Exact and Heuristic Algorithms for the Indel Maximum Likelihood Problem

Abdoulaye Banire Diallo[1,2], Vladimir Makarenkov[2], and Mathieu Blanchette[1*]

[1] McGill Centre for Bioinformatics and School of Computer Science, McGill University, 3775 University Street, Montréal, Québec, H3A 2B4, Canada

[2] Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (Québec), H3C 3P8, Canada

* Corresponding author

Abstract

Given a multiple alignment of orthologous DNA sequences and a phylogenetic tree for these sequences, we investigate the problem of reconstructing the most likely scenario of insertions and deletions capable of explaining the gaps observed in the alignment. This problem, that we called the Indel Maximum Likelihood Problem (IMLP), is an important step toward the reconstruction of ancestral genomics sequences, and is important for studying evolutionary processes, genome function, adaptation and convergence. We solve the IMLP using a new type of tree hidden Markov model whose states correspond to single-base evolutionary scenarios and where transitions model dependencies between neighboring columns. The standard Viterbi and Forward-backward algorithms are optimized to produce the most likely ancestral reconstruction and to compute the level of confidence associated to specific regions of the reconstruction. A heuristic is presented to make the method practical for large data sets, while retaining an extremely high degree of accuracy. The methods are illustrated on a 1Mb alignment of the CFTR regions from 12 mammals.

KEYWORDS: ANCESTRAL GENOME RECONSTRUCTION; INSERTIONS AND DELETIONS; TREE-HMM; ANCESTRAL MAMMALIAN GENOMES, INDEL MAXIMUM LIKELIHOOD PROBLEM

1 Introduction

It has recently been shown that the phylogeny of eutherian mammals is such that an accurate reconstruction of the genome of an early ancestral mammal is possible (Blanchette *et al.*, 2004a). This accurate reconstruction will help on various studies such as adaptation, behavioral changes, functional divergences, etc. (Krishnan *et al.*, 2004). It is also at the core of experimental paleo-molecular biochemistry where sequences of extant taxa are used to predict and resurrect the sequences and functions of ancestral macromolecules (Benner, 2002; Gaucher *et al.*, 2003; Pauling and Zuckerkandl, 1963). The ancestral genome reconstruction procedure involves several difficult steps, including the identification of orthologous regions in different extant species, ordering of syntenic blocks, multiple alignment of orthologous sequences within each syntenic block, and reconstruction of ancestral sequences for each aligned block. This last step involves the inference of the set of substitutions, insertions, and deletions that may have produced a given set of multiply-aligned extant sequences. While the problem of reconstructing substitutions scenarios has been well studied (e.g. (Fitch, 1971) and (Felsenstein, 1981)), the inference of insertions and deletions scenarios has received less attention (but see the seminal contribution of Thorne, Kishino and Felsenstein (Thorne *et al.*, 1991)). Indel evolutionary scenarios are useful for several other problems such as annotating functional regions of extant genomes, including protein-coding regions (Siepel and Haussler, 2004), RNA genes (Rivas, 2005), and other types of functional regions (Siepel *et al.*, 2005). The difficulty of the problem is due in large part to the fact that insertions and deletions (indels) often affect several consecutive nucleotides, so the columns of the alignment cannot be treated independently, as opposed to the maximum likelihood problem for substitutions (Felsenstein, 1981). The reconstruction of the most parsimonious scenario of indels required to explain a given multiple sequence alignment has been shown to be NP-Complete (Chindelevitch *et al.*, 2006) but good heuris-

tics have been developed (Blanchette *et al.*, 2004a; Chindelevitch *et al.*, 2006; Fredslund *et al.*, 2004).

A maximum likelihood reconstruction would be preferable to a most parsimonious reconstruction because it would provide a way of weighing insertions and deletions of various lengths against each other. Moreover, provided an accurate probabilistic model is used, it would be more accurate and would allow to estimate the uncertainty related to certain aspects of the reconstruction. Similarly to statistical alignment approaches (Lunter *et al.*, 2003) (which unfortunately remain too slow for genome-wide reconstructions), we seek to gain a richer insight into ancestral sequences and evolutionary processes. In this paper, we thus focus on the problem we call the *Indel Maximum Likelihood Problem (IMLP)*. It consists of inferring the set of insertions and deletions that has the maximal likelihood, according to some fixed evolutionary parameters, and that could explain the gaps observed in a given multiple alignment. An example of the input and output of this problem is shown in Figure 1. Kim and Sinha (2006) have recently proposed an algorithm for a similar problem, although the range of scenarios handled by their Indelign program is limited to non-overlapping indels.

We emphasize that the problem addressed here assumes that the phylogenetic tree and multiple sequence alignment given as input are correct. The robustness of indel scenarios with respect to alignment and tree accuracy has been previously discussed in (Blanchette *et al.*, 2004a). The more general problem where the alignment in not given as input but has to be found simultaneously with the ancestral sequences (Hein, 1989) is clearly of great interest but appears significantly more difficult and is not addressed here. We refer the reader to (Kim and Sinha, 2006) and (Bray and Pachter, 2004) for interesting first steps in that direction.

Here, we start by giving a formal definition of the Indel Maximum Likelihood Problem. To solve the problem, we use a special type of tree hidden Markov model (tree-HMM), which is a combination of a standard hidden Markov model

and a phylogenetic tree. We show how the most likely path through the tree-HMM leads to the most likely indel scenario and how a variant of the standard Viterbi algorithm can solve the problem. Although the size of the HMM is exponential in the number of extant species considered, we show how the knowledge given by the phylogenetic tree and the aligned sequences allows the state space of the HMM to be considerably reduced, resulting in a practical, yet exact, algorithm. We also present a heuristic algorithm that almost always gives the right solution and can compute the most likely indels scenarios for more than 20 taxa. Thus, our implementations are able to solve large problems on a simple desktop computer and allow for an easy parallelization. Finally, we assess the complexities and accuracies of the presented algorithms on a multiple alignment of twelve orthologous mammalian genomic sequences of ∼1Mb each coming from the CFTR benchmark dataset (ENCODE Project Consortium, 2004).

FIGURE 1 HERE

2 The Indel Maximum Likelihood Problem

In this section we will give a precise definition for the Indel Maximum Likelihood Problem (IMLP). Consider a rooted binary phylogenetic tree $T = (V_T, E_T)$ with branch lengths $\lambda : V_T \to \mathbb{R}^+$. If n is the number of leaves of T, there are $n - 1$ internal nodes and $2n - 2$ edges.

Consider a multiple alignment A of n orthologous sequences corresponding to the leaves of the tree T. Since the only evolutionary events of interest here are insertions and deletions, A can be transformed into a binary matrix, where gaps are replaced by 0's and nucleotides by 1's. Let A_x be the row of the binarized alignment corresponding to the sequence at leaf x of T, and let $A_x[i]$ be the binary character at the i-th position of A_x. Assuming that the alignment A

contains L columns, we add for convenience two extra columns, $A[0]$ and $A[L+1]$, consisting exclusively of 1's.

Definition 1 (Ancestral reconstruction). *Given a multiple alignment A of n extant sequences assigned to the leaves of a tree T, an* ancestral reconstruction *A^* is an extension of A that assigns a sequence $A^*_u \in \{0,1\}^{L+2}$ to each node u of T, and where $A^*_u = A_u$ whenever u is a leaf.*

The following restriction on the set of possible ancestral reconstructions is necessary in some contexts.

Definition 2 (Phylogenetically correct ancestral reconstruction). *An ancestral reconstruction A^* is* phylogenetically correct *if, for any $u, v, w \in V_T$ such that w is located on the path between u and v in T, we have $(A^*_u[i] = A^*_v[i] = 1) \implies (A^*_w[i] = 1)$.*

Requiring an ancestral reconstruction to be phylogenetically correct corresponds to assuming that any two nucleotides that are aligned in A have to be derived from a common ancestor, and thus that all the ancestral nodes between them have to have been a nucleotide. This prohibits aligned nucleotides to be the result of two independent insertions. Assuming that this property holds perfectly for a given alignment A is somewhat unrealistic, but, for mammalian sequences, good alignment heuristics have been developed (e.g. TBA (Blanchette *et al.*, 2004b), MAVID (Bray and Pachter, 2004), MLAGAN (Brudno *et al.*, 2003)) and have been shown to be quite accurate (Blanchette *et al.*, 2004b). In the future, we plan to relax this assumption, but, for now, we will concentrate only on finding phylogenetically correct ancestral reconstructions.

Since we are considering insertions and deletions affecting several consecutive characters, we delimit each operation by the positions s and e in the aligned sequences where it starts and ends. Let x and y be two nodes of the tree, where x is the parent of y. The pairwise alignment consisting of rows A^*_x and A^*_y is divided into a set of regions defined as follows (see Figure 2).

Definition 3 (Deletions, Insertions, Conservations, and Length). *Consider the pairwise alignment of A_x^* and A_y^*, and let $0 \leq s \leq e \leq L+1$.*

- *The region (s,e) is a* deletion *if (a) for all $i \in \{s,\ldots,e\}, A_y^*[i] = 0$, (b) $A_x^*[s] = A_x^*[e] = 1$, and (c) no region $(s',e') \supset (s,e)$ is a deletion (i.e. we only consider regions that are maximal).*
- *The region (s,e) is an* insertion *if (a) for all $i \in \{s,\ldots,e\}, A_x^*[i] = 0$, (b) $A_y^*[s] = A_y^*[e] = 1$, and (c) no region $(s',e') \supset (s,e)$ is an insertion.*
- *The region (s,e) is a conservation if (a) for all $i \in \{s,\ldots,e\}, A_x^*[i] = A_y^*[i]$ and (b) no region $(s',e') \supset (s,e)$ is a conservation.*
- *The* length *of region (s,e) is the number of non-trivial positions it contains: $l(s,e) = |\{s \leq i \leq e | A_x^*[i] \neq 0 \text{ or } A_y^*[i] \neq 0\}|$.*

A pair of binary alignment rows A_x^* and A_y^* can thus be partitioned into a set of non-overlapping insertions, deletions, and conservations.

FIGURE 2 HERE

Definition 4 (Indel scenario). *The indel scenario defined by an ancestral reconstruction A^* is the set of insertions and deletions that occurred between the ancestral reconstructions at adjacent nodes in T.*

All that remains is to define an optimization criterion on A^*. Two main choices are possible: a parsimony criterion or a likelihood criterion.

2.1 The Indel Parsimony Problem (*IPP*)

The parsimony approach for the indel reconstruction problem has been introduced by Fredslund *et al.* (2004) and Blanchette *et al.* (2004a). In its simplest version, it attempts to find the phylogenetically correct ancestral reconstruction

A^* that minimizes the total number of insertions and deletions defined by A^*:

$$indelParsimony(A^*) = \sum_{u,v:(u,v)\in E_T} |\{(s,e) : (s,e) \text{ is a deletion or an insertion from } A^*_u \text{ to } A^*_v\}|$$

The Indel Parsimony Problem is NP-Hard (Chindelevitch *et al.*, 2006). Most authors have studied a weighted version of the *IPP* where the cost of indels depends linearly on their length (affine gap penalty). Blanchette *et al.* (2004a) proposed a greedy algorithm, and good exact heuristics have been developed (Chindelevitch *et al.*, 2006; Fredslund *et al.*, 2004). The limitation of these approaches is that they only give a single solution as output, and provide no measure of uncertainty of the various parts of the reconstruction. In contrast, a likelihood-based approach has the potential of providing a more accurate solution and a richer description of the set of possible solutions.

2.2 Indel Maximum Likelihood Problem

In this section, we define the indel reconstruction problem in a probabilistic framework similar to the Thorne-Kishino-Felsenstein model (Thorne *et al.*, 1992). To this end, we need to define the probability of transition between an alignment row A^*_x and its descendant row A^*_y. This probability will be defined as a function of the probability of the insertions, deletions, and conservations that happened from A^*_x to A^*_y.

Let $P_{DelStart}(\lambda(b))$ be the probability that a deletion starts at a given position in the sequence, along a branch b of length $\lambda(b)$, and let $P_{InsStart}(\lambda(b))$ be defined similarly for an insertion. We assume that these probabilities only depend on the length $\lambda(b)$ of the branch b along which they occur, but not on the position where the indel occurs. A reasonable choice is $P_{DelStart}(\lambda(b)) = 1 - e^{-\psi_D \lambda(b)}$ and $P_{InsStart}(\lambda(b)) = 1 - e^{-\psi_I \lambda(b)}$, for some deletion and insertion rate parameters ψ_D and ψ_I, but our algorithm allows for any other choice of these probabilities. Thus, the probability that none of the two events happens at

a given position, which we call the probability of a conservation, is given by $P_{Cons}(\lambda(b)) = e^{-(\psi_D+\psi_I)\lambda(b)}$. We make the standard simplifying assumption that the length of a deletion follows a geometric distribution, where the probability of a deletion of length k is $\alpha_D^{k-1}(1-\alpha_D)$ and the probability of an insertion of length k is $\alpha_I^{k-1}(1-\alpha_I)$. One can thus see α_D (resp. α_I) as the probability of extending a deletion (resp. insertion). This assumption, necessary to design a fast algorithm, holds relatively well for short indels, but fails for longer ones (Kent *et al.*, 2003). Our algorithm allows the parameters α_D and α_I to depend on the branch b, but the results reported in Section 5 correspond to the case where α_D and α_I were held constant across the tree. The probability that alignment row A_x^* was transformed into alignment row A_y^* along branch b can be defined as follows:

$$\begin{aligned}\Pr(A_y^*|A_x^*,b) = &\prod_{(s,e)\text{: deletion from } A_x^* \text{ to } A_y^*} P_{DelStart}(\lambda(b)) \cdot (\alpha_D^{l(s,e)-1}(1-\alpha_D)) \cdot \\ &\prod_{(s,e)\text{: insertion from } A_x^* \text{ to } A_y^*} P_{InsStart}(\lambda(b)) \cdot (\alpha_I^{l(s,e)-1}(1-\alpha_I)) \cdot \\ &\prod_{(s,e)\text{:conservation from } A_x^* \text{ to } A_y^*} (P_{Cons}(\lambda(b))^{l(s,e)}\end{aligned}$$

This allows us to formulate precisely the problem addressed in this paper:

INDEL MAXIMUM LIKELIHOOD PROBLEM (IMLP):

Given: A multiple sequence alignment A of n orthologous sequences related by a phylogenetic tree T with branch lengths λ, a probability model for insertions and deletions specifying the values of ψ_D, ψ_I, α_D, and α_I.

Find: A maximum likelihood phylogenetically correct ancestral reconstruction A^* for A, where the likelihood of A^* is:

$$L(A^*) = \prod_{b=(x,y)\in E_T} \Pr(A_y^*|A_x^*,b)$$

3 A Tree-Hidden Markov Model

In this section, we describe the tree hidden Markov model that is used to solve the IMLP. A tree-hidden Markov model (tree-HMM) is a probabilistic model that allows two processes to occur, one in time (related to the sequence history in a given column of A), and one in space (related to the changes toward the neighboring columns). Tree HMMs were introduced by Felsenstein and Churchill (1996) and Yang (1996) to improve the phylogenetic models that allows for variation among sites in the rate of substitution, and have since then been used for several other purposes (e.g. detecting conserved regions (Siepel *et al.*, 2005) and predicting genes (Siepel and Haussler, 2004)). Just as any standard HMM (Durbin *et al.*, 1998), a tree-HMM is defined by three components: the set of states, the set of emission probabilities, and the set of transition probabilities.

3.1 States

Intuitively, each state corresponds to a different single-column indel scenario (although additional complications are described below). Given a rooted binary tree $T = (V_T, E_T)$ with n leaves, each state corresponds to a different labeling of the edges E_T with one of three possible events: I (for insertion), D (for deletion), or C (for conservation). The set $\mathcal{S}$ of possible states of the HMM would then be $\mathcal{S} = \{I, D, C\}^{2n-2}$. However, this definition is not sufficient to model certain biological situations (see Figure 3). We will use the '*' symbol to indicate that, along a certain branch $b = (x, y)$, no event happened because there was a base neither at node x nor at node y. This will happen in two situations: when edge b is a descendant of edge b' that was labeled with D (i.e. the base was deleted higher up the tree), and when there exists an edge b' that is not between b and the root and that is labeled with I (i.e. an insertion happened elsewhere in the tree). The fact that these extraneous events can potentially interrupt ongoing events along branch b means that the HMM needs to have a way to remember what event

was actually going on along that branch. This transmission of memory from column to column is achieved by three special labels: I^*, D^*, and C^*, depending on whether the * regions is interrupting an insertion, deletion, or conservation. Thus, we have $\mathcal{S} \subseteq \{I, D, C, I^*, D^*, C^*\}^{2n-2}$. Although this state space appears prohibitively large (6^{2n-2}), the reality is that a number of these states cannot represent actual indel scenarios, and can thus be ignored. The following set of rules specify what states are valid.

Definition 5 (Valid states). *Given a tree $T = (V_T, E_T)$, a state s assigning a label $s(b) \in \{I, D, C, I^*, D^*, C^*\}$ to each branch $b \in E_T$ is* valid *if the two following conditions hold.*

- *(Phylogenetic correctness condition) There must be at most one branch b such that $s(b) = I$.*
- *(Star condition) Let $b \in E_T$, and let $anc(b) \subset E_T$ be the set of branches on the path from the root to b. Then $s(b) \in \{I^*, D^*, C^*\}$ if and only if $\exists b' \in anc(b)$ such that $s(b') = D$ or $\exists b' \in (E_T \setminus anc(b))$ such that $s(b') = I$.*

The number of valid states on a complete balanced phylogenetic tree with n leaves is $O(n \cdot 3^{2n})$ (the number is dominated by states that have a 'I' on a branch leading to a leaf, which leaves all other $2n - 3$ edges free to be labeled with either C^*, D^*, or I^*). Although this number remains exponential, it is significantly better than the 6^{2n-2} valid and invalid states.

3.2 Emission probabilities

In an HMM, each state emits one symbol, according a certain emission probability distribution. In our tree-HMMs, each state emits a collection of symbols, corresponding to the set of characters obtained at the leaves of T when indel scenario s occurs. Intuitively, we can think of a state as emitting an alignment column. The following definition formalizes this.

Definition 6. *Let s be a valid state for tree $T = (V_T, E_T)$ with root r. Then, we define the output of state s as a function $O_s : V_T \rightarrow \{0, 1\}$ with the following recursive properties:*

1. $O_s(root) = \begin{cases} 0, & \text{if } \exists x \in V_T \text{ such that } s(x) = I \\ 1, & \text{otherwise} \end{cases}$.
2. *Let $e = (x, y) \in E_T$, with x being the parent of y. Then,*

$$O_s(y) = \begin{cases} 0, & \text{if } s(e) = D \\ 1, & \text{if } s(e) = I \\ O_s(x), & \text{otherwise} \end{cases}$$

Let C be an alignment column (i.e. an assignment of 0 or 1 to each leaf in T). We then have the following degenerate emission probability for state s:

$$Pr_e(C|s) = \begin{cases} 1 \text{ if } O_s(x) = C(x) \text{ for all } x \in leaves(T) \\ 0 \text{ otherwise} \end{cases}$$

Thus, each state s can emit a single alignment column C. However, many different states can emit the same column.

Missing data In presence of missing characters among the input sequences, the emission probability can be adapted such that the equality between $O_s(x)$ and $C(x)$ is assessed according to 0's and 1's in $C(x)$ only. It is worth noting that missing characters are different to gap noted by $-$. Hence, the presence of missing data increases the number of states for a given column.

3.3 Transition probabilities

The last component to be defined is the set of transition probabilities of the tree-HMM. The probability of transition from state s to state s', $\Pr_t(s'|s)$, is a function of the set of events that occurred along the edges of T. Intuitively, $\Pr_t(s'|s)$ describes the probability of the single-column indel scenario s', given that scenario s occurred at the previous column. This transition probability is

a function of insertions and deletions that started between the two columns, of those that were extended going from one column to the next. Specifically, we have $\Pr_t(s'|s) = \prod_{b \in E_T} \rho(s'(e)|s(e), b)$, where ρ is given in Table 1.

[TABLE 1 HERE]

4 Tree-HMM paths, ancestral reconstruction and assessing uncertainty

We now show how the tree-HMM described above allows us to solve the IMLP. Consider a multiple alignment A of length L on a tree T. A path π in the tree-HMM is a sequence of states $\pi = \pi_0, \pi_1, ..., \pi_L, \pi_{L+1}$. Based on standard HMM theory, we get:

$$\Pr(\pi, A) = \Pr(\pi_0, A_0) \prod_{i=1}^{L+1} \Pr_e(A[i]|\pi_i) \cdot \Pr_t(\pi_i|\pi_{i-1})$$

Figure 3 gives an example of an alignment with some of the non-zero probability paths associated.

FIGURE 3 HERE

Theorem 1. *Consider an alignment A on tree T. Then $\pi^* = argmax_\pi \Pr(\pi, A)$ yields the most likely indel scenario for A, and a maximum likelihood ancestral reconstruction A^* is obtained by setting $A^*_u[i] = O_{\pi^*_i}(u)$.*

Proof. It is simple to show that for any ancestral reconstruction $\hat{A}$ for A, we have $L(\hat{A}) = \Pr(\pi, A)$, where π is the path corresponding to $\hat{A}$. Thus, maximizing $\Pr(\pi, A)$ maximizes $L(\hat{A})$.

4.1 Computing the most likely path

To compute the most likely path π^* through a tree-HMM, we adapted the standard Viterbi dynamic programming algorithm (Durbin *et al.*, 1998). Let $X(i,k)$ be the joint likelihood of the most probable path ending at state k for the i first columns of the alignment. Let $c \in \mathcal{S}$ be the state made of C's on all edges of T. Since the dummy column $A[0]$ consists exclusively of 1's, c is the only possible initial state. For any i between 0 and $L+1$ and for any valid state $s \in \mathcal{S}$, we can compute $X(i,s)$ as follows:

$$X(i,s) = \begin{cases} 1, & \text{if } i = 0 \text{ and } s = c \\ 0, & \text{if } i = 0 \text{ and } s \neq c \\ \Pr_e(A[i]|s) \cdot \max_{s' \in \mathcal{S}}(X(i-1,s') \cdot \Pr_t(s|s')), & \text{if } i > 0 \end{cases}$$

Finally, π^* is obtained by tracing back the dynamic programming, starting from entry $X(L+1,c)$. To ensure numerical stability, we use a log transformation and scaling of probabilities as described by (Durbin *et al.*, 1998).

The running time of a naive implementation of the Viterbi algorithm is $O(|\mathcal{S}|^2 L)$, which quickly becomes impractical as the size of the tree T grows. However, we can make this computation practical for moderately large trees and for long sequences. Even though the number of states is exponential in the number of sequences, most alignment columns can only be generated with non-zero probability by a much more manageable number of states. Given an alignment A, it is possible to compute, for each column $A[i]$, the set S_i of valid states that can emit $A[i]$ with non-zero probability. For instance, an alignment column with only 1's will lead to only one possible state, independently of the number taxa of n. The set S_i can be constructed using a bottom approach presented in Algorithm 1. More states can be discarded by using the fact that the transition probability between most pairs of states is zero. We can thus remove from S_i any state s that is such that the transition to s from any state in S_{i-1} has probability zero. Proceeding from left to right, we get $S'_0 = S_0$, and

$S'_i = \{s \in S_i | \exists t \in S'_{i-1}$ s.t. $\Pr_t(s|t) > 0\}$, where $S'_i \subseteq S_i$. For instance, if, in all states of S_{i-1}, an edge e is labeled by deletion D, then none of the states in S_i can have edge e labeled with C^* or I^*. This yields a large improvement for alignment regions consisting of a number of adjacent positions with a base in only one of the n species and ensures that the algorithm will be practical for relatively large number of sequences (see Section 5).

Algorithm 1 buildValidState(node *root*, C)

Require: *root*: a tree node, C: an alignment column.

Ensure: Set of valid, non-zero probability states for C.

```
1: if root is a leaf then
2:     return list of possible operations according to the character at that leaf
3: else
4:     leftList = buildValidState(root.left, C)
5:     rightList = buildValidState(root.right, C)
6:     return mergeSubtrees(leftList, rightList, root)
7: end if
```

4.2 Assessing uncertainties of the ancestral reconstruction

A significant advantage of the likelihood approach over the parsimony approach is that it allows evaluating the uncertainty related to certain aspects of the reconstruction. For example, it is useful to be able to compute the probability that a base was present at a given position i of a given ancestral node u: $\Pr(A^*_u[i] = 1|A) = \sum_{s \in \mathcal{S}: O_s(u)=1} \Pr(\pi_i = s|A)$. This allows the computation of the probability of making an incorrect prediction at a given position of a given ancestor. The forward-backward is a standard HMM algorithm to compute $\Pr(\pi_i = s|A)$ (see (Durbin *et al.*, 1998) for more details). The optimizations developed for the Viterbi algorithm can be trivially adapted to the Forward-Backward algorithm.

Algorithm 2 mergeSubtrees(StateList $leftList$, StateList $rightList$, node $root$)

Require: $leftList$ and $rightList$: the lists of partial states, $root$: a tree node.

Ensure: Set of valid, non-zero probability states combining elements in $leftList$ and $rightList$.

```
1: mergedList ← emptyList
2: for all partial states l in leftList do
3:    for all partial states r in rightList do
4:       if compatible(l, r) == true  then
5:          m = merge(l, r)
6:          if root == initialroot then
7:             mergedList.add(m)
8:          else
9:             for op ∈ {C, D, I, C*, D*, I*} do
10:               if isPossibleUpstream(m,op) then
11:                  mergedList.add(addAncestorBranch(m,op))
12:               end if
13:            end for
14:         end if
15:      end if
16:   end for
17: end for
18: return  mergedList
```

5 Results of the exact method

Our tree-HMM algorithm was implemented as a C program that is available upon request. The program was applied to a ∼700kb region of the CFTR locus on chromosome 7 of human, together with orthologous regions in 11 other species of mammals: chimp, macaque, baboon, mouse, rat, rabbit, cow, dog,

Rodrigues fruit bat (rfbat), armadillo, and elephant[3] (ENCODE Project Consortium, 2004). This locus is representative of the whole genome, and contains coding, intergenic regions, and intronic regions. The multiple alignment of these regions, computed using TBA (Blanchette *et al.*, 2004b; Miller, 2006), contains 1,000,000 columns. To simplify the calculations, consecutive alignment columns with the same gap structure were assumed to have undergone the same evolutionary scenario and were thus merged into a single "meta-column" we called an alignment *region*. Our alignment consisted of 123,917 such regions. Thus, during the execution of the Viterbi or Forward-Backward algorithm, the states are computed for each region instead of for each individual column, adapting the transition probabilities as a function of the width of each region. The phylogenetic tree used for the alignment and for the reconstruction is shown in Figure 4. The branch lengths are based on substitution rates estimated on a genome-wide basis (Miller, 2006). For illustrative purposes, and similarly to the empirical values obtained by (Kent *et al.*, 2003), the parameters of the indel model were set as follows: $\psi_D = 0.05, \psi_I = 0.05, \alpha_D = 0.9$, and $\alpha_I = 0.9$. However, we find that the ancestral reconstructions and confidence levels are quite robust with respect to these parameters (data not shown).

FIGURE 4 HERE

We first compared the maximum likelihood ancestral reconstruction found using our Viterbi algorithm to the ancestors inferred using the greedy algorithm of Blanchette *et al.* (2004a). Table 2 shows the degree of agreement between the two reconstructed ancestors, for each ancestral node. We observe that both methods agree to a very large degree, with most ancestors yielding more than 99% agreement. The most disagreement concerns the ancestor at the root of

[3] In the case of cow, armadillo, and elephant, the sequence is incomplete and a small fraction of the bases are missing.

the eutherian tree, which, in the absence of an outgroup, cannot be reliably predicted by any method. We expect that in most other cases of disagreement, the maximum likelihood reconstruction is the most likely to be correct, although the opposite may be true in case of gross model violations (Hudek and Brown, 2005).

[TABLE 2 HERE]

The main strength of the likelihood-based method is its ability to measure uncertainty, using the forward-backward algorithm, something that no previous method allowed. Assuming a phylogenetically correct alignment and a correct indel model, the probability that the maximum posterior probability reconstruction is correct is simply given by $\max\{\Pr(A_u^*[i] = 1|A), 1 - \Pr(A_u^*[i] = 1|A)\}$. For example, if $\Pr(A_u^*[i] = 1|A) = 0.3$, then the maximum posterior probability reconstruction would predict $A_u^*[i] = 0$, and would be right with probability 0.7. Figure 5 shows the distribution of this probability of correctness, for each ancestral node in the tree, over all regions of the alignment. We observe, for example, that 98% of the positions in the Boreoeutherian ancestor (the human+chimp+baboon+macaque+mouse+rat, cow+dog+rfbat ancestor, living approximately 75 million years ago), are reconstructed with a confidence level above 99% [4]. The ancestor that is the easiest to reconstruct confidently is obviously the human-chimp ancestor, where less than 0.14% of the regions have a confidence level below 99%. Again, the root of the tree is the node that is the most difficult to reconstruct confidently. Overall, this shows that most positions of most ancestral nodes can be reconstructed very accurately, and that we can identify the few positions where the reconstruction is uncertain.

[4] We need to keep in mind, though, that these numbers assume the correctness of the multiple alignment, as well as that of the branch lengths and indel probability model, so that they do not reflect the true correctness of the reconstructed ancestor.

FIGURE 5 HERE

A potential drawback of the tree-HMM method is that its running time is, in the worst case, exponential in the number of sequences being compared. However, the optimizations described in this paper greatly reduce the number of states that need to be considered at each position, so the algorithm remains quite fast. Our optimized Viterbi algorithm produced its maximum likelihood ancestral predictions on the 12-species, 1,000,000 column alignment in 7 hours on a Powerbook G5 machine, while the forward-backward algorithm produced an output after approximately double of that time. Figure 6 shows the distribution of the number of states that were actually considered, per alignment column. Most alignment columns are actually associated to less than 100 states. However, a small number of columns are associated to a very large number of states (15 regions have more than 100,000 states). Fortunately, these columns are rarely consecutive, so the incurred running time is not catastrophic for small number of species. However, to be applicable to complete genomes and to scale up to the more than 20 mammalian genomes that will soon be available, our algorithm requires further optimizations. These optimizations move away from an exact algorithm, toward approximation algorithms.

FIGURE 6 HERE

6 Heuristic algorithm for the IMLP

For each region i of the alignment and each possible state $s \in S'_i$, the exhaustive method considers all possible states for the next column, even though the Viterbi value $X(i, s)$ of some current state s may be far away from the maximal Viterbi

value at that position, $\max_{s' \in S'_i} X(i, s')$. These states are less likely to be eventually chosen in the best path of the tree-HMM. Hence, to reduce the number of states created and reduce computation time, only states near the maximum Viterbi value are used to compute states for the next column. Thus, for region i, we distinguish between created states S'_i and used states $R_i \subseteq S'_i$, where only the second set will be involved in the creation of the states of the next column and in their Viterbi calculation. For position i, state $s \in S'_i$ is retained in R_i if and only if $\log_2(\frac{\max_{s'} X(i,s')}{X(i,s)}) < t$, for some fixed threshold t. We note that this is equivalent to setting $X(i, s)$ to zero for each $s \in S - R$. A similar heuristic can easily be applied to the Forward-Backward algorithm. If t is sufficiently large, the loss in accuracy should be minimal for both algorithms, as will be shown next.

We computed the indels scenarios of the data sets presented in Section 5 by using different values for the threshold t. The approximate Viterbi algorithm was run using $t = 0, 1, 3, 5, 7, 9, 10, 20, 100$, and $+\infty$. Note that setting $t = 0$ results in a "greedy" algorithm that only considers the maximum Viterbi value at each position, while $t = +\infty$ give the original, optimal Viterbi algorithm. Figure 7 shows the number of states created (average of $|S'_i|$) and used (average of $|R_i|$) for all values of t, as well as the resulting running time. For small values of t, e.g. $t \leq 3$, only a handful of states are used, resulting in a very fast execution (less than 3 minutes). The average of number of states created increases relatively quickly with t, while the number of states used remains quite low (44.34 for $t = 100$). The average number of states created for $t = 20$ is about the same as the average number of states of the exact algorithm (see Figure 7), which shows that the used states are sufficient to give the necessary information to generate most valid states for next columns.

FIGURE 7 HERE

Even though the average number of states created and used for $0 \leq t \leq 5$ is very low, the indels scenarios produced are very similar to the best scenario obtained by the exact method (see Table 3). We note that, for $t = 5$, the agreement with the exact algorithm is more than 99.99% for all the ancestors, while the running time is reduced by a factor of ten, and by a factor of one hundred for $t = 3$. For $t \geq 9$, the heuristic gives the optimal scenario, while still yielding a 5-fold speed-up. All values of t tested gave solutions that agreed with the optimal solution better than the solution produced by the greedy algorithm of (Blanchette *et al.*, 2004a). Finally, we note that, while our optimal Viterbi and Forward-Backward algorithms are limited to 12 to 15 species, our heuristic allows the inference of near-optimal solutions for much larger alignments. When run on a 1,000,000 column alignment of 28 species of vertebrates, our heuristic with $t = 3$ produced a solution in less than two hours. Since the exact algorithm cannot be run on such a large data set, it is difficult to estimate the quality of the solution obtained but, based on our experience on the smaller data set (Table 3), we expect a very high accuracy even at such a stringent cutoff.

TABLE 3 HERE

7 Discussion and Future Work

The method developed here allows predicting maximum likelihood indel scenarios and their resulting ancestral sequences for large alignments. Furthermore, it allows the estimation of the probability of error in any part of the prediction, using the forward-backward algorithm. Integrated into the pipeline for whole-genome ancestral reconstruction, it will improve the quality of the predictions and allow richer analyses. The main weakness of our approach is that it assumes that a phylogenetically correct alignment and an accurate phylogenetic tree are

given as input. While many existing multiple alignment programs have been shown to be quite accurate on mammalian genomic sequences (including non-functional or repetitive regions) (Blanchette *et al.*, 2004b), it has also been shown that a sizeable fraction of reconstruction errors is due to incorrect alignments (Blanchette *et al.*, 2004a). Ideally, one would include the optimization of the alignment directly in the indel reconstruction problem, as originally suggested by Hein (1989). However, with the exception of statistical alignment approaches (Lunter *et al.*, 2003), which remain too slow to be applicable on a genome-wide scale, genomic multiple alignment methods do not treat indels in a probabilistic framework. We are thus investigating the possibility of using the method proposed here to detect certain types of small-scale alignment errors, and to suggest corrections.

When predicting ancestral genomic sequences, it is very important to be able to quantify the uncertainty with respect to certain aspects of the reconstruction. Our forward-backward algorithm calculates this probability of error for each position of each ancestral species. However, errors in adjacent columns are not independent: if position i is incorrectly reconstructed, it is very likely that position $i + 1$ will be wrong too. We are currently working on models to represent this type of correlated uncertainties. This new type of representation will play an important role in the analysis and visualization of ancestral reconstructions.

Finally, it will be important to assess the results given by the heuristic so that the cutoff value t is chosen appropriately for the data at hand. For example, the heuristic could be applied iteratively by increasing the cutoff until a stationary likelihood score is reached. This heuristic will be useful to reconstruct the indel scenarios for data sets containing more than 20 taxa and could be easily applied to the large number of mammalian genomes that are about to be completely sequenced.

8 Acknowledgements

A.B.D. is an NSERC fellow. We thank Éric Gaul, Eric Blais, Adam Siepel, and the group of participants to the First Barbados Workshop on Paleogenomics for their useful comments. We thank Webb Miller and David Haussler for providing us with the sequence alignment data.

Bibliography

Benner, S., 2002. The past as the key to the present: resurrection of ancient proteins from eosinophils. *Proceedings of the National Academy of Science USA* 99, 4760–4761.

Blanchette, M., Green, E. D., Miller, W., and Haussler, D., 2004a. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res* 14, 2412–2423.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D., and Miller, W., 2004b. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* 14, 708–715.

Bray, N. and Pachter, L., 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Research* 14, 693–699.

Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., Sidow, A., and Batzoglou, S., 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* 13, 721–731.

Chindelevitch, L., Li, Z., Blais, E., and Blanchette, M., 2006. On the inference of parsimonious indel evolutionary scenarios. *Journal of Bioinformatics and Computational Biology* 4(3), 721–44.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G., 1998. *Biological Sequence Analysis*. Cambridge University Press.

ENCODE Project Consortium, 2004. The ENCODE (ENCyclopedia of DNA elements) project. *Science* 306, 636–640.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17, 368–376.

Felsenstein, J. and Churchill, G., 1996. A hidden markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* 13, 93–104.

Fitch, W. M., 1971. Toward defining the course of evolution: Minimum change for a specified tree topology. *Systematic Zoology* 20, 406–416.

Fredslund, J., Hein, J., and Scharling, T., 2004. A large version of the small parsimony problem. In *Proceedings of the 4th Workshop on Algorithms in Bioinformatics (WABI)*.

Gaucher, E., Thomson, M., Burgan, M., and Benner, S., 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425, 285–288.

Hein, J., 1989. A method that simultaneously aligns, finds the phylogeny and reconstructs ancestral sequences for any number of ancestral sequences. *Molecular Biology and Evolution* 6(6), 649–668.

Hudek, A. and Brown, D., 2005. Ancestral sequence alignment under optimal conditions. *BMC Bioinformatics* 6:273, 1–14.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D., 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* 100, 11484–11489.

Kim, J. and Sinha, S., 2006. Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment. *Bioinformatics* btl578.

Krishnan, N., Seligman, H., Stewart, C., Jason de Koning, A., and Pollock, D., 2004. Ancestral sequence reconstruction in primate mitochondrial dna: Compositional bias and effect on functional inference. *Molecular Biology and Evolution* 21(10), 1871–1883.

Lunter, G., Miklos, I., Song, Y., and Hein, J., 2003. An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J Computational Biology* 10(6), 869–89.

Miller, W., 2006. Personal communication.

Pauling, L. and Zuckerkandl, E., 1963. Molecular 'restoration studies' of extinct forms of life. *Acta Chem. Scan.* 17, 9–16.

Rivas, E., 2005. Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC Bioinformatics* 6(1), 63.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15, 1034–1050.

Siepel, A. and Haussler, D., 2004. Combining phylogenetic and hidden markov models in biosequence analysis. *J Comput Biology* 11(2-3), 413–28.

Thorne, J., Kishino, H., and Felsenstein, J., 1992. Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.* 34, 3–16.

Thorne, J. L., Kishino, H., and Felsenstein, J., 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol* 33, 114–124.

Yang, Z., 1996. Among-site rate variation and its impact on phylogenetic analysis. *Trends in Ecology and Evolution* 11(9), 367–372.

Tree-HMM edge-transition probabilities

$s(e) \setminus s'(e)$	C	D	I	C^*	D^*	I^*
C	$P_{Cons}(\lambda(b))$	$P_{DelStart}(\lambda(b))$	$P_{InsStart}(\lambda(b))$	1	0	0
D	$(1-\alpha_D)P_{Cons}(\lambda(b))$	α_D	$(1-\alpha_D)P_{InsStart}(\lambda(b))$	0	1	0
I	$(1-\alpha_I)P_{Cons}(\lambda(b))$	$(1-\alpha_I)P_{DelStart}(\lambda(b))$	α_I	0	0	1
C^*	$P_{Cons}(\lambda(b))$	$P_{DelStart}(\lambda(b))$	$P_{InsStart}(\lambda(b))$	1	0	0
D^*	$(1-\alpha_D)P_{Cons}(\lambda(b))$	α_D	$(1-\alpha_D)P_{InsStart}(\lambda(b))$	0	1	0
I^*	$(1-\alpha_I)P_{Cons}(\lambda(b))$	$(1-\alpha_I)P_{DelStart}(\lambda(b))$	α_I	0	0	1

Table 1. Edge transition table $\rho(s'(e)|s(e), b)$. Notice that ρ is not a transition probability matrix, since its rows sum to more than one.

Agreement between maximum likelihood and greedy solutions

Ancestor	**% of agreement**
Mou+Rat	99.8181
Hum+Chi	99.9467
Bab+Mac	99.7275
Mou+Rat+Rab	99.8181
Hum+Chi+ Bab+Mac	99.7157
Hum+Chi+Bab+Mac+Mou+Rat+Rab	99.3901
Cow+Dog	99.917
Cow+Dog+Bat	99.8218
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat	99.0511
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat+Arm	93.6531
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat+Arm+Ele	84.9413

Table 2. Percentage of alignment columns where there is agreement between the ancestor reconstructed by the greedy algorithm of Blanchette *et al.* (2004a) and that predicted by our maximum-likelihood algorithm.

Accuracy of the heuristic Viterbi algorithm

Ancestor	$t = 0$	$t = 1$	$t = 3$	$t = 5$	$t = 7$	$t > 9$
Mou+Rat	0.030	0.012	0.003	0.002	0.001	0
Hum+Chi	0.020	0.004	0.001	0.001	0.001	0
Bab+Mac	0.003	0.003	0.002	0.002	0.002	0
Mou+Rat+Rab	0.160	0.073	0.008	0.003	0.002	0
Hum+Chi+ Bab+Mac	0.060	0.041	0.011	0.002	0.002	0
Hum+Chi+Bab+Mac+Mou+Rat+Rab	0.160	0.070	0.018	0.006	0.004	0
Cow+Dog	0.070	0.032	0.006	0.002	0.001	0
Cow+Dog+Bat	0.080	0.049	0.013	0.002	0.001	0
Hum+Chi+Bab+Mac+Mou+Rat+ Rab+Cow+Dog+Bat	0.170	0.095	0.017	0.005	0.004	0
Hum+Chi+Bab+Mac+Mou+Rat+ Rab+Cow+Dog+Bat+Arm	0.100	0.048	0.010	0.003	0.002	0
Hum+Chi+Bab+Mac+Mou+Rat+ Rab+Cow+Dog+Bat+Arm+Ele	0.010	0.004	0	0	0	0

Table 3. Percentage of alignment columns where there is disagreement between the ancestor reconstructed by the exact maximum-likelihood algorithm and the heuristic with different values for the cutoff t. We emphasize that the numbers quoted are percentages, so, for example, with $t = 0$, the Mouse+Rat ancestor agrees with the optimal solution at 99.97% of the alignment columns.

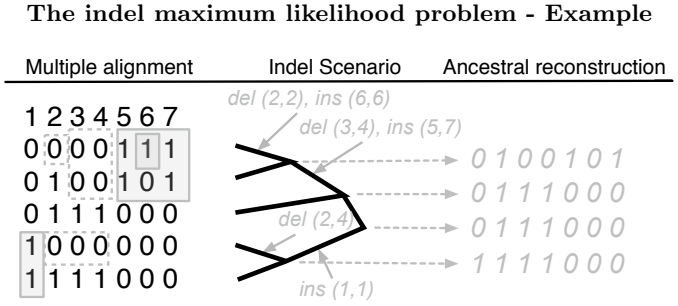

Fig. 1. Example of an input and output to the Indel Maximum Likelihood Problem. The input (in black) consists of the multiple alignment (shown on the left in binary format) and the topology and branch lengths of the phylogenetic tree. The output (in gray and italics) consists of a set of insertions and deletions, placed along the edges of the tree, explaining the gaps (zeros) in the alignment. The dashed (resp. shaded) boxes in the alignment indicate the deletions (resp. insertions) of the scenario shown on the right. This set of operations yields the ancestral reconstruction shown on the right.

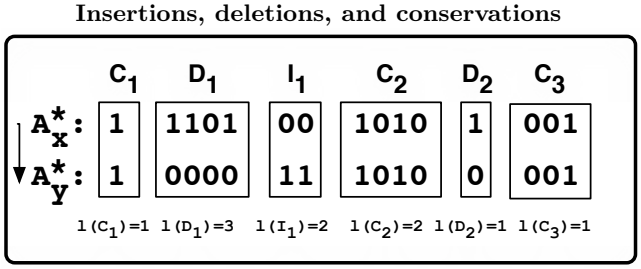

Fig. 2. Example of the partition of a paiwise alignment of A_x^* and A_y^* (where x is the parent of y) into deletions, insertions, and conservations. The length of each operation is given below it.

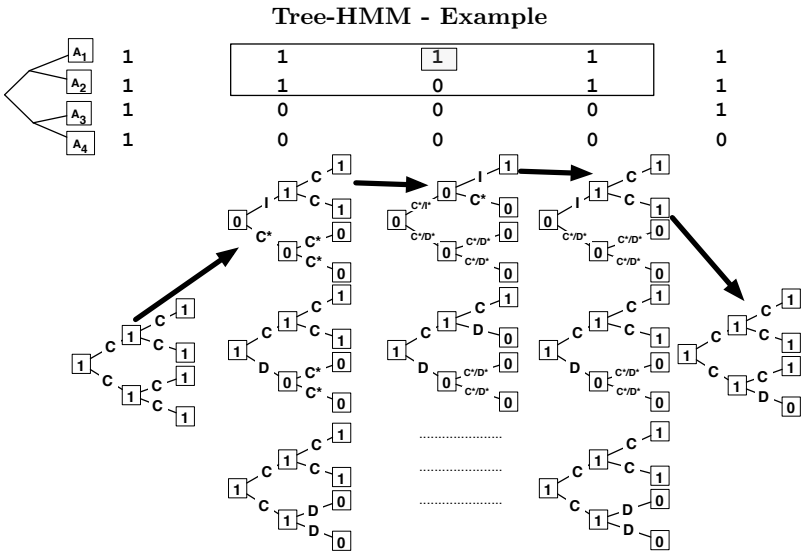

Fig. 3. The set of valid, non-zero probability states associated to the multiple alignment given at the top of the figure. When edges are labeled with more than one character (e.g. C^*, D^*), the tree represents several possible states. For the third column, not all possible states are shown. Arrows indicate one possible path through the tree-HMM. This path corresponds to two interleaved insertions, shown by two boxes in the alignment, illustrating the need for the I^* character.

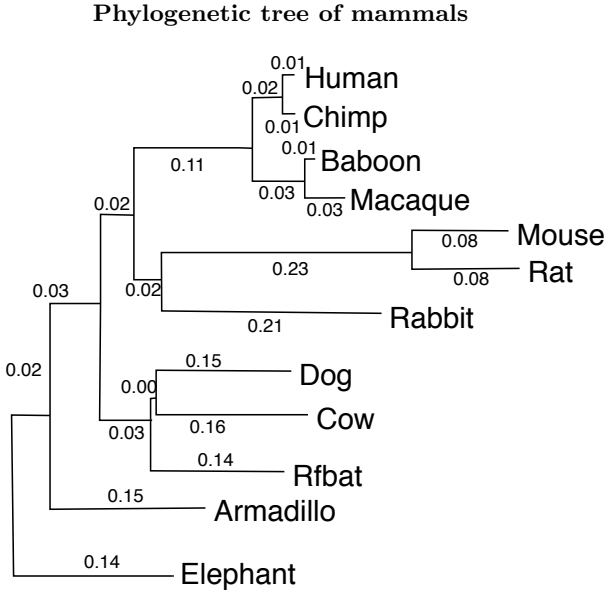

Fig. 4. Phylogenetic tree for the twelve species studied in this paper.

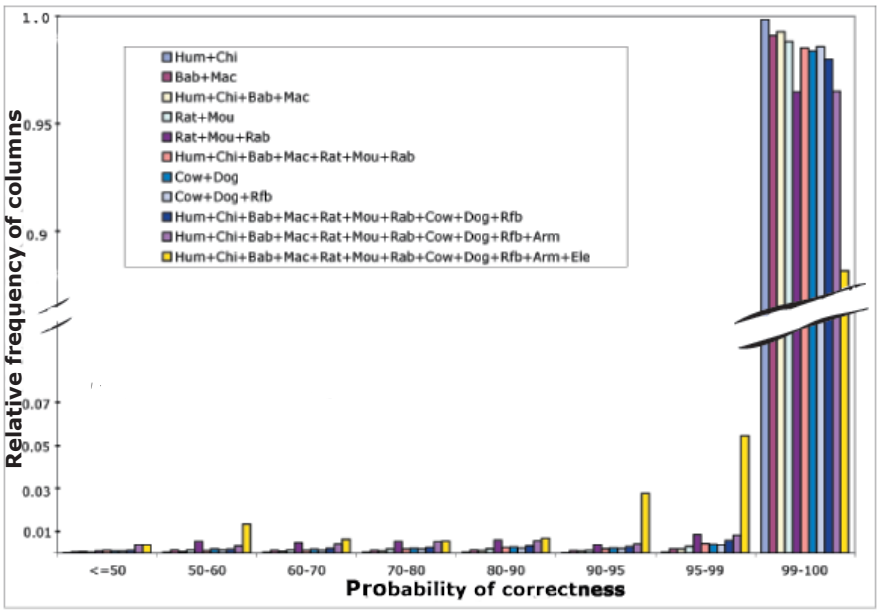

Fig. 5. Distribution of the confidence levels, over all 123,917 alignment regions, for each ancestor. The vast majority of the ancestral positions are reconstructed with a probability of correctness above 99% (assuming the correctness of the alignment).

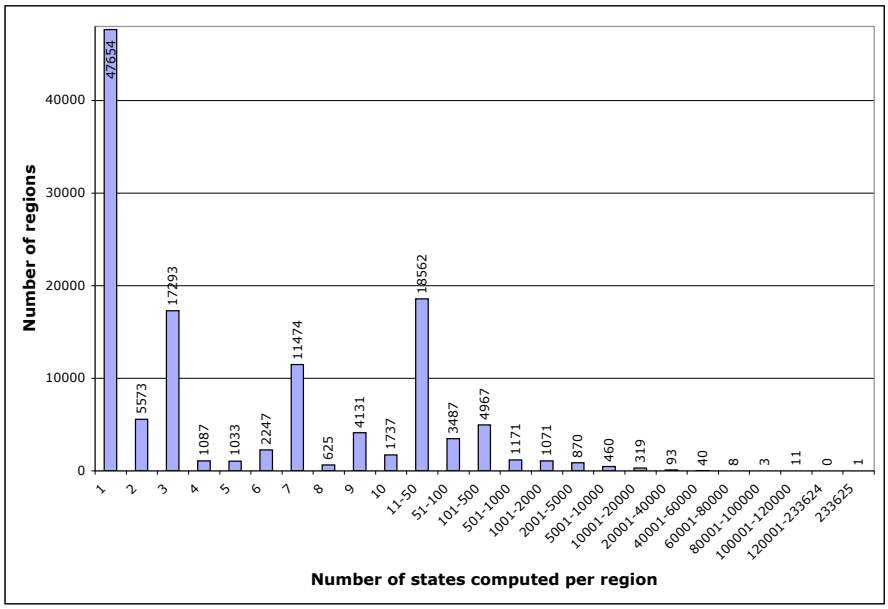

Fig. 6. Distribution of the number of states considered ($|S'_i|$), over all 123,917 regions.

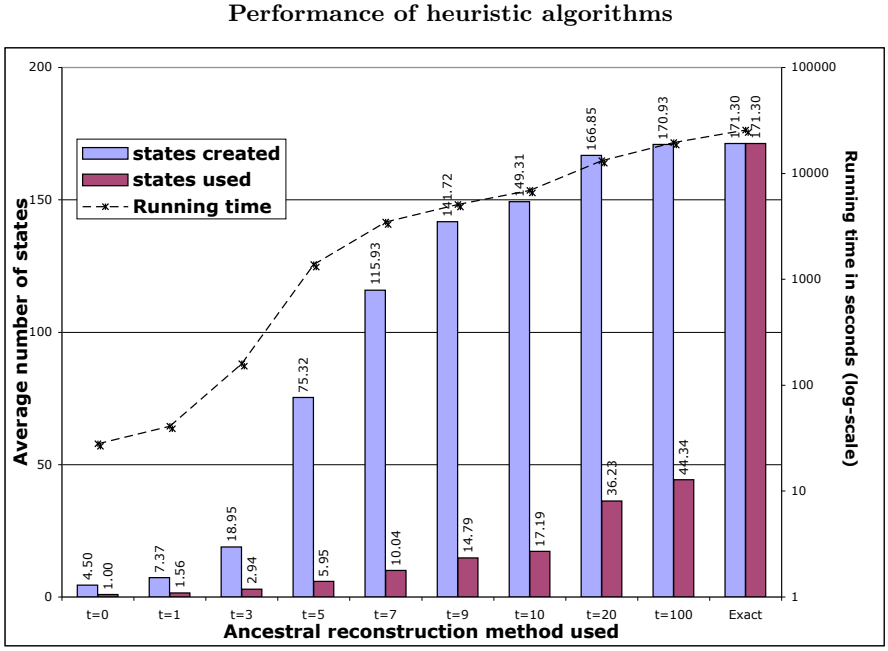

Fig. 7. Average, over all alignment regions, of the number of states created (S'_i) and used (R_i), for the different values of the cutoff t. Running times (in seconds) are plotted with the log-scale shown on the right.