steps have received lot of attention in the past years. However, there currently exists no framework that integrates all of the different steps in an easy workflow. Here, we introduce Ancestors 1.0, a web server allowing one to easily and quickly perform the last three steps of the ancestral genome reconstruction procedure. It implements several alignment algorithms, an indel maximum likelihood solver and a context-dependent maximum likelihood substitution inference algorithm. The results presented by the server include the posterior probabilities for the last two steps of the ancestral genome reconstruction and the expected error rate of each ancestral base prediction.

Availability: The Ancestors 1.0 is available at http://ancestors.bioinfo.uqam.ca/ancestorWeb/.

Contact: diallo.abdoulaye@uqam.ca

1 INTRODUCTION
Ancestral genome reconstruction attempts to predict the DNA sequences of all ancestral species in a given phylogeny according to a multiple sequence alignment of sequences at the leaves. Three of the most important steps in the ancestral genome reconstruction procedure are the computation of a multiple sequence alignment and the prediction of insertions and deletions (indels) and substitutions that may have produced a given set of aligned regions (Blanchette et al., 2004b). Previous methods for ancestral sequence reconstruction are limited to the substitution reconstruction only (Yang et al., 1995); treat gaps the same as nucleotide or amino acids (Koshi and Goldstein, 1996); or treat only indels of length 1 (Edwards and Shields, 2004). Although whole-genome multiple sequence alignment algorithms are now well developed (Blanchette et al., 2004a) and efficient algorithms exist for maximum likelihood and parsimonious inference of substitutions (Felsenstein, 1981; Felsenstein and Churchill, 1996; Liberles, 2001), indel inference has only recently begun to receive attention (Bradley and Holmes, 2007; Chindelevitch et al., 2006; Diallo et al., 2007; Fredslund et al., 2004; Kim and Sinha, 2000). We have recently proposed a statistical framework that enables one to infer the most likely indel scenario, based on fixed indel rate parameters and a given multiple sequence alignment. The developed framework is adequate from small- to medium-scale genomic regions with insertions, deletions and substitutions (Diallo et al., 2007). The maximum likelihood indel scenario is predicted by an exact algorithm and a fast heuristic described in Diallo et al. (2007). Substitutions are predicted using an adaptation of the Felsenstein algorithm (Felsenstein, 1981; Felsenstein and Churchill, 1996). These methods also permit the computation of the uncertainty associated to various aspects of ancestral sequence reconstruction. While the individual algorithmic components of ancestral sequence reconstruction are now available (Blanchette et al., 2008), there exists no easy-to-use system to perform all the steps required with a single click of the mouse. To facilitate the computation of ancestral sequence reconstruction, we introduce the Ancestors 1.0 web server. This server performs the computation of multiple sequence alignments using several widely used algorithms and the inference of maximum likelihood indel and 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.

2 USER INPUTS AND ANCESTORS 1.0 OUTPUTS
The Ancestors 1.0 interface can be divided in two parts. The first concerns the user inputs to the program, whereas the second is dedicated to presenting the obtained results.

2.1 Ancestors 1.0 inputs
The Ancestors 1.0 web form is divided in three parts (Fig. 1). The first part allows users to supply a set of sequences in Fasta format. The sequences could either be aligned or not. Users can choose a method of sequence alignment even though the sequences are already aligned—realigning sequences could help increasing the alignment accuracy. Implemented multiple sequence alignment procedures include popular tools such as Muscle (Edgar, 2004), and TBA (Blanchette et al., 2004a).

The second part allows users to supply a rooted phylogenetic tree (Newick format) that can be used to guide the ancestral reconstruction. Future versions of the program will allow the inference of the phylogenetic tree by various tree reconstruction methods.

*To whom correspondence should be addressed.
methods. The third part concerns the choice of the indel inference approach and the parameters related to the tree-hidden Markov model (HMM) algorithm. Users can either ask to report the most likely indel scenario or the posterior decoding for predicting the presence or absence of a character at each position of the ancestral sequences. In the latter case, the probability of presence of a nucleotide (as opposed to a gap) at any given position of any given ancestral node is reported. Two indel inference methods are available. The exact method builds all the states of the tree-HMM and guarantees accurate results at the cost of longer running times. The heuristic one builds states that are involved in the HMM and guarantees accurate results at the cost of longer running times. The heuristic one builds states that are involved in the best Viterbi or forward-backward paths (Diallo et al., 2007) and provides faster results. Since the initial version was published, several improvements have been made to the program to speed up the running time, handle large data and integrate it into the ancestral reconstruction pipeline. Ancestors 1.0 can currently compute ancestral sequences for more than 30 species of up to a few mega base pair each. For instance, a benchmark CFTR region (1Mb) of 28 mammals has been reconstructed in ~2 min using the heuristics for indel reconstruction with the default parameters.

2.2 Ancestors 1.0 outputs

The Ancestors 1.0 results are made available using various plain text and HTML files. The summary of those files is given as default output. The results include:

1. The set of aligned extant sequences (multiple Fasta format).
2. The results of the indel predictions (the maximum likelihood indel scenario, the posterior probability for each position of each ancestor and a summary of the tree-HMM run).
3. The results of the ancestral sequence prediction, in the form of an alignment of extant and predicted ancestral sequences.

Ancestral sequences are given in plain text file and colored coded according to the level of prediction confidence, as shown in Figure 2. The colored output contains tooltips of the confidence values for each character.

2.3 Documentation, help and source code

At each step of the process, the user can send questions or feedback to the system administrator via e-mail. Sample inputs and outputs are also available. Moreover, there are links to the required formats for the different inputs and an explanation of the different sections of the results. Finally, the C++ source code of Ancestors 1.0 is also available for download.

Conflict of Interest: none declared.

REFERENCES