11

Computational Reconstruction of Ancestral DNA Sequences

Mathieu Blanchette, Abdoulaye Baniré Diallo, Eric D. Green, Webb Miller, and David Haussler

Summary

This chapter introduces the problem of ancestral sequence reconstruction: given a set of extant orthologous DNA genomic sequences (or even whole-genomes), together with a phylogenetic tree relating these sequences, predict the DNA sequence of all ancestral species in the tree. Blanchette et al. ***(1)*** have shown that for certain sets of species (in particular, for eutherian mammals), very accurate reconstruction can be obtained. We explain the main steps involved in this process, including multiple sequence alignment, insertion and deletion inference, substitution inference, and gene arrangement inference. We also describe a simulation-based procedure to assess the accuracy of the reconstructed sequences. The whole reconstruction process is illustrated using a set of mammalian sequences from the CFTR region.

Key Words: Ancestral DNA sequence reconstruction; multiple sequences alignment; mammalian phylogeny; mammalian evolution; substitutions and indels reconstruction; ancestral sequence reconstruction accuracy.

1. Introduction

Following the completion of the human genome sequence, there is now considerable interest in obtaining a more comprehensive understanding of its evolution ***(2–4)***. Patterns of evolutionary conservation are used to screen human DNA mutations to predict those that will be deleterious to protein function and to identify noncoding sequences that are under negative selection, and hence may perform regulatory or structural functions ***(5–7)***. Long periods of conservation followed by sudden change may provide clues to the evolution of new human traits ***(8,9)***. All of these efforts depend, directly or indirectly,

From: *Methods in Molecular Biology: Phylogenomics*
Edited by: W. J. Murphy © Humana Press Inc., Totowa, NJ

on reconstructing the evolutionary history of the bases in the human genome, and hence on reconstructing the genomes of our distant ancestors.

Although some information about ancestral species has been irrevocably lost during evolution, there is still the possibility that large regions of the genomes of ancestral species with many modern descendants can be approximately inferred from the genomes of modern species using a model of molecular evolution. Indeed, it has recently been reported that in the specific case of mammalian evolution, ancestral genome reconstruction was possible to a surprising degree of accuracy ***(1)***.

The ideal target species for a genomic reconstruction is one that has generated a large number of independent, successful descendant lineages through a rapid series of early speciation events. In this case, the problem can be viewed as attempting to reconstruct an original from many independent noisy copies. In the limit of an instantaneous radiation, the accuracy of the reconstruction approaches 100% exponentially fast as the number of copies increases. From the Cretaceous period, a good choice for reconstruction would be the genome of the eutherian ancestor, as this species is believed to have spawned the relatively rapid radiation of the different lineages of modern placental mammals ***(10,11)***. This ancient species also has the added advantage of being a human ancestor, so its reconstruction, however speculative, may shed additional light on our own evolution, perhaps helping to explain features of the human and other modern mammalian genomes.

In this chapter, we describe the set of computational approaches and tools that exist for reconstructing ancestral sequences and for estimating the accuracy of such a reconstruction. This area being relatively new, there is no single tool that performs all the steps involved in the reconstruction. Instead, tools developed by different authors need to be used sequentially. The methods are illustrated on a 1.8-Mb region of mammalian genomes, containing the CFTR gene, sequenced by the ENCODE project ***(12)***.

2. Materials

2.1. Sequence Data

To reconstruct the ancestral sequences, orthologous DNA regions from as many descendants as possible need to be compared. The more orthologous sequences are available, the more accurate the reconstruction will be, provided accurate evolutionary models are used. For vertebrate sequences, a good repository of complete genome sequences is the UCSC Genome Browser (http://genome.ucsc.edu ***[13]***). Besides raw DNA sequences, multiple genome alignments and various types of genome annotation are accessible from the same site.

For the purpose of this chapter, we illustrate the process of ancestral sequence reconstruction using a 1.8-Mb region of the human genome including the *CFTR* gene, together with orthologous regions from 19 other mammals

(available from the UCSC Genome Browser). This deep coverage is not currently available over all the genome, but only for the targeted sequencing of the ENCODE project ***(12)***.

2.2. Phylogenetic Information

An important component of ancestral sequence reconstruction is the knowledge of the phylogenetic relationships among the species being compared. Knowing the correct tree topology and estimating the length of its branches are crucial for an accurate reconstruction, as well as for estimating the accuracy of that reconstruction through simulations. Accepted phylogenetic trees are now available for many sets of species (*see,* e.g., **refs.** ***10,14***). For others, the exact phylogenetic relationships remain unclear and need to be inferred prior to reconstruction, using programs like Phylip ***(15)***, PAUP ***(16)***, or MrBayes ***(17)***. These tools are also necessary to estimate the branch lengths of the phylogenetic tree using a maximum likelihood approach.

2.3. Sequence Annotation

In some cases, functional annotation of extant sequences can be used to obtain more accurate reconstruction of ancestral sequences. This is particularly the case for coding region annotation and repetitive region annotation. For metazoans, a good source of such annotations is the UCSC genome browser and the Ensembl Genome Browser (http://www.ensembl.org).

3. Methods

This section introduces the techniques that have been developed for predicting ancestral DNA sequences based on their extant descendants, and for estimating the accuracy of the reconstruction. We illustrate this reconstruction process (*see also* **Note 1**) and the type of information that can be derived from it using a 1.8-Mb region surrounding the *CFTR* gene in mammals (*see* **ref.** ***1*** and **Note 2** for more details).

3.1. Predicting Ancestral Sequences

The prediction of ancestral genomes can be divided into four main steps. A crucial first step toward the reconstruction is to build an accurate multiple alignments of the extant orthologous sequences, thus establishing orthology relationships among the nucleotides of each sequence. Second, the process of indel reconstruction determines the most likely scenario of insertions and deletions that may have led to the extant sequences. Third, substitution history is reconstructed using a maximum likelihood approach. The last step involves dealing with genome rearrangements (inversions, transpositions, translocations, duplications, and chromosome fusions, fissions, and duplications).

3.1.1. Multiple Sequence Alignment

Given a set of orthologous sequences, the multiple alignment problem consists of identifying (by aligning them together) the sets of nucleotides derived from a common ancestor through direct inheritance or through substitution. Many approaches have been developed to align multiple large genomic regions. Some of the most popular approaches include programs like MAVID ***(18)***, MLAGAN ***(5,19)***, and TBA ***(20)***. All these approaches fall under the category of progressive alignment methods and require the prior knowledge of the topology of the phylogenetic tree that relates the extant sequences compared (*see* **Subheading 2.2.**). The threaded blocks aligner (TBA) program, based on the well established pair-wise alignment program BLASTZ ***(21)***, has been shown to be particularly accurate for aligning mammalian sequences and is thus a tool of choice for ancestral reconstruction for these species. The program is available at http://www.bx.psu.edu/miller_lab/. The multiple sequence alignment problem is discussed in more detail in Chapter 9.

3.1.2. Indel Reconstructing

Given a multiple sequence alignment of the repeat-soft-masked extant sequences and a phylogenetic tree with known topology and branch lengths, the next step consists of predicting, for each ancestral node in the tree, which columns of the alignment correspond to ancestral bases and which correspond to nucleotides inserted after the ancestor. Although the problem of parsimonious indel inference has recently been shown to be NP-Hard ***(22)***, good heuristics have been developed by Fredslund et al. ***(23)***, Blanchette et al. ***(1)***, and Chindelevitch et al. ***(22)***. Currently, the only publicly available program for indel reconstruction is the inferAncestors program based on the greedy approach of Blanchette et al. ***(1)***. This section describes briefly how the program works.

Given a multiple alignment, all the gaps in the alignment are first marked as unexplained. The algorithm iteratively selects the insertion or deletion, performed along a specific edge of the tree and spanning one or more columns of the alignment, which yields the largest number of alignment gaps explained per unit of cost. The number of gaps explained by a deletion is the number of unexplained gaps in the subtree above which the deletion occurs. The number of gaps explained by an insertion is the number of unexplained gaps in the complement of the subtree above which the insertion occurs. The costs can be defined heuristically. The cost of a deletion is given by $1 + 0.01 \log(L) - 0.01b$, where L is the length of the deletion and b is the length of the branch along which the event takes place. The cost of an insertion is given by $1 + 0.01 \log(L) - 0.01b - r$, where r is a term that takes value 0.5 if the repetitive content of the

segment inserted is more than 90%. Once the best insertion or deletion has been identified, its gaps are marked as "explained." This does not preclude them from being part of other indels, but they will not count in their evaluation. Finally, heuristics are used to reduce errors related to incorrect alignment, in particular to reduce the problems caused by two repetitive regions from two distantly related species mistakenly aligned to each other, with other species having gaps in that region.

3.1.3. Substitutions Reconstruction

After having established which positions of the multiple alignment correspond to bases in the ancestor, the inferAncestors program predicts which nucleotide (A, C, G, or T) was present at each position in the ancestor using the standard posterior probability approach ***(24)*** based on a dinucleotide substitution model in which substitutions at two adjacent positions are independent except for CpG, whose substitution rate to TpG is 10 times higher than those of other transitions ***(25)***. This phase of the reconstruction relies on the availability accurate branch-length estimates for the phylogenetic tree, which can be obtained as described under **Subheading 2.2**.

3.1.4. The inferAncestors Program

The inferAncestor program, available from http://www.mcb.mcgill.ca/~blanchem/software, integrates the steps of indel and substitution inference. The algorithm takes as input a multiple alignment in fasta format, together with a phylogenetic tree in New Hampshire format. The program outputs a predicted ancestral sequence for each internal node of the phylogenetic tree. Two other files are outputs, describing the confidence of the prediction made for each base of each ancestral sequence. The first describes the confidence in the prediction of presence or absence of a base at each position of each ancestral sequence. The second describes the confidence of the actual nucleotide (A, C, G, or T) predicted. The inferAncestor program is written in C++ and has been tested on Linux and Mac OS X.

3.1.5. Genome Rearrangements

To complete the inference of ancestral genomes, the ancestral DNA sequences inferred for each block of orthologous sequences need to be ordered into a single, contiguous genome. This problem is made challenging by the presence of genome rearrangements (inversions, transpositions, translocations, and duplications/losses). One of the most popular computer programs for inferring ancestral gene arrangement is MGR (***[26]***, http://www.cse.ucsd.edu/groups/bioinformatics/MGR), which is described in detail in Chapter 10.

3.2. Assessing Reconstruction Accuracy Through Simulations

This section describes a simulation-based method for assessing the accuracy of the reconstructed ancestor. An alternate approach based on retrotransposons is described in ***(1)***.

To assess the reconstructability of ancestral genomic sequences from their extant descendants, the simplest method is to use simulations of sequence evolution. Starting from a known (but synthetic) ancestral sequence, we let the sequence evolve along the branches of the tree until the leaves are reached. The ancestral sequence reconstruction procedure is then applied to the set of simulated leaves, and the prediction made is compared to the known ancestral sequence.

The simulation program Simali (http://www.bx.psu.edu/miller_lab/), based on the Rose program ***(27)***, can be used to mimic the evolution of sequences under no selective pressure. Given a phylogenetic tree, the program simulates sequence evolution by performing random substitutions, deletions, and insertions along each branch, in proportion to its length. The program allows for the insertion of retrotransposons, which is an important source of error in sequence alignment, and thus in ancestral sequence reconstruction.

To assess the reconstructability of ancestral mammalian genomic sequences, Blanchette et al. ***(1)*** performed a series of computational simulations of the neutral evolution of a hypothetical ~50 kb ancestral genomic region into orthologous regions in 20 modern mammals (**Fig. 1**). The simulations are based on the phylogenetic tree inferred by Eizirik et al. ***(10)*** on a set of genes for a large set of mammals. Substitutions follow a context-independent HKY model ***(28)*** with Ts/Tv = 2, $p(a) = p(t) = 0.3$, and $p(c) = p(g) = 0.2$, except that substitution rates of CpG pairs are 10 times higher than other rates ***(25)***. Deletions are initiated at a rate of about 0.056 times the substitution rate, their length is chosen according to a previously reported empirical distribution ***(29)*** that ranges between 1 and 5000 nucleotides, and their starting point is uniformly distributed. Insertions occur randomly according to a mixture model. Small insertions (of size between 1 and 20 nt) occur at half the rate of deletions, their size distribution is empirically determined ***(29)*** and their content is a random sequence for which each nucleotide

Fig. 1. Estimated reconstructability of ancestral mammalian sequences. Average base-by-base error rate in the reconstruction of each simulated ancestral sequence. The error rate shown is the sum of the percentages of bases that are missing, added, or mismatched as a result of errors in the reconstruction, averaged over one hundred simulations of sets of orthologous sequences of length approximately 50 kb. Error rates are given first for all regions, and in parentheses for nonrepetitive regions only. The species names at the leaves only indicate what organisms we simulated; no actual biological sequences were used here. The tree topology and branch lengths are derived directly from Eizirik et al. ***(10)***.

(*Continued on next page*)

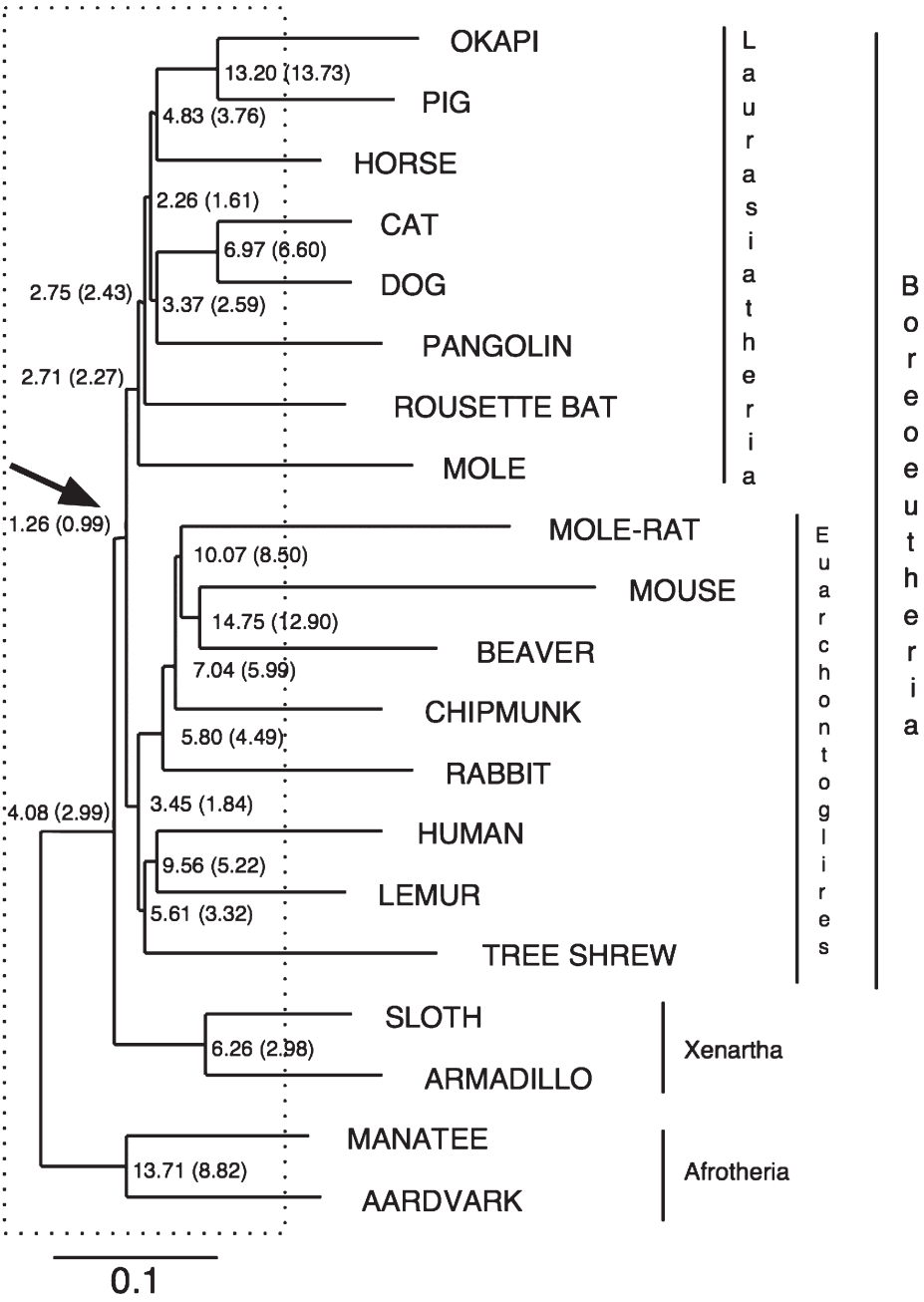
OKAPI
13.20 (13.73)
PIG
4.83 (3.76)
HORSE
2.26 (1.61)
CAT
6.97 (6.60)
DOG
2.75 (2.43)
3.37 (2.59)
PANGOLIN
2.71 (2.27)
ROUSETTE BAT
MOLE
Laurasiatheria
1.26 (0.99)
MOLE-RAT
10.07 (8.50)
MOUSE
14.75 (12.90)
BEAVER
7.04 (5.99)
CHIPMUNK
5.80 (4.49)
RABBIT
4.08 (2.99)
3.45 (1.84)
HUMAN
9.56 (5.22)
LEMUR
5.61 (3.32)
TREE SHREW
Euarchontoglires
Boreoeutheria
SLOTH
6.26 (2.98)
ARMADILLO
Xenartha
MANATEE
13.71 (8.82)
AARDVARK
Afrotheria
0.1

Fig. 1.

is chosen independently from the background distribution. They also simulate the insertion of retrotransposons. For this they used a library of 15 different types of transposable elements chosen to cover the large majority of repetitive elements observed in well studied mammals ***(30)***. The rate of insertion of each repeat varies from branch to branch, so that certain retrotransposons (such as ALUs, SINEs B2, and BOV) are lineage-specific, whereas others (L1, LTR, and DNA) are both present in the sequence at the root of the tree (with a range of decaying level) and can be inserted along any branch. The code and parameters used for our simulations are available with the Simali package.

After generating a set of simulated sequences, the sequences are first soft-repeat-masked using RepeatMasker ***(31)*** and then aligned using one of the methods under **Subheading 3.1.1**. The repeat-masked multiple alignment is then fed into the inferAncestors program, which produces a predictiozn of the ancestral sequence at each internal node of the phylogenetic tree. To compare the actual ancestral sequence generated by simulations to the predicted ancestral sequence, we align them and count the number of missing bases (those present in the actual ancestor but not in the reconstruction), added bases (present in the reconstruction but not in the actual ancestor), and mismatch errors (positions in the reconstruction assigned the incorrect nucleotide). The sum of the rates of all three types of errors, calculated separately at each ancestral node in the phylogenetic tree, is used to estimate the reconstructability of a given ancestor.

In the case of mammalian sequences, Blanchette et al. ***(1)*** used the above simulation-based procedure to show that the sequence of certain mammalian ancestors can be reconstructed with remarkable accuracy. **Figure 1** shows that under this phylogenetic tree with a relatively rapid placental mammalian radiation, the neutral nonrepetitive regions of the Boreoeutherian ancestral genome that have evolved under their simple model should be reconstructable with about 99% base-by-base accuracy from the genomes of 20 present-day mammals. Repetitive regions are not reconstructed as accurately because they are more often involved in misalignments, which can result in incorrect predictions. Nonetheless, even counting errors in repetitive regions, the total accuracy is more than 98%. The simulations suggest that even in the nonrepetitive regions, much of the difficulty of the reconstruction problem lies in the computation of the multiple alignment, as a reconstruction based on the correct multiple alignment derived from the simulation itself (and thus unavailable for actual sequences) had less than half the number of reconstruction errors. Examining reconstructions made using smaller subsets of this set of 20 species, it was found that, including repetitive regions, an accuracy of about 97% can be achieved using only 10 species chosen to sample most major mammalian lineages (**Fig. 2**). Sampling only five of the most slowly evolving lineages yields an accuracy of about 94%. Little is gained with our current reconstruction procedures by adding more than 10 species

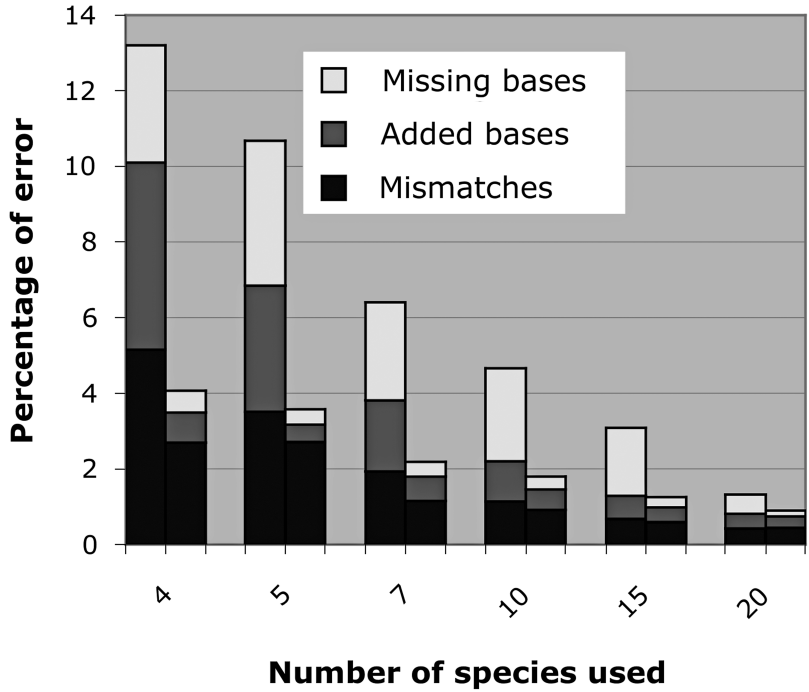

Fig. 2. Estimated reconstructability of the Boreoeutherian ancestor. Fraction of the simulated Boreoeutherian ancestral sequence reconstructed incorrectly as a function of the number of extant species used for the reconstruction. For each number of species used, results are given counting all bases (left columns) and only nonrepetitive bases (right columns). Species are added in the following order: human, cat, chipmunk, sloth, manatee, rousette bat, mole, pig, beaver, tree shrew, horse, pangolin, mouse, armadillo, aardvark, okapi, dog, mole-rat, rabbit, and lemur.

because the risk of misalignment increases, whereas the unavoidable loss of information in the early branches persists.

An alternate approach to assessing the accuracy of a reconstruction is through a pseudo cross-validation procedure. Instead of reconstructing an ancestral sequence based on all the extant sequences available, do so using a (large) subset of these species. Different subsets of species will produce slightly different ancestral reconstructions, and the variability between these reconstructions will give an idea of the expected error rate of the reconstruction that is based on all species.

3.3. Reconstruction of Actual Mammalian Sequences

Blanchette et al. ***(1)*** applied the reconstruction method described above to actual high-quality sequence data from a region containing the human CFTR

locus, using 18 additional orthologous mammalian genomic regions generated by the NISC Comparative Sequencing Program (***[12]***, www.nisc.nih.gov). Simulations on synthetic data like those described above indicate that for the topology and set of branch lengths for these 19 species, the ancestral sequence that can be the most accurately reconstructed based on the sequences available is the Boreoeutherian ancestor, and that neutrally evolving regions of this ancestral genome can be reconstructed with an accuracy of about 96%. On a site-specific basis, simulations suggest that more than 90% of the bases of the predicted ancestor can be assigned confidence values greater than 99%. The reconstructed ancestor and site-specific confidence estimates are available at http://genome.ucsc.edu/ancestors.

Figure 3 illustrates the reconstruction in a noncoding region of the *CFTR* locus that exhibits a typical level of sequence conservation. This region is located in a 32-kb intron of the *CAV1* gene, about 13 kb from the 5′-exon. The bases in this region are relics left over from the insertion of a MER20 transposon sometime prior to the mammalian radiation and are thus unlikely to be under selective pressure.

Notice that despite the fact that the alignment of certain species (in particular, mouse, rat, and hedgehog) appears somewhat unreliable, the inference of the presence or absence of a Boreoeutherian ancestral base at a given position is quite straightforward given the alignment, and so is, to a lesser extent, the prediction of the actual ancestral base itself. The MER20 consensus is shown for comparison. Most positions in which the reconstructed Boreoeutherian ancestral base disagrees with the MER20 consensus are likely owing to substitutions in this MER20 relic that predated the Boreoeutherian ancestor, since the support of the reconstructed base is very strong in the extant species. If the MER20 consensus sequence is used as an outgroup in the reconstruction procedure, only two bases (indicated by a longer arrow) are reconstructed differently, indicating that the reconstructed ancestral sequence is very stable and most of it is likely to be correct.

4. Notes

1. The accuracy of the reconstruction depends crucially on the length of the early branches of the phylogenetic tree. In the context of the ancestral mammalian sequence reconstruction, Blanchette et al. ***(1)*** have shown that if the major placental lineages had diverged instantaneously, they would be able to reconstruct the simulated Boreoeutherian ancestral sequence, including repetitive regions, with less than 1% error. In contrast, if the early branch lengths inferred by Eirizik et al. ***(10)*** turned out to underestimate the actual lengths by a factor of two, the error rate would jump to 3%, and to 6% if they were underestimated by a factor of 4.
2. One of the nonintuitive results presented by Blanchette et al. ***(1)*** is the observation that more ancient ancestral genomes can often be reconstructed more accurately than their more recent descendants. Why exactly is this so? For simplicity, consider

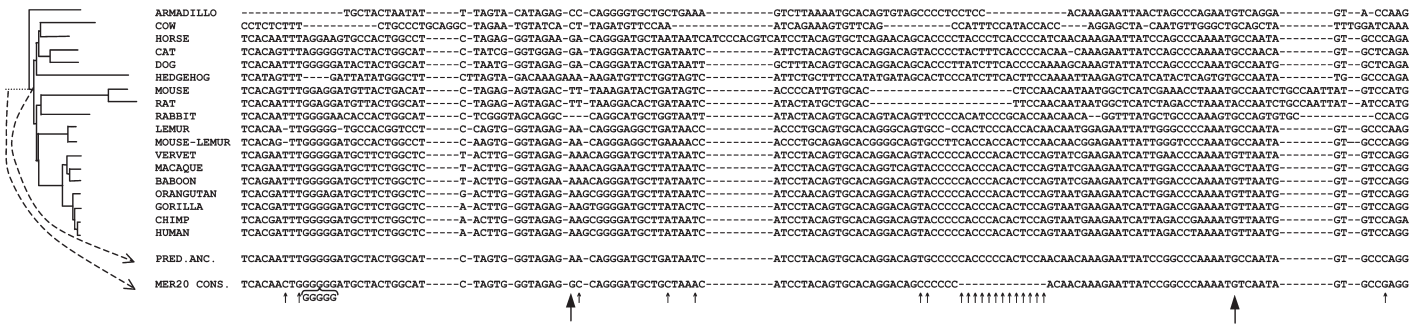

Fig. 3. Example of reconstruction of an ancestral Boreoeutherian sequence based on actual orthologous sequences derived from a MER20 retrotransposon. Arrows indicate positions where the reconstructed ancestor differs from the MER20 consensus. Longer arrows indicate the positions in which the knowledge of the MER20 consensus sequence would have changed the ancestral base prediction. The position of the human sequence displayed is chr7:115,739,755-115,739,899 (NCBI build 34). The alignment of the flanking nonrepetitive DNA (not shown) verifies that the sequences from the different species are in fact orthologous. The tree and branches are derived directly from **ref. *10***.

the case of reconstructing a single binary ancestral character state in the root species (e.g., purine vs pyrimidine at a given site) under a simple model in which the prior probability distribution on the ancestral character is uniform, substitution rates are known, symmetric, homogeneous, and not too high, and the total branch length in the phylogenetic tree from the root ancestor to each of the modern species is the same (i.e., assume a molecular clock). Here each of n modern species has a state that differs from the ancestral one with the same probability $p < 1/2$. If the tree exhibits a star topology, in which each of the modern species derives directly from the ancestor on an independent branch, then it is clear that the maximum likelihood and Bayesian maximum *a posteriori* reconstructions of the ancestral character agree, and the reconstructed state is the one that is most often observed in the n modern species. The probability of an error in reconstruction is:

$$\sum_{k=\lceil n/2 \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

which is at most $\left[4p(1-p)\right]^{n/2}$ (***[32,33]***; Lemma 5, p. 479). This error approaches zero exponentially fast as n increases. The star topology has a kind of "phase transition" where the ancestor becomes highly reconstructible once enough present day sequences are available to compensate for the length of the branches leading back to the ancestor.

In contrast, a nonstar topology such as a binary tree that has the same total root-to-leaf branch length and the same number n of modern species at the leaves has two nonzero length branches from the root ancestor R leading to intermediate ancestors A and B, and information is irrevocably lost along these two branches. No matter how large the number n of modern descendant species derived from A and B, one can do no better at reconstructing the state at R than if one knew for certain the state in its immediate descendants A and B. Even with this knowledge, the accuracy of reconstruction of R from A and B will be strictly less than 100% for all reasonable models and nonzero branch lengths. The reconstruction gets poorer the longer the branch lengths are to A and B. This extends to the case where the ancestor R being reconstructed has a bounded number of independent immediate descendants and to the case where descendants of an earlier ancestor of R (outgroups) are also available. The long branches connecting them to the rest of the tree are why some more recent ancestral sequences in the tree of **Fig. 1** are less reconstructible than the Boreoeutherian ancestor, which acts almost like the root of a star topology (*see* **ref. *34*** for a discussion of optimal tree topologies for ancestral reconstructability).

Acknowledgments

We thank Jim Kent, Arian Smit, Adam Siepel, Gill Bejerano, Elliot Margulies, Brian Lucena, Leonid Chindelevitch, and Ron Davis for helpful discussions and suggestions. A.B.D. was supported by a NSERC postgraduate scholarship. W.M. was supported by grant HG-02238 from the National Human Genome

Research Institute, E.G. was supported by NHGRI, D.H. and M.B. were supported by NHGRI Grant 1P41HG02371 and the Howard Hughes Medical Institute. Finally, we thank the NISC Comparative Sequencing Program for providing multispecies comparative sequence data.

References

1. Blanchette, M., Green, E. D., Webb, M., and Haussler, D. (2004) Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*. **14,** 2412–2423.
2. International Human Genome Sequencing Consortium, Lander, E., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **5, 409**(6822), 860–921 (PMID: 12466850).
3. International Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **5, 420**(6915), 520–562 (PMID: 12466850).
4. Rat Genome Sequencing Project Consortium, Gibbs, R. A., Weinstock, G. M., Metzker, M. L., et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521.
5. Margulies, E. H., Blanchette, M., NISC Comparative Sequencing Program, Haussler, D., and Green, E. (2003) Identification and characterization of multi-species conserved sequences. *Genome Res*. **13**(12)**,** 2507–2518 (PMID: 14656959).
6. Cooper, G. M., Brudno, M., Green, E. D., Batzoglou, S., and Sidow, A. (2003) Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res*. **13**(5)**,** 813–820.
7. Bejerano, G., Pheasant, M., Makunin, I., et al. (2004) Ultraconserved elements in the human genome. *Science* **304**(5675)**,** 1321–1325.
8. Goodman, M., Barnabas, J., Matsuda, G., and Moore, G. W. (1971) Molecular evolution in the descent of man. *Nature* **233,** 604–613.
9. Enard, W., Przeworski, M., Fisher, S. E., et al. (2002) Molecular evolution of FOXP2, a gene involved in speech and language. *Nature* **418**(6900)**,** 869–872.
10. Eizirik, E., Murphy, W. J., and O'Brien, S. J. (2001) Molecular dating and biogeography of the early placental mammal radiation. *J. Hered*. **92**(2)**,** 212–219 (PMID: 11396581).
11. Springer, M. S., Murphy, W. J., Eizirik, E., and O'Brien, S. J. (2003). Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proc. Natl Acad. Sci. U S A* **4, 100**(3)**,** 1056–1060 (PMID: 12552136).
12. Thomas, J., Touchman, J. W., Blakesley, R. W., et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424,** 788–793.
13. Karolchick, D., Baertsch, R., Diekhans, M., et al. (2003) The UCSC genome browser database. *Nucleic Acids Res*. **31,** 51–54.
14. Maddison, D. R. and Schulz K.-S. (ed.) (2004) *The Tree of Life Web Project*. http://tolweb.org
15. Felsenstein, J. (1989) PHYLIP—Phylogeny inference package (Version 3.2). *Cladistics* **5,** 164–166.

16. Swofford, D. L. (2003) *PAUP: Phylogenetic Analysis Using Parsimony*. Sinauer, Sunderland, MA.
17. Huelsenbeck, J. P. and Ronquist, F. (2001) MrBayes: Bayesian inference of phylogeny. *Bioinformatics* **17,** 754–755.
18. Bray, N. and Pachter, L. (2004) MAVID: constrained ancestral alignment of multiple sequences. *Genome Res*. **14,** 693–699.
19. Cooper, G. M., Stone, E. A., Asimenos, G., et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. **15**(7)**,** 901–913.
20. Blanchette, M., Kent, W. J., Riemer, C., et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*. **14**(4)**,** 708–715 (PMID: 15060014).
21. Schwartz, S., Kent, W. J., Smith, A., et al. (2003) Human–mouse alignments with BLASTZ. *Genome Res*. **13**(1)**,** 103–107.
22. Chindelevitch, L., Li, Z., Blais, E., and Blanchette, M. (2006) On the inference of parsimonious indel evolutionary scenarios. *J. Bioinformatics Comput. Biol*. in press.
23. Fredslund, J., Hein, J., and Scharling, T. (2003) A large version of the small parsimony problem. *Lecture Notes in Bioinformatics, Proceedings of WABI'03*. **2812,** 417–432.
24. Yang, Z., Kumar, S., and Nei, M. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141,** 1641–1650.
25. Siepel, A. and Haussler, D. (2003) Combining phylogenetic and hidden Markov models in biosequence analysis. *Proceedings of the 7th Annual International Conference on Research in Computational Molecular Biology*. pp. 277–286.
26. Bourque, G. and Pevzner, P. (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res*. **12**(1)**,** 26–36.
27. Stoye, J., Evers, D., and Meyer, F. (1997) Generating benchmarks for multiple sequence alignments and phylogenetic reconstructions. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. **5,** 303–204 (PMID: 9322053).
28. Hasegawa, M., Kishino, H., and Yano, T. (1985) Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol*. **22**(2)**,** 160–174.
29. Kent, 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. USA* **100**(20)**,** 11, 848–11,489.
30. Jurka, J. (2002) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*. **16**(9)**,** 418–420 (PMID: 10973072).
31. Smit, A. and Green, P. (1999) RepeatMasker, http://ftp.genome.washington.edu/RM/RepeatMasker.html
32. Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc*. **58,** 13–27.
33. Le Cam, L. (1986) *Asymptotic Methods in Statistical Decision Theory*, Springer, New York.
34. Lucena, B. and Haussler, D. (2005) Counterexample to a claim about the reconstruction of an ancestral character states. *Syst Biol*. **54**(4)**,** 693–695.