The Pennsylvania State University

The Graduate School

College of Engineering

RECONSTRUCTING CONTIGUOUS REGIONS OF AN ANCESTRAL GENOME

A Thesis in

Computer Science and Engineering

by

Jian Ma

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2006

The thesis of Jian Ma was read and approved* by the following:

Webb Miller
Professor of Biology and Computer Science and Engineering
Thesis Adviser
Chair of Committee

Hongyuan Zha
Professor of Computer Science and Engineering

Piotr Berman
Associate Professor of Computer Science and Engineering

Wojciech Makalowski
Associate Professor of Biology

Ross Hardison
T. Ming Chu Professor of Biochemistry and Molecular Biology

Raj Acharya
Professor of Computer Science and Engineering
Head of the Department of Computer Science and Engineering

*Signatures on file in the Graduate School.

Abstract

The increasing number of mammalian genome sequences becoming available provides scientists with dramatic opportunities to better understand human evolution. Comparative genome analysis enables us to computationally reconstruct an ancestral mammalian genome by comparing the genomes of living descendants.

Here, we introduce new methods and tools for identifying large-scale rearrangements and reconstructing contiguous ancestral regions. We discuss several critical problems, including the identification of conserved segments that have not been rearranged through evolution of the studied species, the prediction of ancestral order of these conserved segments, and the computational evaluation of the reconstruction.

Using these methods, several analyses have been carried out. In particular, we analyze mammalian genome rearrangements at higher resolution than has been published to date. In the lineages leading to human, mouse, rat and dog from their most recent common ancestor, we identify 1338 conserved intervals over 50Kb in length. Using our algorithm for predicting the ancestral order and orientation of those intervals from their observed adjacencies in modern species, we produce a karyotype map of an early mammalian genome that accounts for 96.8% of the available human genome sequence data. The precision is further increased by mapping inversions as small as 31 bp. We also discuss the biological insights gained from these analyses. Although only a few mammalian genomes are currently sequenced to high precision, our evaluation indicates

that our results are reasonably accurate, and that they will become highly accurate in the foreseeable future.

Table of Contents

List of Tables

List of Figures

Preface

This dissertation is based largely on the paper "Reconstructing contiguous regions of an ancestral genome" published on *Genome Research* (Ma *et al.*, 2006). I co-authored this paper with Louxin Zhang, Bernard B. Suh, Brian J. Raney, Richard C. Burhans, W. James Kent, Mathieu Blanchette, David Haussler, and Webb Miller. Some follow-up studies are also discussed in the dissertation.

The problem of ancestral mammalian karyotype reconstruction is definitely not new. It has been studied for at least three decades. Biologists are trying to solve the jigsaw puzzle through cytogenetic experiments with the so-called chromosome-painting technique. On the other hand, computer scientists are also deeply fascinated by this topic. A fruitful literature can be found. However, a recent Forum in *Genome Research* highlighted the contrast between the reconstructions using cytogenetic and bioinformatic approaches (Froenicke *et al.*, 2006; Bourque *et al.*, 2006).

We developed a set of new computer programs that take advantage of comparative genomics and predicted the karyotype of the Boreoeutherian ancestor (i.e., the last common ancestor of humans, rodents, carnivores, and most other living mammals). Our results show a significant convergence with chromosome painting, though there are still uncertainties that need to be resolved. This work has already drawn attention from the cytogenetic community (Rocchi *et al.*, 2006).

These methods were developed as part of a project, led by David Haussler and Webb Miller, to reconstruct the genome sequence of the last ancestor of Boreoeutherian

ancestor. Some of the methods were further improved during our participation in the rhesus macaque genome analysis consortium.

The dissertation consists of seven chapters. Chapter 1 introduces the biological background of ancestral genome reconstruction and reviews the theoretical work related to genome rearrangement. Chapter 2 discusses the algorithm for reconstructing the ancestral order of conserved segments using adjacency information from living species. Chapter 3 explains the method of constructing the conserved segments that are used as building blocks for the reconstruction algorithm. Chapter 4 presents the results after applying the algorithm to reconstruct the Boreoeutherian ancestral karyotype and gives a detailed comparison with other predictions. Chapter 5 introduces a probabilistic model for reconstructing ancestral adjacencies which is potentially more extensible. Chapter 6 contains some additional work developed during the analysis of rhesus macaque genome. Finally, in Chapter 7, we summarize the dissertation and discuss future directions of ancestral genome reconstruction.

Acknowledgments

What I've been able to accomplish is the direct result of the great people I was fortunate enough to learn from and work with during the past years. Without their help and support, it wouldn't be possible for me to arrive at this stage in my life.

First, I would like to thank my advisor Webb Miller. Webb has helped me and guided me since my first day as a Ph.D. student in the United States. He has led me to this exciting area, introduced me to challenging problems, taught me how to do research, and got me involved in the wider research community. He encouraged me to work independently, but at different stages of my study he always gave me concrete and constructive suggestions. Normally when I went to the lab on weekends, he was always there. Many times when I sent him email at 2 o'clock in the morning, I received his reply instantaneously. Nothing would motivate me more than seeing the advisor was so enthusiastic and hard-working. I am and will always be grateful for his continuous substantial support. His great ability to balance computer science and biology will be a constant source of inspiration to me in my future career.

I would like to thank David Haussler at University of California Santa Cruz, who has given me tremendously insightful ideas. I am also thankful to be able to work with these talented collaborators: Mathieu Blanchette, Louxin Zhang, Jim Kent, Brian Raney, and Bernard Suh. A number of other faculty at Penn State have offered me generous help: Francesca Chiaromonte, Ross Hardison, Piotr Berman, Hongyuan Zha, and Wojciech Makalowski. In addition, I enjoyed many discussions with colleagues in

the group: Bob Harris, John Karro, Minmei Hou, Svitlana Tyekucheva, and David King. I am particularly indebted to Richard Burhans, who carefully went through much of my code.

I cannot appreciate enough the warm support from my friends: my roommates Guilin Chen and Xing Gao; my undergrad classmates Zhifeng Chen, Ming Chen, Li Wei, Qingqing Yuan, Mingxi Wu; my close friends Jing Fu, Jing Zhou, Xilin Zhang.

Finally, I would like to express my deepest heartfelt thanks to my family. I thank my mom Shouzhen Wang and my dad Gang Ma for their unconditional and unreserved love. Without them, none of this would be possible. I thank my wife Hong Shen, who made this stressful process bearable and worthwhile. We always have a good time together, even when we're not together. I dedicate this thesis to them.

To my mother Shouzhen Wang, my father Gang Ma, and my wife Hong Shen.

Chapter 1

Introduction

In this chapter, we first introduce the biological background of chromosome evolution. Then we briefly survey genome rearrangements from a computer science perspective. After reviewing the history of ancestral genome reconstruction, we discuss the motivation of this work.

1.1 Chromosome evolution

A *chromosome* is a threadlike macromolecule of genes and other DNA in the nucleus of a cell. In eukaryotic (as opposed to bacterial) cells, chromosomes are in a linear form. Each chromosome has two arms; the shorter one is called the p arm, while and the longer one is the q arm. *Chromatid* refers to one of the two identical parts of the chromosome after the synthesis phase. Two chromatids are attached at an area called the *centromere*. The *telomere* is the region that appears at either end of a linear chromosome. See Figure 1.1 for detailed illustration.

Different kinds of organisms have different numbers of chromosomes. For example, humans have 23 pairs of chromosomes, 46 in all, dogs have 39 pairs, and mice have 20 pairs. A graphic representation of all the chromosomes in a cell of any species is called a *karyotype*. Karyotype diversity among different species is caused by chromosome rearrangements. Dobzhansky and Sturtevant (1938) reported the observation of

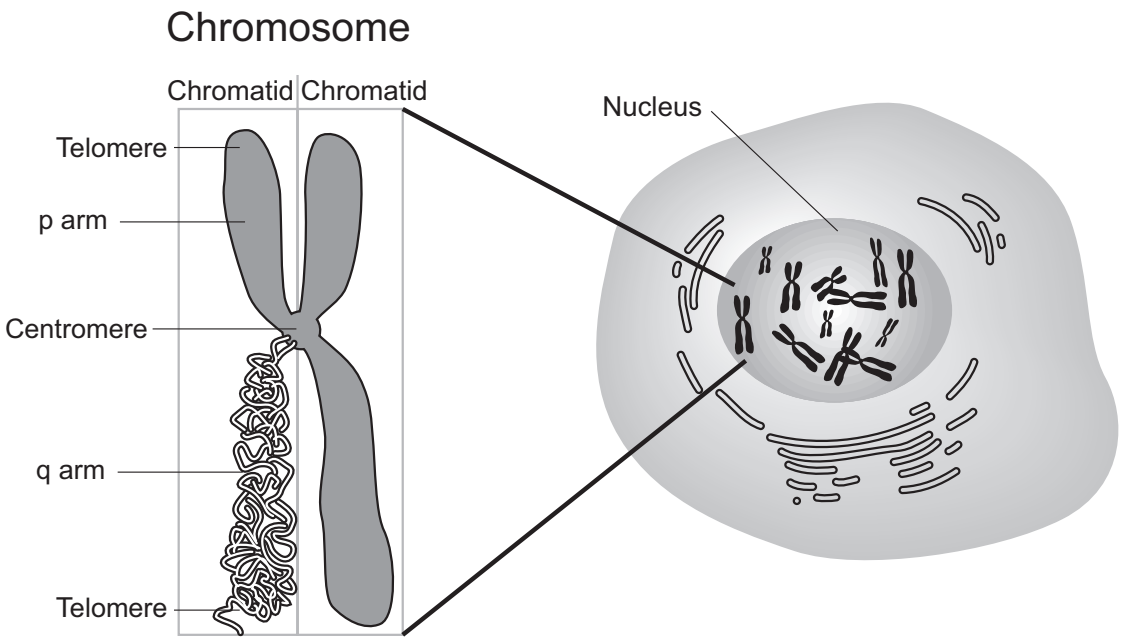

Figure 1.1: Chromosome of eukaryotes. Picture is modified from the National Human Genome Research Institute (NHGRI) "Talking Glossary of Genetic Terms".

inversion events between two *Drosophila* species, pioneering the study of chromosome rearrangement. Since then, many studies have concentrated on understanding the differences between genome architectures from an evolutionary perspective. Researchers have known that there are a number of possible rearrangement operations that have accumulated through evolution. In general, these rearrangements include inversions, translocations, fusions, and fissions.

Figure 1.2 illustrates these four rearrangement operations. In an inversion operation, a genomic segment on one chromosome is reversed and complemented, where complementation refers to the process of altering a string of the letters `A`, `C`, `G` and `T` by everywhere exchanging `A` with `T` and `G` with `C`. In a translocation operation, the end part of one chromosome is swapped with the end of another chromosome. If two chromosomes are joined to form one chromosome, it is called fusion. If a single chromosome

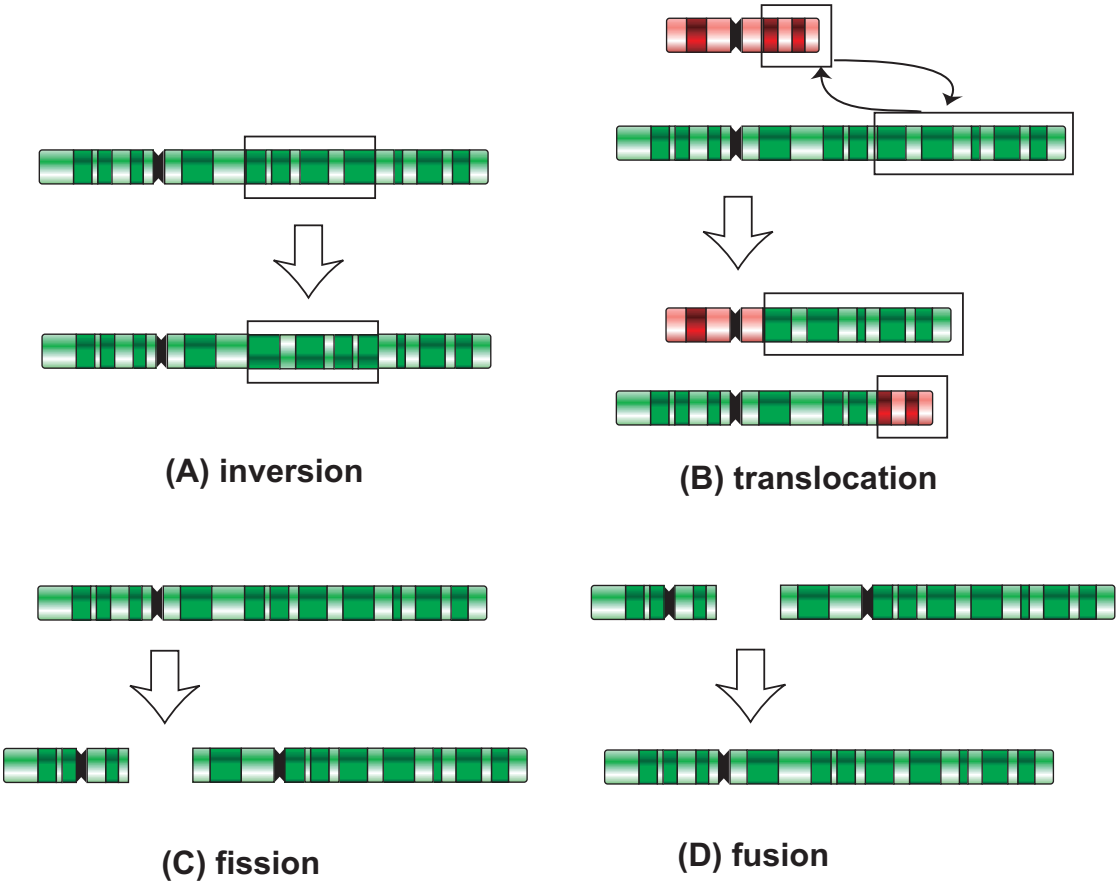

Figure 1.2: Chromosome rearrangements

is broken into two chromosomes, it is called fission. Among these operations, inversions are the most common events in karyotype evolution. For translocations, there are two main types, reciprocal (as shown in Figure 1.2 (B)) and Robertsonian. A Robertsonian translocation involves two chromosomes, where their long arms fuse at the centromere and the remaining two short arms are lost. It has been suggested that Robertsonian translocation also played an important role in mammalian genome evolution (Qumsiyeh, 1994).

Besides the above-mentioned rearrangement operations, chromosome architecture can also be changed by other large-scale operations, e.g. deletions and duplications. All these operations may happen in nested or overlapping form during evolution. As a result, karyotype architectures between different modern species can be highly distinct. A genomic segment in a modern species that is derived from a contiguous ancestral region can be widely scattered to different chromosomes and different positions. Unfortunately, the molecular mechanism responsible for determining the rate, position preference, and the conditions of all these chromosome rearrangements still remains mysterious.

1.2 Genome rearrangement: a computer science perspective

In the general mathematical model of chromosome evolution, a chromosome can be represented as a string of numbers (or permutation), and a genome as a set of these strings, e.g. 1 2 3 4 5 • 6 7 8, where • separates chromosomes. Numbers on a chromosome could be any genomic content, e.g. a single base, a gene, or larger piece of DNA sequence. Numbers may have signs, either + or −, which indicate the relative orientation of the genomic content.

The chromosome rearrangements discussed in the previous section can be interpreted as the following:

- Inversion: 1 2 3 4 5 • 6 7 ⇒ 1 -4 -3 -2 5 • 6 7. In computer science literature, inversion is also called reversal.
- Translocation: 1 2 3 4 5 • 6 7 ⇒ 1 7 • 6 2 3 4 5.
- Fusion: 1 2 3 4 5 • 6 7 ⇒ 1 2 3 4 5 6 7.
- Fission: 1 2 3 4 5 • 6 7 ⇒ 1 2 • 3 4 5 • 6 7.

Overlapping or nested operations form composite operations. For example, 1 2 3 4 5 6 7 can be transformed to 1 -4 -6 • -5 2 3 7 by two overlapping inversions followed by a fission: 1 2 3 4 5 6 7 ⇒ 1 -4 -3 -2 5 6 7 ⇒ 1 -4 -6 -5 2 3 7 ⇒ 1 -4 -6 • -5 2 3 7.

For the past decade, genome rearrangement problems have fascinated the computational biology community. Sankoff pioneered the theoretical study of reversal distance (Sankoff, 1992) and phylogenetic analysis using gene order data (Sankoff *et al.*, 1992). The analysis of the most parsimonious rearrangement scenarios is the central part of theoretical genome rearrangement study, among which sorting by reversal has been studied the most. Sorting by reversals is the problem of converting one permutation into another using the minimum number of reversal operations. The number of reversals is regarded as reversal distance between two permutations.

There are two categories for sorting by reversal problems: unsigned and signed. For the unsigned case, where elements have no relative orientation, it was proved to be NP-hard (Caprara, 1997). However, approximation algorithms do exist, among which the best known is proposed by Berman *et al.* (2002). For signed permutations,

Hannenhalli and Pevzner (1995) gave the first algorithm with polynomial running time. Inspired by this landmark algorithm, more efficient algorithms have been proposed since then (Berman and Hannenhalli, 1996; Kaplan *et al.*, 1997; Bader *et al.*, 2001). The Hannenhalli-Pevzner theory was also improved and implemented for multichromosomal genomes as GRIMM (Tesler, 2002).

Breakpoint distance is another measurement of pairwise rearrangement distance (Watterson *et al.*, 1982). For permutations π_1 and π_2, a breakpoint consists of an ordered pair of elements (e_i, e_j) that appear consecutively in π_1 but where neither (e_i, e_j) nor $(-e_j, -e_i)$ appears in π_2. The breakpoint distance between two permutations can be measured in linear time.

Phylogenetic analysis using genome rearrangement is based on methods for measuring rearrangement distances. A typical problem can be described as: Given three signed genomes A, B, and C, as well as the distance measurement d, infer a median genome M, such that $\sum d = d(A, M)+d(B, M)+d(C, M)$ is minimal. This is also known as the Median Problem for signed genomes, and it is NP-hard for all known distance measurements (Caprara, 1999), including reversal distance and breakpoint distance. Moreover, for the Breakpoint Median Problem, it is NP-complete. Sankoff and Blanchette (1998) reduced the Breakpoint Median Problem to the Travelling Salesman Problem and implemented a heuristic solution, called BPAnalysis. BPAnalysis was later improved by GRAPPA (Moret *et al.*, 2001). The latest version of GRAPPA also implements reversal distance; indeed, part of the code for distance computation in GRAPPA is used in MGR (Bourque and Pevzner, 2002). MGR adopts heuristic approaches to measure reversal distance and is applicable to multichromosomal genomes. The heuristic approach that

MGR uses is to choose "good" reversals to reduce the computational complexity. So far, MGR is probably the only program that has been applied to the real data of multichromosomal mammalian genomes. Pevzner and colleagues have used MGR to do a series of rearrangements studies among the complete genomes of human, mouse, rat, and some other mammals (Pevzner and Tesler, 2003; Bourque *et al.*, 2004, 2005; Murphy *et al.*, 2005).

1.3 Ancestral genome reconstruction: a brief overview

All placental mammals living today are descended from an ancient species that lived about 125 million years ago (MYA). As the result of speciation events and many significant changes in each lineage, we see remarkable differences among living placental mammals, both genetic and morphological. However, since these species are descended from a common ancestor, they all have inherited specific DNA sequences from the ancestral genome.

The increasing number of mammalian genome sequences becoming available provides many different versions of the original genome that can be compared, thereby giving us the opportunity to computationally reconstruct an ancestral mammalian genome. Computer simulations have shown (Blanchette *et al.*, 2004a) that the genome sequence of the so-called Boreoeutherian ancestor (see Figure 1.3; a few mammals, including elephants and armadillos, branched off from the human lineage a bit earlier) can be computationally predicted at high accuracy within most euchromatic intervals that are free of large-scale rearrangements, given adequate data from living mammals.

For instance, when sequences from 20 appropriately chosen mammalian species are available, we expect that over 98% of the reconstructed nucleotides will be identical to the corresponding ancestral base. Because all mammals have experienced large-scale chromosomal rearrangements since their last common ancestor, there is tremendous karyotype diversity among modern species. In order to determine regional correspondence, it is important to analyze and restore these rearrangements to infer a partition of each genome into intervals where nucleotide-level reconstruction methods can be applied.

In fact, the problem of ancestral mammalian karyotype reconstruction has been studied for quite a long time. As early as 1970, Todd (1970) suggested that the mammalian ancestor probably had 7 chromosomes. A few years later, Matthey (1972) proposed a completely different number, 24. Due to the development of comparative gene mapping and cytogenetic methods, biologists have been trying to solve the jigsaw puzzle with more powerful tools by combining these different techniques. However, the number of chromosomes is still not fixed, and different numbers in the ancestral eutherian karyotype have been suggested, e.g. 22 (Yang *et al.*, 2003), 24 (Chaudhary *et al.*, 1998; Froenicke *et al.*, 2003, 2006), and 25 (Murphy *et al.*, 2001; Richard *et al.*, 2003). Even though there is no consensus of the number of chromosomes in the ancestral eutherian karyotype, some joins have been widely confirmed, e.g. Hsa3/Hsa21 and Hsa14/Hsa15. ("Hsa" refers to a human chromosome.)

Currently, chromosome painting is the main experimental technique (surveyed by Wienberg (2004) and Froenicke *et al.* (2006)), in which fluorescently labeled chromosomes from one species are hybridized to chromosomes from another species. Although the requirement of optical visibility means that the cytogenetic approach can recognize

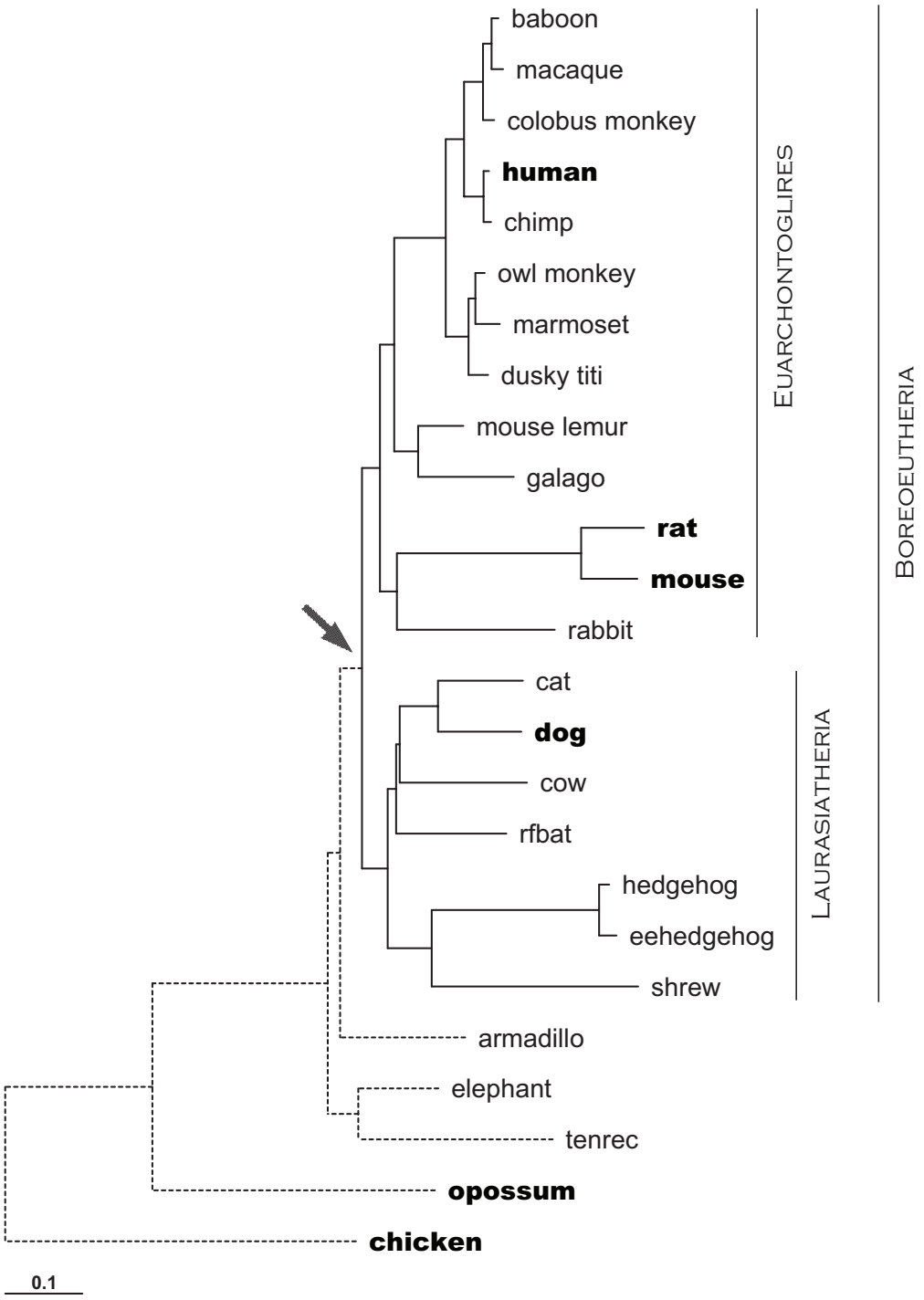

Figure 1.3: Position of Boreoeutherian ancestor. Species used in the analysis in later chapters are highlighted. Tree is inferred from the ENCODE data freeze (Jan 2006), provide by Elliott H. Margulies.

only rearrangements with conserved segments longer than 4Mb (Froenicke *et al.*, 2006) and cannot identify intrachromosomal rearrangements (Wienberg, 2004), the chromosomal painting approach has the advantage that data are available for over 80 mammals (50 primates).

On the other hand, computer scientists have also tried to reconstruct the ancestral genome architecture using bioinformatic algorithms in a parsimony framework based on certain distance measurement. Among them, the heuristic algorithm MGR (Bourque and Pevzner, 2002) is the most successful tool so far that has been applied to real biological data. Relying on the MGR program, Murphy *et al.* (2005) estimated the rearrangement rates using Radiation Hybrid data in the lineages leading to human, mouse, rat, cat, cattle, dog, pig and horse, and predicted that the Boreoeutherian ancestor had 24 chromosomes.

However, the obvious contrast between the most recent reconstructions using cytogenetic and bioinformatic approaches was highlighted in a recent Forum in Genome Research (Froenicke *et al.*, 2006; Bourque *et al.*, 2006). It showed that several interchromosomal joins proposed by MGR were not supported by cytogenetic experiments, while bioinformatics approach analyzed the rearrangements in much higher resolution.

1.4 Motivation

Determining the ancestral order and orientation of large genomic regions is one of the main steps toward our ultimate goal of a nucleotide-level reconstruction of the Boreoeutherian ancestral genome. The conflicting results from different existing methods encourage us to develop powerful bioinformatic tools that make more biological sense.

As we reviewed before, most computational methods for rearrangement analysis assume a pre-defined set of operations, e.g. inversions, but the frequencies or cost functions of operations are still elusive. Currently there is no widely accepted model for genome rearrangement, compared with a number of solid and convincing substitution models. An ideal model for rearrangements will account for the frequencies, length distributions, and positional heterogeneity of each operation. That these rearrangement parameters are species-specific makes this problem even harder.

Therefore, the motivation of this study is to explore approaches that are independent of models and essentially more efficient, but still biologically realistic. By taking advantage of genome-wide sequence alignment and comparative genomics, we observed that adjacencies of genomic content in modern species can be used to infer the ancestral adjacencies. When the adjacencies are available, the other task is to connect these genomic content into contiguous regions.

Adjacencies of genomic content (e.g. genes) have been used as a binary character to infer phylogeny in a parsimony framework (as surveyed in Savva *et al.* (2003)). However, in a different context, where the phylogeny is known, our objective is to predict the ancestral order and orientation based on adjacencies in modern genomes. Consider an end of a conserved genomic segment that does not correspond to a human telomere or centromere. How can we identify the segment that was adjacent in the ancestral genome? If the segment that is currently adjacent in human is identical to the one that is adjacent in dog (but a different segment is adjacent in mouse and rat), the most parsimonious assumption is that the first and second segments were adjacent in the ancestral genome (and that a disruption occurred in the rodent lineage at this genomic position).

If the same segment is adjacent to the chosen segment in human, mouse and rat, but not in dog, we need more information to confidently predict the ancestral configuration, since there is a chance that the dog adjacency is ancestral and that the breakage occurred on the short branch from the human-dog ancestor to the human-rodent ancestor (see Fig. 1.3). To help resolve these cases, we can add outgroup information, e.g. opossum and chicken sequence. If the outgroup information does not resolve the issue (by agreeing with either the human adjacency or the dog adjacency), we assume the more likely scenario, i.e., that the break occurred in the lineage leading to dog.

We generalize these observations and develop a set of new computer programs that are discussed in the following chapters. Using these methods, we propose a karyotype prediction of Boreoeutherian ancestor. Our results show a significant convergence with chromosome painting, though there are still uncertainties that need to be resolved.

Chapter 2

The algorithm for reconstructing CARs

2.1 Preliminaries

This chapter is organized as follows. Given information about adjacencies between genomic contents in each modern species (including outgroups), we generalize the observations from sequence alignments and develop a computational procedure for predicting the order and orientation of genomic segments in the ancestor, based on observed adjacency relationships in the modern genomes. To get a clean and precise statement of the problem, we formalize it using graph theory. We develop an algorithm that identifies a most-parsimonious scenario for the history of each individual adjacency, though the whole-genome prediction is not guaranteed to optimize traditional measures like the number of breakpoints. We introduce weights to the graph edges to model the reliability of each adjacency. Finally, we propose a greedy heuristic approach to look for vertex-disjoint paths in the graph, which represent the contiguous ancestral regions.

Let us start with some important definitions that will appear in later discussions.

An **element** is any member of a fixed set containing N genomic building blocks, which could be genes, orthologous genomic segments, or any other atomic evolutionary units.

We identify these elements with the integers 1 through N. Each element appears in the targeted ancestral genome and in some or all of the modern genomes. A **chromosome** of a modern or ancestral genome consists of a list of elements, each with a sign (orientation) that is either positive (+) or negative (−); thus a chromosome's entries are just non-zero integers between $-N$ and N. Currently, we do not allow duplication, i.e. an element cannot appear more than once in the chromosomes of a given genome.

The **reverse complement** of a chromosome is obtained by reversing the list and flipping the sign of each entry. If modern genome g contains element i, then the **predecessor** $p_g(i)$ is defined as the signed element that immediately precedes i on the same chromosome relative to the original orientation. In the opposite orientation, $p_g(-i)$ immediately precedes $-i$ in the reverse complement of the same chromosome. We set $p_g(i) = \phi_A$ if i appears first on a chromosome. The **successor** $s_g(i)$ of i is defined analogously; we set $s_g(i) = \phi_Z$ if i appears last on a chromosome.

For instance, let g have the chromosome $(1\ -4\ -3\ 5\ 2)$. Then in the positive orientation, we have: $p_g(1) = 0$, $p_g(2) = 5$, $p_g(-3) = -4$, $p_g(-4) = 1$, $p_g(5) = -3$, while $s_g(1) = -4$, $s_g(2) = 0$, $s_g(-3) = 5$, $s_g(-4) = -3$, $s_g(5) = 2$. In the opposite orientation, we have: $p_g(-1) = 4$, $p_g(-2) = 0$, $p_g(3) = -5$, $p_g(4) = 3$, $p_g(-5) = -2$, while $s_g(-1) = 0$, $s_g(-2) = -5$, $s_g(3) = 4$, $s_g(4) = -1$, $s_g(-5) = 3$.

Given a phylogenetic tree and the set of chromosomes for each modern (i.e., leaf) species, we want to determine a set of lists of signed elements that closely approximates the chromosomes of the species corresponding to the root of the tree. For our purposes, the reverse complement of an ancestral chromosome is entirely acceptable.

To emphasize the fact that our reconstruction of ancestral chromosomes may be incomplete, i.e., find only parts of chromosomes, we call each of the constructed lists a **contiguous ancestral region**, or **CAR**.

2.2 The algorithm

Our approach is inspired by Fitch's method (Fitch, 1971), which was originally used to infer minimum character changes in a specified tree topology. For that problem, one is given a phylogenetic tree and a letter for every position in each leaf of the tree (corresponding to the contents of orthologous sequence sites). The problem is to infer the ancestral letters (corresponding to internal nodes of the tree), so as to minimize the number of substitutions, i.e., differences between the letters at each end of an edge in the tree.

The algorithm works sequentially, in two stages (see Felsenstein, 2003, chap. 2). For each position, in a bottom-up fashion, it first determines a set M_π of candidate nucleotides at each node π in the tree according to the following rule: if π is a leaf, M_π just contains its nucleotide character; otherwise, if π has children τ and φ, then M_π equals to $M_\tau \cap M_\varphi$ or $M_\tau \cup M_\varphi$ depending on whether M_τ and M_φ are disjoint or not. I.e.,

if $\| M_\tau \cap M_\varphi \| \neq 0$
  then $M_\pi \leftarrow M_\tau \cap M_\varphi$
  else $M_\pi \leftarrow M_\tau \cup M_\varphi$

where $\| X \|$ denotes the number of items in the set X.

Then, in a top-down fashion, it assigns a character b_π from M_π to π according to the following rule: Let ρ be the parent of π; if the character b_ρ assigned to ρ belongs to

M_π, then, $b_\pi = b_\rho$. Otherwise, set b_π to be any character in M_π. Although character assignment in this second stage may not be unique, any assignment gives an evolutionary history with the minimum number of substitution events.

The rationale behind Fitch's method is as follows. If the character b_π belongs to both children of π, then an optimal strategy for labeling nodes in the subtree rooted at π is to put b at each of π, τ and φ, and label the subtrees of τ and φ optimally. If there is no such b, then the strategy is to put a character of either in M_τ or M_φ at π, pay for one substitution to reach the other child, and optimally label the two subtrees.

See Figure 2.1 for an example. The characters at leaves are given. Then we do a postorder tree traversal and create sets in the internal nodes until we reach the root. In this example, the ancestral nucleotide `A` will give us the minimum number of substitutions, which is 2, for this position.

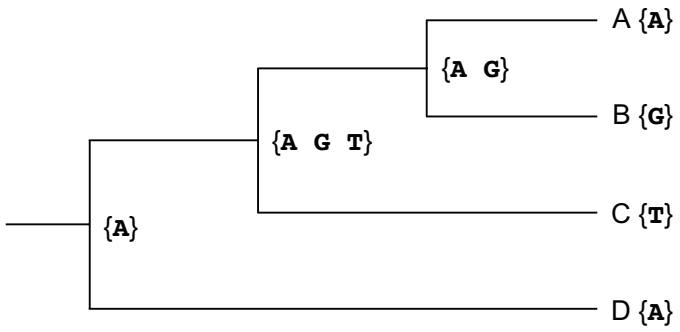

Figure 2.1: An example of Fitch's algorithm

In our case, we deal with sequences of signed integers, rather than characters of nucleotides or amino acids, and instead of keeping track of letters at a particular sequence position, we track the elements for each of the immediately adjacent positions. For instance, consider keeping track of the set of possible elements that follow a fixed

orthologous element in a most-parsimonious evolutionary scenario. In the genome that corresponds to node π, element i could be followed by any element that follows i in both τ and φ, without requiring any rearrangements on the branches leading from π to its children. Otherwise, i can be followed by any element that follows i in one of π's children, at the cost of a chromosomal break next to i along the branch leading from π to the other child. This is all closely analogous to the case of substitutions, as sketched above, though besides the new complication of keeping track of two sets at each node, we must deal with reverse complements.

Thus, for any genome g, we associate with each element i two sets of signed elements, denoted $P_g(i)$ and $S_g(i)$, giving potential predecessors and successors of i relative to chromosomes of g. If g is a modern genome, $P_g(i) = \{p_g(i)\}$ and $S_g(i) = \{s_g(i)\}$, for each i. If g does not contain i, then both sets are empty.

The procedure GET-PREDECESSOR-SUCCESSOR(R) constructs $P_g(i)$ and $S_g(i)$ for each element i of every ancestral genome g in the tree rooted at R.

```
GET-PREDECESSOR-SUCCESSOR(π)
 1  if t is non-leaf node
 2    then GET-PREDECESSOR-SUCCESSOR(τ)
 3         GET-PREDECESSOR-SUCCESSOR(φ)
 4         for i ← −N to N (i ≠ 0)
 5               do if ‖ P_τ(i) ∩ P_φ(i) ‖≠ 0
 6                     then P_π(i) ← P_τ(i) ∩ P_φ(i)
 7                     else P_π(i) ← P_τ(i) ∪ P_φ(i)
 8                  if ‖ S_τ(i) ∩ S_φ(i) ‖≠ 0
 9                     then S_π(i) ← S_τ(i) ∩ S_φ(i)
10                     else S_π(i) ← S_τ(i) ∪ S_φ(i)
```

However, the root of the tree is not always the target genome we want to eventually reconstruct, which means we have outgroup information. We treat outgroups in a consistent manner. We first infer $P_R(i)$ and $S_R(i)$ in the common ancestor R of all the species, including outgroups. Then we propagate $P_R(i)$ and $S_R(i)$ down the tree until we reach the target ancestor α. During the propagation process, suppose O and A are ancestor and descendant on one branch, respectively. For each element i, we adjust the inferred predecessor set $P_A(i)$ of i at the node A as follows: if $P_O(i)$ and $P_A(i)$ share common elements, we just take them as the predecessor set of i at A; otherwise, $P_A(i)$ is unchanged. We do the same for $S_A(i)$ during the propagation. The whole process can be summarized as pseudo-code ADJUST-ANCESTOR. We assume that the path from the root to the target ancestor has already been recorded.

ADJUST-ANCESTOR(R, α)

```
1   u ← R
2   v ← R.next
3   while u ≠ α
4         do for i ← −N to N (i ≠ 0)
5               do if ‖ P_u(i) ∩ P_v(i) ‖≠ 0
6                     then P_v(i) ← P_u(i) ∩ P_v(i)
7                  if ‖ S_u(i) ∩ S_v(i) ‖≠ 0
8                     then S_v(i) ← S_u(i) ∩ S_v(i)
9           u ← v
10          v ← v.next
```

It is useful to construct a **predecessor graph** G_g^P and a **successor graph** G_g^S for each genome g. In digraph $G_g^P = (V, E)$, $|V| = 2N$ where each element i corresponds to two nodes, i and $-i$, and the set of directed edges is:

$$E(G_g^P) = \{(u, v) \mid u \in P_g(v)\} \tag{2.1}$$

Similarly, in digraph $G_g^S = (V, E)$, $|V| = 2N$, and:

$$E(G_g^S) = \{(u, v) \mid v \in S_g(u)\} \tag{2.2}$$

Here, (u, v) denotes an arc directed from u to v. Note that an edge in G_g^P is *from* the predecessor, while an edge in G_g^S is *to* the successor.

For instance, let g have the chromosome (1 -4 -3 5 2). Then G_g^P and G_g^S are as shown in Figure 2.2 and Figure 2.3, respectively.

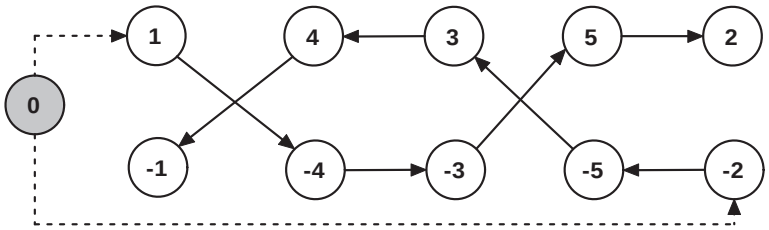

Figure 2.2: A predecessor graph G_g^P

In general, there is a strong symmetry between the two graphs, as given by the identity:

$$P_g(i) \;=\; \{-j : \; j \in S_g(-i)\} \tag{2.3}$$

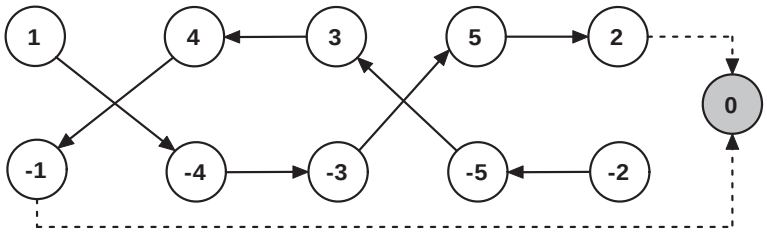

Figure 2.3: A successor graph G_g^S

With the way that we will draw the graphs, this means that G_g^P can be obtained from G_g^S by inverting the graph and flipping the orientation of its edges.

At the ancestor, α, we have predecessor graph G_α^P and successor graph G_α^S. In order to find CARs, for each element s, we need to find a unique predecessor p, such that s is the unique successor of p. We first intersect G_α^P and G_α^S, producing $G = G_\alpha^P \cap G_\alpha^S$, to retain edges that are not connecting to either of the endpoints, ϕ_A and ϕ_Z. Then special care is taken to add endpoint edges, basically retaining all the endpoint edges that appear in both G_α^P and G_α^S. All three graphs have the same set of $2N$ nodes. G's edges are:

$$E(G) = \Big\{E(G_\alpha^P) \cap E(G_\alpha^S)\Big\} \cup \Big\{(\phi_A, v) \mid (\phi_A, v) \in E(G_\alpha^P)\Big\} \cup \Big\{(u, \phi_Z) \mid (u, \phi_Z) \in E(G_\alpha^S)\Big\} \tag{2.4}$$

Those edges indicate consistent predecessor and successor relationships that are supported by the tree and the modern genomes.

A node i of G can be involved in the three kinds of ambiguity depicted in Figure 2.4. In (a), i has several incoming edges. In (b), i has several outgoing edges. In (c), i forms a cycle with j, where each node j satisfies $indegree(j) = outdegree(j) = 1$. (If a more complex cycle exists, then some node falls in either case (a) or case (b)).

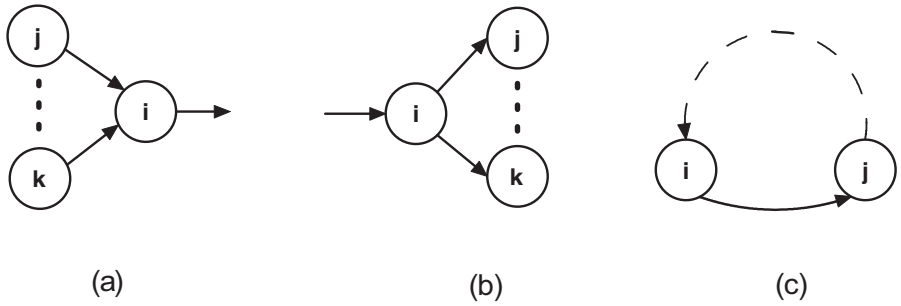

Figure 2.4: Three potential ambiguous cases in the intersection graph G

If none of these ambiguous cases is present, the graph itself forms the set of paths that covers all the nodes. When ambiguity exists, we need to resolve the ambiguity and choose appropriate directed edges to form CARs. We assign a weight to each of the directed edges in the remaining graph using the following approach.

- For an edge (i, j), if $outdegree(i) = 1$ and $indegree(j) = 1$ (in other words, it is not among one of the incoming edges of case (a) nor it is among one of the outgoing edges of case (b)), we set $w_{\alpha}(i, j) = 1$.

- Otherwise, the corresponding weight $w_{\alpha}(i, j)$ is determined by:

$$w_{\alpha}(i, j) = \frac{D_L \cdot w_R(i, j) + D_R \cdot w_L(i, j)}{D_L + D_R} \tag{2.5}$$

  where D_L and D_R are the branch lengths to the left child and right child; $w_L(i, j)$ and $w_R(i, j)$ are the edge weights on left child and right child, respectively. On a leaf genome, if (i, j) is present in the predecessor graph, we set $w(i, j) = 1$, otherwise $w(i, j) = 0$. This kind of edge weight can also be determined by a postorder traversal. Note that if an edge (i, j) is involved in ambiguous case (a)

or (b), $w(i,j) < 1$. The greater the value of an edge (i,j), the more plausible it is that i and j should be joined. The underlying assumption is that rearrangement is more likely to happen on longer branches (if branch lengths are consistent with rearrangement distances).

Our goal is to connect elements into the longest possible CARs that are consistent with the observed data. The problem can be transformed into looking for vertex-disjoint paths that cover all the nodes in the digraph G with the maximum weight. Here we also allow degenerate paths, where there is only one node. The simplified version of this problem when all the edge weights are the same, say 1, is equivalent to the Minimum Path Cover Problem, i.e., finding the minimum number of vertex-disjoint paths covering all the nodes in the digraph. The minimum path cover problem was proved to be NP-hard (Boesch and Gimpel, 1977).

Here, we use a greedy heuristic approach to achieve an approximate solution. See the algorithm of FIND-CARS below. We first sort the edges by weight. The greedy approach always tries to add the heaviest edge to the resulting path set. When an edge (i,j) is being added to the path set, we make sure: (1) indegree of j is 0; (2) outdegree of i is 0. Otherwise, we just discard that edge and choose the next available one. The process is performed until no edge can be added. However, this process doesn't guarantee there will be no cycle in the path set. We need a final step (line 6) to detect and break cycles. We use the depth-first-search algorithm to detect cycles in graph G then remove the edge with the smallest weight. The remaining paths in G correspond to the contiguous ancestral regions we want.

FIND-CARS(G)

1 Sort edges by weight in descending order.
2 Create a new graph C, $V(C) = V(G)$ and $E(C) = \emptyset$
3 **for** each available $(i, j) \in E(G)$, in order of edge weight
4 **do if** $outdegree(i) = 0$ and $indegree(j) = 0$
5 **then** Add edge (i, j) and $(-j, -i)$ to $E(C)$
6 Break cycles in C.
7 Return

However, the greedy algorithm is just an approximation algorithm. We can prove from the worst-case counterexample shown in Figure 2.5 that the approximation ratio is 3.

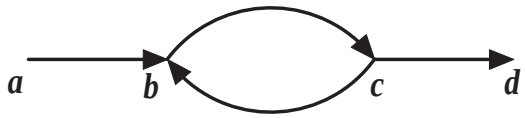

Figure 2.5: The worst case counterexample that greedy approach doesn't produce the optimal solution. $w(a, b) = w(b, c) = w(c, d) = k$ and $w(c, b) = k + \epsilon$, $1 < \epsilon \ll k$. FIND-CARS will return path $c \rightarrow b$ and degenerate paths a and d. Since $3w(c, b) > w(a, b) + w(b, c) + w(c, d)$, at worst, the greedy result is about $\frac{1}{3} \times OPTIMAL$.

Although in theory the approximation ratio is 3, in practice we usually can achieve a very good approximation, especially when the ratio of breakpoint reuse is low, say 10%. As we will discuss in the Chapter 4, we observed from our real data that most nodes in the graph have only one outgoing edge.

In the resulting graph after the greedy procedure, all ambiguous cases of (a) and (b) will be resolved. However, we haven't dealt with ambiguous case (c), when a cycle

is formed. In fact, we can prove that if there is a cycle, the weight of each edge in that cycle is 1. Therefore, we can simply discard an arbitrary edge to break the cycle.

When adding edges into an existing path, particular care is needed to avoid putting j and $-j$ in the same CAR. In addition, we add both (i, j) and its symmetric version, $(-j, -i)$. For each path found by this approach, a symmetric path in the opposite orientation is also found, since we have nodes for both i and $-i$. The two paths correspond to the same CAR, and we choose one of them.

In outline, the whole Infer-Cars algorithm can be described as follows, where $\mathcal{T}$ denotes the phylogenetic tree, $\mathcal{G}$ denotes the collection of modern genomes (leaves of $\mathcal{T}$), N is the number of ancestral elements, and α is the target ancestor.

Infer-Cars$(\mathcal{G}, N, \mathcal{T}, \alpha)$

1 $\mathcal{C} \leftarrow$ empty set of CARs
2 $R \leftarrow root(\mathcal{T})$
3 **for** each modern genome g in $\mathcal{G}$
4 **do for** $i \leftarrow -N$ **to** N $(i \neq 0)$
5 **do** $P_g(i) \leftarrow \{p_g(i)\}$ and $S_g(i) \leftarrow \{s_g(i)\}$
6 Get-Predecessor-Successor(R)
7 Adjust-Ancestor(R, α)
8 Get graph G according to Equation 2.4
9 Find-CARs(G)
10 **return** $\mathcal{C}$

Suppose we are given S modern genomes. There are $2S - 1$ nodes in the phylogenetic tree, including leaf and internal nodes. The postorder traversal of Get-Predecessor-Successor takes $O(S)$ steps. At each step, the intersection or union

operation for each element takes $O(S)$ operations at most, since there are at most S elements in the predecessor set or successor set. Thus GET-PREDECESSOR-SUCCESSOR runs in $O(S^2N)$ time. The running time of ADJUST-ANCESTOR is also $O(S^2N)$.

Time for the operation in line 8 in INFER-CARS is bounded by $O(SN)$. Because there are S genomes, there are at most S possible elements in each $P_\alpha(i)$ or $S_\alpha(i)$. Hence, each node (except special endpoint nodes ϕ_A and ϕ_Z) in the predecessor graph G_α^P or successor graph G_α^S has at most S outgoing edges. Consequently, the number of directed edges in either graph is $\leq 2SN$. The operations for endpoints are both bounded in $O(N)$. Therefore, the operation is bounded in $O(SN)$.

For FIND-CARS, the sorting procedure takes $O(SN \log N)$, and the running time of line 3 - 5 in FIND-CARS is $O(SN)$. The cycle-breaking step takes $O(SN + N)$. Hence the total running time for FIND-CARS is $O(SN \log N)$.

Therefore, the entire running time for the FIND-CARS algorithm can be bounded by $O(S^2N + SN \log N)$.

2.3 A detailed example

Here, we explain the algorithm using a detailed example.

The predecessor graphs of A, B, and C can be obtained directly from the leaf genomes; see Figure 2.7, 2.8, and 2.9. There are two special nodes representing the beginning and the end of a chromosome. The predecessor graphs of internal nodes D and E are as shown in Figure 2.10 and 2.11. The predecessor graph of root F is shown in Figure 2.12. Figure 2.13 is the result after E is adjusted using F. The corresponding successor graph for E is shown in Figure 2.14. Then we create the intersection of the

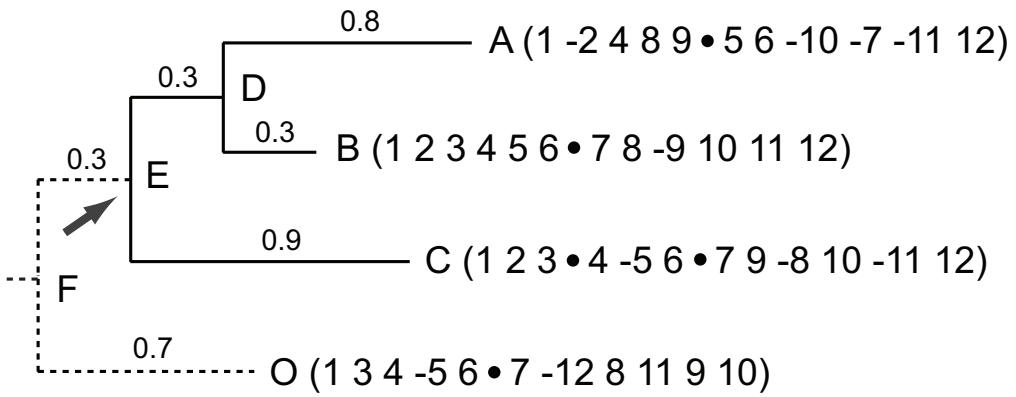

Figure 2.6: The phylogeny of genomes A, B, C, E, and O. Our target ancestor is E, and O is the outgroup. The bullet symbol, •, separates chromosomes. Branch lengths are above each branch.

predecessor and successor graphs in 2.13 and 2.14, giving the graph in Figure 2.15. Note that in this step we actually do not intersect edges connecting the beginning or the end of the chromosome.

In Figure 2.15, there are ambiguous cases for nodes 7, 8, 9, 10. We then assign weights to edges recursively using the approach discussed above. For example, $w(7,8) = w(-8,-7) = 0.54$. We have $w_A(7,8) = 0$, $w_B(7,8) = 1$, $w_C(7,8) = 0$, and $w_D(7,8) = \frac{0.8}{0.3+0.8} = 0.72$. So $w_E(7,8) = \frac{0.72\times 0.9}{0.3+0.9} = 0.54$. Note that edges of weight 1 are not shown in the picture.

Then we sort all the edges by weight and add them to the graph until every node in the graph has a unique predecessor and successor. The final edges are indicated by the dark edges in Figure 2.16. The paths come in pairs, which corresponds to the two orientations of each CAR. We select one path from each pair, obtaining for example CARs (1 2 3 4 -5 6) and (7 8 -9 10 -11 12).

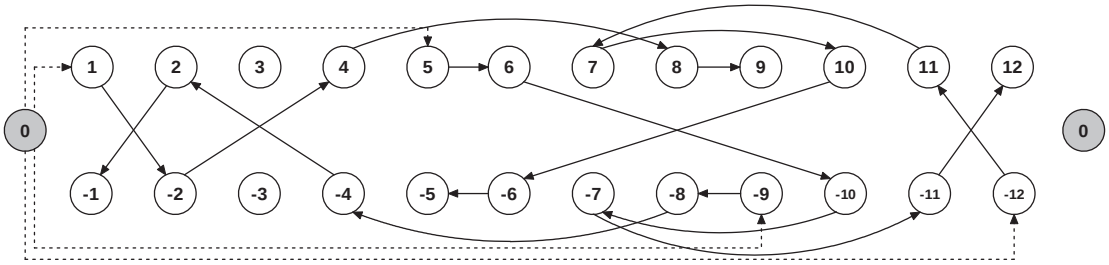

Figure 2.7: Predecessor graph of A

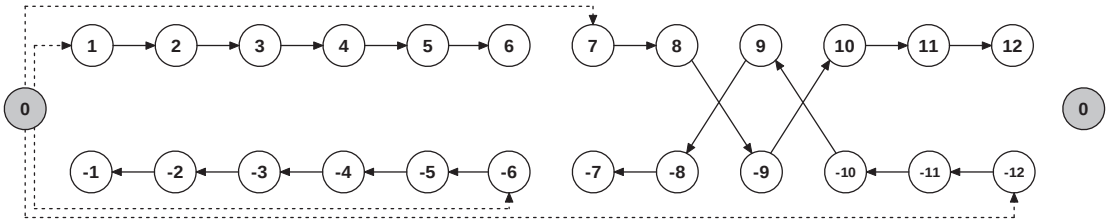

Figure 2.8: Predecessor graph of B

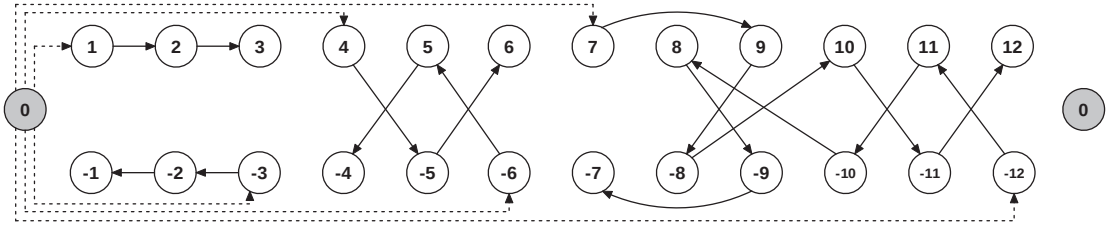

Figure 2.9: Predecessor graph of C

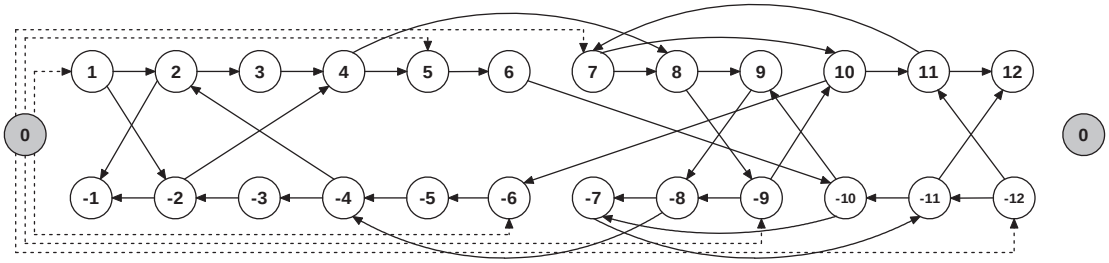

Figure 2.10: Predecessor graph of D

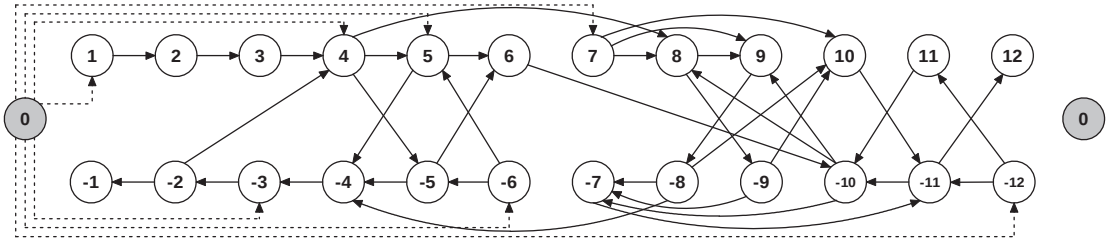

Figure 2.11: Predecessor graph of E

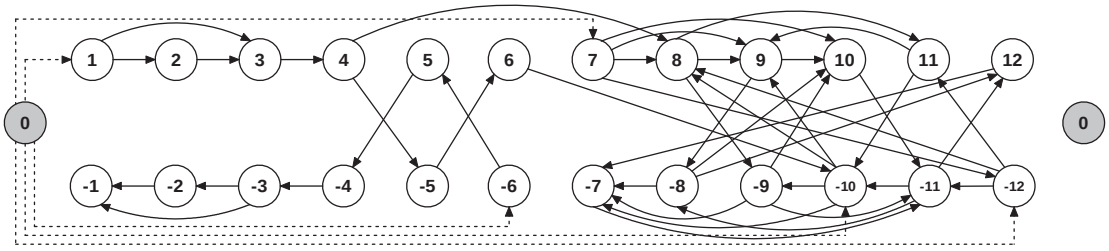

Figure 2.12: Predecessor graph of F

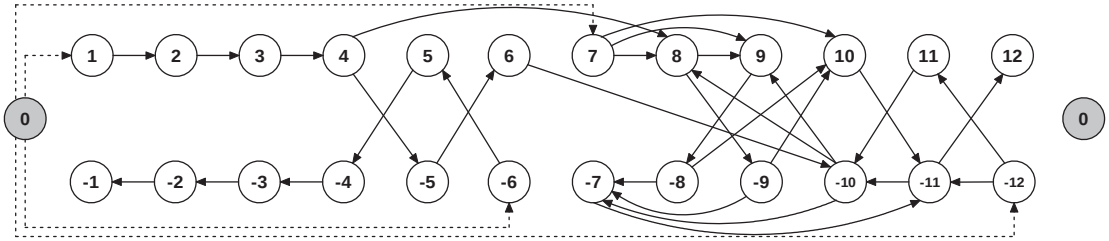

Figure 2.13: Predecessor graph of E after being adjusted by F

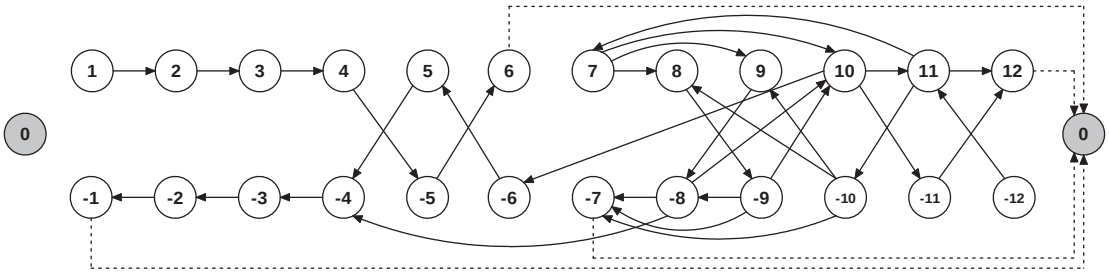

Figure 2.14: Successor graph of E

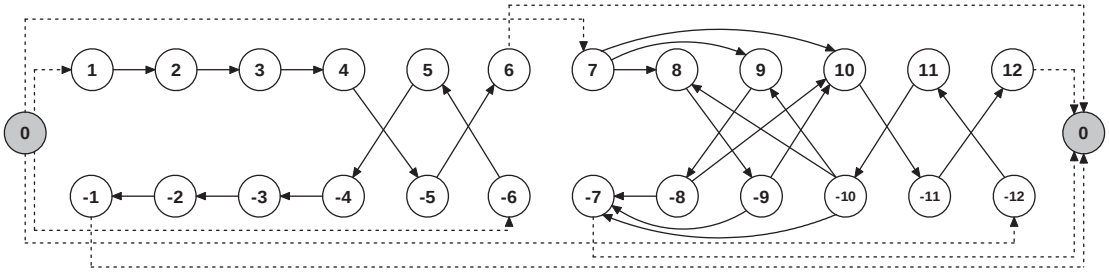

Figure 2.15: Intersection of the predecessor graph and successor graph of E

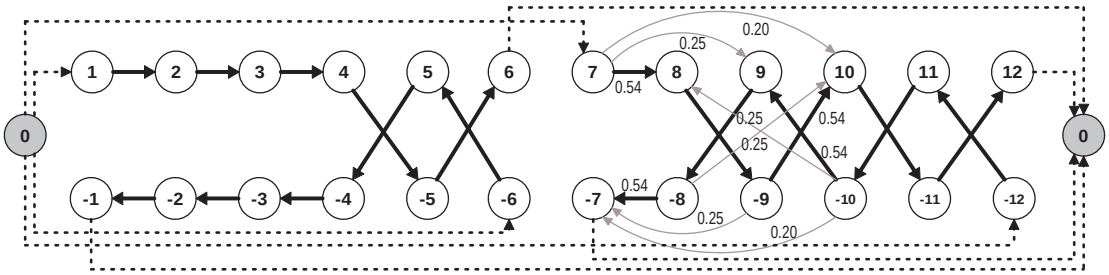

Figure 2.16: The resulting CARs

2.4 Discussion

The traditional definition of breakpoint can be described as (see the definition in (Sankoff and Blanchette, 1998)): If two elements are adjacent in genome $A = a_1...a_n$ but not adjacent in genome $B = b_1...b_n$, they determine a breakpoint between A and B. In addition, a_1 is considered to be adjacent to the "start" of genome A and a_n is adjacent to the "end", so there can be a breakpoint before a_1 or after a_n.

Here, we propose a variant definition of breakpoint distance in order to handle multichromosomal situations. Suppose genomes A and B share n common elements, located in p chromosomes and q chromosomes, respectively. Then, there are a total of $n + p$ adjacencies in A and $n + q$ adjacencies in B.

Using ϕ to denote the beginning and end of a chromosome (i.e. the special elements), we assign a score c_k $(k = 1, \ldots, n + p)$ to every adjacency $(a_i a_j)$ in A:

$$c_k = c(a_i, a_j) = \begin{cases} 0 & \text{if } (a_i a_j) \text{ or } (-a_j - a_i) \text{ is in } B; \\ \frac{1}{2} & \text{if } a_i = \phi \text{ or } a_j = \phi \text{ and both } (a_i a_j) \text{ and } (-a_j - a_i) \text{ are not in } B; \\ 1 & \text{otherwise.} \end{cases}$$

Then the **breakpoint distance** between A and B is defined as:

$$d(A, B) = \sum_{k=1}^{n+p} c_k \tag{2.6}$$

For example, $A = (1\ 2 \bullet 3\ -4\ 5)$ and $B = (5\ 3 \bullet 1 \bullet 2\ 4)$ (the bullet symbol, $\bullet$, separates chromosomes). In A, we have $c(\phi, 1) = 0$, $c(1, 2) = 1$, $c(2, \phi) = 0.5$, $c(\phi, 3) = 0.5$, $c(3, -4) = 1$, $c(-4, 5) = 1$, $c(5, \phi) = 0.5$, therefore $d(A, B) = 4.5$.

In our method, we consider predecessor changes and successor changes of each element independently. Suppose we have two genomes, A and B. Assume each element i (except ϕ) in the genome g has a predecessor $p_g(i)$ and a successor $s_g(i)$. We set $\mathcal{P}(A, B)$ to be the number of i where $p_A(i) \neq p_B(i)$, and $\mathcal{S}(A, B)$ to be the number of i where $s_A(i) \neq s_B(i)$. We can see that:

$$\mathcal{P}(A, B) + \mathcal{S}(A, B) = \frac{1}{2} d(A, B) \tag{2.7}$$

Since $\mathcal{P}(A, B) = \mathcal{P}(B, A)$ and $\mathcal{S}(A, B) = \mathcal{S}(B, A)$, it follows that $d(A, B) = d(B, A)$.

We let $T = (V, E)$ be the given phylogenetic tree. Suppose there are S genomes with N elements. Without loss of generality, we denote the S leaf genomes as $V_1, V_2, ..., V_S$. The phylogenetic tree T can be regarded as a *full binary tree* where each node is either a leaf or has two children. If a full binary tree has S leaf nodes, it is easy to prove that there are $S - 1$ non-leaf nodes. We denote the non-leaf nodes as $V_{S+1}, V_{S+2}, ..., V_{2S-1}$. Also, we let $V_{2S-1} = \alpha$, which corresponds to the target ancestral genome. In the tree, we associate genome g_i to node V_i. Using this notation, an optimal reconstruction for sequences corresponding to a phylogenetic tree with root α would satisfy the condition:

$$\sum_{(V_i, V_j) \in E(T)} d(g_i, g_j) \text{ is the minimum.}$$

The GET-PREDECESSOR-SUCCESSOR procedure gives us predecessor and successor sets of each element that guarantees the minimum changes over the evolutionary tree. However, this is for the local changes because we look at each element independently. If we look globally throughout the genome, the median problem condition cannot always be satisfied, especially when there are ambiguous cases in the ancestor. Further analysis is necessary to investigate in which conditions our algorithm will produce an optimal solution in terms of breakpoints.

The discussion so far has implicitly assumed that there are no duplicated elements in the genome. Here *duplicated* refers to two elements that have the same absolute value (relative to the permutation numbers). See the following example in Figure 2.17.

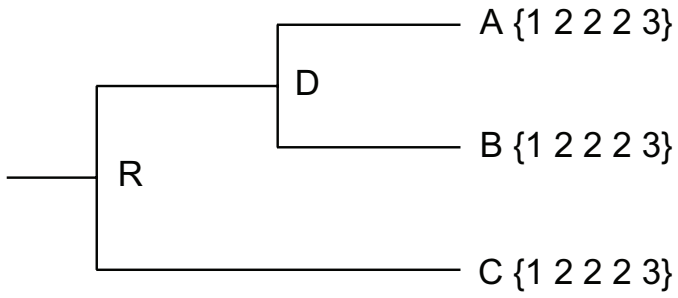

Figure 2.17: An example that INFER-CARS algorithm fails

In R, GET-PREDECESSOR-SUCCESSOR gives us $P_R(1) = \{0\}$, $P_R(2) = \{1, 2\}$, $P_R(3) = \{2\}$, $S_R(1) = \{2\}$, $S_R(2) = \{2, 3\}$, $S_R(3) = \{0\}$. In the intersection of G_R^P and G_R^S, there is an undesired self-cycle on 2. Obviously, the correct answer should be (1 2 2 2 3).

A plausible solution is to determine when the duplication happened.

1. If a duplication happened after speciation, then it is lineage-specific. We can remove the target copy of the duplication and retain the source copy, since the target copy is not present in the ancestor.

2. If a duplication happened before speciation, then both copies should be present in the ancestor. We then determine the 1-to-1 orthologous relations and reassign the numbers to elements. In other words, we regard paralogous copies as different elements. Only the orthologous elements between two genomes will have the same number. In the above example, the duplication happened before speciation and we assign the ortholog relations, yielding 2, $2'$ and $2''$ in all the genomes.

Chapter 3

Constructing conserved segments

3.1 Previous work

Identifying the genomic content that signed permutations can represent has always been an essential problem in studying genome rearrangements. Nadeau and Taylor (1984) first introduced the term "conserved segment" to name a genomic segment with preserved gene orders that are not disrupted by rearrangements between species. In the past decade, using comparative gene mapping to find orthologous gene loci as the evolutionary markers played an important role in testing algorithms and understanding rearrangement scenarios. However, although this approach works well in small genomes, e.g. virus genomes (Hannenhalli *et al.*, 1995) or mitochondrial genomes (Blanchette *et al.*, 1999), reliable gene annotation and orthology assignment in the entire mammalian genome are technically difficult to achieve (Sankoff and Nadeau, 2003), partly because of the great number of duplicated genes existing in mammals. Also, the large non-coding regions throughout the whole genome further complicate this problem.

Pevzner and Tesler (2003) proposed that the GRIMM-Synteny algorithm partition the genomes into segments which tolerate a certain amount of local microrearrangements. They called these segments "synteny blocks". These synteny blocks were constructed based on high-scoring anchors in genomic alignments. For the pairwise case, the algorithm first finds a set of nonoverlapping anchors. Then a graph is formed using

anchors as nodes. Finally a clustering procedure identifies connected components in the anchor graph and clusters these anchors into pairwise synteny blocks. The multi-way synteny blocks are created based on the intersection of all combinations of two-way anchors, as discussed in Bourque *et al.* (2004). The GRIMM-Synteny algorithm improved the resolution and precision for whole genome rearrangement study. However, it still suffered from relatively low coverage of the genomes when more species, esp. distantly related species, are under consideration. For example, when the algorithm was used to construct synteny blocks for human, mouse, rat, and chicken, the coverage is 52% of human genome at 100Kb resolution (Bourque *et al.*, 2005). In Murphy *et al.* (2005), the GRIMM-Synteny algorithm covers only 48% of human genome at 120Kb resolution when human, mouse, rat, cat, dog, pig, and cow are included as descendant species.

For the purpose of high-resolution and high-coverage ancestral genome reconstruction, we developed an approach relying on whole genome alignment nets (Kent *et al.*, 2003) to partition the genomes into putative orthology blocks. Conserved segments are created by postprocessing these orthology blocks. Moreover, when making orthology blocks, we treat outgroup species differently than descendant species in order to further increase the coverage.

3.2 Alignment chains and nets

To predict segments of the ancestral genome, we start with pairwise alignment nets (Kent *et al.*, 2003), downloaded from the UCSC Human Genome Browser (Kent *et al.*, 2002) (http://genome.ucsc.edu/). Nets are created from chains. Chains are derived from Blastz alignments (Schwartz *et al.*, 2003) and the chaining algorithm (Zhang *et al.*,

1994). The Blastz alignments are chained together to incorporate large gaps. These gaps can be in either species or simultaneously in both species. Moreover, chains skip over inversions, transpositions, and duplications. Therefore, chains can represent widely scattered sequences in the two modern species that are descended from the same region in the common ancestor without large rearrangements.

Before making nets, chains are sorted according to descending chain score. Then a program tries to cover the whole genome using these sorted chains one at a time, throwing out part of the current chain as necessary to fit into regions not already covered by higher-scoring chains. This process is performed until no more chains can be added to the resulting set covering the genome. If a chain covers bases that are in a gap in a previously taken chain, it is annotated as a lower level chain of the previous chain. Finally, the hierarchy structure of chains creates nets. Thus, a net identifies putative orthologous genomic segments between two genomes, in each of which no large-scale rearrangements occurred since the last common ancestor. This structure minimizes the difficulties caused by interspersed repeats, retroposons, and microrearrangements in identifying conserved regions.

3.3 Orthology Blocks and Conserved Segments

We developed a program to process the nets. We split nets if necessary to guarantee that they never contain an indel (insertion or deletion) of length exceeding a chosen threshold, e.g. 50Kb. Based on nets, we progressively construct sets of genomic intervals called, respectively, *orthology blocks*, *conserved segments*, and eventually *CARs*. Each set

contains pairwise orthologous genomic intervals, one from each species under consideration, which for this study means human (Lander *et al.* (2001); build hg18, March 2006), mouse (Waterston *et al.* (2002); build mm8, Feb. 2006), rat (Gibbs *et al.* (2004); build rn3, June 2003), and dog (Lindblad-Toh *et al.* (2005); build canFam2, May 2005). Since the intervals in a given set are orthologous, the set corresponds to a genomic interval in the last common ancestor of those species. We categorized the set according to restrictions on the kinds of large-scale evolutionary operations predicted to have happened in the lineages leading to the modern species, as summarized in Table 3.1.

Name	Species	From ancestor to descendants
net	2	No large rearrangements or indels
orthology block	N	No large rearrangements or indels
conserved segment	N	No large rearrangements
contiguous ancestral region (CAR)	N	Arbitrary rearrangements or indels

Table 3.1: Types of orthologous-interval sets discussed in this dissertation

The methods for constructing orthology blocks and conserved segments are illustrated in Figure 3.2. Figure 3.2(A) shows human-mouse nets. Four mouse intervals are depicted, as ordered and oriented by the orthologous human segments. The second and third mouse intervals are actually adjacent (and appropriately orientated) on a mouse chromosome, and the intervening bases, if any, do not align to human; this is depicted by a thin line connecting the representations of those intervals.

Figure 3.2(B) shows the human-mouse, human-rat and human-dog nets for a segment of the human sequence, and illustrates the creation of orthology blocks. A

dashed line between orthology blocks lies half-way between two intervals that are adjacent relative to human. For instance, the line at human position 2 in Figure 3.2 is midway between two mouse intervals. When the gap between two adjacent intervals in one species overlaps a gap relative to another species, as at position 1, we use the point half-way between the larger of the interval end-points to the left and the smaller of the end-points on the right. We discard all orthology blocks that cover less than 50Kb of human, because experiments showed that they tend to be unreliable (e.g., aligned segments are not always clearly orthologous). For this reason we describe our orthology blocks as having "50-Kb resolution".

Since duplications occurred after speciation produced paralogous sequences that are not derived directly from the ancestral genome, we also filter out large lineage-specific duplicated segments to simplify the reconstruction process. Since we use human as the reference species and we assume the pairwise net retains reciprocal-best alignments, in practice we remove duplications on two intervals of the human lineage. On the evolutionary branch leading from the human-rodent ancestor to human, if two segments A and B in the human genome are aligned to the same regions C in mouse, rat, dog, we use the flanking regions of A and B to infer the original copy. Assume the alignment between A and C stops at the endpoints of A, while the alignment between B and C extends into B's flanking regions; then we keep B as the original copy and remove the duplicated copy, A. See Figure 3.1 for an illustration. On the lineage leading from the human-dog ancestor to the human-rodent ancestor, we do a similar analysis. Application of this process created 3171 genomic intervals, which include 92.36% of the available human genome sequence.

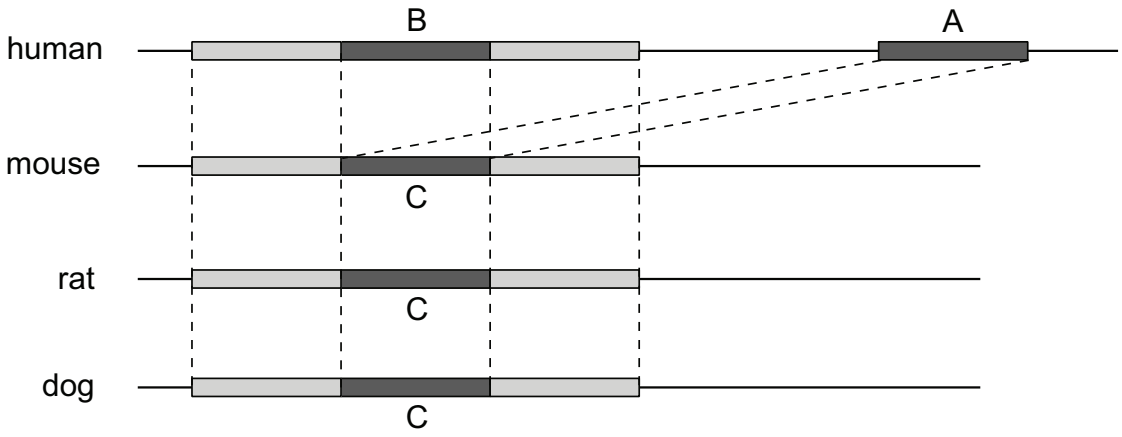

Figure 3.1: A human duplication happened after human-rodent ancestor

As shown in Figure 3.2(C), we fuse runs of consecutive orthology blocks whenever the order and orientation of these blocks are conserved in each of the contemporary genomes. In terms of the convention used in Figure 3.2(A), this means that for all non-human species, the boundary between the blocks is crossed either by a net or by a thin line. The results of the fusion process are independent of the order that fusions are performed. We call each resulting union of blocks a conserved segment. The concept of conserved segment here is slightly different from what was proposed by Nadeau and Taylor (1984) because we are dealing with the whole genomic sequences instead of particular genes.

We found 1338 conserved segments, each containing an average of around 2.4 orthology blocks, which include 94.81% of the available human genome sequence. To help the process of inferring adjacencies between conserved segments in the ancestral sequence, we add nets from outgroup species to the conserved segments (Fig. 3.2(C)). The intervals in a conserved segment from an outgroup species are not required to be

consecutive on the same chromosome. In this way, we increase the coverage of descendant species and also take into account the adjacency information from the outgroup species.

In the implementation of the above approach, both nets and chains are used extensively. In summary, the procedure can be described as:

- Get pairwise nets longer than 50Kb, using human as the reference.
- Partition the descendant genomes into blocks according to the pairwise nets. In this step, chains are used to get precisely aligned bases.
- Remove blocks that appear to be the target copy of duplication on human lineage. Also remove blocks that are smaller than 50Kb after the partitioning process.
- Add outgroup species to the blocks. The set of orthology blocks is thus created.
- If the order and orientation of the blocks are conserved across the descendant species, fuse them into conserved segments.

Figure 3.3 shows the length distributions of orthology blocks and conserved segments across the whole genome.

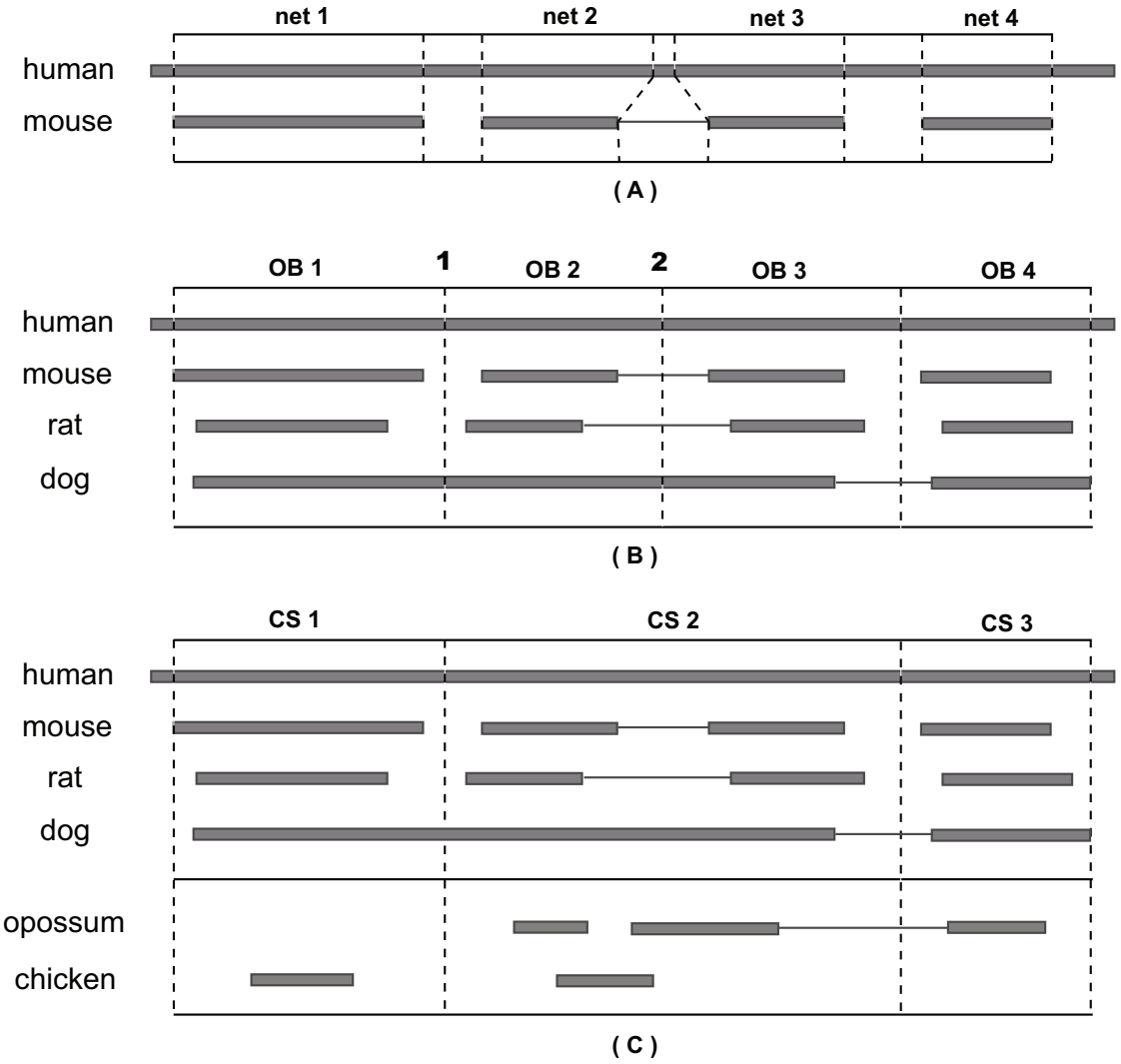

Figure 3.2: Nets, ortholog blocks, and conserved segments. (A) Nets. Human is the reference species. The line between intervals indicates that a genomic interval of zero or more unaligned bases exists in the non-reference species between the adjacent intervals (see text). (B) Orthology Blocks. (C) Conserved Segments, including outgroup nets. The order and orientation of OB2 and OB3 are conserved in all four species, so we merge them into a conserved segment.

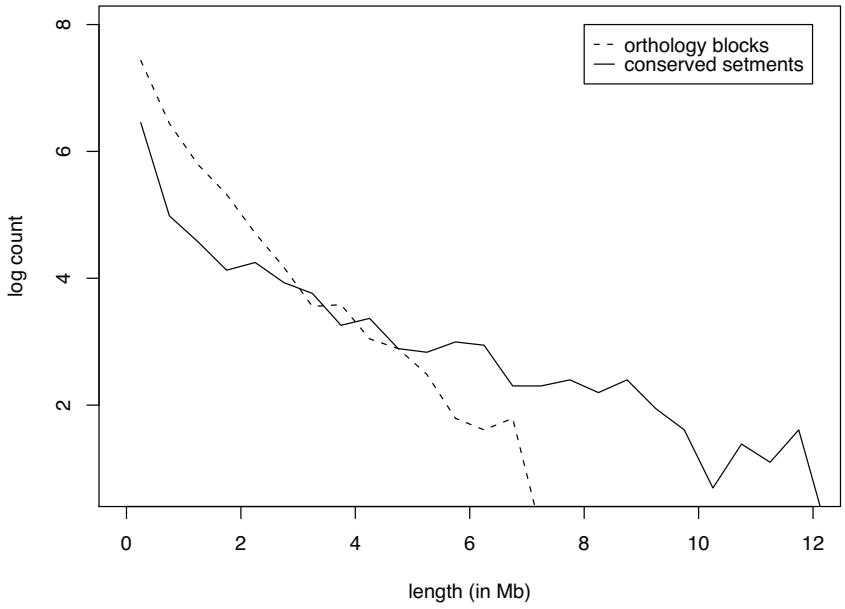

Figure 3.3: Length distribution of orthology blocks and conserved segments. Both orthology blocks and conserved segments are grouped into bins of 500Kb. Counts scaled by natural logarithm are plotted against lengths (in Mb).

Chapter 4

CARs in the Boreoeutherian ancestor

4.1 Overview

Using computer simulations, Blanchette *et al.* (2004a) showed that the genome sequence of the Boreoeutherian ancestor (Fig. 1.3) can be computationally predicted at high accuracy within most euchromatic intervals that are free of large-scale rearrangements, given adequate data from living mammals. For instance, when sequences from 20 appropriately chosen mammalian species are available, we expect that over 98% of the reconstructed nucleotides will be identical to the corresponding ancestral base. Because all mammals have experienced large-scale genomic rearrangements since their last common ancestor, in order to determine regional correspondence we analyze these rearrangements to infer a partition of each genome into intervals where nucleotide-level reconstruction methods can be applied.

To predict large-scale relationships among modern and ancestral genomes with sufficient accuracy for our needs, we have applied our methods discussed in Chapter 2 and Chapter 3. Conserved genomic segments were identified directly from freely available data, and we chose 50Kb as the threshold for making conserved segments. Then they were analyzed by a new computer program, Infer-Cars, which is based largely on the algorithm in Chapter 2. We also estimated the accuracy of our results, compared them with published analysis, and explore the biological properties of rearranged regions. The

computer software described herein and details of our predictions for human, mouse, rat, and dog are freely available at http://www.bx.psu.edu/miller_lab/car/.

4.2 The karyotype of Boreoeutherian ancestor

We found 29 CARs in total from the data we used. If we add the human sequence between conserved segments that are adjacent in both human and the ancestor (since a nucleotide-level reconstruction can include those intervals), 96.8% of the available human genome sequence is included. In Figure 4.1 we also use some chromosome painting results to combine CARs (leaving gaps in the figure) into our prediction of the genome structure of the Boreoeutherian ancestor. In Figure 4.1, black tick marks indicate joins with relatively weak support. For example, the left-most tick mark on CAR 1 (which corresponds to human chr21 and chr3) shows a predicted ancestral adjacency between two conserved segments (conserved segments 238 and 239 in the on-line materials) that are adjacent in human, mouse, and rat, but not in dog and the outgroups. Numbers above bars indicate the corresponding human chromosomes. Black tick marks below the bars indicate ambiguous joins (Fig. 2.4 in Chapter 2). Our predicted CARs are colored and ordered to facilitate comparison with Froenicke *et al.* (2006). Gaps between CARs are joins suggested by Froenicke *et al.* (2006). Diagonal lines within each block show the orientation and position in the human chromosome (Bourque *et al.*, 2006).

Table 4.1 shows the number of conserved segments involved in each of 29 CARs, lengths of each CAR, and corresponding parts in mouse, rat, and dog, which tells us how each contiguous ancestral region has been scattered among chromosomes in human, mouse, rat, and dog.

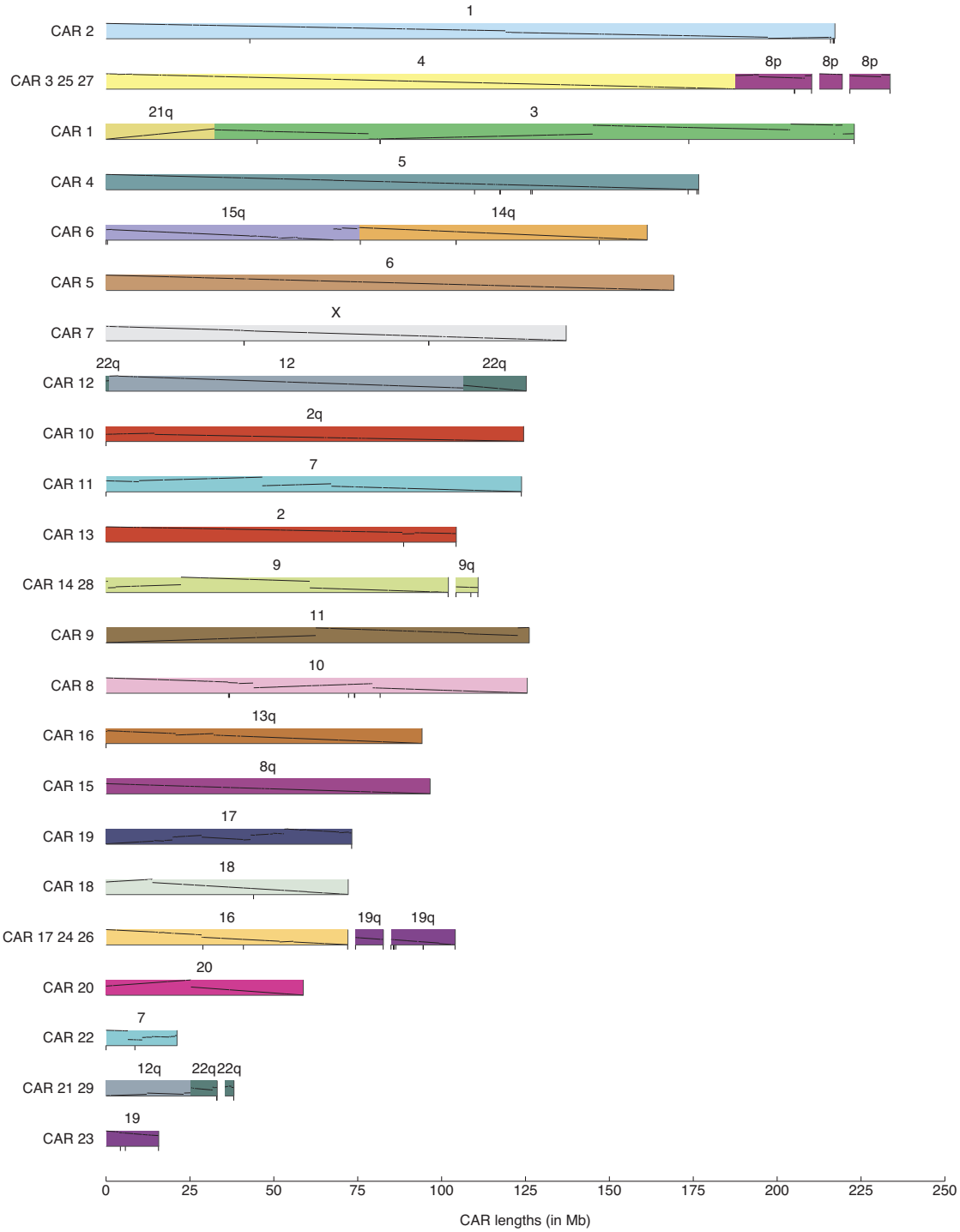

Figure 4.1: Map of the Boreoeutherian ancestral genome.

CAR	Bases covered (Mb)	# of conserved segments	Human	Mouse	Rat	Dog
1	225.16	111	21 3	10 17 16 3 9 14 11 6	20 11 2 8 16 9 15 4	31 33 23 34 20
2	219.22	98	1	4 6 3 5 1 13 8 11 7	5 4 2 14 13 17 19 10 1	5 2 9 15 6 17 7 38 4 14 8 16
3	208.27	96	4 8	5 6 3 8	14 4 2 19 16	3 13 6 15 32 19 25 16
4	175.49	96	5	13 15 17 1 18 11	1 17 2 9 18 10	34 4 2 3 11
5	166.60	76	6	13 17 14 1 9 4 10	17 20 15 9 8 5 1	35 12 1
6	161.57	68	15 14	2 9 7 14 12	3 8 1 15 6	30 13 3 15 8
7	140.61	90	X	X	X 15	X
8	128.75	70	10	13 2 18 8 6 14 10 19 7	17 19 4 16 15 20 1	2 4 28 26
9	126.79	50	11	9 7 2 19	8 7 1 3	5 21 18
10	125.76	45	2	1 18 2	9 18 13 3	19 36 37 25
11	124.12	66	7	11 6 13 9 5 12	14 4 17 8 6	16 18 14
12	123.57	53	22 12	6 16 15 10 8	4 11 7 19	27 3 10 15
13	106.48	49	2	12 5 17 11 6 2 1 10	6 14 4 3 9 20	17 10
14	101.40	40	9	13 4 19 2	17 5 1 3	1 11 9
15	97.05	38	8	16 1 4 3 13 15	11 5 2 7	29 13
16	93.57	54	13	14 5 3 8 1	15 12 2 16 9	25 22
17	75.69	41	16	11 17 16 7 8	10 1 19	6 15 2 5
18	73.14	35	18	18 17 5 1	18 9 13	1 7
19	72.91	47	17	11	10	9 5
20	58.45	11	20	2	3	23 24
21	33.58	16	12 22	5 11 10	12 19 14 20	26
22	26.21	24	7	5	12	6
23	22.27	22	19	7	1	1
24	19.40	15	19	10 8 17 9	7 12 8 19 16	20
25	11.58	3	8	8 14	16 15	25
26	8.08	5	19	7	1	1
27	6.84	9	8	8	16	37 16
28	6.25	4	9	13	17	1
29	2.71	6	22	16 10	11 20	26

Table 4.1: Number of conserved segments involved in each of 29 CARs

In our reconstruction, we first infer the predecessor set for the human-chicken common ancestor. Then we used it to adjust the human-opossum common ancestor. And finally, we used the human-opossum ancestor to adjust the human-dog common ancestor. In order to estimate the breakages on each lineage, we also reconstructed the intermediate ancestral genomes, i.e. the rodent ancestor and human-rodent ancestor. Boreoeutherian adjacencies were propagated to the intermediate ancestors. Based on that, we were able to estimate the number of chromosomal breakages that happened on each of the lineages. See Figure 4.2 for details. We categorized the breakages into inter-chromosomal breakages and intrachromosomal breakages. If conserved segments i and j are adjacent in the ancestral genome but not in the descendant genome, then we call the break interchromosomal if i and j are on different chromosomes in the descendant, and intrachromosomal otherwise. We suspect that many of the predicted intrachromosomal breaks in rat are assembly artifacts.

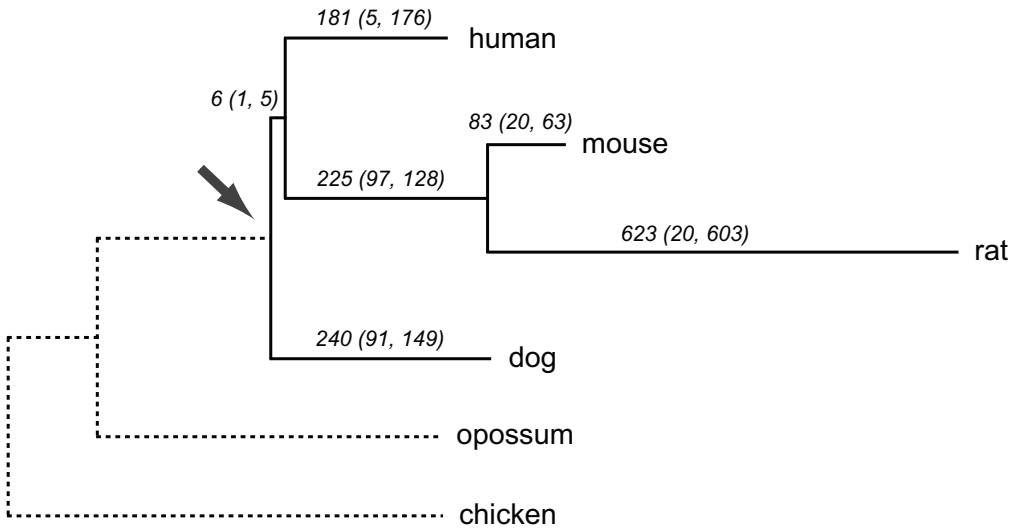

Figure 4.2: Estimated number of chromosomal breakages on each lineage. Breakages are categorized as (interchromosomal, intrachromosomal)

4.3 Reliability of predicted ancestral adjacencies

If an ancestral adjacency is unambiguously determined by our data, we refer to these adjacencies as *strongly supported*. Otherwise, we measure the reliability of predicted ancestral adjacencies according to Equation 2.5, and we call them *weakly supported*. Among 1367 ancestral adjacencies, 77 are weakly supported (only 5.7%). Also, 24 of these weakly supported joins are actually supported by human, mouse, and rat, which are very likely to be the ancestral state in the Boreoeutherian ancestor. We need more outgroup information to confirm the data. A detailed classification can be found in Table 4.2. These ancestral adjacencies with low reliability have been identified in Figure 4.1 with black tick marks.

supported by species	# of adjacencies
human, mouse, and rat	24
human, mouse/rat	5
mouse and rat	3
human	22
dog	3
none	20
total	77

Table 4.2: Weakly supported ancestral adjacencies. The 20 weakly supported adjacencies that appear in none of the species are all endpoint adjacencies.

In fact, all the inconsistent joins between our prediction and other publications we have found so far are weakly supported by our data. For example, our prediction shows that Hsa16 was conserved as a whole chromosome in the ancestor. However, some publication supports that Hsa16 was split in two pieces (16q and 16p) in the eutherian

ancestor (Murphy *et al.*, 2001; Froenicke *et al.*, 2006). Furthermore, Misceo *et al.* (2003) hypothesized that chr16 was separated into two parts in the primate ancestor and then were fused before the Catarrhini ancestor, i.e. the common ancestor of human and old world monkeys. In our result, the first weakly supported join on CAR17 actually corresponds to the split on Hsa16 that was reported by another study (the centromere of Hsa16), suggesting that there could be other possible configurations.

4.4 Identification of small inversions

Within each conserved segment, we also identified in-place inversions that are too small to create a new orthology block. Because we currently lack a good outgroup species (such as a well-assembled elephant or armadillo genome sequence), it was frequently difficult to confidently predict ancestral orientation. In ambiguous cases, we assumed that human is in the ancestral orientation relative to the immediately flanking regions, in part because the human assembly is more accurate than the others. However, this means that inversions in the human lineage are currently under-estimated.

Using human as the reference, we look at the pairwise alignments between human and other species. The inference method is also similar to Fitch's parsimony algorithm. But instead of using nucleotide or amino acid characters, we use notations `+`, `-`, and `?`, where `+` represent positive orientation, `-` represents negative orientation, `?` indicates the gap, i.e. there is no alignment for that species. We then recursively determine the ancestral orientation of each microinversion segment. As mentioned before, we assume human has the ancestral orientation if ambiguity exists. The last step is to place the inversions onto specific lineages.

This inference method assigns 856 inversions to human, 3210 to mouse, 3067 to rat, and 4924 to dog. Among the 4924 inversions assigned to the dog lineage, only 703 were confirmed by an outgroup (e.g., opossum agreed with human); many of the remaining 4221 will be resolved by better outgroup data, e.g., from elephant. Figure 4.3 shows the length distributions of observed inversions assigned to each species. Among human inversions, 29 were assigned to the short branch leading from the Boreoeutherian ancestor to the human-rodent ancestor. The shortest inversions we found are 87 bp (in human), 31 bp (in mouse), 36 bp (in rat), and 34 bp (in dog). Figure 4.4 gives a detailed map of CAR 16, including small inversions. The detailed coordinates of these inversions and the tools for identifying them are freely available from http://www.bx.psu.edu/miller_lab/car/.

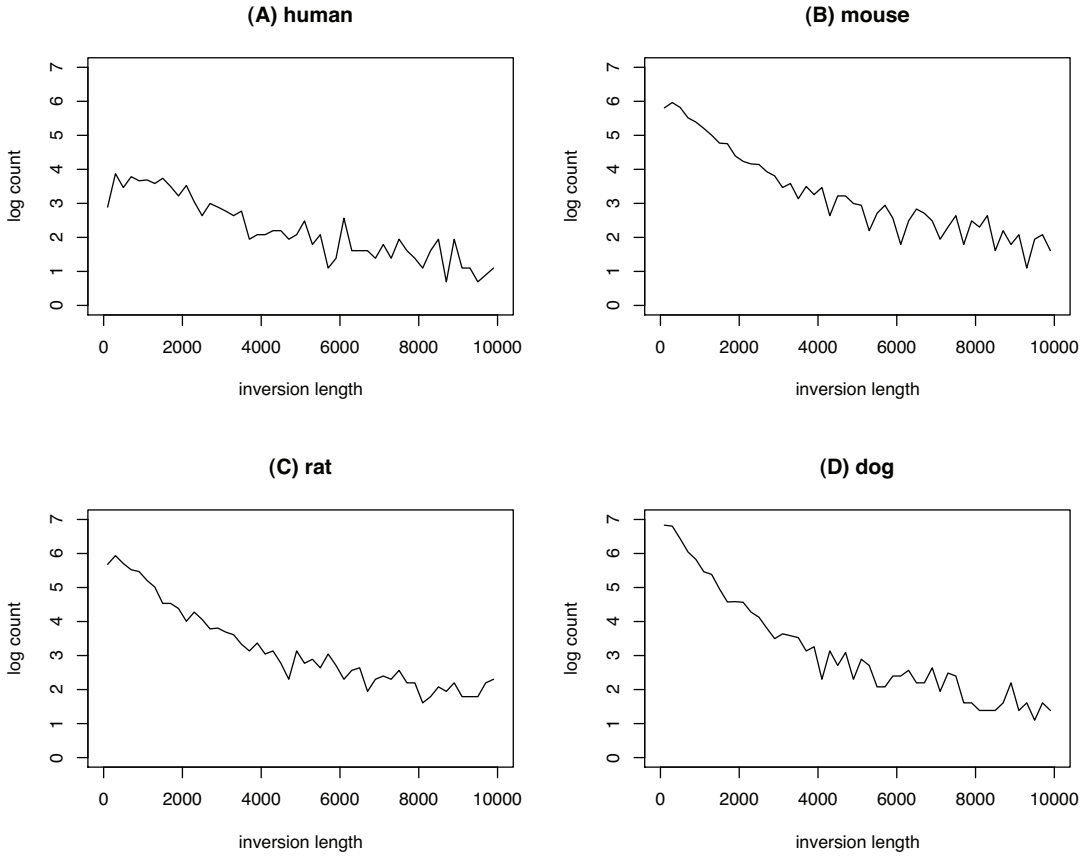

Figure 4.3: Length distribution of predicted inversions. Inversions of lengths exceeding 10Kb are not represented in the plots. Inversions lengths are grouped into bins of 250 bp.

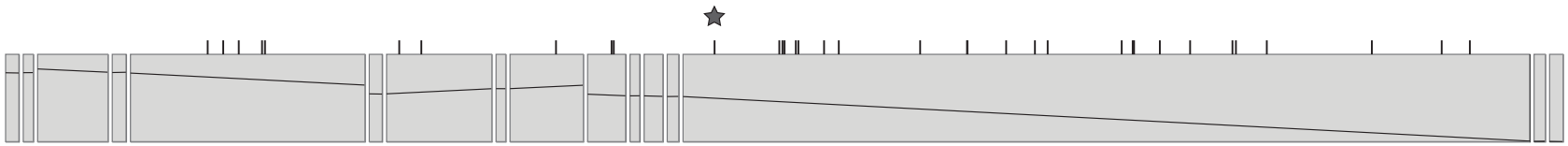

Figure 4.4: Detailed map of human chromosome 13q onto CAR 16. In CAR16, there are 16 large-scale rearrangement-free pieces. In the picture, vertical lines above each piece indicate positions of small in-place inversions. The star indicates an inversion (around *hg18.chr13:57,380,591-57,383,765*) that happened on the branch from the Boreoeutherian ancestor to the human-rodent ancestor. See Figure 4.5.

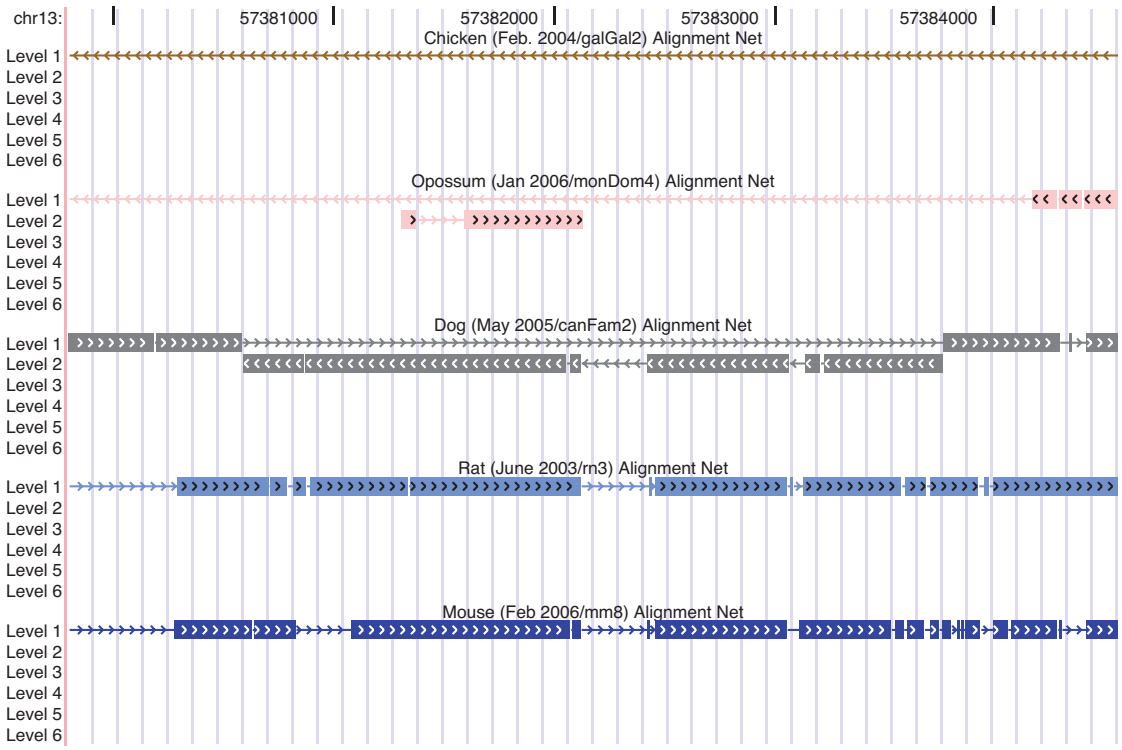

Figure 4.5: A micro-inversion that happened on the branch from the Boreoeutherian ancestor to the human-rodent ancestor

4.5 Properties in breakpoint regions

Of 1309 pairs of conserved segments that were predicted to be adjacent in the Boreoeutherian ancestor, 149 (11%) were separated by events in at least two independent lineages (12 were separated in three lineages). When we omit rat (because of the potential assembly problems indicated in Figure 4.2), we find breakpoint re-use for 57 of 742 (8%). The ratio of breakpoint reuse we found is lower than what was reported in Murphy *et al.* (2005). One reason is that we use higher resolution to partition the genomes. Another reason is that we only use four Boreoeutherian descendant species in our study. Some of

the breakpoints identified using dog and rodent might be reused in other Boreoeutherian descendants.

Simulations, described in the next section, suggest that this frequency of breakpoint reuse is approximately what one would expect if breakage was equally likely for every genomic position, but a careful analysis is beyond the scope of this study. See Peng *et al.* (2006), Sankoff (2006), and references therein for an introduction to the long-standing debate about validity of the uniform breakage model.

We inspected intervals around breakpoints in the human sequence, looking for properties that might help explain why breaks occur at some positions but not others. We used 50Kb intervals centered on the end of a conserved segment where the adjacent segments in the ancestor and human differ. Breaks that occurred only in the human lineage were treated separately from those that were reused in another species, giving 16.97 Mb of human-specific breakpoint regions and 11.96 Mb around reused breakpoints. For small inversions in human, we used 1Kb intervals centered on each endpoint, covering 1.60 Mb. Our observations are summarized in Table 4.3. GC content around breakpoints is slightly higher than the genome average, but not as elevated as reported for dog chromosomal breaks by Webber and Ponting (2005)). The breakpoint regions are substantially enriched for RefSeq genes, consistent with what Murphy *et al.* (2005) observed in larger (about 1 Mb) regions around breakpoints. The density of SINEs is also much higher than average. Finally, we observed that a large amount of DNA (41.72%) in human-specific breakpoint regions is in human segmental duplications (Bailey *et al.*, 2001).

	Human-specific breakpoints	Reused breakpoints	Human short inversions	Genome-wide average
GC content (%)	43.58	42.81	39.82	40.91
Segmental Duplication (%)	41.72	17.89	3.94	5.24
Gene density (genes/Mb)	20.45	21.99	–	8.58
Repeats density (%)	52.88	54.30	44.23	48.58
SINE	18.58	16.60	14.30	13.60
LINE	19.97	23.04	16.87	21.32
LTR	9.15	9.92	7.53	8.61
DNA	2.48	2.74	3.56	3.01
Others	2.70	2.00	1.97	2.04

Table 4.3: Genomic content of breakpoint regions. Gene density is not given for inversions, because we used regions too short to give meaningful results. Repeats were identified with RepeatMasker, and segmental duplications were obtained from the UCSC Genome Browser segmental duplication track.

4.6 Evaluation

Computer simulation of sequence evolution based on certain substitution models has been successfully used to evaluate different multiple sequence alignment programs (Blanchette *et al.*, 2004b; Ovcharenko *et al.*, 2005). The program simulates neutral evolutionary processes starting from a hypothetical ancestral sequence. It records the true relationship among the generated sequences, which can be regarded as the true alignment. Finally, the agreement for alignments produced by alignment programs can be measured. Since the program keeps track of the changes between ancestor and descendants, it was also used to evaluate the performance of nucleotide-level reconstruction (Blanchette *et al.*, 2004a).

We used a similar kind of computer simulations to inject realism into the analysis. Instead of simulating nucleotide substitutions and indels, we evolved the hypothetical genome with rearrangements. However, the lack of a well-founded model and theory of large-scale genome evolution makes the simulation more difficult.

Here, we regard genomes as signed permutations and evolve these numbers by rearrangements. At the beginning, all parameters in the simulation program were set purely based on empirical estimation. Then these parameters were further tuned according to our observed data.

We employed a realistic evolutionary tree with branch lengths based on substitution frequencies. Starting with a hypothetical human-chicken ancestor having 6000 orthology blocks and 25 chromosomes, we simulated inversions, translocations, fusions and fissions along each branch. Rearrangements were distributed as 90% inversions, 5% translocations, 3.75% fusions, and 1.25% fissions. We modeled lengths of inverted blocks with a Gamma distribution, with shape and scale parameters $\alpha = 0.7$ and $\theta = 500$, respectively. Since we required inversion lengths not to exceed 50 blocks, we truncated the Gamma distribution at 50 (probabilities for shorter lengths are renormalized). In addition, we also allowed each branch to have its own adjustment parameter for each operation, to account for the differences among branches.

The simulated genomes produced by this approach are consistent with actual mammalian genomes in terms of number of conserved segments, number of breakpoints, chromosome count etc. Some important features of the simulated data are compared with what were seen in real data in Table 4.4.

	# of conserved segments	breakpoint distance			% of breakpoint reuse
		H-M	H-R	H-D	
Real data	1338	564.5	1059	452	11.38
Simulated data	1375.64 (39.87)	559.02 (30.31)	1038.39 (37.94)	434.72 (25.24)	11.74 (0.92)

Table 4.4: Comparison between our simulated data and real data. Simulated statistics are the average from 50 simulated datasets. The standard deviations of numbers in simulated data are in the parentheses.

For the species shown in Figure 4.2 we repeated the simulation 50 times, in each case running our program for inferring CARs on the resulting dataset and comparing the predicted adjacencies with the known (simulated) ones. For determining the success rate, we considered only the ancestral joins that were broken in at least one lineage, since the unbroken joins will be found by essentially any procedure.

Using human, mouse, rat and dog, with opossum and chicken as outgroups (Fig. 4.2), the frequency of correctly predicted adjacencies was 98.96% (SD=0.39) for the Boreoeutherian ancestor, 98.37% (SD=0.55) for the human-rodent ancestor, and 97.07% (SD=1.01) for the mouse-rat ancestor. Note that as for inference of nucleotides Blanchette *et al.* (2004a), the prediction accuracy is higher for the Boreoeutherian ancestor than for some younger ancient genomes.

We also reconstructed the Boreoeutherian ancestor without using opossum and chicken; the accuracy decreased to 97.40% (SD=0.69). If we retain the outgroups but leave out rat, the accuracy drops to 98.29% (SD=0.67). However, if chimp, cow, and macaque are included in the reconstruction, the simulation indicates that joins in the Boreoeutherian ancestor are computed with 99.34% (SD=0.29) accuracy.

4.7 Comparison with other reconstructions

A comparison with other reconstructions identifies which part of the ancestral genome can be confidently reconstructed, and highlights regions where further investigation is needed.

Our predicted CARs agree well with predictions from chromosome painting with respect to interchromosomal operations (compare Figure 1 in Froenicke *et al.* (2006) and our Figure 4.1). Of 8 strongly supported interchromosomal breaks in human predicted by Froenicke *et al.* (2006), our reconstruction agrees with 5, see Table 4.5. In the prediction of Murphy *et al.* (2005), joins of Hsa16q/Hsa19q and Hsa16p/Hsa7 were reconstructed with weak support, using data from cat and pig, and cat and cow, respectively, all of which were unavailable for computing CARs. Also, in both cases, the species (cat and pig or cat and cow) come from the same superorder, so they did not provide strong evidence as to the Boreoeutherian ancestral state. Therefore, more species are needed. As discussed before, whether or not Hsa16 was separated into 16q and 16p in the Boreoeutherian ancestor is still questionable. The endpoints of the relevant conserved segments for these two joins are all connected with ambiguity in our prediction, suggesting other possible scenarios which could be joins of Hsa16q/Hsa19q and Hsa16p/Hsa7. The join of Hsa16p/Hsa7 corresponds to the centromere of Hsa16 (the first weakly supported join on CAR17) and the left side of CAR22 in the picture. We think the small sample size of Boreoeutherian descendant genomes in the current study is likely responsible for the failure to confirm these adjacencies.

The predictions of Murphy *et al.* (2005), which were mainly based on the MGR algorithm (Bourque and Pevzner 2002), are more or less similar to Froenicke *et al.* (2006)'s

Comparison	Froenicke et al. 2006	Murphy et al. 2005	Our method	Included in Fig.4.1	Comments
# of species	> 80	8	4		We also used 2 outgroups
Coverage of human genome	–	48%	96%		
Resolution	4Mb	120Kb	50Kb		
inter/intra	only inter	both	both		
Hsa4a/Hsa8p	+	+ (weak)	+ (strong)	+	Join (379,-653). Supported by mouse,rat,dog, chicken
Hsa4b/Hsa8p	+	+ (weak)	-	-	
Hsa21/Hsa3	+	+ (strong)	+ (strong)	+	Join (-1212,229). Supported by mouse,rat,dog, chicken
Hsa15/Hsa14	+	+ (weak)	+ (strong)	+	Join (-994,968). Supported by mouse,rat,dog, opossum
Hsa10p/Hsa12a	+ (weak)	-	-	-	
Hsa12a/Hsa22a	+	-	+ (strong)	+	Join (909, 1239). Supported by mouse,rat,dog, opossum,chicken
Hsa12b/Hsa22b	+	+ (weak)	+ (strong)	+	Join (-1231,910). Supported by mouse,rat,dog
Hsa16q/Hsa19q	+	+ (weak)	-	+	
Hsa7b/Hsa16p	+	+ (weak)	-	-	
Hsa1/Hsa22a	-	+ (weak)	-	-	
Hsa5/Hsa19p	-	+ (weak)	-	-	
Hsa2pq/Hsa18	-	+ (weak)	-	-	
Hsa1q/Hsa10q	-	+ (weak)	-	-	
Hsa20/Hsa2	-	+ (weak)	-	-	

Table 4.5: Comparison with Froenicke *et al.* (2006) and Murphy *et al.* (2005). The naming convention of human chromosomes regions follows Figure 1 in Froenicke et al. 2006. + indicates the method made that join, - otherwise. If the join was made by our method, we also list the corresponding conserved segment numbers and show which species support that join.

result. However, five of Murphy *et al.* (2005)'s putative weakly supported interchromosomal breaks (Hsa1/Hsa22, Hsa5/Hsa19, Hsa2/Hsa18, Hsa1/Hsa10, Hsa2/Hsa20) are not supported by chromosome painting data. Our program made none of these joins, which is in agreement with Froenicke *et al.* (2006).

To further examine the differences among reconstruction algorithms, we ran our CAR-building program on Murphy *et al.* (2005)'s dataset with 307 conserved segments from their supplementary materials. Table 4.6 compares our predicted ancestral joins with theirs.

	Total number of joins based on Murphy et al.'s data	Two methods agree
All types of joins	338	287
All non-endpoint joins	276	255
Strongly supported non-endpoint joins	246	242
Weakly supported non-endpoint joins	30	13
All endpoint joins	62	32
Strongly supported endpoint joins	37	29
Weakly supported endpoint joins	25	3
Non-human-consecutive joins	9	7

Table 4.6: Results of running our CAR-building program on data from Murphy *et al.* (2005). Endpoint joins have two cases: (1) the join between the first conserved segment in a chromosome and the beginning of the chromosome; (2) the join between the last conserved segment in a chromosome and the end of the chromosome. Non-human-consecutive joins refer to two consecutive elements in human that are not consecutive in the ancestor, indicating a breakpoint in human.

Chapter 5

Reconstructing CARs in a likelihood framework

5.1 Introduction

The ancestral adjacency inference algorithm proposed in Chapter 2 is based on the parsimony of predecessor and successor changes through evolution. Although the application of the algorithm to real data seems to be satisfactory (Chapter 4), the model itself is by no means very extensible. In this chapter, we introduce a probabilistic method for reconstructing ancestral order. The essential part of this method is to predict the posterior probability of an adjacency occurring in the ancestor based on an extended Jukes-Cantor model for breakpoints. This is our initial attempt trying to extend the model to handle more sophisticated evolutionary operations.

Recall the definitions of predecessor and successor. If modern genome g contains element i, then the predecessor $p_g(i)$ is defined as the signed element that immediately precedes i on the same chromosome relative to the original orientation. In the opposite orientation, $p_g(-i)$ immediately precedes $-i$ in the reverse complement of the same chromosome. We set $p_g(i) = \phi_A$ if i appears first on a chromosome. The successor $s_g(i)$ of i is defined analogously, setting $s_g(i) = \phi_Z$ if i appears last on a chromosome.

Now we define the adjacency based on $p_g(i)$ and $s_g(i)$. If $p_g(j) = i$ and $s_g(i) = j$, we call (i, j) an **adjacency** in g, i.e. $Adj_g(i, j) = 1$. Given a phylogenetic tree and the

set of chromosomes for each modern species, our goal is to determine a set of lists of signed elements that closely approximates the chromosomes of the species corresponding to the target ancestor in the tree.

Assume that the model for adjacency changes is reversible. If the target ancestral genome we want to reconstruct is not the root of a phylogenetic tree, we create a new tree by transforming the original tree to have the target ancestral genome as the root. See Figure 5.1 for example.

In Figure 5.1, A is the target ancestral genome. We reroot the tree by adding a new node A' on branch aA and make branch length $t_{A'a} = t_{aA}$ and $t_{A'A} = 0$. If the rearrangement rate is the same on each branch, which refers to parameter α in our following discussion, we can remove the node e and make $t_{af} = t_{fe} + t_{ea}$. But here we do not remove node e in the rerooted tree, considering that the rearrangement rates on ea and ef might be different. Rerooting the tree enables us to treat the outgroups in a symmetrical manner.

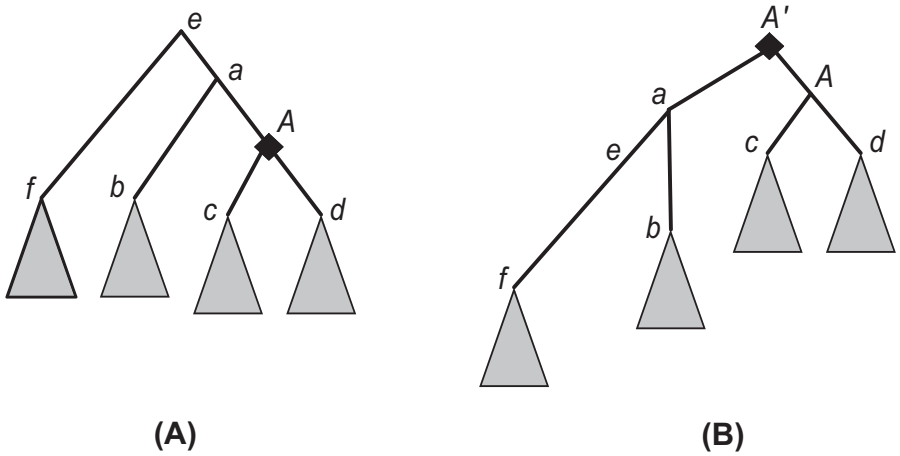

Figure 5.1: Reroot the tree. (a) is the original phylogenetic tree. A is the target ancestral genome. We can reroot the tree into (b) by adding a new node A'. In (b), branch length $t_{A'A} = 0$, $t_{A'a} = t_{aA}$

5.2 General framework

Suppose the parent genome is π and its two children are τ and φ. We want to measure the posterior probability that an ancestral configuration X appears in π, i.e. $P(X \text{ in } \pi|D_{\pi})$. Here X is an ancestral configuration that we want to reconstruct. D_{π} represents all the observed data in all leaves of the subtree rooted by π.

Suppose there are m mutually exclusive possible cases $X_1, X_2, \ldots, X_m$ for configuration X. We can calculate the posterior probability of each X_i occurring the ancestor, i.e. $P(X_i \text{ in } \pi|D_{\pi})$, using the Bayes' theorem:

$$P(X_i \text{ in } \pi|D_{\pi}) = \frac{P(D_{\pi}|X_i \text{ in } \pi)P(X_i \text{ in } \pi)}{\sum_{j=1}^{m} P(D_{\pi}|X_j \text{ in } \pi)P(X_j \text{ in } \pi)} \tag{5.1}$$

If we assume the prior probabilities $P(x_i \text{ in } \pi)$ are the same, then:

$$P(X_i \text{ in } \pi|D_{\pi}) = \frac{P(D_{\pi}|X_i \text{ in } \pi)}{\sum_{j=1}^{m} P(D_{\pi}|X_j \text{ in } \pi)} \tag{5.2}$$

The likelihood of the form $P(D_{\pi}|X_i \text{ in } \pi)$ can be calculated recursively in a post-order traversal fashion:

$$\begin{aligned} P(D_{\pi}|X_i \text{ in } \pi) &= P(D_{\tau}|X_i \text{ in } \pi)P(D_{\varphi}|X_i \text{ in } \pi) \\ &= \sum_{j=1}^{m} P(D_{\tau}|X_j \text{ in } \tau)P(X_j \text{ in } \tau|X_i \text{ in } \pi) \\ &\quad \times \sum_{k=1}^{m} P(D_{\varphi}|X_k \text{ in } \varphi)P(X_k \text{ in } \varphi|X_i \text{ in } \pi) \end{aligned} \tag{5.3}$$

where $P(X_j \text{ in } \tau | X_i \text{ in } \pi)$ represents the probability of changing from configuration X_i to X_j when evolving from π to τ. The base case, where π is the leaf, is:

$$P(D_\pi | X_i \text{ in } \pi) = \begin{cases} 1 & \text{if } X_i \text{ in leaf } \pi \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

Ideally, if configuration X represents the whole ancestral genome in π, then we can get the globally optimal reconstruction. However, it is generally difficult to achieve global optimum due to the high computational complexity. Our approach is to predict locally optimal configurations for predecessor and successor relationships in order to predict the possibility of each adjacency occurring in the ancestor. Finally, we use another procedure (same as FIND-CARS in Chapter 2) to figure out an approximately global optimal solution for the whole-genome reconstruction.

5.3 Extended Jukes-Cantor model for adjacency

For each element, we distinguish predecessor and successor relationships. We use the Extended Jukes-Cantor Model for breakpoints (Sankoff and Blanchette, 1999) to model predecessor and successor changes. The overall assumption is that a genome π with n elements will evolve through a series of rearrangement operations (e.g. inversion, translocation, fusion, fission) with unknown proportions.

We first consider the successor relationship. We assume that the successor of element f is changed from g to h over a time unit with the same probability α for all $h \neq f, -f, g$. Since f could be followed by ϕ_Z when it is the last element on a

chromosome, there are altogether $2n-2$ such changes possible ($2n+1-3=2n-2$). Also, the probability that g remains as the successor of f is $1-(2n-2)\alpha$.

Suppose π evolves into τ along a branch with time t. We use $P_i(s(f)=g)$ to denote the probability that $s(f)=g$, i.e. g is the successor of f after time i, for $g \neq f, -f$, $i=0,1,\ldots,t$. Then for any i, we have,

$$P_{i+1}(s(f)=g) = (1-(2n-2)\alpha)P_i(s(f)=g) + \alpha(1-P_i(s(f)=g)) \tag{5.5}$$

Equivalently,

$$P_{i+1}(s(f)=g) - P_i(s(f)=g) = \alpha - \alpha(2n-1)P_i(s(f)=g) \tag{5.6}$$

If we approximate the discrete-time process by a continuous model, we can rewrite the above equation as:

$$\frac{dP_y(s(f)=g)}{dy} = \alpha - \alpha(2n-1)P_y(s(f)=g) \tag{5.7}$$

We solve the above first-order linear differential equation, for any $0 \leq i \leq t$

$$P_i(s(f)=g) = \frac{1}{2n-1} + \left(P_0(s(f)=g) - \frac{1}{2n-1}\right) e^{-(2n-1)\alpha i} \tag{5.8}$$

Therefore, using $s_\pi(f)=g$ to denote the event that the successor of f is g in π, we have,

$$P(s_\tau(f)=g | s_\pi(f)=g) = \frac{1}{2n-1} + \frac{2n-2}{2n-1} e^{-(2n-1)\alpha t}, \tag{5.9}$$

$$\text{since } P_0(s(f) = g) = P(s_\pi(f) = g) = 1.$$

Similarly, for any $h \neq f, -f, g$ in genome τ,

$$P(s_\tau(f) = h | s_\pi(f) = g) = \frac{1}{2n-1} - \frac{1}{2n-1} e^{-(2n-1)\alpha t} \tag{5.10}$$

Therefore, along branch $\pi\tau$, the probability that the successor remains the same is:

$$P(s_\tau(f) = g | s_\pi(f) = g) = \frac{1}{2n-1} + \frac{2n-2}{2n-1} e^{-(2n-1)\alpha_{\pi\tau} t_{\pi\tau}} \tag{5.11}$$

The probability that the successor of f changed is

$$P(s_\tau(f) = h | s_\pi(f) = g) = \frac{1}{2n-1} - \frac{1}{2n-1} e^{-(2n-1)\alpha_{\pi\tau} t_{\pi\tau}} \tag{5.12}$$

where $h \neq f, -f, g$.

Similarly, if we assume the probability that the predecessor of f is changed from g to h over a time unit is also α. We also have two probabilities for predecessor changes along branch $\pi\tau$:

$$P(p_\tau(f) = g | p_\pi(f) = g) = \frac{1}{2n-1} + \frac{2n-2}{2n-1} e^{-(2n-1)\alpha_{\pi\tau} t_{\pi\tau}} \tag{5.13}$$

and

$$P(p_\tau(f) = h | p_\pi(f) = g) = \frac{1}{2n-1} - \frac{1}{2n-1} e^{-(2n-1)\alpha_{\pi\tau} t_{\pi\tau}} \tag{5.14}$$

where $h \neq f, -f, g$.

Since we do not model insertion and deletions, if an element f is not in leaf node τ, we imagine f is hidden in τ and is adjacent to an unknown element. At the last branch leading to the leaf node, we let $P(f \notin \tau | s_{\pi}(f) = g)$ equal formula (5.12), i.e.

$$P(f \notin \tau | s_{\pi}(f) = g) = \frac{1}{2n-1} - \frac{1}{2n-1} e^{-(2n-1)\alpha_{\pi\tau} t_{\pi\tau}} \tag{5.15}$$

where τ is leaf, $\pi = parent(\tau)$

Then we can calculate the posterior probabilities for each successor and predecessor relationships in the ancestral genome π according to equation (5.2), $P(p_{\pi}(g) = f | D_{\pi})$ and $P(s_{\pi}(f) = g | D_{\pi})$. Since $p_{\pi}(g) = f$ and $s_{\pi}(f) = g$ imply that $Adj_{\pi}(f, g) = 1$, the posterior probability of adjacency (f, g) is:

$$P(Adj_{\pi}(f,g) = 1 | D_{\pi}) = \begin{cases} P(p_{\pi}(g) = f | D_{\pi}) P(p_{\pi}(g) = f | D_{\pi}) & \text{if } f = \phi_A \\ P(s_{\pi}(f) = g | D_{\pi}) P(s_{\pi}(f) = g | D_{\pi}) & \text{if } g = \phi_Z \\ P(p_{\pi}(g) = f | D_{\pi}) P(s_{\pi}(f) = g | D_{\pi}) & \text{otherwise} \end{cases} \tag{5.16}$$

To make the computation more efficient, especially when there are many leaf genomes with large numbers of elements, we can consider possible predecessor or successor relationships in the ancestor only if the adjacency appears at least once in the leaf genomes. For instance, if there are n elements and s species, the possible cases of successor of element i in the ancestor are at most s. We can calculate only the posterior probability of these possible successor relationships associated with i to speed up the

computation of recursive steps in equation (5.2). In other words, if a successor relationship $s(f) = g$ does not appear in any of the leaf genomes, we let $P(s_\pi(f) = g|D_\pi) = 0$. The overall result will be sightly different from the full version, but still will achieve a good approximation.

5.4 From ancestral adjacency to ancestral order

Once we have the posterior probability of each adjacency, $P(Adj_\pi(f, g) = 1|D_\pi)$, in the ancestral genome, we construct an **adjacency graph** G for π. In digraph $G = (V, E)$, $|V| = 2n + 2$ where each element i corresponds to two nodes, i and $-i$, as well as ϕ_A and ϕ_Z. The set of directed edges is: $E(G) = \{(u, v)|Adj_\pi(u, v) = 1\}$. Here, (u, v) denotes an arc directed from u to v. Node ϕ_A only has outgoing edges and ϕ_Z only has incoming edges. We also associate weights representing the probability of that adjacency to corresponding edge, i.e. $w(u, v) = P(Adj_\pi(u, v) = 1|D_\pi)$.

Here, we use a greedy heuristic approach to achieve an approximate solution. See the algorithm FIND-CARS in Chapter 2.

5.5 Result

We first estimated the parameter α for each lineage, given the branch lengths (Fig. 5.2). We then predicted 26 CARs in the human-dog common ancestor. We were able to further combine them into 23 putative ancestral chromosomes using information from chromosome painting data.

We compared the result to the predicted human-dog ancestor in Chapter 4. They are almost the same, except that we made a few more joins that were uncertain in the

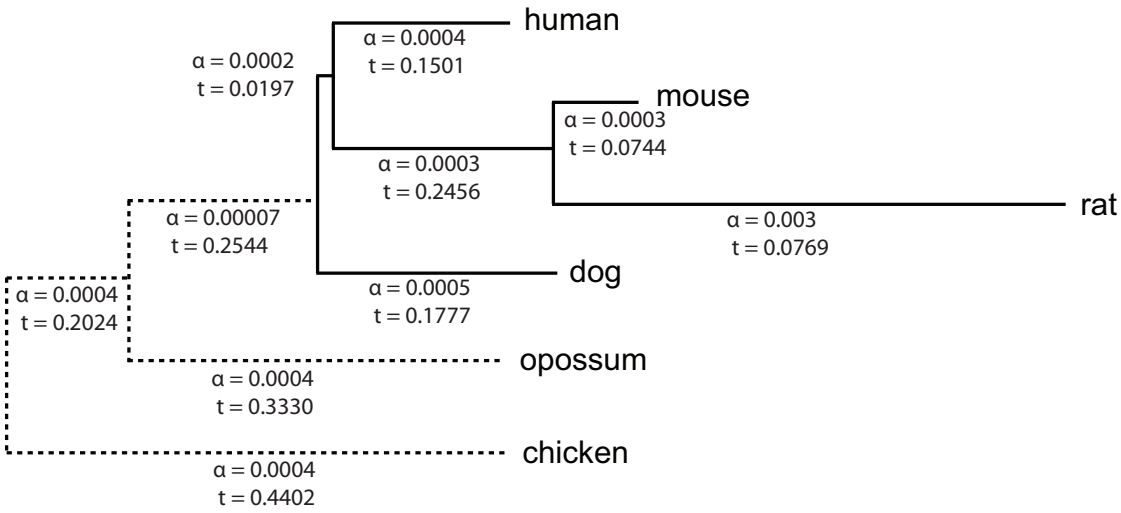

Figure 5.2: Numbers below each branch give the branch length t and parameter α in the extended Jukes-Cantor model. Branch lengths in the picture are proportional to the number of chromosomal breakages. The assembly artifacts probably make the rat branch much longer than the mouse branch.

result in Chapter 4. However, slight differences still remains when compared to the cytogenetic result (Froenicke *et al.*, 2006), e.g., the joins of Hsa16/Hsa19 and Hsa16/Hsa7. We hope to resolve these differences when more genomes are available.

5.6 Discussion

In this chapter, we introduced a probabilistic model for reconstructing ancestral genomic order. The essential part of this method is to predict the posterior probability of an adjacency occurring in the ancestor based on an extended Jukes-Cantor model for breakpoints. The result of our probabilistic method is quite similar to our previous result based on a parsimony approach. However, this adjacency model is extensible to incorporate insertion, deletion, and duplications if we can infer the history of them. But reconstructing the history of these operations, especially for duplications, is still difficult.

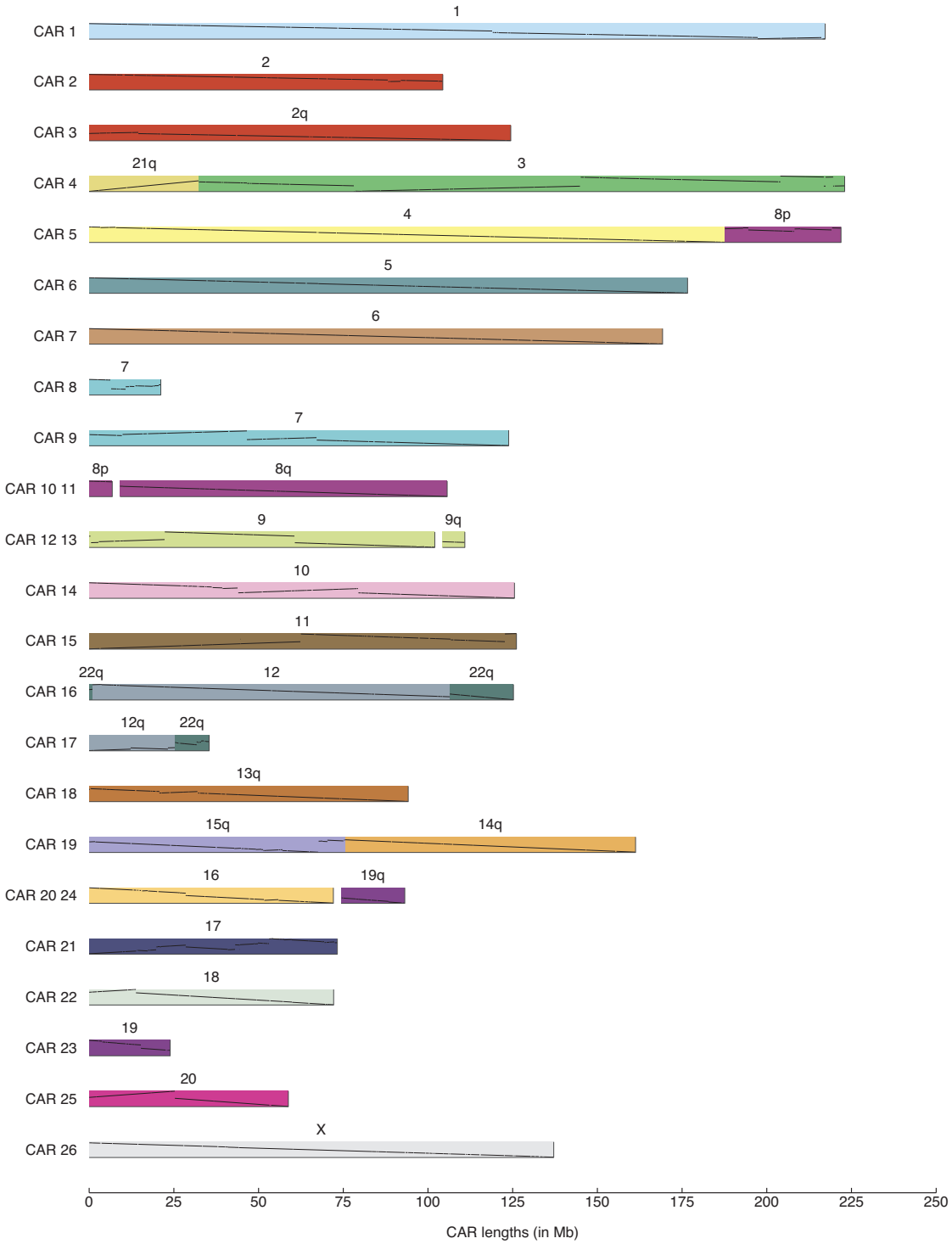

Figure 5.3: Predicted CARs in Boreoeutherian common ancestor using probabilistic model

For insertion and deletion, we want to know exactly when the insertion or deletion event happened. The configuration x represents the presence or absence of a conserved segment. Assume that the probability of changing from presence to absence after a time unit is ω and the branch length leading from π to τ is t. Then the probability of a deletion event of an element a along branch $\pi\tau$ is:

$$P(a \notin \tau | a \in \pi) = \frac{1}{2} - \frac{1}{2}e^{-2\omega t} \tag{5.17}$$

while that for the configuration remaining unchanged is:

$$P(a \in \tau | a \in \pi) = \frac{1}{2} + \frac{1}{2}e^{-2\omega t} \tag{5.18}$$

If we assume the probability of changing from absence to presence after a time unit is also ω, then the probability of an insertion event of an element a along branch $\pi\tau$ is:

$$P(a \in \tau | a \notin \pi) = \frac{1}{2} - \frac{1}{2}e^{-2\omega t} \tag{5.19}$$

and for the configuration remaining unchanged is:

$$P(a \notin \tau | a \notin \pi) = \frac{1}{2} + \frac{1}{2}e^{-2\omega t} \tag{5.20}$$

Therefore, we can calculate the posterior probabilities of $P(a \notin \pi | D_{\pi})$ and $P(a \in \pi | D_{\pi})$ and determine if a is in the ancestral genome π. If a is predicted not to be in the ancestor, we remove a in all the leaf genomes in order to infer adjacencies.

For the adjacency model, we currently only use one parameter, α, indicating the probability of adjacency changing, under the assumption that the proportion of each operation is unknown. One possible improvement of the adjacency model is to distinguish inversion and other operations due to the fact that inversions happen more often. In order to capture this feature, analogous to two-parameter model for substitutions, we can use different probabilities for these two categories, say β and α and $\alpha < \beta$. Again, there are altogether $2n-2$ possible changes that will replace the successor of element f. But among these $2n-2$ changes, $n-1$ are the result of inversion (i.e. a's successor g is replaced by $-h$, which was h in the original configuration, no matter whether h is positive or negative), and $n-1$ is the result of other operations. Thus, after a time unit, the probability that a's successor is changed is $(n-1)\alpha + (n-1)\beta$. However, estimating parameters α and β is nontrivial.

Chapter 6

Reconstructing the Catarrhini ancestor

6.1 Introduction

Except for great apes (e.g. chimp, gorilla, and orangutan) and gibbons, the animals most closely related to humans are Old World monkeys (superfamily Cercopithecidea, see Figure 6.1), which include rhesus macaque. The human lineage separated from the ancestors of chimp 6-7 MYA (Chimpanzee Sequencing and Analysis Consortium, 2005), while the human and ape lineage diverged from Old World monkeys about 23 MYA (Raaum *et al.*, 2005). Rhesus macaque has been widely recognized as an excellent model organism in biomedical research. With its complete whole genome sequence, it provides us with the opportunity to reconstruct the Catarrhini common ancestor (i.e. the common ancestor of human and Old World monkeys) so as to obtain more insights into primate genome evolution. Also, rhesus macaque is a very informative outgroup for predicting the human-chimp ancestor. This will help identify which parts of the human genome are conserved in primates since the divergence with rodents and what kinds of changes are more recent.

In this chapter, we discuss the reconstructions of the human-macaque and human-chimp ancestors. This reconstruction enables us to track the karyotype evolution on the human lineage since human-macaque divergence. In particular, we introduce methods for progressive ancestral genome reconstruction. We developed an improved version of

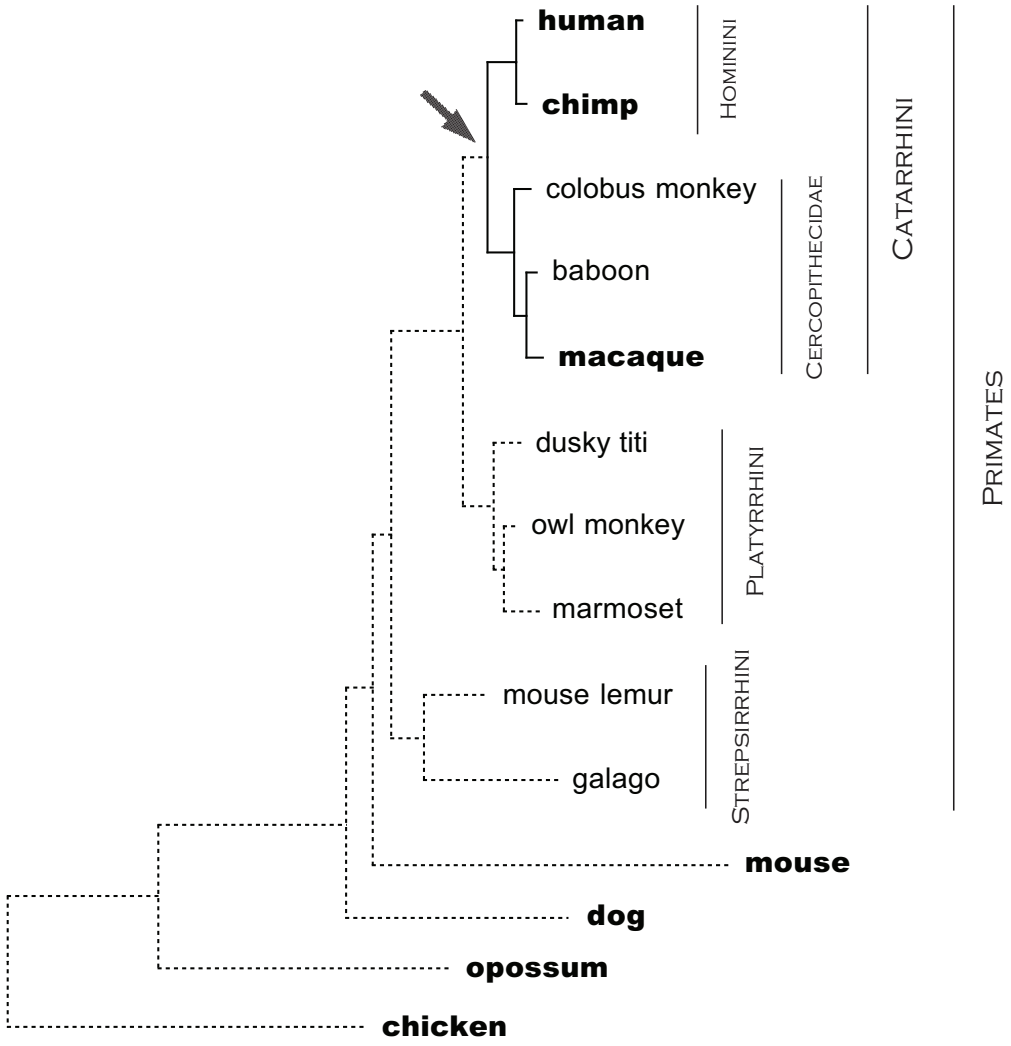

Figure 6.1: Position of Catarrhini common ancestor. Species used in the reconstruction are highlighted.

the INFER-CARS program, called INFER-CARS-PRO, which simultaneously reconstructs each ancestral genome along a particular lineage.

6.2 Progressively reconstructing every intermediate ancestor

Our objective is to reconstruct CARs in the human-macaque common ancestor and at the same time to simultaneously reconstruct the human-chimp common ancestor. In Chapter 3, we introduced the method of building conserved segments for a particular ancestral genome based on alignment nets. We require that every conserved segment should contain one genomic piece from each species. Therefore, when we are reconstructing the human-chimp ancestor, every conserved segment contains pieces from human and chimp, but does not necessarily contain a counterpart from macaque. However, for the human-macaque ancestor, the set of conserved segments previously constructed for the human-chimp ancestor needs to be refined. It turns out that we can progressively reconstruct every intermediate ancestor along a particular lineage by walking up the tree.

Based on this idea, we developed a program called INFER-CARS-PRO. The whole process can be summarized as:

- We first generate the orthology blocks for the human-chimp ancestor, using macaque as well as mouse, dog, opossum, and chicken as outgroups.
- Conserved segments are created for the human-chimp ancestor, and CARs are inferred.

- We walk up the tree to the human-macaque ancestor. By promoting macaque to be a descendant (instead of an outgroup species),, we further partition the human-chimp orthology blocks into human-macaque orthology blocks. Undesired blocks (smaller than some threshold or a duplication target) are discarded.

- Conserved segments are created for the human-macaque ancestor, and CARs are inferred.

Obviously, this process can be continued if we also want to reconstruct more ancient genomes along the human lineage, e.g. the human-mouse ancestor.

Figure 6.2 illustrates an example of how the orthology blocks are progressively constructed. In Figure 6.2 (A), we have 3 orthology blocks. In orthology block 2, there are pieces a, b, c from different chromosomes in macaque. Figure 6.2 (B) shows the orthology blocks in human-macaque ancestor. By refining the orthology blocks obtained in human-chimp ancestor, we are able to partition this region into 5 blocks. However, macaque piece c is too small and that particular block is not eligible to form a single orthology block. Finally, we obtain 4 orthology blocks in human-macaque ancestor.

6.3 Result

To reconstruct both the human-macaque and human-chimp ancestors, we used genomes of human (build hg18, March 2006), chimp (build panTro2, March 2006), and rhesus (build rheMac2, January 2006), downloaded from UCSC Genome Browser. Mouse, dog, opossum, and chicken genomes were also utilized as outgroups.

We choose 100Kb as the threshold retain conserved segments. After running Infer-CARs-Pro, we end up with 84 conserved segments in the human-chimp ancestor

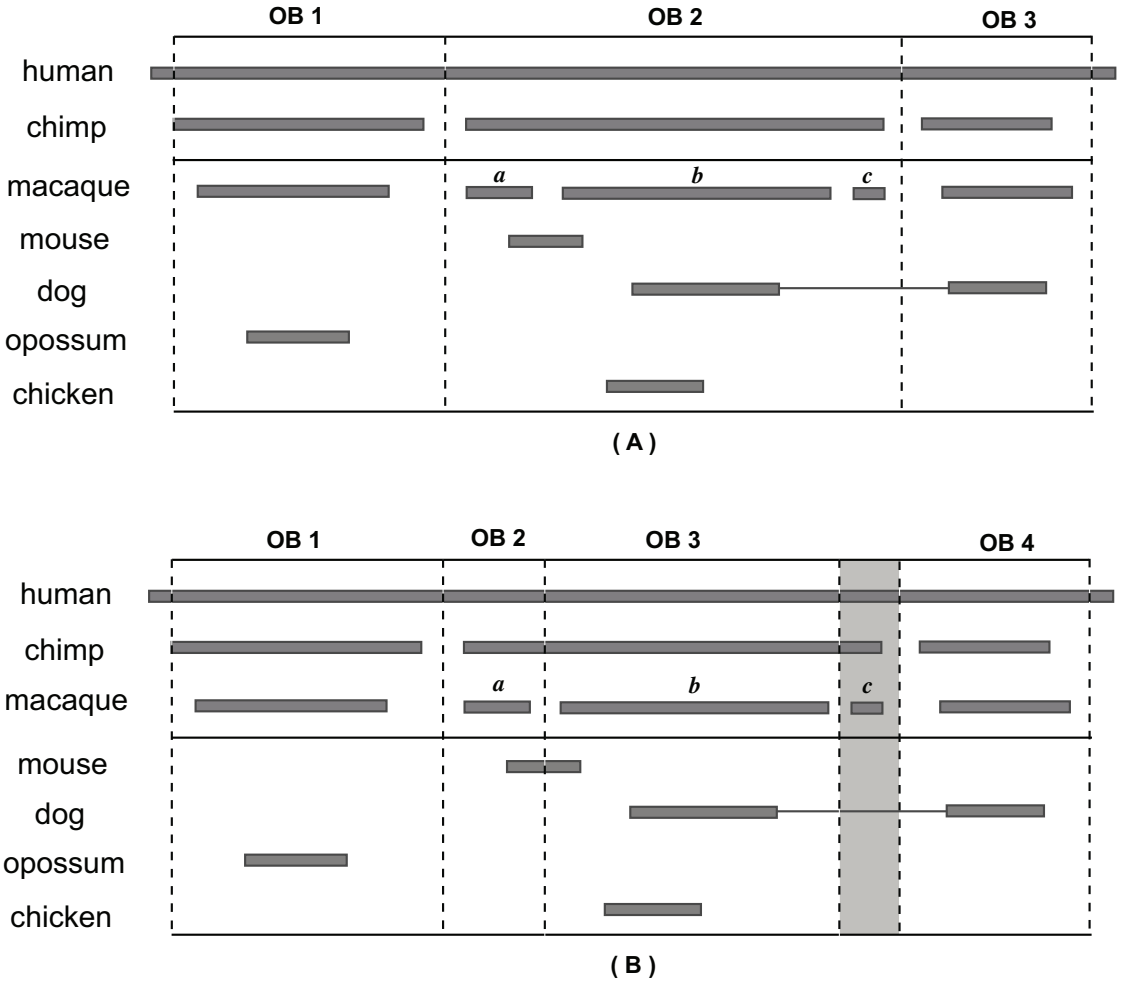

Figure 6.2: Progressively constructing orthology blocks. (A) In the human-chimp ancestor, we have 3 orthology blocks in this example. Pieces a, b, c are from different chromosomes in macaque. (B) In the human-macaque ancestor, we further partition the orthology blocks obtained in (A) into 5 blocks. But the block which includes macaque piece c is smaller than the threshold, so it is discarded (shaded in the picture). Thus, there are 4 orthology blocks in the human-macaque ancestor.

and 231 conserved segments in the human-macaque ancestor. We also get 27 CARs and 31 CARs in the human-chimp and human-macaque ancestor, respectively. From the literature, we know that there are two major interchromosomal rearrangements along the human lineage after divergence from macaque (Wienberg, 2004). One is the fission of Hsa14/Hsa15 after the human-macaque ancestor but before the human-chimp ancestor. The other is the fusion that formed Hsa2, which happened after human-chimp divergence. From our data, we only correctly predict the Hsa2 fusion in human-macaque ancestor. However, we reflect other configurations in Figure 6.3 and Figure 6.4. A closer New World monkey and more Old World monkeys might help to recover these joins.

We also run our programs that identify small inversions. We find 154 small inversions in human after human-chimp divergence; 253 inversions happened after divergence from macaque but before divergence from chimp. Also, there are 353 chimp-specific inversions and 2024 rhesus-specific inversions. But the number in rhesus is over-estimated now because we currently do not have a good outgroup.

6.4 Some additional work

Hierarchically refining the reconstruction resolution

The resolution of the reconstruction reported in the previous section is 100Kb. A much higher resolution reconstruction (say 10Kb or 20Kb) is crucial for fully understanding the karyotype evolution after the divergence between human and Old World monkeys. But the biggest issue is that when we increase the resolution we will be picking up more questionable joins, which decrease the overall reconstruction accuracy.

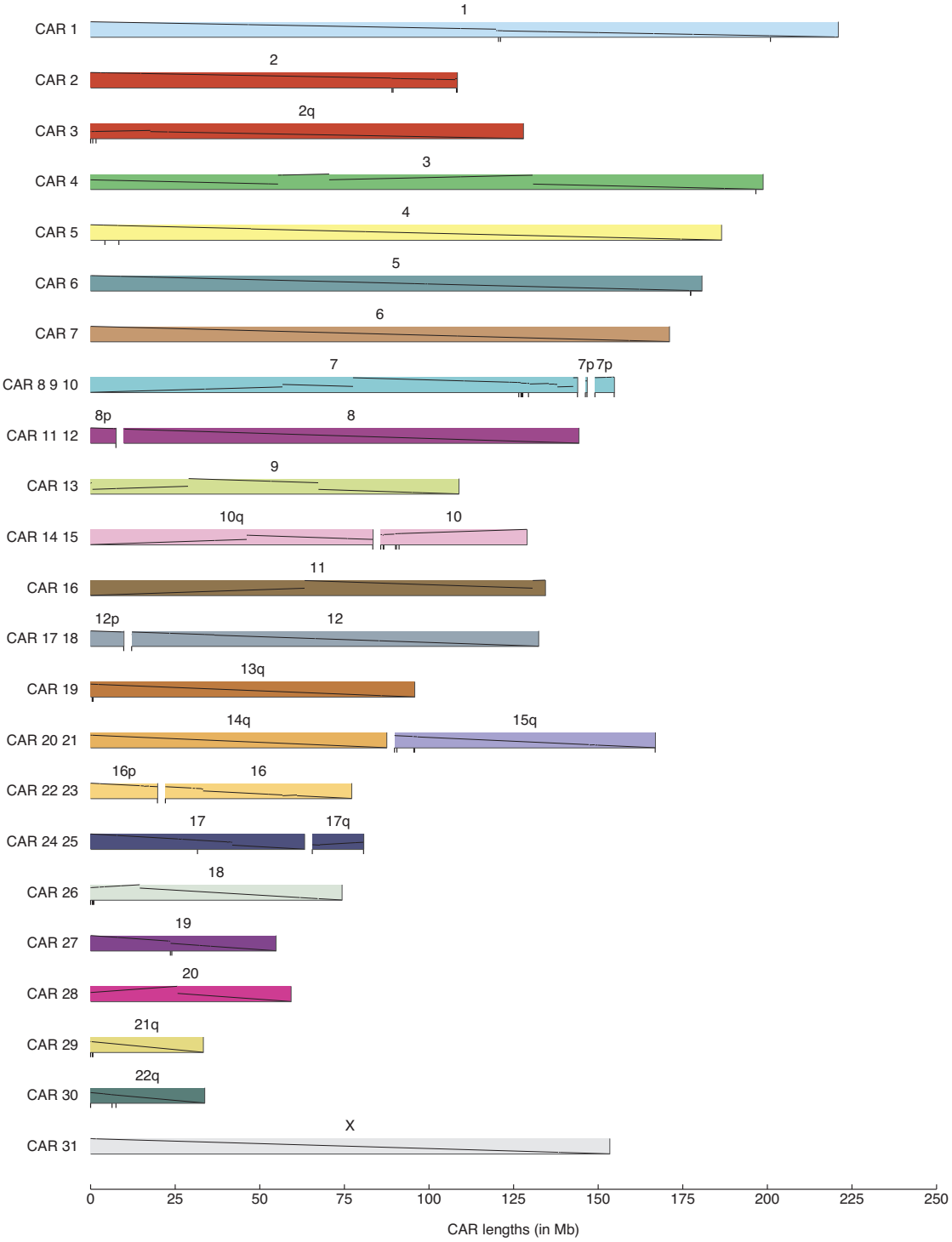

Figure 6.3: Map of the Catarrhini ancestral genome

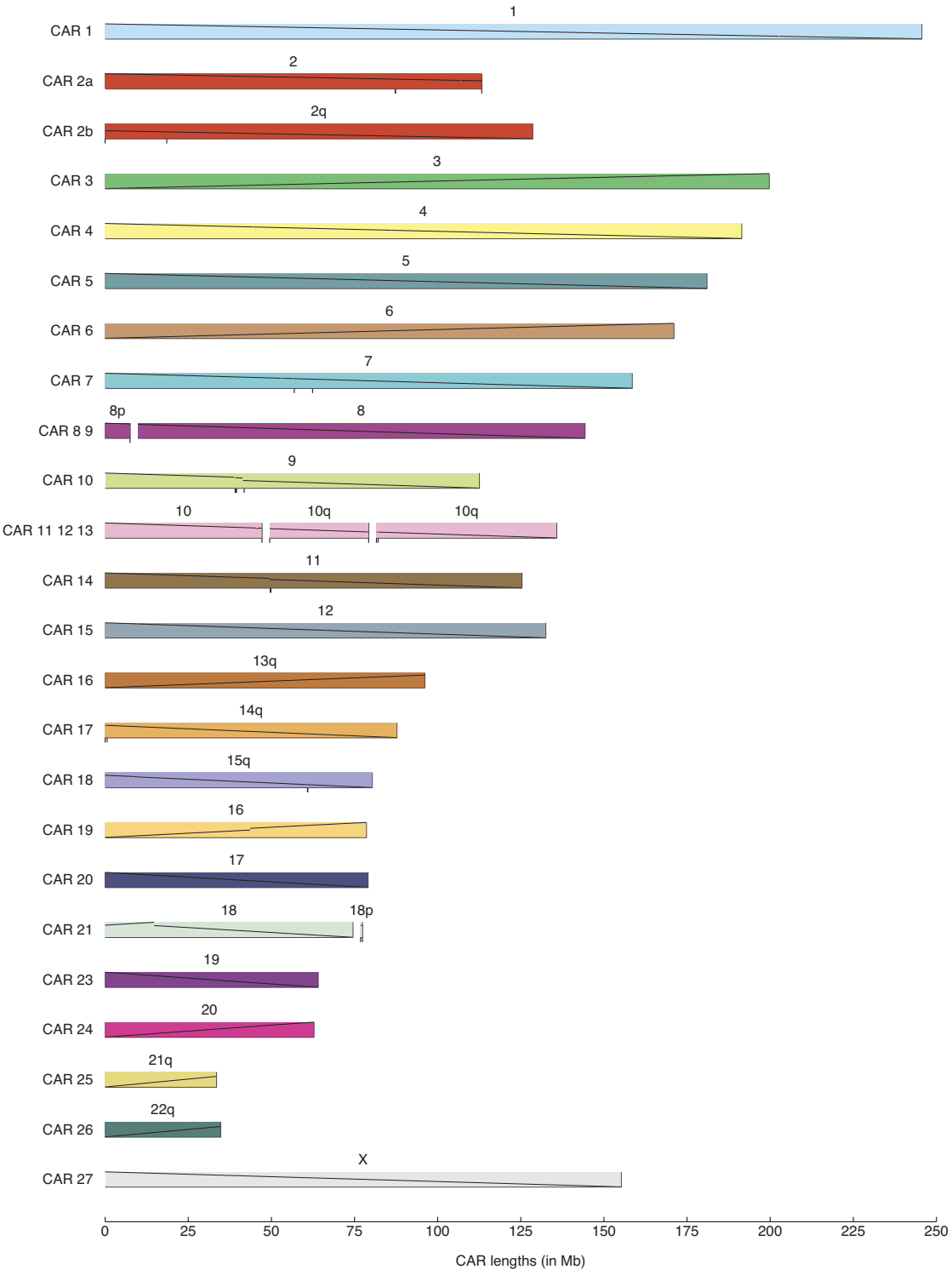

Figure 6.4: Map of the Hominini ancestral genome

An ongoing project is to hierarchically refine the reconstruction results. We first reconstruct the ancestor in a low resolution, for instance, 300Kb. The reconstructed joins in low resolution tend to be more reliable. Then we increase the resolution to 50Kb. It is certain that more rearrangements will appear in the result. For some reason a single CAR in 300Kb will become several CARs in 50Kb resolution. We then use 300Kb reconstruction as a reference to reassemble CARs in 50Kb. Basically, we break the 50Kb CARs in the positions of joins disagreeing with 300Kb CARs and assemble these CARs according to the 300Kb CARs. But if a join in the 300Kb CARs happens to be a weakly supported join, then special care is needed to determine if we should reassemble the 300Kb CARs. This is actually a zoom-in process and eventually we will be able to identify rearrangement in very high precision. However, it is still challenging because there are only a few primate species have been fully assembled. Our initial experiment showed that there are too many questionable joins in 10Kb resolution. We expect to have a more reliable reconstruction when more primate genomes are available.

Identifying potential misassemblies

In the reconstructed CARs, we can observe some odd configurations that look like transposition cause by a series of overlapping rearrangements in the macaque, see Figure 6.5 for example. This whole human chr20 region is aligned to rhesus chr10. But there is a macaque piece accidently transposed into the middle. The outgroups mouse and dog confirm that this is an rearrangement in macaque. From another perspective, it is also reasonable to suspect that this is an assembly artifact caused by misplacement of some contig.

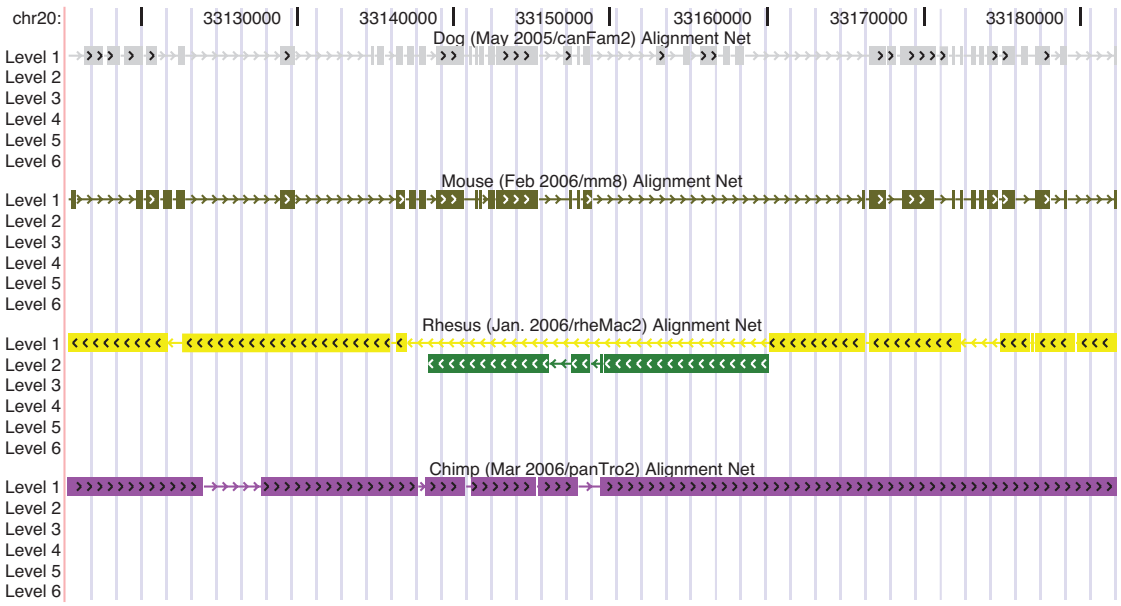

Figure 6.5: A potential assembly error in rhesus macaque. Around *hg18.chr20: 33,137,635-33,159,993*. The odd piece from rhesus chr13 is *rheMac2.chr13: 27,333,416-27,356,282*

Suspicion aroused from this kind of odd pieces forced us to computationally look for potential assembly errors in large scale. The program reported 26 such cases by looking at the reconstructed human-macaque ancestor in 20Kb resolution. Comparing these 26 regions to mate-pair chains of rhesus fosmids mapped to human and rhesus, 11 of the regions appear to most likely be assembly artifacts and the remaining 15 are ambiguous based on the mate-pair chain data (experimental results from R. Alan Harris at Baylor College of Medicine). The example shown in Figure 6.5 was confirmed as an assembly artifact instead of a transposition in rhesus.

More computational and experimental work is needed to look for other types of potential assembly artifacts. It is also important to have a systematic approach to assign reliability to each breakpoints, indicating how likely it is to be an artifact, as opposed to true biology.

Chapter 7

Conclusion and future work

7.1 Summary

In this study, we develop new methods and computer programs for reconstructing contiguous ancestral regions. The key goal of the algorithm is to predict the ancestral orders from modern adjacencies. Based on whole-genome alignment nets, we successfully construct conserved segments having high coverage and high resolution.

Using these methods, we compare the completely sequenced genomes of human, mouse, rat, and dog, utilizing opossum and chicken as outgroups, to reconstruct the karyotype organization of the Boreoeutherian common ancestor. We analyze mammalian genome rearrangements at higher resolution than has been published to date. We identify 3171 intervals, covering about 92% of the human genome, within which we find no rearrangements larger than 50Kb in the lineages leading to human, mouse, rat and dog from their most recent common ancestor. Combining intervals that are adjacent in all contemporary species produces 1338 segments that may contain large insertions or deletions, but that are free of chromosome fissions or fusions as well as inversions or translocations over 50Kb in length. We combine the results from our algorithm with data from chromosome painting experiments to produce a map of an early mammalian genome that accounts for 96.8% of the available human genome sequence data. The precision is further increased by mapping inversions as small as 31 bp. Analysis of

the predicted evolutionary breakpoints in the human linage confirms certain published observations but disagrees with others. Although only a few mammalian genomes are currently sequenced to high precision, our computer simulations indicate that our results are reasonably accurate.

In addition, we introduce a probabilistic method that reconstructs contiguous ancestral regions. Although the result for the Boreoeutherian ancestor does not differ much from the original parsimony based approach, the method can be potentially extended to handle large insertion, deletion, and duplications that we currently ignore. We further improve the programs for constructing conserved segments and progressively reconstruct both human-chimp and human-macaque ancestors simultaneously.

7.2 Future directions

A number of additional mammals are already being sequenced "at low redundancy" for the purpose of identifying human regions under negative selection for substitutions (Margulies *et al.*, 2005). The resulting sequence data are extremely useful for predicting ancestral nucleotides. However, they lack the long-range contiguity needed for accurate identification of large-scale evolutionary events. We look forward to the day when a high-accuracy assembly of, say, elephant or armadillo provides us with an ideal outgroup for large-scale reconstruction of the Boreoeutherian ancestral sequence.

Our computational methods need further refinement. In particular, we anticipate improvements in the handling of large duplications and deletions. Additional progress may be possible through the modeling of other evolutionary events, such as gene conversion or expansion/contraction of short tandem repeats caused by strand slippage.

The inconsistencies between our prediction and chromosome painting results indicate that additional investigation is required. It seems obvious that cytogenetic and BAC analysis will be important to resolve the questionable joins predicted by our method. A systematic approach to incorporate information from other sources is needed to improve our computational predictions.

It is necessary to improve the partitioning process to incorporate regions where the ancestral DNA sequence has been deleted in human. Right now, we only use human as the reference genome to do the segmentation. In fact, a considerable amount of DNA has been deleted on the human lineage after human-rodent speciation. For this purpose, we can consider using other species (e.g. mouse, rat, dog) as reference too, giving the resulting conserved segments with more coverage.

More advances are needed to facilitate simultaneous reconstruction of ancestral genomes at all internal nodes of the phylogenetic tree, not only along a certain lineage. In other words, we expect to be able to reconstruct all the intermediate ancestors, e.g. the human-chimp and mouse-rat ancestors, by walking up the tree.

An accurate conceptual model of large-scale evolutionary events will be critical for successful reconstruction of ancestral genomes. A well-founded theoretical model of chromosome evolution will be an important complementary evaluation of the reconstruction.

None of the bioinformatic reconstruction methods to date attempt to predict the ancestral centromeres. Actually, the evidence of centromere repositioning further complicates this problem (Ventura *et al.*, 2004). It will be interesting and important to explore this direction.

A number of challenges remain before the genome sequences of mammalian ancestors can be accurately predicted at nucleotide resolution. An integrated approach is needed in order to increase the accuracy. However, we believe that this goal is an appropriate focus for our twin aims of understanding the evolutionary history of every position in the human genome and of providing an Internet resource that optimally organizes and presents the ever-expanding wealth of mammalian and vertebrate sequence data in the context of the ancestral sequence they have in common.

Bibliography

Bader, D. A., Moret, B. M., and Yan, M., 2001. A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J Comput Biol*, **8**(5):483–491.

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., and Eichler, E. E., 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res*, **11**(6):1005–1017.

Berman, P. and Hannenhalli, S., 1996. Fast sorting by reversal. In Hirschberg, D. S. and Myers, E. W., editors, *CPM*, volume 1075 of *Lecture Notes in Computer Science*, pages 168–185. Springer.

Berman, P., Hannenhalli, S., and Karpinski, M., 2002. 1.375-approximation algorithm for sorting by reversals. In *ESA '02: Proceedings of the 10th Annual European Symposium on Algorithms*, pages 200–210, London, UK. Springer-Verlag.

Blanchette, M., Green, E. D., Miller, W., and Haussler, D., 2004a. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*, **14**(12):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., *et al.*, 2004b. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, **14**(4):708–715.

Blanchette, M., Kunisawa, T., and Sankoff, D., 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. *J Mol Evol*, **49**(2):193–203.

Boesch, F. T. and Gimpel, J. F., 1977. Covering points of a digraph with point-disjoint paths and its application to code optimization. *J ACM*, **24**(2):192–198.

Bourque, G. and Pevzner, P. A., 2002. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res*, **12**(1):26–36.

Bourque, G., Pevzner, P. A., and Tesler, G., 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res*, **14**(4):507–516.

Bourque, G., Tesler, G., and Pevzner, P. A., 2006. The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction. *Genome Res*, **16**(3):311–313.

Bourque, G., Zdobnov, E. M., Bork, P., Pevzner, P. A., and Tesler, G., 2005. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. *Genome Res*, **15**(1):98–110.

Caprara, A., 1997. Sorting by reversals is difficult. In *RECOMB*, pages 75–83.

Caprara, A., 1999. Formulations and hardness of multiple sorting by reversals. In *RECOMB*, pages 84–94.

Chaudhary, R., Raudsepp, T., Guan, X. Y., Zhang, H., and Chowdhary, B. P., 1998. Zoo-FISH with microdissected arm specific paints for HSA2, 5, 6, 16, and 19 refines known homology with pig and horse chromosomes. *Mamm Genome*, **9**(1):44–49.

Chimpanzee Sequencing and Analysis Consortium, 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**(7055):69–87.

Dobzhansky, T. and Sturtevant, A. H., 1938. Inversions in the chromosomes of *Drosophila Pseudoobscura*. *Genetics*, **23**(1):28–64.

Felsenstein, J., 2003. *Inferring Phylogenies*. Sinauer Associates.

Fitch, W. M., 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Zool*, **20**:406–416.

Froenicke, L., Caldes, M. G., Graphodatsky, A., Muller, S., Lyons, L. A., Robinson, T. J., Volleth, M., Yang, F., and Wienberg, J., 2006. Are molecular cytogenetics and bioinformatics suggesting diverging models of ancestral mammalian genomes? *Genome Res*, **16**(3):306–310.

Froenicke, L., Wienberg, J., Stone, G., Adams, L., and Stanyon, R., 2003. Towards the delineation of the ancestral eutherian genome organization: comparative genome maps of human and the African elephant (Loxodonta africana) generated by chromosome painting. *Proc Biol Sci*, **270**(1522):1331–1340.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., Scott, G., Steffen, D., Worley, K. C., Burch, P. E., *et al.*, 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**(6982):493–521.

Hannenhalli, S., Chappey, C., Koonin, E. V., and Pevzner, P. A., 1995. Genome sequence comparison and scenarios for gene rearrangements: a test case. *Genomics*, **30**(2):299–311.

Hannenhalli, S. and Pevzner, P. A., 1995. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. In *STOC*, pages 178–189. ACM.

Kaplan, H., Shamir, R., and Tarjan, R. E., 1997. Faster and simpler algorithm for sorting signed permutations by reversals. In *SODA '97: Proceedings of the eighth annual ACM-SIAM symposium on Discrete algorithms*, pages 344–351, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

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**(20):11484–11489.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D., 2002. The human genome browser at UCSC. *Genome Res*, **12**(6):996–1006.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.*, 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822):860–921.

Lindblad-Toh, K., Wade, C. M., Mikkelsen, T. S., Karlsson, E. K., Jaffe, D. B., Kamal, M., Clamp, M., Chang, J. L., Kulbokas, E. J. r., Zody, M. C., *et al.*, 2005. Genome

sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**(7069):803–819.

Ma, J., Zhang, L., Suh, B. B., Raney, B. J., Burhans, R. C., Kent, W. J., Blanchette, M., Haussler, D., and Miller, W., 2006. Reconstructing contiguous regions of an ancestral genome. *Genome Res*, . Published online before print September 18, 2006.

Margulies, E. H., Vinson, J. P., Miller, W., Jaffe, D. B., Lindblad-Toh, K., Chang, J. L., Green, E. D., Lander, E. S., Mullikin, J. C., and Clamp, M., *et al.*, 2005. An initial strategy for the systematic identification of functional elements in the human genome by low-redundancy comparative sequencing. *Proc Natl Acad Sci U S A*, **102**(13):4795–4800.

Matthey, R., 1972. Chromosomes and evolution. *Triangle*, **11**(3):107–112.

Misceo, D., Ventura, M., Eder, V., Rocchi, M., and Archidiacono, N., 2003. Human chromosome 16 conservation in primates. *Chromosome Res*, **11**(4):323–326.

Moret, B. M. E., Wyman, S. K., Bader, D. A., Warnow, T., and Yan, M., 2001. A new implmentation and detailed study of breakpoint analysis. In *Pacific Symposium on Biocomputing*, pages 583–594.

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A., Bourque, G., Tesler, G., Auvil, L., Beever, J. E., Chowdhary, B. P., Galibert, F., Gatzke, L., *et al.*, 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science*, **309**(5734):613–617.

Murphy, W. J., Stanyon, R., and O'Brien, S. J., 2001. Evolution of mammalian genome organization inferred from comparative gene mapping. *Genome Biol*, **2**(6):REVIEWS0005.

Nadeau, J. H. and Taylor, B. A., 1984. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci U S A*, **81**(3):814–818.

Ovcharenko, I., Loots, G. G., Giardine, B. M., Hou, M., Ma, J., Hardison, R. C., Stubbs, L., and Miller, W., 2005. Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res*, **15**(1):184–194.

Peng, Q., Pevzner, P. A., and Tesler, G., 2006. The fragile breakage versus random breakage models of chromosome evolution. *PLoS Comput Biol*, **2**(2):e14.

Pevzner, P. and Tesler, G., 2003. Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. *Genome Res*, **13**(1):37–45.

Qumsiyeh, M. B., 1994. Evolution of number and morphology of mammalian chromosomes. *J Hered*, **85**(6):455–465.

Raaum, R. L., Sterner, K. N., Noviello, C. M., Stewart, C.-B., and Disotell, T. R., 2005. Catarrhine primate divergence dates estimated from complete mitochondrial genomes: concordance with fossil and nuclear DNA evidence. *J Hum Evol*, **48**(3):237–257.

Richard, F., Lombard, M., and Dutrillaux, B., 2003. Reconstruction of the ancestral karyotype of eutherian mammals. *Chromosome Res*, **11**(6):605–618.

Rocchi, M., Archidiacono, N., and Stanyon, R., 2006. Ancestral genomes reconstruction: an integrated, multi-disciplinary approach is needed. *Genome Res*, . Published online before print October 19, 2006.

Sankoff, D., 1992. Edit distances for genome comparisons based on non-local operations. In Apostolico, A., Crochemore, M., Galil, Z., and Manber, U., editors, *CPM*, volume 644 of *Lecture Notes in Computer Science*, pages 121–135. Springer.

Sankoff, D., 2006. The signal in the genomes. *PLoS Comput Biol*, **2**(4):e35.

Sankoff, D. and Blanchette, M., 1998. Multiple genome rearrangement and breakpoint phylogeny. *J Comput Biol*, **5**(3):555–570.

Sankoff, D. and Blanchette, M., 1999. Probability models for genome rearrangement and linear invariants for phylogenetic inference. In *RECOMB*, pages 302–309.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B. F., and Cedergren, R., 1992. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proc Natl Acad Sci U S A*, **89**(14):6575–6579.

Sankoff, D. and Nadeau, J. H., 2003. Chromosome rearrangements in evolution: From gene order to genome sequence and back. *Proc Natl Acad Sci U S A*, **100**(20):11188–11189.

Savva, G., Dicks, J., and Roberts, I. N., 2003. Current approaches to whole genome phylogenetic analysis. *Brief Bioinform*, **4**(1):63–74.

Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W., 2003. Human-mouse alignments with BLASTZ. *Genome Res*, **13**(1):103–107.

Tesler, G., 2002. GRIMM: genome rearrangements web server. *Bioinformatics*, **18**(3):492–493.

Todd, N. B., 1970. Karyotypic fissioning and canid phylogeny. *J Theor Biol*, **26**(3):445–480.

Ventura, M., Weigl, S., Carbone, L., Cardone, M. F., Misceo, D., Teti, M., D'Addabbo, P., Wandall, A., Bjorck, E., de Jong, P. J., *et al.*, 2004. Recurrent sites for new centromere seeding. *Genome Res*, **14**(9):1696–1703.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.*, 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**(6915):520–562.

Watterson, G., Ewens, W., Hall, T., and Morgan, A., 1982. The chromosome inversion problem. *J Theor Biol*, **99**:1–7.

Webber, C. and Ponting, C. P., 2005. Hotspots of mutation and breakage in dog and human chromosomes. *Genome Res*, **15**(12):1787–1797.

Wienberg, J., 2004. The evolution of eutherian chromosomes. *Curr Opin Genet Dev*, **14**(6):657–666.

Yang, F., Alkalaeva, E. Z., Perelman, P. L., Pardini, A. T., Harrison, W. R., O'Brien, P. C. M., Fu, B., Graphodatsky, A. S., Ferguson-Smith, M. A., and Robinson, T. J.,

et al., 2003. Reciprocal chromosome painting among human, aardvark, and elephant (superorder Afrotheria) reveals the likely eutherian ancestral karyotype. *Proc Natl Acad Sci U S A*, **100**(3):1062–1066.

Zhang, Z., Raghavachari, B., Hardison, R. C., and Miller, W., 1994. Chaining multiple-alignment blocks. *J Comput Biol*, **1**(3):217–226.

Vita

Jian Ma was born in Shanghai, China. He received his Bachelor's degree in 2000 and his Master's degree in 2003, both from the Department of Computer Science at Fudan University, Shanghai, China. In August 2003, he joined the Ph.D. program in the Department of Computer Science and Engineering at the Pennsylvania State University. Since then, he has been working in Professor Webb Miller's lab at the Center for Comparative Genomics and Bioinformatics. After his Ph.D. study, he will work as a postdoc with Professor David Haussler at University of California Santa Cruz.