The Path of the Blind Watchmaker: A Model of Evolution

Andrew Anthony Poggio

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2011-26
http://www.eecs.berkeley.edu/Pubs/TechRpts/2011/EECS-2011-26.html

April 6, 2011

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
06 APR 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Path of the Blind Watchmaker: A Model of Evolution	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California at Berkeley,Electrical Engineering and Computer Sciences,Berkeley,CA,94720**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Evolution has been described by Dawkins as a blind watchmaker due to its being unconscious and random but selective and able to produce complex forms. Evolution from an early, primitive organism (the Last Universal Common Ancestor of all life, LUCA) to Homo sapiens is the most dramatic biological process that has taken place on Earth and knowledge of it is important to understanding many aspects of biology including disease prevention and treatment. We claim that computational biology has now reached the point that astronomy reached when it began to look backward in time to the Big Bang. Our goal is look backward in biological time, and to begin to describe, in more detail, LUCA and the evolution from LUCA to us. This is the path of the blind watchmaker. This thesis presents a novel dataset of ancestral, genome sequences that we have reconstructed. These ancestors serve as reference species for our models. We develop a sequence evolution model that reflects biological processes more accurately than prior work and apply it to the ancestral genome dataset. This model uses empirical mutation probabilities for scoring alignments and includes inversion mutations. The results of this model describe the mutations that must have taken place during the evolution of our reference species. We then apply the sequence evolution results to our population evolution model. This model uses a dynamic set of subpopulation pools with related but distinct, mutating genomes reproducing sexually and asexually, and subject to speciation effects, selection pressures, and environmental carrying capacity limitations. The results of this model are population size estimates, evolution duration estimates, and identification of critical evolution parameters and estimates for their values. We present the results of these models along with evidence for some tantalizing, if speculative, discoveries along the path. This work also reveals significant opportunities for further efforts in silico, in vitro, and in vivo.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**150**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

The Path of the Blind Watchmaker:
A Model of Evolution

by

Andrew Anthony Poggio

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

In

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David Patterson, Chair
Professor Adam Arkin
Professor Brent Mishler
Professor Christos Papadimitriou

Spring 2011

The dissertation of Andrew Anthony Poggio, titled The Path of the Blind Watchmaker: A Model of Evolution, is approved:

Chair	Date
	Date
	Date
	Date

University of California, Berkeley
Spring 2011

The Path of the Blind Watchmaker:
A Model of Evolution

by
Andrew Anthony Poggio

Abstract

The Path of the Blind Watchmaker:

A Model of Evolution

by

Andrew Anthony Poggio

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor David Patterson, Chair

Evolution has been described by Dawkins as a blind watchmaker due to its being unconscious and random but selective and able to produce complex forms. Evolution from an early, primitive organism (the Last Universal Common Ancestor of all life, LUCA) to Homo sapiens is the most dramatic biological process that has taken place on Earth and knowledge of it is important to understanding many aspects of biology including disease prevention and treatment.

We claim that computational biology has now reached the point that astronomy reached when it began to look backward in time to the Big Bang. Our goal is look backward in biological time, and to begin to describe, in more detail, LUCA and the evolution from LUCA to us. This is the path of the blind watchmaker.

This thesis presents a novel dataset of ancestral, genome sequences that we have reconstructed. These ancestors serve as reference species for our models. We develop a sequence evolution model that reflects biological processes more accurately than prior work and apply it to the ancestral genome dataset. This model uses empirical mutation probabilities for scoring alignments and includes inversion mutations. The results of this model describe the mutations that must have taken place during the evolution of our reference species.

We then apply the sequence evolution results to our population evolution model. This model uses a dynamic set of subpopulation pools with related but distinct, mutating genomes reproducing sexually and asexually, and subject to speciation effects, selection pressures, and environmental carrying capacity limitations. The results of this model are population size estimates, evolution duration estimates, and identification of critical evolution parameters and estimates for their values.

We present the results of these models along with evidence for some tantalizing, if speculative, discoveries along the path. This work also reveals significant opportunities for further efforts in silico, in vitro, and in vivo.

Chair Date

Acknowledgements and Dedication

I would like to thank all of the people who helped to make this thesis possible. Raphi Rom, Greg Papadopoulos, Doug Engelbart, Jay Singer, and J.J. Garcia-Luna, all encouraged me to go back to school and complete my doctorate. They also all wrote recommendation letters for my readmission to UC Berkeley after I had been carefully reminded by the administration that readmission was not automatic. I would like to thank my SRI International colleagues, Carolyn Talcott, Pat Lincoln, Steven Eker, Al Valdez, Ken Nitz, Ashish Gehani, Keith Laderoute, and Merrill Knapp, for their consulting on topics ranging from biological mutation details to sequence alignment algorithms to statistical inference. I would also like to thank my SRI management, Carolyn Talcott and Pat Lincoln, for their support and for providing me with extraordinary professional flexibility. I would like to thank SRI International, Sun Microsystems, the Office of Naval Research, and the Defense Research Projects Agency for their financial support of my research. I would like to thank the University of California, Berkeley for being a world class research institution that is willing admit graduate students of any age. I would like to thank my thesis committee, Adam Arkin, Brent Mishler, and Christos Papadimitriou, for their advice throughout my research. Finally, I would like to thank my advisor, Dave Patterson, for encouraging me to complete my doctorate, supporting, motivating, and advising me throughout the research process, and remaining optimistic when completion may have seemed doubtful.

I wish to dedicate this thesis to my family. My late mother and father, Marie and Charles Poggio, were unable to attend college themselves but always supported my higher education in every way. My children, Maria and Tony Poggio, sacrificed many an occasion with Dad so that I could complete this research; it is my fervent hope that this and related research benefit their generation in ways that my generation can only imagine. Finally, I would like to dedicate this thesis to my wife, Mamie. Her encouragement, assuming of my family responsibilities while maintaining her own, and providing me with opportunities to work on this thesis when other activities beckoned, are gifts to me beyond price. Without her, this thesis would not have been possible.

Table of Contents

List of Figures

List of Tables

List of Equations

1 Introduction

The evolution from a very primitive organism to Homo sapiens, as put forth by Darwin [1], is the most dramatic, biological process that has taken place on Earth. In his popular book *The Blind Watchmaker* [2], Dawkins states that a stone being found in a field needs no explanation as it can be assumed to be formed by natural processes. However, a watch so found, as it has mechanical complexity and apparent purpose, cannot be so simply explained – if there is a watch then there must have been a watchmaker. In the context of biology, he uses the widely varying capabilities of bat echolocation as an example of biological complexity that is in need of explanation. He further argues effectively that evolution through natural selection (an unconscious and natural but complex and selective process analogous to a blind watchmaker) is the source of this biological complexity.

While evolution is a blind watchmaker to be sure, there has nonetheless been a path that it has taken, unchosen and unplanned, that has resulted in the current genome of each existing species. Beginning with an early genome, the path consists of the ordered set of mutations that have taken place in the genome over subsequent generations. This path was not followed because it did not pre-exist; the evolutionary process itself, in fact, created it.

When astronomy reached a critical mass of theory, technology, and observational data, astronomers were able to look backward in time, and describe the primordial Big Bang and the changes in the universe between the Big Bang and the present. We claim that computational biology has reached a critical mass of theory (mutation mechanisms, phylogenetic and sequence alignment algorithms), technology (genome sequencers and fast, multicore computers), and observational data (DNA sequence and protein data, mutation rates) that enables us to now look backward in time, and to begin to describe LUCA in more detail and evolution's path from LUCA to us.

In this research, our goal is to further illuminate the path of the blind watchmaker.

1.1 Motivation

There is a fundamental scientific value to a more comprehensive understanding of evolution over the substantial periods of time that this research pursues. Such understanding should enable:

1. More accurate homology (gene relationships)
2. More accurate phylogeny (species relationships)
3. Evolution prediction (under some circumstances)

Beyond the fundamental scientific value of this research, we have, substantial practical motivation for this effort. In particular, we note that pathogens, primarily in the form of viruses and bacteria, continue to plague us. In 2005, 4.9 million people were newly infected with HIV and 3.1 million were killed by it [3]. The avian flu virus H5N1 has a fatality rate of greater than 50% [4] and "ordinary" flu causes 250,000 to 500,000 deaths per year worldwide [5].

In the bacterial realm, increased use of antibiotics in developed countries has caused an increase in antibiotic-resistant bacteria. Vancomycin has become known as a drug of last resort and there was a 20-fold increase in Vancomycin-resistant bacteria in some hospitals from 1987-1993 [6].

A fundamental challenge in preventing and treating these pathogenic infections is that the pathogens evolve, changing in ways that reduce or eliminate the effects of our countermeasures. Viruses evolve into new strains such that the antibodies produced due to vaccines are ineffective. Bacteria evolve into new strains such that they are resistant to existing antibiotics. These organisms evolve over time and it would be advantageous for us to be able to predict their evolutionary pathways.

Cancer is a disease of mutations [7]. There are six essential alterations in cell physiology that collectively dictate malignant growth: self-sufficiency in growth signals, insensitivity to growth-inhibitory (antigrowth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained ability to increase blood supply for nutrients (angiogenesis), and movement allowing tissue invasion (metastasis). All of these alterations are caused by mutations. According to CDC data [8], cancer was the second leading cause of death in the United States in 2009. In addition cancer, like pathogenic bacteria and viruses, acquires drug resistance through mutation. Once again, it would be advantageous to both predict evolutionary pathways to enable malignant growth and the pathways that confer drug resistance.

With an appropriately comprehensive evolution model that would allow us to make predictions about future evolution, we could potentially:

1. Be prepared with effective disease treatments in advance of need.
2. Create treatments that act on multiple targets such that anticipated mutation events would not confer resistance to all of the treatment targets.

Long available geologic and fossil data provide ample evidence of the gross process of evolution on Earth. Though the fossil record contains many gaps among different organisms as well as gaps between possible ancestors and descendents [9], this data indicates a coarse temporal relationship among different species as well as allowing a measurement of the duration from the origin of one species to the origin of a subsequent one. The species, between whose origins a duration may be measured, may have widely separated origins, e.g. cyanobacteria and H. sapiens. Over the past two decades, developments in three areas have provided a significant source of relevant data that is independent of the fossil record:

1. Computational biology algorithms – the intellectual convergence of biology and mathematics has resulted in a comprehensive set of algorithms for the creation and manipulation of biological data. Shotgun sequencing and genome alignment [10], used to sequence a partial or complete genome, is exemplary. Phylogenetic, sequence alignment, and sequence reconstruction algorithms have also made significant progress.

2. Computer technology – the capabilities of computer components, and therefore computer systems, have been on an exponential growth curve along several dimensions. Among them are instruction execution rate, bus bandwidth, memory capacity, disk capacity, and network bandwidth. This exponential curve, known as Moore's Law [11], has resulted in a doubling of capability every 12 to 36 months.
3. Biological laboratory instrumentation – partially due to the positive effects of items 1. and 2. above, certain relevant laboratory instrumentation, e.g. genome sequencers, has also been on an exponential curve in terms of capability. At the initiation of the Human Genome Project, the cost to sequence a human genome was on the order of US$1,000,000,000; the National Institutes of Health is now funding technology expected to lead to a cost of US$100,000 in five years and a cost of US$1000 in ten years [12].

As a result of these developments, there is a rapidly growing mass of new data relevant to the process of evolution on earth. In particular, the National Library of Medicine announced in 2006 that the three leading public repositories of sequence data had reached a total of 100 Gigabases from over 165,000 organisms [13]. Appropriate application and analysis of this body of data can provide us with the broader, more comprehensive understanding of evolution that we seek.

1.2 Background

As described in [14], the hereditary information of all life on earth is carried by DeoxyriboNucleic Acid (DNA) in one or more long strands known as chromosomes. DNA is a double-stranded helix, consisting of a series of nucleotide base pairs. Each base pair consists of one of two purines, adenine (a) or guanine (g), paired with one of two pyrimidines, thymine (t) or cytosine (c). The base a always pairs with t and c always pairs with g; the elements of a pair are known as complements. Each of the two strands in a DNA molecule has a polarity that is evident from one end (the 5' end) to the other end (the 3' end). The polarity of a strand is opposite its paired strand. During cell division, the means of cellular reproduction, the cell's DNA is duplicated.

Cells use proteins to perform nearly all their functions, including reproduction and metabolism. Synthesizing proteins is thus a central function of cells. Some sequences of a cell's DNA code for the amino acid sequences of proteins that a cell produces. Many such sequences of DNA, known as coding regions, are preceded by a regulatory region that controls the circumstances under which the protein(s) defined by the coding region is produced. The combination of the regulatory region, if present, together with the coding region (and for Eukaryotes additional, interspersed sequences of noncoding DNA called introns [14]; the coding regions are called exons) is known as a gene.

The process of synthesizing a protein from a gene has two steps. Once the process begins, typically controlled by a regulatory region, the first step is known as transcription and it occurs on the DNA molecule. During transcription, one strand of the coding region of a gene is transcribed into complementary RiboNucleic Acid (RNA). RNA is a single-stranded molecule similar in structure and content to a single strand of DNA.

The second step in protein synthesis is known as translation and it occurs in a cellular organelle known as the ribosome. During translation, the sequence of nucleotides in an RNA molecule, produced during transcription, is translated into the sequence of amino acids that

constitute the protein. There are four types of RNA nucleotides (that correspond to the four types of DNA nucleotides) while there are 20 amino acids. Thus, a one-to-one correspondence between RNA nucleotides and protein amino acids is not possible. The RNA nucleotides are translated by the ribosome as triples – each combination of three RNA nucleotides (known as a codon) is translated into one of the 20 amino acids or as an indication to stop the protein synthesis. Since there are 64 combinations of RNA and 20 amino acids plus the protein stop code, the coding is redundant – some codons translate into the same amino acid and are said to be synonymous.

In biological evolution, three components must be present for evolution by natural selection to take place [15]:

1. Each subsequent generation must inherit the characteristics of the previous generation. In biology, this inheritance is carried from one generation to the next by duplicate copies of DNA.
2. A mutation mechanism must act to make changes to individuals, which can also be inherited. In biology, these mutations change the DNA directly.
3. The environment must be such that it causes differential selection on individuals. This differential arises due to differing characteristics of individuals and causes an increase or decrease in the population size of their progeny due to changes in longevity or fecundity.

With these components present, evolution by natural selection not only capable of taking place, it must and will take place unless otherwise prevented. We observe such evolution actually taking place in the development of antibiotic resistance among Staphylococcus aureus in hospitals and we observe innumerable instances of evidence of such evolution in every sequenced genome.

Many different types of mutations occur when DNA is duplicated during mitosis (cell reproduction via division); these are discussed in depth in chapter 6. The most common type of mutation is the substitution of one nucleotide for another. Other common types of mutations are the insertion or deletion of 1 or more DNA nucleotides. A comprehensive list of mutations is given in chapter 6.

Any mutation may have an attendant phenotypic effect, i.e. one that affects the physical characteristics of the organism and the way in which it interacts with the environment. When phenotypic effects occur, there may in turn be natural selection pressure either positive or negative, or the phenotypic effect may be neutral. In the case of no phenotypic effect or a neutral phenotypic effect, the mutation is said to be neutral [16].

1.3 Hypothesis

Our hypothesis is that through the application of algorithms, existing and novel to this work, to available sequence, phylogenetic, and related data, a novel model of evolution may be created that will further our knowledge of evolution. We will create such a model of evolution, produce a computer-based implementation of it, run the model with a range of inputs, collect and analyze the results it produces, and draw conclusions about the evolution process.

1.4 Contributions

In this research, we have made the following contributions to computational biology as applied to the understanding of evolution:

1. [Model] A long-term, evolution model based on a wide swath of sequence, phylogenetic, mutation, growth, and related data and the application of existing and novel relevant algorithms. The model is actually a metamodel composed of two distinct models, a sequence evolution model and a population evolution model. This model provides an ongoing, computational test bed for evolutionary theory, as new theoretical and empirical results can be applied to it and their validity tested or consistency verified.
2. [Algorithms] We produced algorithms in two areas:
    a. Sequence evolution modeling using alignment with inversion mutations, mutation probabilities for scoring alignments, and multiple alignments per sequence pair.
    b. Population modeling using a dynamic set of subpopulation pools with related but distinct, mutating genomes reproducing sexually and asexually, and subject to speciation effects, selection pressures, and environmental carrying capacity limitations.
3. [LUCA Genome] Completed first phase of LUCA genome project. Determined gene sequences for LUCA and other LUCA reference species.
4. [Sequence Evolution Model Results]:
    a. Inversions identified as an important mutation mechanism, though typically ignored in sequence alignments.
    b. Discovery that nonhomologous genes are more likely to evolve from coding sequence than from random sequence.
    c. Evidence that some coding sequences more likely to evolve to multiple, nonhomologous genes than other coding sequence. Speculation about universal source sequences.
    d. Estimate of required mutations that occurred in the evolution from LUCA to Homo sapiens.
5. [Population Evolution Model Results]:
    a. Discovery of 4 epochs in the population evolution from one reference species to the next.
    b. Discovery of new population evolution parameter, mating radius.
    c. Analysis of population evolution process sensitivity to multiple parameters of the process.
    d. Identification and analysis of the four most significant parameters in population evolution.
    e. Improved understanding of population evolution process over time.
    f. An independent measurement of long-term evolution duration consistent with that obtained via geologic and fossil duration-related data.
    g. Discovery of the speciation ratchet.
6. [Software]:
    a. To make use of our model, we created a computer-based implementation, the Blind Watchmaker Path (BWMPath) software.
    b. To complete our model runs in an acceptable timeframe by using multiple computer systems at varying locations, we created a simple, distributed processing implementation, Dropbox Distributed Processing (DDP).

2 A simple evolution model

2.1 Structure of the simple model

We assume that there was a primitive organism whose origin is relatively near to the beginning of life on earth and which is ancestral to Homo sapiens as well as all other life on the planet. This organism is known as the Last Universal Common Ancestor (LUCA) [17]. As shown in Figure 1, LUCA reproduced and mutated until eventually one of the LUCA offspring first mutated in some way, occasionally beneficial with respect to natural selection, and, most importantly, that made it different from other LUCA progeny and a more recent ancestor of ours than LUCA – we refer to this individual as More Recent Ancestor 1 (MRA_1) and the mutation as a transitional mutation. Note that the majority of transitional mutations are not speciation (creation of a new species) events, though some are such.

MRA_1's mutation proliferated in the population over multiple generations as it and its progeny reproduced; by definition we know it didn't die out without producing offspring. At some point, one of the MRA_1 offspring again mutated in a way that made it different from MRA_1 and a more recent ancestor of ours than MRA_1 – we refer to this individual as MRA_2. This process repeats up to some MRA_N where N is an integer to be determined that is very likely large.

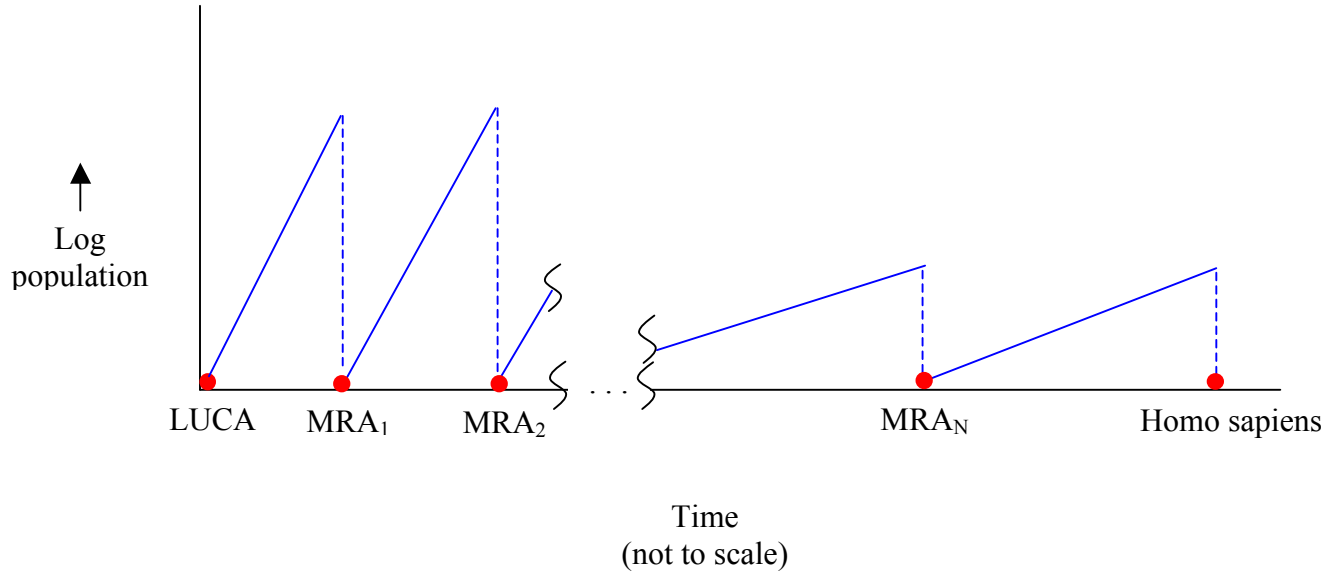

Figure 1: Simple model population timeline

Note in that in Figure 1, as organism complexity increases, e.g. evolves from a unicellular to a multicellular organism, the rate of population growth slows and peak population size is diminished. This is due to the longer, more complex reproductive process. In the simplest unicellular organisms, reproduction consists of a single mitosis. Organism complexity increases over time until reproduction requires meiosis and massive mitosis in combination with cellular differentiation.

Note also that in reality the leading edge of each triangle is neither necessarily smooth nor is monotonically increasing. For a mutation that is detrimental, despite being a transitional mutation, the leading edge of the related triangle will at some times have a negative first derivative, indicating a diminishing population. By definition, however, the progeny of each MRA_i will not die out until an MRA_{i+1} has been produced.

Finally, the expected integral of each triangle is to a first order approximation the inverse of the mutation rate. In particular, if we demand a specific transitional mutation at any given time, the expected number of individuals that must be produced before that specific mutation occurs is the inverse of the rate of that mutation. That is, the lower the rate of the mutation that transitions a predecessor MRA to its successor MRA, the more the predecessor MRA must reproduce and the higher its total population will become before the successor MRA is produced.

2.2 Duration calculation using simple model

To serve as an application and to provide evidence that our approach has merit, we use this simple model to calculate an estimate of evolution's duration. We begin with population calculations. The total number of individuals N produced by time t starting from a single individual is

$$N = (1 + \text{growth rate})^t.$$

Taking the log of both sides of the equation we have

$$\log N = \log (1 + \text{growth rate})^t$$
$$= t * \log (1 + \text{growth rate})$$

and solving for t we obtain

$$t = \log N / \log(1 + \text{growth rate})$$

Equation 1: Population time

The mutations rates we use will be mutations per site per generation; these rates are often abbreviated μ in the literature. The expected number N of individuals that must be produced in order to have a mutation at a particular site is

$$N = 1 / \text{mutation rate}$$

To calculate the time required for this mutation to take place, we substitute into Equation 1

$$t = \log (1 / \text{mutation rate}) / \log(1 + \text{growth rate})$$

Equation 2: Evolution time

As a proxy for LUCA, we use cyanobacteria, which are the source of fossils over 3 billion years old based on the geologic record [18]. For this simplified model, rather than do a comparison, e.g. alignment, of any cyanobacteria, H. sapiens, or intermediate genomes, we take an approach that is limited to point mutations [19]. That is we assume that, for a transitional mutation, a base at a specific site must change from the ancestral, incorrect base to the correct base (Specific Site, Correct Base: SSCB). We further assume that gene duplication, ploidy increases (duplication of an entire genome), and other mechanisms that increase genome size provide sufficient raw DNA for mutation and do so at a sufficient rate such that it is not a significant factor in the model. Thus, in this simple model we do not further consider mutation mechanisms beyond substitutions that increase genome size nor do we allow additional time for them.

For the mutation rate, we cannot use a rate determined by molecular clock models based on fossil data, as this would not allow us to determine a duration for evolution independent of fossil data. Instead, we use a fixed rate experimentally determined for C. elegans [20], an organism whose complexity is intermediate between cyanobacteria and Homo sapiens:

$$\text{mutation rate} = (13/30) * 2.1*10^{-8} \text{ mutations per base pair per generation}$$

We further assume that only the biologically effective portion of the genome has to be correct, i.e. must have mutated from the LUCA bases to their corresponding Homo sapiens bases. We define the biologically effective portion of the genome to be the coding portion plus the highly conserved, noncoding portion under the assumption that this latter portion has biological significance, e.g. contains gene regulatory regions, as evidenced by its being conserved. We use an estimate of the coding portion of the human genome of approx. 2% [21]. In addition, we use an estimate of 2% for the noncoding portion of the human genome that is highly conserved [22]. Thus, we use a total of 4% as the portion of the human genome that evolved from LUCA by changing to the correct Homo sapiens nucleotides.

For this point mutation model, we consider only base substitution mutations. Of the mutations noted above, 13 out of 30 were substitutions. The C. elegans substitution mutation rate is thus $9.1*10^{-9}$ mutations per base pair per generation as described above. We very strictly require that an SSCB change be from the incorrect base to the one correct base out of the three remaining bases. Thus, we use a rate of

$$\text{mutation rate}_{SSCB} = \text{mutation rate} / 3$$
$$= 3*10^{-9} \text{ SSCB mutations per base pair per generation}$$

Note that with unicellular organisms, a generation consists of a single cell division. For multicellular organisms, e.g. sexually reproducing organisms, there are many mitotic divisions between a gamete in one generation and a progeny gamete in the next. Thus we consider mutations per generation.

The reciprocal of mutation rate_{SSCB} is $3.3*10^{8}$. Thus, a population of one must produce an expected number of $3.3*10^{8}$ progeny in order for an SSCB mutation to occur. This equals the expected integral of each individual triangle in Figure 1.

Cyanobacteria have a growth rate of approx. 1 / day.

$$t = \log (3.3*10^{8}) / \log (1 + 1)$$
$$= 28 \text{ days} = .077 \text{ years}$$

Thus, cyanobacteria produce one SSCB mutation every 28 days or .077 years. The 2004 mean human growth rate for the western Sahara is 2.1% annually [23].

$$t = \log (3.3*10^{8}) / \log (1 + .021)$$
$$= 944 \text{ years}$$

Thus the time for a SSCB mutation in Homo sapiens is 944 years. Strictly speaking, no further SSCB mutations need take place in H. sapiens as our species is the final reference species for our model. This time is appropriate, however, for complex, relatively slowly reproducing species that are recent ancestors of ours.

The human genome consists of approx. $3.2*10^{9}$ base pairs [24]. The biologically effective portion, 4% as described above, is $1.3*10^{8}$ base pairs. Starting with our proxy for LUCA, cyanobacteria, this is the number of SSCB mutations are needed to evolve to H. sapiens; this is also the number of MRAs in our simple model.

To calculate evolution's duration, we must know the growth rate of all of the MRAs. For this simple model, we use the mean of cyanobacteria and H. sapiens rates. Since growth rate is multiplicative (we multiply the current population by 1 plus the growth rate to calculate the new population), we calculate the geometric mean of cyanobacteria and human times

$$\text{expected time per mutation} =$$
$$= 8.5 \text{ years}$$

Thus the mean expected time for one SSCB mutation is 8.5 years and the time t_{total} for complete evolution is

t_{total} = expected time per mutation * number of mutations
= 8.5 * $1.3*10^8$
= $1.1*10^9$ years

This is within an order of magnitude of the geologic and fossil data result of approx. $3.5*10^9$ years [18]. This order-of-magnitude agreement between the measurement of evolutionary duration provided by this simple model and the measurement provided by geologic and fossil data offers evidence that our overall approach to the problem has merit.

3 Comprehensive evolution model overview

The simple model is built on a significant number of significant assumptions and simplifications. To the extent that an assumption or simplification is not, in fact, a valid one, it may induce a severe limitation in the simple model. The approach taken in this research has been to create a comprehensive evolution model derived from the simple model by

1. Analyzing each assumption in the simple model and either confirming its validity and retaining it in our comprehensive model or recognizing that it is invalid and eliminating it in our comprehensive model, which will have the effect of alleviating any limitation it induced.
2. Increasing the set of reference species from two (LUCA and Homo sapiens) to many.
3. Increasing the range of data applied to each reference species, for example, in terms of genome sequence, population model, and mutation type.

We begin with an analysis of the simple model, followed by a summary of the comprehensive model.

3.1 Simple model analysis

The simple model described in chapter 2 makes a number of implicit and explicit assumptions and simplifications. Some are valid and their validity will be examined in this section. The remaining ones do not sufficiently reflect reality to apply in our comprehensive evolutionary model and we need to devise means to alleviate them. Our analysis all of these follow.

All bases must change. The first assumption in the simple model is that during the evolution from LUCA to Homo sapiens, all bases (including those that are added over time as the genome grows in size) have to change – none are correct in their ancestral form. By this we mean that every site in a biologically effective region of the Homo sapiens genome had to change from its ancestral nucleotide in LUCA to its current nucleotide. For example, consider a particular site in an exon (a protein coding region in a gene) in the human genome that is an A. In the simple model, we assume its ancestral nucleotide in LUCA was C, T, or G.

This assumption is clearly too strict. One relevant fact is that 24% of point mutations are synonymous for exons – while a single base changes, the amino acid produced upon translation

of the codon (triplet of nucleotides that corresponds to an amino acid in the protein made by the gene) containing it does not [19]. In addition, no attempt at any sort of genome alignment or other comparison between cyanobacteria (our proxy for LUCA) and Homo sapiens was made in the simple model. Such a comparison would identify conserved regions between the two genomes [22], thus reducing the need for bases in those regions to change. Thus, the simple model assumes that evolution took a nearly worst-case path to Homo sapiens, a significantly less unlikely event than a best-case path.

The comprehensive model quantified these conserved regions by using multiple reference species which act as proxies for various MRAs and did an alignment between the genomes of adjacent reference species.

One specific base must change. The simple model assumes that each mutation must be the base at one specific site. That is, in order to progress from MRA_i to MRA_{i+1}, one specific base, out of all of the existing bases in the genome, must change. In reality, at any given time there will typically be multiple possible changes that could take place that would represent progress from the current MRA to its successor. The comprehensive model alleviated this assumption by all possible potential changes (mutations) for MRA transitions at any given time point.

All mutations are point mutations. The simple model limits mutations to substitutions (single nucleotide changes) only. A substantial number of additional mechanisms have been observed, from simple base insertion or deletion to duplication of entire genomes through ploidy change [19]. The comprehensive model considered all mutation mechanisms relevant to the specific reference species characteristics.

No time is needed for genome size increase. The simple model considers substitutions only and that genome size increase mutations, e.g. gene duplication, provide sufficient "raw material" for mutation. No additional time is allocated for genome size increase mutations. The comprehensive model considered relevant genome size increase mutation mechanisms.

Mutation rate is constant. The simple model assumes the mutation rate is constant and uses the experimentally determined rate of an organism, C. elegans, which is intermediate in complexity between LUCA and Homo sapiens. We know that mutation rate is a function of several factors:

1. Species [25]: To alleviate the species-variability of mutation rate, our comprehensive model used rates appropriate to our reference species to the extent possible.
2. Environment [26]: As detailed empirical evidence for the relevant environmental characteristics during the time periods we considered is not generally available, we continued to use constant mutation rates.
3. Genomic context [27]: Highly conserved regions, which have relatively low effective mutation rate, are assumed to be not only biologically effective but, in fact, so significant that mutations in these regions prevent viability or reproduction. However, we claim that the fundamental, spontaneous mutation rate continues to apply. This fundamental, spontaneous rate is masked in measuring allele frequencies, as any mutated allele (alternative form of a gene) is absent from the population as a result of its not being reproduced due to the lack of viability or reproduction of any individual containing the mutated allele.

4. Mutation type: Among point mutations, for example, transitions are sometimes more common than transversions [28]. For our comprehensive model, we used experimentally measured rates when available to refine our results.

MRAs do not regress. The simple model assumes that MRAs do not regress, that is, they do not lose any of their previous SSCB changes through further mutations at those same locations. In reality, a fundamental, spontaneous mutation rate applies at all locations and so regression of MRA transition mutations is inevitable. Our comprehensive model includes MRA regression.

A previous MRA never overtakes a current MRA. The simple model assumes once a new MRA_{i+1} has been produced, no member of the MRA_i population will overtake it. For example, consider a population of MRA_i's. One of them mutates and becomes MRA_{i+1A}. Before its progeny can produce an MRA_{i+2}, a different member of the MRA_i population mutates to become MRA_{i+1B} and one of its progeny mutates to become MRA_{i+2}. Our comprehensive model makes no assumptions about relative evolutionary progress of individuals.

Independence from exogenesis. Exogenesis is the hypothesis that life developed at a location elsewhere, came to Earth by some means, and was the origin of life on Earth. One form of exogenesis is Panspermia [29], the hypothesis that some form of life is ubiquitous in the universe and that this form of life came to Earth, serving as the basis for life here. If exogenesis is correct, we postulate that either that form of life was LUCA or was ancestral to LUCA. Since both the simple and the comprehensive model begin with LUCA, our both models are independent of the factual nature of exogenesis.

A simple population model is accurate. In the simple model, the population model is fully characterized by a single parameter, growth rate, and each MRA reproduces at this rate. Further, growth rate is based on a geometric average between cyanobacteria and human. In reality an accurate population models must be based on a number of parameters:

1. A most significant factor in a population model is the species itself. Between cyanobacteria and Homo sapiens we note a factor of more than 10^4 in difference in the time needed to produce one SSCB change.
2. For many species, the effective population size N_e is smaller than the census population size [30] under some conditions.
3. Some MRA mutations will be beneficial due to improvement of reproduction and/or survival rates; they increase the fitness of the MRA. Those MRAs and their progeny who inherit the beneficial mutation(s) will have a higher growth rate than the remainder of the population.
4. Growth requires external resources and population growth may be limited by environmental carrying capacity [31].

The comprehensive model had multiple population pools and took into account all of the above parameters.

A small fraction of the human genome is biologically effective. The simple model assumes that 4% of human genome is biologically effective [21 22] and ignores any need for mutation

outside of the biologically effective regions. Our comprehensive model made this same assumption; in the future it could be adjusted if and when evidence of biological effects from DNA outside of these regions becomes available.

3.2 A comprehensive model of evolution and its application

In the previous sections, we presented a vision for a unique model of evolution, presented the results of a simple model, analyzed the assumptions it made, and discussed how those assumptions would be retained or eliminated (along with their induced model limitations) in our comprehensive model. Our goal for this research has been to create the comprehensive evolution model, apply it to our set of reference species, and analyze the model results. We now describe this comprehensive model and its application. This description is divided into the set of steps we performed for the research.

3.2.1 Reference species

Our first step was to choose a set of reference species that represent points on the lineage from LUCA to Homo sapiens. Reference species were presumed root ancestors of relevant phylogenetic trees, created using reconstruction algorithms [32]. The first of these reference species is LUCA. The last reference species is, of course, Homo sapiens. For each reference species, actual, relevant genome sequence data must be available and we must estimate the fraction that is biologically effective if it differs from Homo sapiens.

3.2.2 Reference species genome reconstruction

In the simple model described above and the subsequent discussion, we focused completely on evolution at the level of individual nucleotides. It is possible to observe evolution at a number of other levels, among them codon, exon, gene, and gene family. We needed to decide at which of these levels to focus our efforts in the comprehensive model. Since mutations actually occur to nucleotides, we continued to use nucleotide level data while adding codon-level information. An interesting observation is that, given that evolution may be said to operate at all of these levels and that the mutation component of evolution is a stochastic process, evolution may have a fractal geometry [33] aspect to it.

In addition, we had to make a determination of the size of the genome fractions to consider. The ideal would have been to consider whole genomes as this provides a maximum of information for our model. However, considerations of available data, computational resource constraints, and time constraints precluded the use of entire genomes. We thus needed to determine which genome portions to consider and which existing species to use to reconstruct the ancestral, reference species genome.

With these determinations, we used existing tools and our BWMPath software (described in section 10.1.1) to reconstruct the ancestral, reference species genomes.

3.2.3 Mutations

We define a mutation [19] [25] [34] as an alteration of a genome's DNA sequence with respect to the expectation based on the individual's parent(s) genome. Since we are interested in inherited mutations, we consider only germline mutations, which can be inherited by offspring, and ignore somatic mutations.

Our tasks with respect to mutation mechanisms were threefold:

1. Identify all relevant mutation mechanisms. Note that for our purposes we interpreted the term mutation broadly so that it included mechanisms such as horizontal gene transfer.
2. Characterize each reference species with respect to individual mutation mechanisms. Not all species are capable of every mutation mechanism. For example, species that do not reproduce sexually are not capable of unequal crossover.
3. Obtain rates for each mutation type from the literature. Some of these were experimentally determined. We did not perform laboratory experiments to obtain these rates but rather leveraged the existing work of biologists in this field, for example [25].

3.2.4 Sequence evolution model

We then evolved each reference species genome into that of its successor reference species using our evolution model. A common form of modeling this evolution is by a genome alignment using an edit table as shown in Figure 5 and Figure 6. This form of alignment considers only single nucleotide changes, i.e. base insertions, deletions, and substitutions. For our purposes, we must consider the entire suite of mutation mechanisms that have taken place. Some whole genome alignment tools, e.g. MUMmer [35], consider a wider range of mutations but assume that the genomes being aligned are those of closely related species. We expect our reference species to be relatively distantly related. In addition, typical alignment tools find a single alignment that is optimal based on a scoring criterion. Due to the stochastic nature of evolution, it is unlikely to take an optimal path; thus, we determined multiple, probable paths. Due to the limitations of existing alignment tools, we created our own tool as part of the BWMPath software.

In general, we discovered several possible sets of mutations that would transform the genome of a given reference species into that of its successor reference species. We call each such set a mutation path as it represents a path through a generalized edit table and also a path that evolution may have taken. For example, consider hypothetical, adjacent reference species$_i$ and reference species$_{i+1}$ with differing genome segments genome$_i$ and genome$_{i+1}$ where

genome$_i$ = GAGTCC

and

genome$_{i+1}$ = GGACTA

One mutation path from genome$_i$ and genome$_{i+1}$ is simply five single base substitutions, changing all but the first base in genome$_i$ to the corresponding base in genome$_{i+1}$. Another path is an inversion, in which the sequence is reversed and each nucleotide replaced with its complement, of genome$_i$ followed by a single base substitution of the sixth base G with an A. The set of mutation paths between two adjacent reference species constitutes a Hidden Markov Model (HMM) of evolution between the two species [36 37] as shown in Figure 2. We classify the model as a hidden one in that we cannot observe the sequence of states that evolution moved through while progressing from one sequence to a successor sequence.

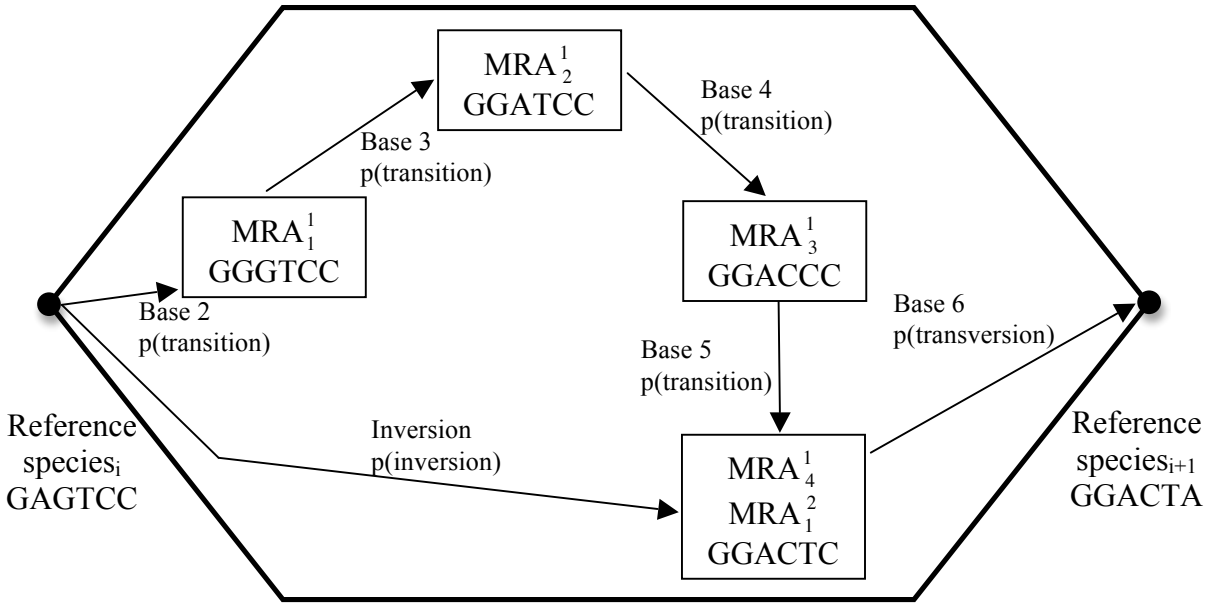

Figure 2: Hidden Markov model of sequence evolution

In the hidden Markov model, the vertices between two adjacent reference species are all of the MRAs between them. Note that MRA numbering is now more complex: MRA numbering is relative to both a specific mutation path and to a specific position between a pair of reference species. In particular, MRA_j^i is the *j*th MRA on mutation path *i* between two reference species. The edges connecting the vertices are labeled with the type of mutation and the probability of that mutation which equals its mutation rate. In our example, note that while there are two paths, they share a common vertex because that MRA has the same genome on both paths, i.e. the 4th MRA on path 1 is the same as the 1st MRA on path 2, i.e. $MRA_4^1 \equiv MRA_1^2$. For convenience, reference species *i* is MRA_0 for each path in the hidden Markov model and reference species *i*+1 is the last MRA on each path.

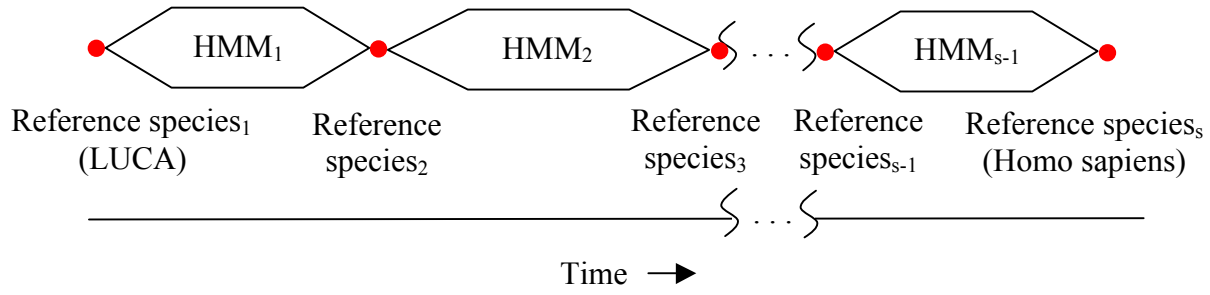

Figure 3: Multiple, hidden Markov model of evolution

We show our comprehensive, hidden Markov model of evolution in Figure 3. It is clear at a glance that it has changed substantially from our simple model. Our first reference species is unchanged as LUCA and our final reference species is unchanged as Homo sapiens; between these are all of the remaining, intermediate reference species *i*. In total, these comprise the reference species selected earlier. Between adjacent pairs of reference species are the hidden Markov models as shown in Figure 2, represented as simplified hexagons in Figure 3.

All of our MRAs are now integral to the hidden Markov models. Each transition of an MRA_j^i to an MRA_{j+1}^i still represents a mutation on a mutation path between a pair of reference species but now there are multiple paths between each pair. We may interpret a wider hexagon in Figure 3 as containing paths whose probabilities are lower than those in narrower hexagons; this also implies that a greater expected number of individuals must be produced between that pair of reference species.

The primary results from our sequence evolution model were the mutation types, lengths and counts in the evolution of one reference species to the next.

3.2.5 Population evolution model

An important goal of our comprehensive model was an understanding of how a population of individuals evolves over time from one reference species to the next [38 39]. Our requirements for such a model were unmet in existing tools, so we created a novel population model as part of the BWMPath software. This model uses a dynamic set of subpopulation pools each containing related but distinct, independently mutating genomes. The individuals reproduce sexually and/or asexually as appropriate to their species and the effects of interbreeding are modeled. The populations are subject to speciation effects, selection pressures, and environmental carrying capacity limitations.

The primary results of this model were population size estimates, evolution duration estimates, and identification of critical evolution parameters and estimation of their values.

4 Reference species

To construct our comprehensive model, we defined a set of reference species that were successive precursors to Homo sapiens. By having intermediate reference species between LUCA and Homo sapiens, we provide a greater quantity of structured (e.g. taxonomy) and raw (e.g. nucleotide sequence) data to our model with the expectation that the model will be correspondingly more accurate in its results. As with LUCA, our intermediate reference species are primarily theoretical constructs, though we know that they must have existed in the past and that they are now extinct. Each of the members of this set of reference species must be individually characterized with respect to phylogenetic relationships, genome content, mutation mechanisms and rates, and population models. The data that makes up these characterizations are the fundamental inputs to the comprehensive model.

4.1 Reference species selection

Initially we defined two of our reference species, the end points in our model: LUCA and Homo sapiens. We then selected several intermediate reference species. Our process was to select a clade to represent the reference species that lies at the root of the clade's phylogenetic tree. A subset of the leaf nodes was then chosen in such a way as to balance the number of species representing a subclade within the clade. These leaf nodes, their sequences, and their phylogenetic relationship were later used to reconstruct the reference species' genome.

Since our research has a significant speculative component, we chose to use data sources in which there is a high degree of trust. We began with the Universal Protein Resource (Uniprot) [40] data, described as the gold standard in protein sequence and related data. We used Uniprot's taxonomic data to guide our intermediate reference species selections.

An important property of our reference species is that they have sufficiently distinct characteristics that there are substantial differences in their genomes. This is to ensure that our evolution model has substantial differences to analyze as it determines mutation occurrences between two adjacent reference species. This suggests a relatively small number of reference species.

LUCA	root clade containing all of life
Eukaryota	structured cells with a nucleus
Metazoa	sexual reproduction, motility, ingesting other organisms
Chordata	active animals with bilaterally symmetric bodies
Craniata	heads with a skull
Vertebrata	contains a vertebral column
Euteleostomi	contains bones
Mammalia	air-breathing, neocortex region in the brain
Eutheria	placental mammals
Primates	large brains
Haplorrhini	vision is primary sense
Catarrhini	diurnal, social groups
Hominidae	capacity for language, simple culture
Homo	

Table 1: Clades containing Homo sapiens

Table 1 lists all of the clades containing Homo sapiens, starting from the clade that represents all of life – we refer to this as the LUCA clade. The sources for this data are [40 41]. In this table each clade is a superset of the clades listed below it. The second column of the table lists significant characteristics that distinguish the clade from its siblings.

We concluded that many of the clade distinctions, e.g. the distinction between Catarrhini and Hominidae, are not substantial enough for our purposes. For this reason, we would not choose both Catarrhini and Hominidae as reference species. In the following sections, we discuss our reference species selections, the reasons for their selection, and their salient characteristics.

4.2 Last universal common ancestor (LUCA)

As discussed in [17], biologists grew confident that there must be a common, universal ancestor to all existing life on earth as the genetic code was deciphered during the 1960s and, as additional species were examined throughout the decade and indeed up to the present time, found to be universal in biological organisms. Since we would like to examine the process of evolution over a substantial period of time, we wish our first reference species to be one taken from relatively early in the history of life on our planet. This quest for an early beginning must be weighed against the need for considerable information about each of our reference species, including this, our first. Since LUCA by definition has significant commonality with all species, we may infer much information about it, using sequence data from existing species, and applying genomic reconstruction techniques and phylogenetic techniques. Thus, we choose LUCA to be our first reference species as it provides an optimum balance among antiquity, universality, and comprehensive knowledge that may be relatively confidently inferred.

There was likely life on earth prior to LUCA. In fact the exogenesis conjecture, discussed earlier, presupposes life (or significant components critical to it such as ribosomes, the cellular organelle that translates messenger RNA into protein) predating that on earth and arriving here by extraterrestrial transport, e.g. comets. Panspermia [29] posits a galaxy ubiquitous with life and that this formed the basis of life here on earth and elsewhere. Since our

model doesn't attempt to describe evolution prior to LUCA, as described earlier it is independent of life and/or its components that predate LUCA.

For most of the twentieth century, it was thought that the fundamental division of life was between Eukaryota (structured cells with a nucleus, e.g. animals such as Homo sapiens) and Prokaryota (relatively unstructured cells without a nucleus, e.g. bacteria such as Escherichia coli) [42]. LUCA would then have been the species after which eukaryotes and prokaryotes split into two separate clades. However, a division was found among the prokaryotes that was both ancient (approx. 3 billion years ago) and unique: the species in one of the divisions were thermophiles, optimally growing at temperatures above 50˚ C. The prokaryotes were then split into two domains: Bacteria and the newly identified Archaea. The matter of whether Eukaryotes are more closely related to Archaea or Bacteria is the subject of current controversy. In fact, there is debate as to which of the three domains, Archaea, Bacteria, or Eukaryota, is most ancient. While it is tempting to assume that the one of the two prokaryote domains with simpler, unstructured cells came first, there is evidence that Eukaryotes are the most ancient and the Archaea and Prokaryotes are more recent simplifications of Eukaryotes.

In [43], research on the nature of LUCA is surveyed extensively and an overall assessment is described which suggests a number of primary LUCA characteristics. Many are diametrically opposed to earlier LUCA characterizations and the veracity of each of these characteristics remains controversial. The most salient of them are:

1. LUCA was a complex protoeukaryote rather than a simple prokaryote.
2. LUCA had an RNA genome rather than a DNA genome.
3. LUCA was a mesophile, growing optimally at moderate temperatures, rather than a thermophile.
4. Horizontal gene transfer played a minor role subsequent to LUCA compared to more likely events, such as gene duplication within a species's genome (creating paralogs), subsequent speciation, and finally differential loss of paralogs, that lead to similar phylogenetic anomalies.
5. Reductive evolution produced the prokaryote domains Bacteria and Archaea.

Several of these characteristics are relevant to our evolution model. LUCA being a complex protoeukaryote rather than a prokaryote has implications for our population model, in particular population growth rate. With their greater complexity and slower cellular machinery, e.g. transcription mechanisms, even single cell eukaryotes in general have a much slower growth rate than prokaryotes.

LUCA having an RNA genome [44] increases the complexity of our modeling its evolution into a reference species with a DNA-based genome. It is clearly the case that the hypothesis of an RNA LUCA is significantly controversial. In [45], Delaye et al argue that the monophyletic (single source) origin of ribonucleotide reductase suggests a DNA-based LUCA. In addition, they note that high fidelity duplication of RNA would be problematic for a large RNA genome. Finally, it is clear that an RNA LUCA implies that DNA-based genomes evolved independently at least twice – an unlikely but not impossible set of events. We believe that, taken together, this argues strongly for a DNA-based LUCA and our evolution model makes this assumption. To adopt the RNA-based LUCA assumption, we could have simply mapped our DNA-based LUCA genome into the corresponding RNA bases; this would have had trivial effect on our results.

Finally, the characteristic of relative rarity of horizontal gene transfer subsequent to LUCA's genesis will affect the mutations component of our evolution model.

Beyond the relevance of these certain characteristics, we do not weigh in on these controversies but continue simply to define LUCA as the most recent, common ancestor of all three domains.

4.3 Last universal common ancestor of Eukaryota (LUCAEukaryota)

As described in [41 46], Eukaryota have a number of characteristics that distinguish them from their sibling clades, Archaea and Bacteria. Among the salient characteristics are:

1. A cytoskeleton consisting of microtubules and microfilaments. Microtubules are hollow rods that undergo continual assembly and dissaembly within the cell. They determine cell shape, function in cell movements including locomotion, and separate chromosomes during mitosis. Microfilaments are smaller but similar to microtubules and also determine cell shape. They allow movement of the cell surface enabling cells to engulf particles and divide.
2. An endomembrane system composed of related membranes within the cytoplasm. These include endoplasmic reticulum, Golgi bodies, vacuoles, lysosomes, peroxisomes, and the nuclear membrane.
3. A primary genome of multiple, linear (as opposed to circular) chromosomes within a nuclear membrane. Some species have multiple nuclei. During mitosis, the genome is replicated and the copies segregated.
4. Mitochondria organelles that usually contain a genome distinct from the primary one. Mitochondria have diverse functions, e.g. aerobic respiration and synthesis of small molecules such as amino acids.

Many, but not all, Eukaryotes also have the following characteristics:

1. They form multicellular organisms.
2. They reproduce sexually.

It is these latter characteristics, distinguishing Eukaryota from LUCA, that motivate us to select LUCAEukaryota as one of our reference species. In particular, sexual reproduction augments the set of mutations that occur universally with mutations that occur only in sexually reproducing species; this is further described in chapter 6. This substantial, additional set of mutations is sufficient reason to make LUCAEukaryota a reference species.

4.4 Last universal common ancestor of Metazoa (LUCAMetazoa)

Metazoa are commonly known as animals. They share the following characteristics [47-50]:

1. They are heterotrophic, engulfing external food and digesting it in an internal chamber.
2. They lack rigid cell walls. Instead, metazoan cells are surrounded by an extracellular matrix composed of collagen and elastic glycoproteins.
3. They are capable of motion during at least one life stage.

4. They have a developmental stage effectively defined by regulatory DNA. Plants have this, too, but their development is significantly different.

This last characteristic of Metazoa, regulatory DNA controlled development, is of special interest for our purposes. Animals as different as worms, flies, and humans use similar sets of genes to define their development from egg cell fused with spermatozoon to adult organism. A small number of intercellular signaling pathways, defined by the genome and incorporating multiple proteins, are used repeatedly in different organisms at different times during development. These pathways ultimately create the differentiated cells, organized into tissues and organs, in the adult organism. The wide differences in adult organisms of Metazoan species arise to a large extent from differences in the regulatory DNA of the genes that define these pathways.

It is primarily this last characteristic, regulatory DNA controlled development, that motivates us to make LUCAMetazoa a reference species. This is due to the substantial functional impact this regulatory DNA has on the Metazoan genome.

4.5 Last universal common ancestor of Mammalia (LUCAMammalia)

Mammals share the following characteristics [51-53]:

1. They are warm-blooded and air breathing.
2. They are vertebrates.
3. Mammalian mothers nourish their young with milk or a similar, secreted substance.
4. Their brains have a neocortex region.

The characteristic of mammals having a neocortex is of special interest for the purpose of reference species selection. There is a large diversity in morphological patterns and organization of the neurons that occurred during the evolution of the neocortex. In addition, there are a variety of neuron types with unique regional distributions. Finally there are a number of structural proteins, calcium-binding proteins, and neurotransmitters in the neocortex. In aggregate, the development of the neocortex has had not only a selective advantage for mammals but also a functional impact on the mammalian genome. As a result, we chose LUCAMammalia as a reference species.

4.6 Homo sapiens

In our evolution model, we needed an early biological species for our initial reference species and we chose LUCA to be our initial reference species. Similarly, we needed an final reference species. Our final reference species should be one that has recently evolved to allow our model to examine evolution over a large span of time. Our final reference species must be fully sequenced to allow us to model the evolution of its sequence from earlier reference species. Finally, the ideal final reference species should have a complex genome to provide a wide variety of sequence content to model.

Species	*Common name*	*Mbp*
Colostethus marchesianus	poison dart frog	6700
Homo sapiens	human	3200
Mus musculus	mouse	2400
Drosophila melanogaster	fruit fly	120
Arabidopsis thaliana	mouse-ear cress	100
Caenorhabditis elegans	roundworm	78
Saccharomyces cerevisiae	yeast	12
Escherichia coli	bacteria	4.6
H. influenzae	bacteria	1.8

Table 2: Genome sizes

Homo sapiens satisfies well each of these criteria. We evolved some 300,000-400,000 years ago, which is extremely recently in evolutionary time [54]. We were fully sequenced at the conclusion of the Human Genome Project [55]. Our genome is not the largest but is substantial; see Table 2 [56 57], which lists some genome sizes in millions of base pairs (Mbp). As important, it contains the full complement of eukaryotic complexity in addition to content unique to our species. Thus, Homo sapiens was chosen as our final reference species. An additional advantage to this choice is that there is a natural curiosity about our origins and this research may serve to more fully satisfy that curiosity.

5 Reference species genomes

Having selected our reference species, we then reconstructed the genomes of these Homo sapiens ancestors. We begin with a discussion of the theory of and algorithms applied to ancestral genome reconstruction. We then describe the process we used for our reconstructions. Finally, we discuss the reconstructed genomes we used for our model.

5.1 Ancestral genome reconstruction background

Using sequence information from existing species and the phylogenetic relationships among those species, one can infer and reconstruct sequence information for ancestral species [58-66]. There are three phases in the process of ancestral genome reconstruction:

1. Sequence alignment
2. Phylogenetic tree creation
3. Ancestral sequence reconstruction

Sequence alignment identifies which sites in a set of sequences correspond to each other. Phylogenetic tree creation identifies the evolutionary relationships among the sequences. Ancestral sequence reconstruction takes the information from the previous two phases and determines which character (nucleotide or amino acid) is most likely present at each site in the sequence being reconstructed. Each phase is described below.

5.1.1 Sequence alignment

The process of determining which nucleotides in two or more sequences correspond is called sequence alignment [67].

```
cgaaagcggcgttccgaccttcagcggggccatggatggactgt
 |||||  ||  ||||||||||||||  ||  ||     | |       |
agaaagtggtgttccgaccttcagaggagctggaggt---tatt
```

Figure 4: Sequence alignment example

Figure 4 contains a portion of the alignment between a LUCA gene coding region, reconstructed as part of the research effort for this thesis, and its Homo sapiens ortholog. The first line of the figure is the LUCA nucleotide sequence and the last is the Homo sapiens sequence. Letters in

the two sequences that are arranged directly over one another are said to be matched, i.e. the alignment process has inferred that they are corresponding sites in the sequences. If two matched letters are identical, the match is known as an identity and a vertical bar connects them in Figure 4. Two dissimilar letters indicate that one of the sequences has undergone a substitution. A site where one or more sequences has a "-" indicates an insertion or deletion (indel). At these sites, a sequence with a "-" may have undergone a deletion or a sequence with a letter may have undergone an insertion. With no additional information, it is not possible to distinguish between these two cases. In our alignments, we know which sequence is ancestral and we place this sequence above the descendent sequence. Thus, a "-" in the upper sequence indicates an insertion in the descendent and a "-" in the lower sequence indicates a deletion in the ancestor.

Any two sequences have a number of possible alignments that is exponential in their length. As a result, much effort has been expended in determining "best" alignments in a computationally tractable manner [68-73]. Best alignments must be determined by a scoring method based on the probabilities of identities, substitutions, and indels in sequences.

An example of a simplistic alignment score, derived from [67], is the following. Assume an alignment with i identities, s substitutions, and d indels. If the probabilities of individual identities, substitutions, and indels are p, q, and r respectively, then the probability P of the alignment is:

$$P = p^i q^s r^d.$$

We define S' by the log likelihood of P:

$$S' = \log P = i(\log p) + s(\log q) + d(\log r).$$

Finally, we derive score S from S':

$$S = i - s\,\mu - d\,\delta,$$

where μ and δ are the total substitution and indel probabilities normalized against the identity probability, respectively. The substitution probability has been empirically shown to be higher than the indel probability. To summarize this scoring approach, identities increase a score while substitutions and indels decrease a score. It is frequently the case that, rather than using the terminology of the score of an alignment, the distance between two sequences is described. The distance between two sequences is larger if their alignment the score is smaller.

With an ability to score different alignments against each other to determine the best one, we now address the problem of constructing the alignments in the face of their exponentially large quantities. The typical approach is the use of edit graphs [74].

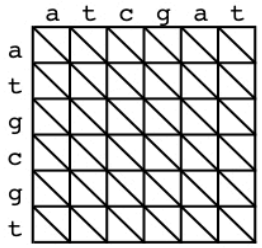

Figure 5: Edit graph

Figure 5 is an edit graph for two sequences of identical length, though edit graphs may be used for sequences of dissimilar lengths. The first sequence is written along the left side of the edit graph and the second sequence along the top. Any path through the edit graph from top left corner to bottom right corner, following diagonal, vertical, or horizontal moves through cells, represents an alignment. A diagonal move through a cell indicates a match: a solid line represents an identity and a dashed line represents a substitution. A downward move represents a nucleotide the first sequence has that the second sequence lacks while a rightward move represents a nucleotide the second sequence has that the first sequence lacks. These two latter moves cover any indel.

It is clear that any two sequences may be trivially aligned via a path that first moves to the bottom of the graph, indicating the second sequence lacks all of the nucleotides in the first sequence and then moves to the far right of the graph, indicating the first sequence lacks all of the nucleotides in the second sequence. Since there are only indels and no identities or substitutions in such a path, it has a minimal score and represents a worst-case alignment.

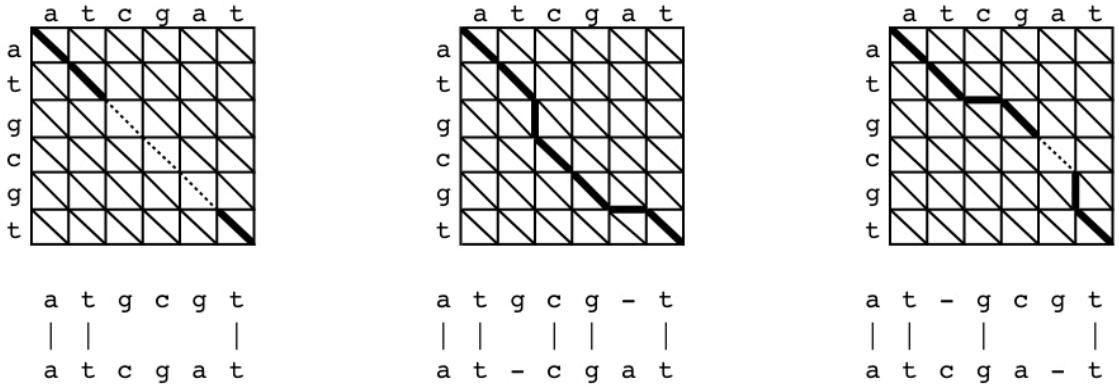

Figure 6(a, b, c): Paths through edit graphs

Figure 6 shows three paths through the same edit graph and their corresponding sequence alignment. The first line in the sequence alignment is the source sequence with '-'s indicating insertion locations; these bases are missing in the source sequence. The next line indicates identity matches between the two sequences with a '|' and substitution matches with a space. The final line is the target sequence with '-'s indicating deletion locations, these bases are missing in the target sequence.

In Figure 6(a), the path contains only matches, three identities and three substitutions. The path in Figure 6(b) contains five matches, all identities, and two indels, one in each sequence. Finally, the path in Figure 6(c) contains five matches, four identities and one

substitution, and two indels, one in each sequence. The paths are ordered left to right from best score to worst. Due to the relatively low probability of indels, paths without them score best and so (a) has the highest score. While paths (b) and (c) have identical numbers of matches and number of indels, path (b) scores better than (c) because all of its matches are identities while one of the matches in (c) is a substitution.

A common approach to identifying high-scoring paths through an edit graph is the application of dynamic programming [75-77]. Dynamic programming is a method of solving complex problems by breaking them down into simpler steps, in particular as a series of recursive steps. The task of finding a path through an edit graph may be recursively broken down into finding a path through smaller and smaller edit graphs that are a subsets of the entire edit graph.

Dynamic programming benefits from multiple techniques that reduce the path search space of an edit graph. Branch and bound techniques can put bounds on path scores, eliminating entire branches of paths through an edit that fall outside of the established bound from further consideration. In addition, heuristics can also limit the search space at the cost of not finding the globally highest scoring path. One example of such a heuristic is to limit the search to paths near the diagonal of the edit graph as these will contain more matches. A full description is given in Waterman [67]

It is frequently the case in this thesis, and in general, that an alignment must be found among more than two sequences; this is known as the multiple sequence alignment problem [78]. The issue of an exponentially large search space that is found when aligning two sequences is further exacerbated when aligning multiple sequences. One approach to multiple sequence alignment is to consider the mutation distance of each site in each sequence to that of the same site in a common ancestor.

Another approach is to let the total score for an alignment of n sequences to be the sum of the alignment scores of the n*(n-1)/2 pairs; this is known as a sum of pairs score. It is possible to simply compute an upper bound on the sum of pairs score and so significantly reduce the search space.

5.1.2 Phylogenetic tree creation

The problem of phylogenetic tree creation is illustrated in Figure 7.

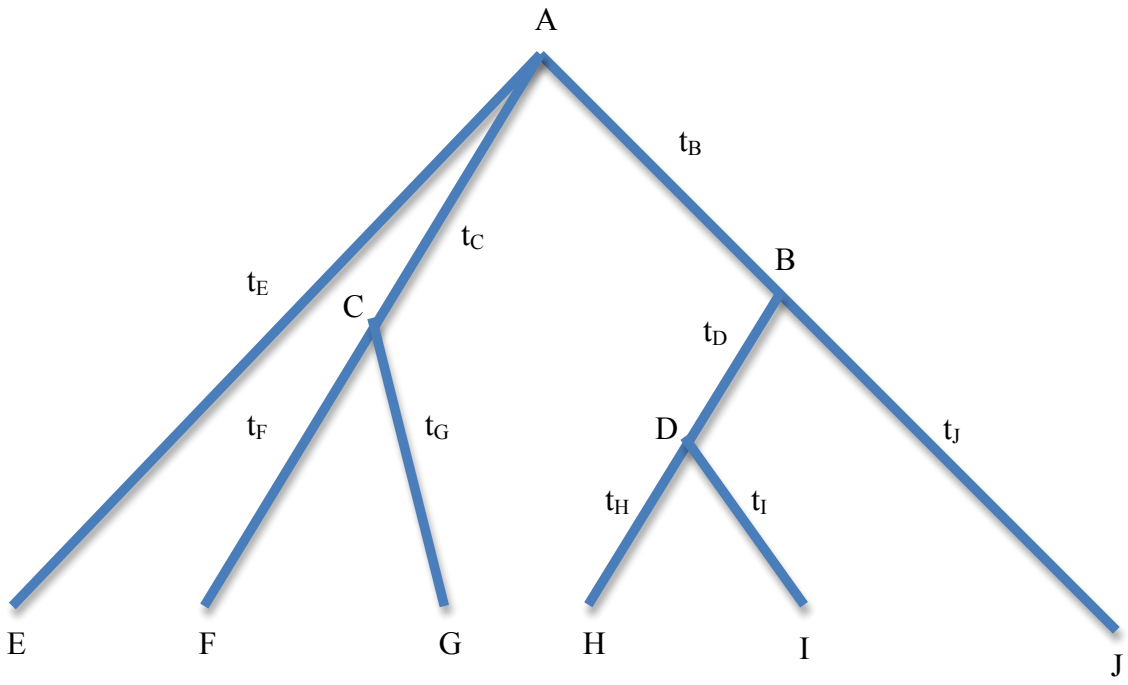

Figure 7: Phylogenetic tree example

In the tree depicted in Figure 7, nodes are labeled with capital letters. Each of the nodes represents a species with an associated genome sequence. The leaf nodes (E-J) represent existing species (referred to as operational taxonomic units – OTUs) and their sequence information is obtained from sequencing the actual DNA of a species sample. The remaining nodes (A-D) are ancestral species (referred to as hypothetical taxonomic units – HTUs), of which one or more we wish to infer the associated sequence information. The t_B through t_J represent the length of the edge from a node to its predecessor; they typically represent the inferred time taken for the sequence to have evolved from its predecessor sequence.

Inferring a phylogenetic tree [64 79 80] is done by means of an estimation procedure -- given that there is incomplete information (only the sequence information at the leaf nodes, representing existing species, is known), one can only produce a best estimate of the tree based on some objective criterion. There are a variety of algorithmic approaches to this problem, which essentially fall into one of two categories: maximum parsimony (MP) and maximum likelihood (ML).

Maximum Parsimony

In the maximum parsimony category of phylogenetic tree reconstruction, [81 82], the objective criterion for evaluating trees is the number of evolutionary changes, e.g. nucleotide insertions, deletions, or substitutions in the case of DNA sequences. Trees constructed using maximum parsimony have a minimum number of such evolutionary changes. For a given set of data, it may be that more than one tree has the same minimum number of changes and so there is no unique solution.

Each sequence in the set has an identical number of sites; this simplifies our maximum parsimony algorithm as we need only consider nucleotide substitutions rather than substitutions, insertions, and deletions combined. Each site in the known sequences is either:

- invariant, if all OTUs have the identical value at the site, or
- variable, if at least one OTU has a different value at the site compared to the other OTUs

Variable sites are further distinguished as being:

- informative, if the site is able to distinguish among potential phylogenetic tress based on the number of evolutionary changes (parsimony) required to produce the variations observed at the site, or
- uninformative

To create a maximum parsimony tree, we first identify the variable sites and, from these, choose the informative sites. Then, for each possible phylogenetic tree containing all of the OTUs, we calculate the minimum number of evolutionary changes needed to produce the variations observed at the site. Finally, we sum the number of changes required at each informative site for each possible tree. The tree (or trees) with the minimum number of required changes is the maximum parsimony tree.

Determining the number of evolutionary changes (substitutions) required can be accomplished using Fitch's algorithm [81]. We need only consider variable sites. For an informative site, the value(s) at an HTU site is the intersection of the values of the same site at all of its immediate descendents if the intersection is nonempty. If the intersection is empty, it is the union of the values of the immediate descendents. When a union is required, a nucleotide substitution must have occurred at this site. Thus the number of unions equals the number of evolutionary changes required to produce the variation at the site that is observed in the descendents.

For an uninformative site, the number of changes is simply equal to the number of different nucleotides observed at the site minus one. We subtract one because, for example, if there are two variations then there has been one change that changes one variant into the other. The total number of changes at both informative and uninformative sites is called the tree length and is our parsimony measure.

When there are only a small number of OTUs, we can exhaustively enumerate the trees and calculate the tree length of each. However, the number of possible trees grows rapidly with the number of OTUs. With ten OTUs, the number of rooted trees exceeds 30 million. Exhaustive methods are impractical when the number of OTUs is significant.

Two methods are applied to reduce the tree search space. The first is branch-and-bound [83]. At any given point in the search, the tree length of the highest parsimony tree found so far serves as an upper bound and is compared to the current tree length for the tree being calculated. Once the tree length of the tree being calculated exceeds upper bound, it may be abandoned as the current upper bound tree is shorter and so has higher parsimony.

The second method is the application of heuristic approaches to prune the tree search space. In such approaches, only a tractable subset of all possible trees is considered. The essential principle is that an initial tree is constructed by an appropriate method such as neighbor-joining [84]. In neighbor-joining, pairs of OTUs are found that minimize total branch

length. Similar trees (by an appropriate objective measure of similarity) are then examined to see if a shorter tree can be found. If one is found, then it serves as a new starting point for the examination of similar trees. The trial iterations are terminated when at a certain trial threshold we fail to find a shorter tree. These heuristics approaches tradeoff a guarantee of finding the most parsimonious tree against acceptable compute time.

Maximum Likelihood

In the ML category of phylogenetic tree reconstruction [85-88], the likelihood of a given tree is the probability of observing the data given the tree and a model of mutations used in constructing the tree, for example a substitution-only model. This is written L = P(data|tree) which says that L, the likelihood of the tree, equals the probability of the data given the tree. ML methods explore the search space of possible trees to find the one with the highest likelihood.

An example adapted from [79] illustrates the principles. Assume we have the sequences of four OTUs and we wish to determine the ML tree for them.

	Site 1	Site 2
OTU1	C	G
OTU2	C	G
OTU3	A	G
OTU4	G	G

Table 3: OTU sequence data

In Table 3, the sequence data at two sites in our four OTUs is presented. In Figure 8, one of the three possible, unrooted trees that may be constructed from four OTUs is shown. In this tree, the two, internal nodes are identified as HTU5 and HTU6.

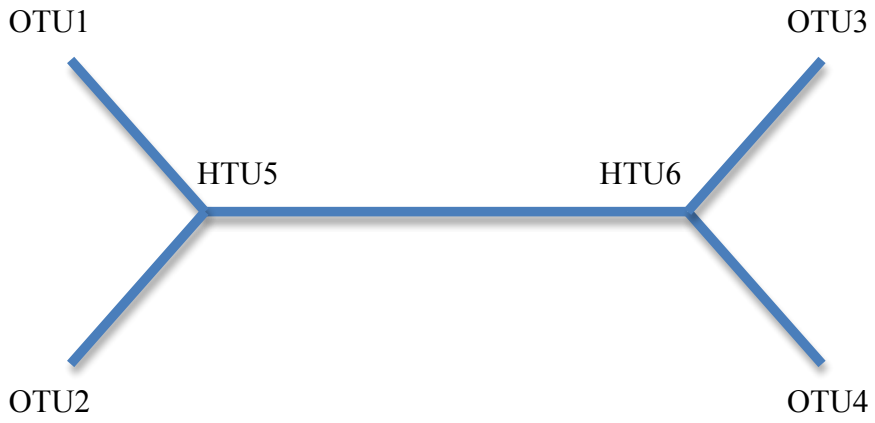

Figure 8: Maximum likelihood tree example

We first consider site 1. At this site, The OTUs have data C, C, A, and G respectively. The internal nodes of the tree can each have any of A, C, T, or G since any nucleotide can mutate into

any other. Thus, for the sequence data at the internal nodes we must consider $4^2 = 16$ possibilities. Some of these combinations are significantly more likely than others. For example, C at HTU5 and G at HTU6 is more likely than G at HTU5 and C at HTU6 as the former requires one-quarter the number of mutations compared to the latter. However, none of the possibilities has a zero probability and so the entire group of 16 must be considered. Thus, for this tree the likelihood of observing the data that we do observe in the OTUs is the sum of the 16 independent probabilities. That is, the likelihood of site 1 given the tree example in Figure 8 is

$$L_{(1)} = \sum_{i \in \{A,C,T,G\}} \sum_{j \in \{A,C,T,G\}} P \left(\begin{array}{ccc} C & & A \\ & i \quad j & \\ C & & G \end{array} \right)$$

Equation 3: Phylogenetic tree likelihood

Note that probabilities depend on nucleotide mutation rates, for example, the rate that a C mutates to a G. With the ML methods, branch lengths are estimated together with tree topology.

ML models have emerged as superior to MP models as numerous studies [89] have noted the disadvantages of the latter models.

5.1.3 Ancestral sequence reconstruction

With existing sequences aligned and their relationships captured in phylogenetic trees, the final phase in ancestral genome reconstruction is the actual sequence reconstruction [59 62-64 66]. As with phylogenetic tree creation, ancestral sequence reconstruction has two common approaches, maximum parsimony and maximum likelihood. The following descriptions of the two approaches are adapted from [60].

Maximum parsimony

As with maximum parsimony in phylogenetic tree reconstruction, maximum parsimony in ancestral sequence reconstruction attempts to minimize change. In this case, it seeks to minimize the number of changes needed for the ancestral sequence, being reconstructed at the root of a phylogenetic tree, to evolve to the leaf sequences.

The Fitch algorithm [81] proceeds as follows. The tree is traversed in post order, assigning a set of nucleotides at each node, beginning with the internal nodes one level up from the leaf nodes – the set at each leaf node is simply the nucleotide at that site in the leaf sequence. The set at an internal node is the intersection of the sets of its descendents if non-empty or the union otherwise. If the union is assigned, one change is counted in the reconstruction. The total change count is the total number of union sets used in the reconstruction. At the root, if the set consists of more than one nucleotide, then multiple, equally parsimonious, reconstructed sequences exist.

Maximum likelihood

For an explanation of the ML approach to ancestral sequence reconstruction, we refer back to Figure 8 and Equation 3 in the explanation of the ML approach in phylogenetic tree creation. The ML approach in ancestral sequence reconstruction is very similar and the likelihood

equations are analogous. For tree creation, the goal is to find most probable tree topologies and branch lengths. For sequence reconstruction, the goal is to find the most probable nucleotide at a given site in an ancestral sequence.

5.2 Ancestral genome reconstruction process

With the reference species selected, we then proceeded with the genome reconstruction process. We decided to focus our model at the nucleotide level (as opposed to higher levels of genome abstraction such as genes or gene families) since that is where the mutations we model actually took place. We did take into consideration codon synonymy (the fact that several different codons code for the same amino acid and so are functionally equivalent) where appropriate. The first phase of this process was to determine the orthologs among the existing species for which the reference species is an ancestor. The determination of the orthologs requires protein data and so we consider here only coding DNA sequences. The second phase was to determine the DNA sequence of each ortholog of each reference species. These two phases are described below.

5.2.1 Ortholog group selection

Once again, we began with the Universal Protein Resource (Uniprot) [40] data. We used Uniprot's taxonomic data and considered only those entries that are species. This excludes, e.g., all but one of the variants of bacterial species that are known as strains. We also included only those entries for which a lineage (the nested set of clades which contains the species) is specified so that we could properly represent our chosen clades.

The Uniprot taxonomic data indicates the three major clades in biology: Archaea, Bacteria, and Eukaryota. These clades are also referred to as domains or kingdoms. The Archaea clade contains just five subclades, some of which have only a single species. When we reconstructed the genomes of our reference species, we wanted equal weighting from the clades that comprise our reference species. This prevents any clade comprised of a large number of species from dominating the reconstructed reference sequence result. Since some of Archaea's five clades only contain a single species, we allowed only five species to represent each of the other two clades. Where there was a choice of subclades, we chose the most populous ones under the assumption that they are the most studied and so are likely to provide the most accurate data.

Our first approach to finding genes in common used the Uniprot gene data. Remaining conservative with our data selections, we considered only genes that have been reviewed for accuracy. Analyzing this data with BWMPath, we found only six genes in common within the species of the Archaea clade. This small number of genes would likely further diminish with the addition of more clades. However, since this approach yielded too few common genes in Archaea for our purposes even without additional clades, we did not pursue it further.

We continued with a different approach using gene homology, again using a conservative data source, the Orthologous MAtrix (OMA) project [90 91]. Genes are described as homologous if they have a common ancestor gene and thus are related through one of two evolutionary mechanisms:

1. A gene may be duplicated in a species and the two copies may evolve independently; these homologous genes are paralogs.

2. A speciation event, where a new species divides itself from an existing species, may occur; the corresponding genes in each species' genome are homologous and are known as orthologs.

For our research, we needed to find orthologs among the clades of our reference species and the OMA data is ideal for this purpose. The OMA project has generated a total of approx. 500,000 ortholog groups from over 4 million proteins. The following algorithm generated the ortholog groups:

1. Protein sequences were obtained from GenBank [92] and Ensembl [93]. Full alignments were performed between all pairs of sequences using dynamic programming [67], calculating similarity based on a comprehensive amino acid substitution matrix [94]. Highly similar pairs form the initial set of potential orthologs.
2. Stable pairs of sequences are then identified. A stable pair of sequences, x in genome X and y in genome Y, is one in which x is more similar to y than any other sequence in Y and y is more similar to x than any other sequence in X. Since the similarity measures are estimates, variance is also estimated and a 95% confidence interval is used.
3. The stable pairs identified in the previous step will include paralogs as well as orthologs. In order to detect that a pair of sequences, x in X and y in Y, are paralogs, an exhaustive search is made of all other genomes Z for homologs of x and y that could indicate a paralogous relationship [95]. Such paralogous pairs are deleted from the set and the remaining pairs are described as verified pairs.
4. A graph is then constructed with the sequences at the vertices and edges connecting verified pairs with edges weights being the sequence similarity. Maximum weight cliques are identified in the graph and these cliques make up the OMA ortholog groups.

The groups identified by the above algorithm comprise the central database of the OMA project.

```
558  NEQGADK
METMA01264     PYRHO01249     PYRAB01059     SULTO02195
CHICK09725     HUMAN26738     MOUSE02520     ERWCT01784
CANFA06539     PYRKO00317     BOVIN07909     PANTR01190
SULAC00094     STAHJ02450     MONDO04964     MACMU00889
METBF00912     CIOSA02016     RABIT07485     DASNO02879
SHESR01720     ORNAN08520     HAES101486     SPETR05182
MYOLU09455     CAVPO02632     THEPD00118     SHESW01757
METS300095     SHEB502213     SHEB802397     SHEPC02173
SHEB902510     GEOUR04234     HORSE07527     METM500006
NEMVE25830     HAES200504     TURTR08192     PTEVA07112
CAEJA06662     CHOHO10123     TAEGU16266     SULIA01981
SULIY02129     SULIL02055     SULIM01910     SULIN00635
MACEU01389
```

Figure 9: OMA entry for group 558

An OMA database entry for group 558, which is included in our LUCA genome, is shown in Figure 9. The entry begins with the group number followed by a "fingerprint", which is a

sequence of contiguous amino acids that uniquely identify the protein. Following these are the 49 OMA protein ids for the orthologs that comprise the group. Each id consists of a five-letter identifier for the species and a five-digit number for the protein of that species. Some species are divided into different strains, each with its own unique identifier; in our ortholog group selection process, we considered at most one strain per species based on the available taxonomy information. For example, HUMAN26378 is a Homo sapiens deoxyribonuclease protein. Since Homo sapiens is our terminal reference species, we required any group we select to have a HUMAN protein.

558	NEQGADK		
	Eukaryota		
		HUMAN26738	Homo sapiens
		NEMVE25830	Nematostella vectensis
		CAEJA06662	Caenorhabditis japonica
		CIOSA02016	Ciona savignyi
		CHICK09725	Gallus gallus
		MONDO04964	Monodelphis domestica
		ORNAN08520	Ornithorhynchus anatinus
		BOVIN07909	Bos taurus
		MACEU01389	Macropus eugenii
		DASNO02879	Dasypus novemcinctus
		TAEGU16266	Taeniopygia guttata
		MOUSE02520	Mus musculus
		MYOLU09455	Myotis lucifugus
		CHOHO10123	Choloepus hoffmanni
		HORSE07527	Equus caballus
		CAVPO02632	Cavia porcellus
		TURTR08192	Tursiops truncatus
		PTEVA07112	Pteropus vampyrus
		MACMU00889	Macaca mulatta
		SPETR05182	Spermophilus tridecemlineatus
		PANTR01190	Pan troglodytes
	Bacteria		
		ERWCT01784	Erwinia carotovora subsp. atroseptica
		SHESR01720	Shewanella sp. (strain MR-7)
		SHESW01757	Shewanella sp. (strain W3-18-1)
	Archaea		
		METMA01264	Methanosarcina mazei
		SULTO02195	Sulfolobus tokodaii
		PYRHO01249	Pyrococcus horikoshii
		PYRKO00317	Pyrococcus kodakaraensis
		SULAC00094	Sulfolobus acidocaldarius
		PYRAB01059	Pyrococcus abyssi

Table 4: Ortholog group 558 proteins and species

We needed to select a number of ortholog groups for each reference species. Each reference species is at the root of a clade containing multiple subclades. For each ortholog group, our goal is to have the same number of species represent each subclade in a given clade, so that each subclade has equivalent effect on the reconstructed genome. For example, Table 4 lists the species for group 558. Note that the Bacteria subclade has two entries that appear to be different strains of the same species. However, the taxonomy data indicates that they are different species and since we consistently apply this information, we treat them as different species. The distinction between species is challenging for non-sexually reproducing species and this taxonomic result is simply an example of that challenge. The Bacteria subclade has the least number of species, three, of all of the subclades. Hence, we must select only three species from each of the other subclades to meet our balance goal. This aspect of the selection process is described below.

In determining the number of groups for the clade of a reference species, there is a tradeoff between the number of subclades included and the minimum number of species required for each subclade. An increase in the subclade count that represents in a group decreases the number of applicable groups. An increase in the minimum number of species required for each subclade also decreases the number of applicable groups. As we sought to maximize diversity in terms of subclades, species, and ortholog groups, we explored the subclade and minimum required species parameter space for each reference species to determine the number of applicable groups.

To explore this tradeoff, we used BWMPath, iterating over a range of subclade counts and number of species required per subclade, to search for groups meeting each set of parameters. For subclade count *c*, the *c* most populous subclades in the clade were used. For each clade count, the number of required species was increased until no groups were found. For a given parameter set of subclade count *c* and required number of species *s*, BWMPath reads the entire OMA group set and selects groups that:

1. Contain the specified *c* subclades
2. Contain at least *s* species from each subclade
3. Contain a Homo sapiens protein since that is our final reference species

Once an appropriate compromise among subclade count, number of required species per subclade, and number of applicable groups was selected, the specific species (and therefore proteins) applied had to be selected. Since the required species parameter is a minimum, many groups and subclades within groups have more than the required species available and we needed to select which species to use in this case, constrained by having an equal number of species for each subclade. To make these species choices, we chose to maximize diversity within the subclade.

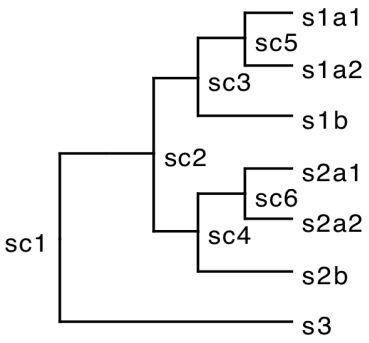

Figure 10: Species tree

Any given species has a location in the overall taxonomic tree. Figure 10 shows such a tree with species s1a1 to s3 at the leaf nodes and subclades sc1 to sc6 at intermediate nodes. We use Dendroscope to view our trees; it is described in section 10.1.8. One approach to maximizing diversity is to choose the set of species that have the greatest distance sum. Distance between two species could be measured by determining the nearest, common subclade of two species and summing the number of edges from each species to the common subclade. In our example, s1a1 and s2a1 would have a distance of six while s1a1 and s3 would have a distance of five. This distance measure indicates that s1a1 is closer to s3 than it is to s2a1. However, the subclade divisions, e.g. sc3 and sc5, are somewhat arbitrary in number and location; thus distances measured in this way tend to give arbitrary results – s1a1 is actually more closely related to s2a1 than it is to s3.

We note that s1a1 and s2a1 have a nearest common subclade, sc2, which is closer than s1a1 and s3's nearest common subclade, sc1. We thus defined a closeness measure between two species that counts the number of subclades the species have in common. Species s1a1 and s2a1 have two subclades in common, sc1 and sc2; s1a1 and s3 have one subclade in common, sc1. Thus s1a1 and s2a1 are closer than s1a1 and s3. To maximize diversity in our final species selection, we chose species that have the minimum closeness sum.

We note that we could create a phylogenetic tree of all of the species in a group and base our diversity selections on that tree. However, such a tree would necessarily be based on only a single sequence for each species, the ortholog for that group. We chose instead to use the overall taxonomic tree that is much more broadly based.

To assemble a diverse set of species for a given subclade, we first included in the set any required species. Our only required species is Homo sapiens, which of course is present in at most one subclade. We then added species in order of decreasing diversity, which is equivalent to increasing closeness. Initially when the set is empty, we added the species that is least close to all of the other species. If there were already species contained in the set, we added the species that was least close to the species already contained in the set. The genes in Table 4 are listed in order of decreasing diversity within their respective subclades.

A given group in a given reference species must have had at least the minimum required species in each subclade to be included in that reference species' group selection. However, the group may have more than the minimum in each subclade. In this case, the maximum number of proteins in the group was used within the constraint that the same number must be used in each subclade.

At this point, we searched for common groups identified among our references species (we did not include Homo sapiens here since it is the final reference species and has no ortholog

groups but only ortholog genes). We found that there were none. This was due to the differing group inclusion criteria used for each reference species. Yet we knew that LUCA must have orthologs with each of the later reference species and we desired to model the evolution between such orthologs.

group 558		
	Eukaryota	
		HUMAN26738
		NEMVE25830
		CAEJA06662
	Bacteria	
		ERWCT01784
		SHESR01720
		SHESW01757
	Archaea	
		METMA01264
		SULTO02195
		PYRHO01249

Table 5: Group selected proteins

Table 5 lists the proteins used for LUCA group 558. Three proteins are used for each subclade. We can see that group 558 has orthologs in the LUCAEukaryota reference species, namely HUMAN26738, NEMVE25830, and CAEJA06662. To include group 558 in the more recent reference species LUCAEukaryota, we lifted it to the more recent reference species by reconstructing it for LUCAEukaryota using only the three Eukaryota proteins rather than all nine proteins used in reconstructing it for LUCA. In this way, only sequences from species within the subclade were used for reconstructing the reference species sequences.

Our sequence determination method, described below, required a minimum number of sequences to produce valid results; this minimum is four sequences in total; this includes species from all relevant subclades. In cases where there were not enough species in the lifted group's subclade to meet the minimum number of sequences requirement, we added enough, maximally diverse species from the subclade to meet the minimum. In the case of lifting group 558 to LUCAEukaryota, we needed to add one more species to make the minimum requirement. We added CIOSA02016, shown in Table 4. In a case where there were not enough species available to meet the minimum requirement for valid results, we omitted the group.

For each reference species, we lifted all of its specific groups to all later reference species. In each case, the lifted group included only species contained in the later reference species subclade and maximal diversity was maintained.

With our subclades, species, and proteins selected, we used sequence data from the OMA project to make sequence files, one for each group. Each group sequence file for each reference species contained the sequences for all of the selected proteins for that group in that reference species. These files served as input to the second phase of this process.

5.2.2 Sequence determination

The second phase of the ancestral genome reconstruction process is sequence determination, which attempts to recreate the nucleotide sequences of our chosen reference species. Once again, we took a conservative approach by including only those entries for which the vast

majority, 90%, of the DNA sequence is specifically known – this excluded a few entries that include too large a fraction of unknown sites, signified by X's in the sequence data.

After considerable experimentation with a number of existing computational biology tools for sequence alignment, phylogenetic tree creation, and ancestral genome reconstruction, we settled on a "pipeline" of tools, mediated and augmented by BWMPath, for our sequence determination. The tools selected for the pipeline were those that technical reviews and our own experience indicated were most accurate in the context in which they were applied. As is typical with pipelines, in our pipeline each stage takes input from the previous stage and produces output for the next stage.

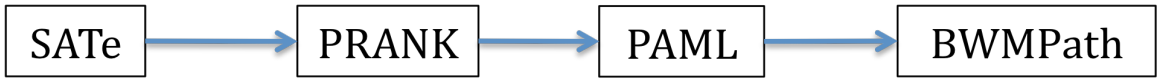

Figure 11: Sequence determination pipeline

Figure 11 depicts the pipeline we used for sequence determination. The sequences for each ortholog group determined in phase one for each reference species were the input to the initial stage of the pipeline; the reference species sequence for that ortholog was output from the final stage of the pipeline. Each group in each reference species was passed through the pipeline.

The first stage of our pipeline was Simultaneous Alignment and Tree estimation (SATe); it si described in section 10.1.5. SATe is the highest performing tool we found to both align sequences and create the phylogenetic tree relating the sequences. Its input was the set of sequences for an ortholog group from one of the reference species produce by BWMPath. It output both an alignment of the sequences and the phylogenetic tree. We used the phylogenetic tree in the next stage of our pipeline. The alignment produced by SATe was not ideal for our purposes and was discarded.

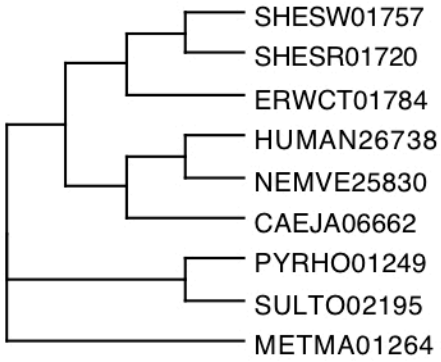

Figure 12: LUCA group 558 phylogenetic tree

Figure 12 shows the LUCA group 558 phylogenetic tree produced by SATe. This group contains three species (proteins) from each of three subclades. As expected, the three proteins from a subclade are all on the same branch of the tree. For example, the Eukaryotes Homo sapiens (HUMAN26738), Starlet sea anemone (NEMVE25830), and Japanese round worm (CAEJA06662) are all on the same branch.

The next stage in our pipeline was PRANK; it is described in section 10.1.7. PRANK took as input the set of sequences from an ortholog group and the phylogenetic tree produced by SATe. PRANK produced a sequence alignment and ancestral genome sequences. In the latter, PRANK excels at correctly inferring indels.

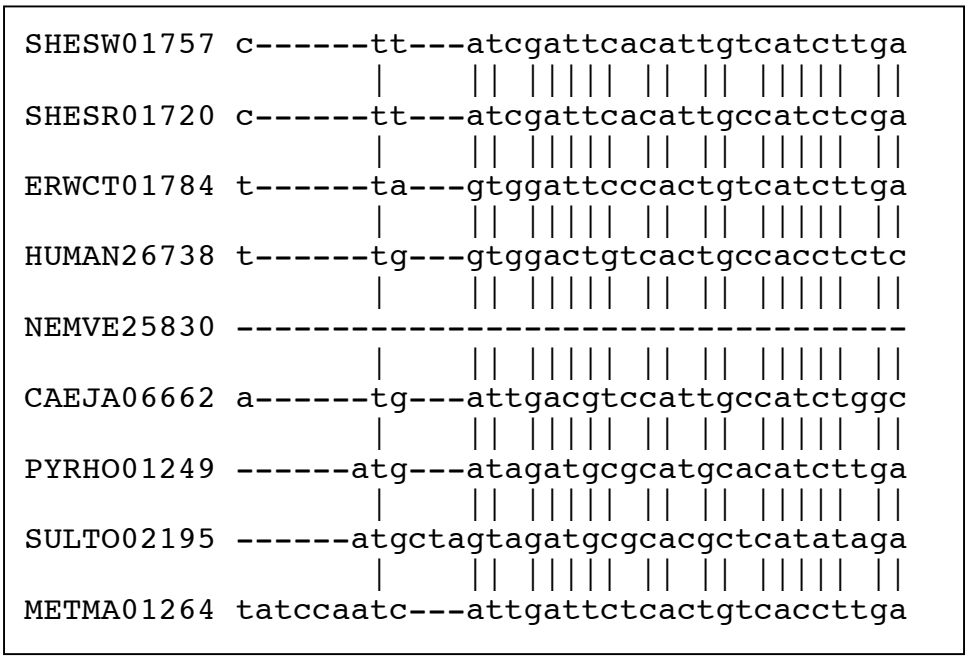

Figure 13: PRANK group 558 sequence alignment

In Figure 13, we show a portion of the PRANK alignment for LUCA group 558. Vertical bars indicate sites where there is 66% or higher agreement on which nucleotide is present.

We used both the sequence alignment and the ancestral sequence at the root of the phylogenetic tree (which is the sequence for our reference species) produced by PRANK.

The next stage in our pipeline was Phylogenetic Analysis by Maximum Likelihood (PAML); it is described in section 10.1.6. PAML took as input the SATe tree and the PRANK alignment. PAML produced as output a sequence alignment and ancestral genome sequences. With the latter, PAML excels at correctly inferring which nucleotide is present at a given site. We used the root sequence of the ancestral sequences from PAML for input to the next stage.

The final stage in our pipeline was performed by BWMPath; it is described in section 10.1.1. The PRANK root ancestral sequence had the most accurate indels, while the PAML root ancestral sequence had the most accurate nucleotides. BWMPath created a combined root ancestral sequence by, for each potential site, including the nucleotide from the PAML sequence if the PRANK sequence indicated that the site is present.

PRANK attggatgggtatccatgc----------tg
PAML **atggaatacaaacccatgc**ggtccttttttatttatctgcttggaaattgctg**ct**
Combo atggaatacaaacccatgcct

Figure 14: Combining PRANK and PAML ancestral sequences

In Figure 14, we show a portion of the ancestral sequence for LUCA group 558 produced by PRANK and a portion produced by PAML as well as the sequence that resulted from combining the PRANK and PAML results. At the sites where the PRANK sequence had a nucleotide and so indicated that that site was present in the ancestral sequence, the corresponding site in the PAML sequence, shown in bold, was included in the combination ancestral sequence. At the

sites where the PRANK sequence showed a "-", indicating the site was absent in the ancestral sequence, no nucleotide was included in the combination ancestral sequence.

5.2.3 Parallel processing

In total there were over 400 ortholog groups among the reference species, each of which needed to be run through the sequence determination pipeline at least once. Most groups were run through the pipeline multiple times as the ortholog group selection and sequence determination processes were refined, and all groups were run through the final version of the pipeline. Initial pipeline runs on a contemporary model Apple Macbook Pro laptop demonstrated that running a single group through the pipeline just once could take as much as 8 hours. Fortunately, our ancestral genome reconstruction process is what computer scientists describe as an embarrassingly parallel problem, because each group may be processed independently of all of the others. Thus it is especially amenable to the application of parallel processing, where multiple processors operate on different parts of the overall problem simultaneously. In the case of the sequence determination process, a problem part is simply running a single group through the pipeline.

In a desire to complete this research prior to the conclusion of the current geologic epoch, we designed and programmed a suitable parallel, distributed computing framework. We call our framework Dropbox Distributed Processing (DDP); it is described in section 10.1.2. DDP was used to process problem parts in parallel; in the case of sequence determination, DDP accomplished this by running multiple instances of the pipeline simultaneously with each pipeline instance applied to a single group at a time. DDP provided coordination services, e.g. telling the processors which reference species to process, and mutual exclusion assurance, i.e. allowing at most a single processor to apply the pipeline to any given ortholog group and excluding all other processors from doing so.

This effort was entirely successful. DDP provided linear speedup of the ancestral genome reconstruction pipeline. At the peak of available processing resources, we had a total of 55 processors (actually cores, each used independently) simultaneously running our code in the DDP framework at three disparate locations.

While DDP was recognized as a necessity and then created during the reference species genomes phase of this research, it was also used for all subsequent compute-intensive tasks. We would not be writing this thesis today (it would be at a much later date) if not for the productivity increase provided by DDP.

5.3 LUCA genome

Considerable effort has been expended on characterizing LUCA's putative genome [17 42-44 96-99]. We noted in section 4.2 that LUCA may have had either a DNA or an RNA genome and that we used the former assumption in our model, though adopting the latter assumption would have had trivial effect. Forterre and Poole [17 98] conclude that LUCA's genome was complex and contained elements otherwise unique to Eukaryota. This implies that Archaea and Bacteria evolved their genetically simpler state through extensive gene loss.

Glansdorff et al [43] conclude that LUCA's genome was redundant. Phylogenetic inferences on its metabolism produce a sophisticated result that can be taken as an indication of generalized genetic redundancy. Further, it is likely that most LUCA cells possessed more than a single copy of essential genes as there would be positive selection for this characteristic given the primitive, error-prone division mechanism in LUCA.

Using a minimum gene set approach, Koonin [97] noted that only about 60 proteins, mostly involved in the translation of RNA to proteins, e.g. ribosomal proteins, are universal in all species. His estimate is that LUCA had 500-600 genes. Inferring ancestral gene content using a large number of extant genome sequences and their phylogenetic tree, Ouzounis [99] estimates just over 1000 genes in LUCA.

In our research, we did not attempt to define a comprehensive genome for LUCA. Rather, we sought to identify those genes that LUCA had in common with all three domains of life and those it had in common with the subclades that represent each of our reference species. These latter genes are lifted to the later reference species as described in section 5.2.1.

The LUCA subclades are shown in Table 6, which was generated by BWMPath. The number following a subclade name indicates the total number of species contained in the clade. For the sake of brevity, subclades containing only a single species are not listed; the species totals in the table are correct, but will not sum correctly due to the omission of the single species subclades. Subclades within a subclade are shown indented and below the containing subclade; they are listed in order of descending number of contained species.

LUCA	6937		
	Eukaryota	6303	
		Metazoa	3941
		Viridiplantae	1916
		Fungi	255
		Alveolata	65
		Rhodophyta	40
		stramenopiles	38
		Euglenozoa	13
		Amoebozoa	8
		Haptophyceae	7
		Cryptophyta	6
		Rhizaria	4
		Diplomonadida	2
		Parabasalia	2
		Heterolobosea	2
	Bacteria	575	
		Proteobacteria	288
		Actinobacteria	78
		Firmicutes	76
		Cyanobacteria	33
		Bacteroidetes	26
		Tenericutes	16
		Chloroflexi	6
		Fusobacteria	6
		Aquificae	5
		Spirochaetes	5
		Thermotogae	5
		Chlorobi	5
		Deinococcus-Thermus	3
		Planctomycetes	3
		Acidobacteria	3
		Verrucomicrobia	3
		Chlamydiae	3
		Elusimicrobia	2
	Archaea	59	
		Euryarchaeota	40
		Crenarchaeota	15
		Thaumarchaeota	2

Table 6: LUCA clade

Table 7, also generated by BWMPath, presents information on the groups in the LUCA clade. The first column lists the number of subclades being considered. For the LUCA clade, this number is a constant three, as we wanted to include all of the subclades. The second column lists the minimum number of species that we require to represent each subclade. The third column lists the number of ortholog groups found in the OMA data given the number of subclades and the minimum required number of species representing each of the subclades. For example, in the first row with three subclades, we required only a single species to represent each subclade and

we found 101 groups. In the second row, we required a minimum of two species instead of one to represent each subclade and the number of groups found was reduced from 101 to 62.

# subclades	*# species*	*# groups*	*product*
3	1	101	303
3	2	62	372
3	**3**	**35**	**315**
3	4	26	312
3	5	20	300
3	6	16	288
3	7	12	252
3	8	9	216
3	9	7	189
3	10	7	210
3	11	5	165
3	12	3	108
3	13	2	78
3	14	2	84
3	15	2	90
3	16	2	96
3	17	1	51
3	18	0	0

Table 7: LUCA ortholog groups

One approach to selecting which row in the table to choose for our LUCA genome is to maximize the product of subclades, species, and groups; this product is shown in column four. Such a criterion tends to maximize diversity while providing a relatively large number of groups. Based solely on this criterion, we would have selected the second row as it has a maximal product of 372. However, we augmented this criterion by also requiring that no individual factor in the product be too small. Row 2 has only two species / subclade while later rows have a greater number of species; however, as we increase the number of required species, the number of groups found decreases. Row 3, in bold, offers a good compromise for the LUCA genome, requiring at least three species per subclade while still offering 35 groups in total. For the LUCA genome, we used the parameters in row 3: all three of the LUCA subclades with at least three species representing each subclade.

The sequence for each of these LUCA genes that is a homolog of a group was reconstructed from the gene sequences in the group. For example, Table 5 lists the nine gene sequences used to reconstruct LUCA's homolog to the ortholog group 558 genes. In each case for LUCA, the same number of genes from each subclade, three or more, was used in the reconstruction. The reconstruction itself was effected by using the selected group sequences as input to our genome determination pipeline. For group 558, the sequences for nine genes listed in Table 5 were the pipeline input. The output of the pipeline was the sequence of the LUCA homolog to the ortholog group. After removing some problematic ones, our LUCA genome has a total of 33 genes, one for each ortholog group.

5.4 LUCAEukaryota genome

In comparison to the quantity that has been applied to LUCA's genome, research on LUCAEukaryota's (and later reference species) genome is scant. Dacks and Doolittle [100] note that it would be possible to reconstruct the evolution of LUCAEukaryota if some existing Eukaryotes retained features of LUCAEukaryota and if we could identify which Eukaryotes these were, but that it is not clear the either situation is true. They conclude that the best approach is to apply information from many Eukaryote genome sequences to the problem, but do not further pursue this research. Arisue et al [101] used an ML analysis of 22 genes to determine the root of the LUCAEukaryota phylogenetic tree but did not reconstruct the genome.

As with the LUCA genome, we sought not to reconstruct a comprehensive LUCAEukaryota genome, but to determine its orthologs to genes in a subset of its subclades in a conservative, balanced manner. Table 8 lists the Eukaryota subclades and species counts. Again, subclades containing only a single species are not listed.

Eukaryota	6303		
	Metazoa	3941	
		Chordata	2364
		Arthropoda	1049
		Mollusca	183
		Cnidaria	82
		Echinodermata	71
		Annelida	71
		Nematoda	39
		Porifera	25
		Platyhelminthes	24
		Sipuncula	5
		Ctenophora	3
		Bryozoa	3
		Nemertea	2
		Nematomorpha	2
		Rotifera	2
		Hemichordata	2
		Brachiopoda	2
		Acanthocephala	2
		Echiura	2
	Viridiplantae	1916	
		Streptophyta	1865
		Chlorophyta	51
	Fungi	255	
		Dikarya	226
		Fungi incertae sedis	12
		Microsporidia	7
		Neocallimastigomycota	3
		Chytridiomycota	3
		Blastocladiomycota	2
		Glomeromycota	2
	Alveolata	65	
		Dinophyceae	24
		Apicomplexa	22
		Ciliophora	18
	Rhodophyta	40	
		Florideophyceae	32
		Bangiophyceae	7
	stramenopiles	38	
		PX clade	20
		Bacillariophyta	8
		Oomycetes	5
	Euglenozoa	13	
		Kinetoplastida	10
		Euglenida	3
	Amoebozoa	8	
		Mycetozoa	3
		Archamoebae	2
		Tubulinea	2
	Haptophyceae	7	
		Isochrysidales	2
		Coccolithales	2
	Cryptophyta	6	
		Pyrenomonadales	4
		Cryptomonadales	2
	Rhizaria	4	
		Cercozoa	3
	Diplomonadida	2	
		Hexamitidae	2
	Parabasalia	2	
	Heterolobosea	2	
		Schizopyrenida	2

Table 8: Eukaryota clade

With LUCAEukaryota and its successor reference species, there is a tradeoff between the number of subclades included and the minimum number of species required to represent each clade. The results of exploring this tradeoff using BWMPath are shown in Table 9. The number of groups obtained goes down with the product of number of subclades and minimum required number of species. For example, two subclades and four species yield 426 groups. Increasing the subclade number to three reduces the number of groups to 289; similarly, increasing the number of required species from four to five with two subclades reduces the number of groups to 71.

# subclades	*#species*	*# groups*	*product*
2	1	1445	2890
2	2	1121	4484
2	3	818	4908
2	4	426	3408
2	5	71	710
2	6	0	0
3	1	976	2928
3	2	740	4440
3	3	549	4941
3	4	289	3468
3	**5**	**45**	**675**
3	6	0	0
4	1	251	1004
4	2	0	0
5	1	0	0

Table 9: LUCAEukaryota ortholog groups

Since no ortholog groups are found when the subclade count is five (or greater) and Eukaryota has significantly more than five subclades, it is clear from these results that it is not possible to include all of the Eukaryota subclades. Using the augmented criterion that was applied to the LUCA clade, we choose a total of three subclades with at least five species representing each subclade, shown in bold in Table 9. Removing some problematic groups, this results in 43 groups for LUCAEukaryota, a tractable number and similar to the number of LUCA groups. In addition, we lift the 33 groups in LUCA's genome to LUCAEukaryota. This results in a total of 76 genes in LUCAEukaryota's genome.

5.5 LUCAMetazoa genome

As with the LUCAEukaryota genome, research on the LUCAMetazoa genome is scant. In early work, Lake [102] concluded that Metazoa is a monophyletic (sharing a unique, common ancestor) taxon by examining ribosomal RNA sequences. Larroux et al [103] concluded that LUCAMetazoa had a cluster of specific genes involved in the development of overall body shape but did not research beyond this cluster.

Table 10 lists the Metazoa subclades and species counts. Again, subclades containing only a single species are not listed.

Metazoa	3941		
	Chordata	2364	
		Craniata	2350
		Urochordata	13
	Arthropoda	1049	
		Hexapoda	785
		Crustacea	141
		Chelicerata	117
		Myriapoda	6
	Mollusca	183	
		Gastropoda	96
		Bivalvia	63
		Cephalopoda	17
		Polyplacophora	5
		Scaphopoda	2
	Cnidaria	82	
		Anthozoa	62
		Hydrozoa	9
		Scyphozoa	6
		Cubozoa	5
	Echinodermata	71	
		Eleutherozoa	71
	Annelida	71	
		Polychaeta	57
		Clitellata	14
	Nematoda	39	
		Chromadorea	36
		Enoplea	3
	Porifera	25	
		Demospongiae	23
		Calcarea	2
	Platyhelminthes	24	
		Cestoda	9
		Trematoda	8
		Turbellaria	6
	Sipuncula	5	
		Sipunculidea	5
	Ctenophora	3	
		Cyclocoela	2
	Bryozoa	3	
		Gymnolaemata	3
	Nemertea	2	
		Anopla	2
	Nematomorpha	2	
		Gordioida	2
	Rotifera	2	
	Hemichordata	2	
		Enteropneusta	2
	Brachiopoda	2	
		Linguliformea	2
	Acanthocephala	2	
	Echiura	2	

Table 10: Metazoa clade

As with LUCAEukaryota, we used BWMPath to determine the number of groups that match a range of subclade count and minimum required species parameters. These results are shown in Table 11.

# subclades	*# species*	*# groups*	*product*
2	1	3807	7614
2	2	2740	10960
2	3	2137	12822
2	4	1575	12600
2	5	921	9210
2	**6**	**256**	**3072**
2	7	0	0
3	1	2399	7197
3	2	0	0
4	1	1648	6592
4	2	0	0
5	1	0	0

Table 11: LUCAMetazoa ortholog groups

Using our augmented selection criterion, we choose a total of two subclades with at least six species representing each clade. This results in 256 groups for LUCAMetazoa, a number significantly larger than that for the earlier clades. In order to have comparable numbers of groups for each reference species, from the 256 groups we select the 43 groups with the most proteins. This is similar to the number of groups in LUCAEukaryota. Choosing the groups that are the most populous gives us the widest possible sets of proteins from which to draw. Our LUCAMetazoa genome then has 43 genes plus 76 genes lifted from earlier reference species for a total of 119 genes.

5.6 LUCAMammalia genome

As with our earlier, intermediate reference species, little published research has been performed on the LUCAMammalia genome sequence. Bourque et al [104] analyzed the gross genomic architecture in terms of rearrangements of LUCAMammalia sequences over time. Goodman et al [105] analyzed the history of the mammalian beta-globin gene family and Graves and Watson [106] researched the evolution of mammalian sex chromosomes, which are highly conserved.

Table 12 lists the Metazoa subclades and species counts. Again, subclades containing only a single species are not listed.

Mammalia	843		
	Eutheria	760	
		Laurasiatheria	404
		Euarchontoglires	321
		Afrotheria	21
		Xenarthra	14
	Metatheria	80	
		Diprotodontia	33
		Dasyuromorphia	22
		Didelphimorphia	15
		Peramelemorphia	5
		Paucituberculata	3
	Monotremata	3	
		Tachyglossidae	2

Table 12: Mammalia clade

Again we use BWMPath to determine the number of groups found based on subclade and species parameters. These results are shown in Table 13.

# subclades	*# species*	*# groups*	*product*
1	1	20625	20625
1	2	20138	40276
1	3	19035	57105
1	4	18137	72548
1	5	17489	87445
1	6	17015	102090
1	7	16587	116109
1	8	16192	129536
1	9	15782	142038
1	10	15361	153610
2	1	13747	27494
2	**2**	**7382**	**29528**
2	3	0	0
3	1	6474	19422
3	2	0	0

Table 13: LUCAMammalia ortholog groups

Using our augmented selection criterion, we choose a total of two subclades with at least two species representing each clade. This results in 7382 groups for LUCAMammalia, the largest number of groups yet. Once again we reduce the number of groups in order to have comparable numbers of groups for each subclade. As with LUCAMetazoa, for LUCAMammalia we select the 44 most populous groups out of the 7382 groups in total. Together with 119 genes lifted from earlier reference species, our LUCAMammalia genome has 163 genes.

5.7 Homo sapiens genome

As an existing species, there was no need to reconstruct the Homo sapiens genome but merely to sequence it. We do need to identify a subset of the genome to model. We included in our HS

genome all of the orthologs lifted from all of the earlier reference species. In addition, we need to include a set of genes that are unique to HS, i.e. not present in the earlier reference species, just as we have done for the other reference species.

To determine which genes are unique to HS, we used the OMA data to search for human genes that have no orthologs. These are genes that not only do not belong to any groups, but also have no pairwise orthologs. We found that there are a total of 875 such HS genes. To reduce this total, we omitted those for which no function is known and found that 68 genes remained. We also omitted those whose function is similar to that of another gene in this set. There remained 39 genes with known, unique function.

We might expect these genes to provide functions unique to our species, e.g. a function related to neocortex development. However, that is not typically the case. Some exemplary functions for these genes are:

1. structural constituent of ribosome
2. nucleic acid binding
3. DNA binding
4. ATP binding
5. RNA binding

These broad functions are common throughout the clades. Since we observe that a significant majority of the HS unique genes have an unknown function, we conclude that among them are a number whose function we associate with uniquely HS characteristics.

Our selected 39 genes appear to be unique to Homo sapiens. Together with 163 lifted genes from earlier reference species, we have a total of 202 genes in our Homo sapiens genome.

For our population model, we required additional information about the HS genome. We needed its effective length and again used $1.3*10^8$ bases from chapter 2.

We also required the fraction of HS genes that are homologous to LUCAMammalia. The approach we took was to use Mus musculus (house mouse) as a proxy for LUCAMammalia. The mouse genome analysis [107] presents several relevant results: over 90% of the mouse and human genomes have corresponding regions of synteny (genes occurring on the same chromosome), approximately 80% have a single ortholog in the HS genome, and less than 1% have no homolog in the HS genome. It seems likely that the actual value is between 80% and 99% homologs between LUCAMammalia and HS; we used the approximate mean of 90%, which is also consistent with the synteny result.

6 Mutations

The nucleotides in a strand of DNA are typically replicated without error during the process of cell division (reproduction). In some instances, an error will occur; such errors in the DNA copying process are known as mutations. When a mutation occurs in any cell in an asexually reproducing species or any cell line leading to a gamete (including gametes themselves) in a sexually reproducing species, it is passed on to the next generation and so may be a source of variation upon which natural selection can act. This is a central tenant of evolutionary theory: errors in DNA duplication are the source of individual variation upon which natural selection acts.

The salient characteristic of biological mutation is that it is a relatively rare occurrence. The assessment in [20] measured the empirical rate for substitutions (one nucleotide changing to another) to be $9.1*10^{-9}$ mutations per base pair per generation. Thus, observing mutations typically requires substantial time spans.

Alberts et al [108], Futuyma [19], and Gascuel [34] in aggregate provide good coverage of mutation types. Mutation rates tend to be measured in a relatively piecemeal fashion, probably owing to the typically lengthy time period required for such experimental results; Denver et al [20], Drake et al [25], Rosenberg and Hastings [26], Zang and Yoder [28] , and Barrick et al [109] are exemplary. In the case of Barrick et al, multiple generations of E. coli, a rapidly reproducing bacteria species, were sampled over a period of 20 years.

In our model we assumed spontaneous mutation rates are relatively independent of location, e.g. spontaneous rates are essentially identical in highly conserved regions and in rapidly evolving regions. The primary reason that observed rates vary as a function of location is because mutations in highly conserved regions prevent viability or reproduction and therefore do not fix and do not appear in the general population, as they are culled in no more than a single generation. Hence, conserved regions appear to have a low mutation rate only because organisms with mutations in such regions do not proliferate in the population.

There has been considerable effort devoted to the neutral theory of evolution, primarily by Kimura [110], which states that the rate of evolution in terms of nucleotide substitutions is sufficiently high that most mutations must be nearly neutral with respect to natural selection. If that were not the case and approximately half of all mutations induced phenotypes for which there is significant negative selection pressure, many fewer individuals would proliferate and extinction would be a much more common event. There is significant evidence in favor of the neutral theory [111] as well as against it [112]. In our model, we made little in the way of assumptions with respect to the neutral theory. We assumed only that transitional mutations,

those that transition an MRA to its successor MRA, are not sufficiently deleterious to preclude the successor MRA from proliferating to the point where its successor is produced.

We organized mutation types into four categories as shown in Table 14.

Mutation Type	Genome size change	Sexual species
Intragenic mutation		
Single nucleotide substitution		
change of one base to another		
Nucleotide insertion	x	
insertion of new base(s)		
Nucleotide deletion	x	
deletion of existing base(s)		
Chromosome segment inversion		
loop in chromosome formed, then reversed (and complemented)		
Microsatellites		
repeat count variation in small, repeated sequences	x	
DNA segment rearrangement		
Crossover		x
normal meiotic crossover		
Unequal crossover	x	x
crossover between imperfectly aligned chromosomes		
Chromosome fission		
chromosome breaks into two		
Chromosome fusion		
two chromosomes fuse into one		
Chromothripsis		
rare, massive DNA rearrangements		
Transposable elements		
moving sequences, carrying transposase gene, sometimes nearby genes		
Gene gain and loss		
Chromosome gain and loss	x	
gain or loss of an entire chromosome		
Polyploidy	x	x
change in the number of whole sets of chromosomes		
Horizontal transfer		
Horizontal gene transfer	x	
bacteria through plasmids; other species through retroviruses		

Table 14: Mutations

6.1 Empirical Mutations

As described above, significantly deleterious mutations tend not to be observed. That is, mutations that result in early cell fatality will not be observed by typical methods. Thus, the rates described here represent lower bounds. Early fatality mutations have the effect of lowering the observed growth rate of a species and so are treated in our population evolution model as an effect on growth rates.

6.1.1 Transitions and Tranversions

As described earlier, for the simplest mutation type a single, different nucleotide is substituted, typically during mitosis, for the original nucleotide. Two types of substitutions are identified, transitions and transversions. If the process were unbiased, we would expect transitions to happen half as often as transversions, since of the three substitutions that can occur to a given nucleotide, one is a transition and two are transversions. A bias has often been noted in mutation rates, with transitions happening more frequently than expected in the purely unbiased case [28 113]. However, this bias is not universal but rather is species specific. Keller et al [114] investigated four genome sequences of Podisma pedestris, a grasshopper species with a very large genome, and found no significant differences between transition and transversion rates.

In the cases where this is such a bias, the Kimura two-parameter matrix model [115] may be applied. In this model, transitions occur at rate α per site and transversion occur at a rate of β per site. The transition rate bias for this model is $\alpha/(2\beta)$. With this model, Kimura derived the probability of observing a transition $P(t)$ or transversion $Q(t)$ at a site in two sequences separated by time t:

$$P(t) = \tfrac{1}{4} + \tfrac{1}{4} e^{-8\beta t} - \tfrac{1}{2} e^{-4(\alpha+\beta)t}$$

and

$$Q(t) = \tfrac{1}{2} - \tfrac{1}{2} e^{-8\beta t} .$$

From these equations, it is clear that

$$\lim_{t\to\infty} P(t) = 1/4$$

and similarly,

$$\lim_{t\to\infty} Q(t) = 1/2 .$$

Hence for large t, the transition bias, $P(t)/Q(t)$, approaches the value in the unbiased case, $\tfrac{1}{4} / \tfrac{1}{2} = \tfrac{1}{2}$. Since our model involves a wide variety of species for which transition bias is not likely to be universal and since our reference species are comparatively widely separated in time, we assume the transition bias for our model to be ½, i.e. an unbiased transition model.

6.1.2 Overall mutation rate

Drake et al [25] surveyed mutation rates over virii and all biological domains. They found rates as high as 1/genome per generation in virii. They defined effective genome size as the portion of a genome where mutations are likely to be deleterious. For microbes, the effective genome size is approximately equivalent to the total genome size, while for Eukaryotes the effective genome size was much less than the total size. They found a somewhat unexpected constant value: for all biological domains, a mutation rate of 1/300 per cell division per effective genome was found.

Of course, the mutation rate per site per generation varies widely as both the number of cell divisions per generation and the genome size vary substantially: the number of cell divisions per generation vary from 1 for bacteria to as many as 600 in Homo sapiens males [116] and

genome size varies varies over 7 orders of magnitude, from approximately 10^5 to 10^{12} bases [117]. Effective genome size likely varies over a comparable magnitude.

There is evidence for elevated mutation rates in response to environmental factors [26]. Since we are modeling evolution over long time periods under conditions we cannot accurately replicate, we do not further consider these environmentally induced rate variations.

6.1.3 Substitutions

Substitutions, where a single base changes to a different one, are the cause of Single Nucleotide Polymorphisms (SNPs). Kimura [115] described a mathematical model for substitutions but did not provide empirical data of the sort we require.

Barrick et al [109] determined E. coli substitution rates over a significant period of time – 40,000 generations over nearly 20 years. The E. coli were grown with glucose as a limiting nutrient. Samples were sequenced at generations 2,000, 5,000, 10,000, 15,000, 20,000, and 40,000. Substitutions and other mutations were observed. A mutator mutation occurred sometime after generation 20,000 that caused a significant increase in the substitution rate. We calculated total rates that include the increased, mutator rates. They observed a substitution rate of $6.84*10^{-9}$ per generation per site.

Denver et al [20] observed C. elegans mutating in a benign environment to ensure all but the most deleterious mutations accumulated over time. This assessment appears correct as they noted a decline in fitness over time. Their observations covered 396 generations. They observed substitutions and indels. They observed a substitution rate of $9.1*10^{-9}$ per generation per site.

Eyre-Walker and Keightley [118] determined mutation rates in hominids by assuming that the synonymous mutation rate in coding regions is equivalent to the mutation rate. They noted substitution differences between human and chimpanzee genomes, and human and gorilla genomes. We converted their mutations/year measures into mutations/generation measures using their generation times. They observed substitution rates of $3.33x10^{-8}$ for Homo sapiens, $2.44x10^{-8}$ for Pan troglodytes, and $2.46*10^{-8}$ for Gorilla gorilla, all per generation per site.

Nachman and Crowell [119] investigated the rate of mutation in Homo sapiens by comparing pseudogenes in Homo sapiens and Pan troglodytes. They estimated the substitution rate to be $2.5*10^{-8}$. Roach et al [120] also observed Homo the sapiens substitution rate by comparing the sequence of a small family. They estimated the rate to be $1.1*10^{-8}$.

Hudson et al [121] noted that bacterial substitution rates may vary with respect to adjacent nucleotides and chromosome location. In our model, we use mean mutation rates independent of both adjacent nucleotides and chromosome location as both types of information are difficult to discern reliably in long-term evolution models.

Lynch et al [122] did a long-term mutation-accumulation experiment with Saccharomyces cerevisiae. They observed a substitution rate of $3.3*10^{-10}$ per site per generation.

Haag-Liautard et al [123] did a long-term mutation-accumulation experiment with Drosophila melanogaster. They observed a substitution rate of $5.68*10^{-9}$ per site per generation.

6.1.4 Insertions and Deletions (Indels)

Barrick et al [109] also observed indels. They observed a total insertion rate of $2.27*10^{-10}$ per generation per site and a rate of $4.32*10^{-11}$ for insertions of length 1. They also found a total deletion rate of $1.40*10^{-10}$ and a rate of $5.40*10^{-11}$ for deletions of length 1.

Denver et al [20] also observed indels as well as substitutions with indels being a majority of the mutations. They observed a total insertion rate of $9.1*10^{-9}$ and a total deletion rate of $2.8*10^{-9}$ per generation per site.

Jeffreys et al [124] found high mutation rate, tandem-repeat minisatellites (repeated sections of DNA consisting of a short series of bases 10–60 bases in length). Mutation rates as high as $5*10^{-2}$ per generation were observed. Evidence of such high mutability is limited to Homo sapiens and does not appear to occur in coding regions of the genome. As a result, we do not consider these mutations further.

Weber and Wong [125] observed short, tandem repeats (STRs) in Homo sapiens. They found very high rates of mutation in repeat number of the STRS with a mean of $1.2*10^{-3}$ per site per generation. The majority of the changes (91%) were the gain or loss of a single repeat unit. This result does not appear to be broadly applicable and we do not consider it in our model.

Fan et al [126] observed indel patterns in mammalian genomes. They used multiple alignments of 19 mammalian species to determine two things. First, they tabulated the ratio of deletions to insertions and found a mean of 2.09. Second, they fit the observations to a power law $f_k = \text{coefficient} * k^{-power}$ where f_k is the probability of an insertion or deletion with gap length k and coefficient and power are derived from fitting the observations to the power law. They found mean values of $f_k = .52 * k^{-1.48}$ for deletion and mean values of $f_k = .47 * k^{-1.38}$ for insertion.

Graur and Wen-Hsiung [127] found the ratio of nonsynonymous substitution to synonymous substitution in mammalian coding sequences to average .74/3.51 where the units are substitutions per site per 10^9 years. They also found the substitution rate for untranslated regions to average 2.03 per site per 10^9 years. Since mammalian generation time varies over approximately two orders of magnitude, we cannot effectively convert this result to a substitution rate per generation per site.

Haag-Liautard et al [123] observed an insertion rate of $9.8*10^{-10}$ per site per generation in Drosophila melanogaster. The observed deletion rate was $1.59*10^{-9}$ per site per generation.

6.1.5 Inversions

Inversions reverse the order of a contiguous set of bases in a sequence. Since DNA is directional, inversions also have the effect of complementing the reversed bases; for example, inverting the sequence GCAAAC results in the sequence GTTTGC, with the reversed but uncomplemented sequence CAAACG, on the opposite DNA strand. Single inversions can have significant implications for genome function, e.g. they can cause hemophilia [128].

Barrick et al [109] found an inversion in their mutation rate study. They observed an inversion rate of $5.40*10^{-12}$ per generation per site. Kelchner and Wendel [129] observed inversions as short as 4 bases in length in chloroplast genomes. Contrastingly, long inversions (>1000 bases) are comparatively widely reported [130-132].

Feuk et al [133] observed a distribution of lengths in human genome inversions. In our model, we fit this distribution to a power law, analogous to the power law found by Fan et al [126] for indels.

6.1.6 Microsatellites and homopolymeric runs

Microsatellites [134] are short (typically 2-4 bases in length) contiguous, repeated sequences. Homopolymeric runs are a single base version of microsatellites. Repeat count variation of

microsatellites is a common mutation. We treat these simply as insertions or deletions as appropriate.

6.1.7 Normal and unequal crossover

Sexually reproducing species have a process of genetic recombination that occurs occasionally during meiosis. This crossover process results in new combinations of DNA sequences in each chromosome [135].

During normal crossover, DNA lost from a chromosome is exactly matched by the DNA it gains from the other chromosome in the pair and vice versa. While this offers an opportunity to gain advantageous genes, it offers an equal opportunity to loss them or to gain deleterious genes. From this perspective, normal crossover is neutral in effect and we do not consider it further.

Unequal crossover occurs when the exchange of DNA material is not symmetric: one chromosome is increased in size while the other in the pair is reduced by the same amount. We modeled unequal crossover as we would any other insertion or deletion as appropriate.

6.1.8 Chromosome fission and fusion

Chromosome fission occurs when a single chromosome divides to become two chromosomes. Chromosome fusion occurs when two distinct chromosomes fuse to become a single one. We found a specific reference only to fusion. Ijdo et al [136] found evidence that Homo sapiens chromosome 2 was produced by the fusion of two ancestral ape chromosomes. They concluded that head-to-head, telomere-telomere fusion had occurred. These mutations have little or no effect on the genes contained on the fissioned chromosome or the fused chromosomes and we do not consider them further.

6.1.9 Chromothripsis

Massive rearrangements, termed chromothripsis, in a one-off event have been observed for cancer cells [137]. In these events, tens to hundreds of sequence rearrangements occur. We do not further consider these rare events that likely typically lead to fatality.

6.1.10 Transposable elements

Transposable elements (transposons) [135] are mobile genetic elements that have only modest target site selectivity, i.e., they are typically able to insert themselves at any DNA site. They move at a rate of $1*10^{-5}$ in bacteria. We model them as a deletion from the source site and an insertion at the target site as appropriate.

6.1.11 Chromosome gain and loss

Chromosome gain and loss occurs when an entire chromosome is gained or lost in the genome of a cell. We found specific references only to chromosome gain.

Crow [116] examined mutation rates based on prevalence of mutation-induced diseases. He found that males have a higher mutation rate than females and that the male rate increases with age with an accelerating rate. He also found that .3% of liveborns are aneuplid (have an abnormal number of chromosomes); the most common aneuploidy is trisomy 21, an additional copy of chromosome 21 leading to Down's Syndrome.

Hook [138] looked at Homo sapiens chromosome abnormalities as a function of maternal age. In particular, they found trisomy of some chromosomes to vary from approximately .2% at maternal age of 15 to 15% at maternal age 49.

In our model, we treated chromosome gain as mechanism of producing additional DNA material.

6.1.12 Polyploidy

According to Ramsey and Schemske [139], Polyploidy, defined as the possession of three or more complete sets of chromosomes, is an important feature of chromosome evolution in many eukaryote taxa. Yeasts, insects, amphibians, reptiles, and fishes are known to contain polyploid forms, and recent evidence of extensive gene duplication suggests that the mammalian genome has a polyploid origin. They performed a broad literature review and concluded that the studies reviewed provide insights into the process of polyploid formation in natural populations, but caution that further research in natural populations is needed to test the findings.

Ramsey and Schemske distinguished between autoploidy in which all sets of chromosomes come from the same species and alloploidy in which the sets of chromosomes come from different species. They found an autoploidy rate of $1.05*10^{-4}$.

Wolfe and Shields [140] found evidence for whole genome duplication (tetraploidy) in yeast (Saccharomyces cerevisiae).

In our model, we treat polyploidy as a mechanism of producing additional DNA material.

6.1.13 Horizontal gene transfer

Horizontal gene transfer (HGT) is any process whereby an organism incorporates genetic material from a source other than a parent organism. This includes retroviral infection, bacterial transformation where a bacterial cell absorbs external DNA, and other processes.

Dagan and Martin [141] looked at HGT in the Prokaryote domain. Using current distributions of genes across genomes, a reference tree, and a model that assumes the distributions are due to gene loss only, Dagan concludes that more than 66% of prokaryotic genomes have been affected by HGT.

Zhaxybayeva et al [142] evaluated HGT between cyanobacteria and other prokaryotes. Out of a large number of datasets, 23% suggested HGT to or from cyanobacteria.

Choi and Kim [143] evaluated Horizontal Gene Transfer (HGT) across all of the biological domains. They used a method to identify HGT among all curated protein domain families in the Pfam database [144]. More than 50% of Archaea have evidence of Horizontal Gene Transfer (HGT); 30-50% of Bacteria did the same, but less than 10% of Eukaryota show such evidence. They suggest that HGT will have very little effect on phylogenetic tree construction if large sequence lengths (whole genomes or large numbers of common genes) are used.

Kunin et al [145] reconstructed the microbial phylogeny network using four distinct reconstruction methods. They found that the bulk of gene transfer occurred vertically as opposed to horizontally. They calculated a mean ratio of HGT events to vertical transfers of 5.35%.

Glansdorff et al [43]suggest that HGT may not have been so common during the time after LUCA. They suggest that loss of paralogs just prior to phylogenetic tree bifurcation may explain many events attributable to HGT.

Costas [146] states that approximately 8% of the human genome comes from retroviruses and that most are from human endogenous retroviruses (HERVs). He suggests that the retroviral

sequence may have impact in regulatory regions near coding sequence, but does not suggest that this is a significant effect.

HGT clearly had a significant role in the evolution of Prokaryotes, though it was perhaps not as prevalent as some suggest. However, it seems to have had a very minor role in Eukaryote evolution and perhaps even in the evolution of LUCA itself if LUCA was in fact a Eukaryote. HGT is therefore not further considered in our sequence evolution model.

6.2 Model mutations

There is significant evidence [147] that the substitution rate variation among sites is a gamma distribution [148]. However, in our model we use empirically derived averages as these represent the available results and more detailed data is unlikely to be obtainable given the lengthy time frames that we consider.

In our work, we modeled four major evolutionary steps: LUCAEukaryota from LUCA, LUCAMetazoa from LUCAEukaryota, LUCAMammalia from LUCAMetazoa, and HUMAN from LUCAMammalia. For each step, we derived the relevant mutation rates, using rates from appropriate species if available, taken from the empirical data described above. For example, in deriving rates to use in LUCAEukaryota from LUCA evolution, we used E. coli rates. Similarly, in deriving rates for HUMAN from LUCAMammalia evolution, we used primate rates.

For unchanged nucleotides, we used a probability of 1.0 as the probability of any mutation at all is so small. While the actual probability is clearly slightly less than 1, this is a very good approximation. For substitution rates, we used the data in Table 15. The first column indicates the major evolutionary step. The second column indicates the actual species relevant to the evolution step; these species are listed between their encompassing reference species. The third column indicates the empirical probability of the mutation per site per generation. The last column indicates the mean probability used in our model. In cases where there are no relevant species, the rate for the previous reference species was used. The rate mean is divided by 3 because only 1/3 of substitutions result in the correct base being produced.

Reference species evolution	Relevant species	Rate	Rate mean divided by 3
LUCA			
	E. coli	6.84E-09	
LUCAEukaryota from LUCA			**2.28E-09**
	S. cerevisiae	3.30E-10	
LUCAMetazoa from LUCAEukaryota			**1.10E-10**
	C. elegans	9.10E-09	
	D. melanogaster	5.68E-09	
LUCAMammalia from LUCAMetazoa			**2.46E-09**
	Primates	2.37E-08	
HUMAN from LUCAMammalia			**7.89E-09**

Table 15: Substitution rates

We modeled indels of length one or greater. Three approaches from three different sources are most relevant:

1. *Fixed probability:* Empirical evidence from Barrick et al [109] and Denver et al suggest that length 1 indels are most common and longer indels have a lower probability independent of length.
2. *Affine probability:* The study by Liu et al [72] concluded that affine gap costing for indels yielded the most accurate results. Gap costs are given by cost = $c_0 + c_1 * k$ where k is the gap (indel) length, c_0 is the gap-open cost and c_1 is the gap extend cost. In affine costs, $c_0 > 1$. Affine costs seem to model the cost of breaking and rejoining the DNA strand plus the increasing cost of rejoining it at an increasing distance from the original break.
3. *Power law probability:* The study by Fan et al [126] concludes that indel probability is most accurately modeled by a power law $P_k = c_0 * k^{-p}$ where P_k is the probability of an indel of length k, c_0 is the probability of and indel of length 1, and p is the power applied to the length with $p > 1$ indicating that the probability declines with length.

Using fixed probability in our model caused a unrealistically large number of large indels. In the worst cases, nearly the entire source genome was deleted and the entire target genome inserted. Using power law probability in our model, even with the comparatively small powers from Fan et al [126], only the very shortest indels were found as longer ones were too improbable to occur. It may be that the power law observed by Fan et al is primarily an artifact of the scoring that was used in their alignments.

For our model indel rates, we used the observed rates for the relevant species combined with the affine probability in Liu et al [72]. In particular, we used $c_0 + c_1 * k$ where c_0 is the relevant species rate for an indel of length 1, and c_1 was the ratio of length 1 to length 2 probabilities. These were done separately for insertions and deletions.

Reference species evolution	Relevant species	Rate	Rate mean
LUCA			
	E. coli	4.32E-11	
LUCAEukaryota from LUCA			**4.32E-11**
LUCAMetazoa from LUCAEukaryota			**4.32E-11**
	C. elegans	9.10E-09	
	D. melanogaster	9.80E-10	
LUCAMammalia from LUCAMetazoa			**5.04E-09**
HUMAN from LUCAMammalia			**5.04E-09**
		Ratio	
LUCA			
	E. coli	4.00	
LUCAEukaryota from LUCA			**4.00**
LUCAMetazoa from LUCAEukaryota			**4.00**
	C. elegans	9.00	
LUCAMammalia from LUCAMetazoa			**9.00**
HUMAN from LUCAMammalia			**9.00**

Table 16: Insertion rates

The insertion rates are shown in Table 16. It is formatted similarly to Table 15 with the addition of the ratio parameters. We note that insertions have an additional aspect that deletions do not. When bases are deleted through a mutation, the correct bases are deleted by definition. However, when there is an insertion, it is unlikely that the correct bases, i.e. the bases in that subsequence of the target, are inserted. We used a more accurate model where we assumed that random bases are inserted and then substitutions take place to correct the random bases. As a given base has a ¾ chance of being incorrect, we added a probability of .75 * k * the relevant substitution rate to the overall probability of each insertion, where k is the length of the insertion.

The deletion rates are shown in Table 17. They were determined in a way entirely analogous to the insertion rates.

Reference species evolution	Relevant species	Rate	Rate mean
LUCA			
	E. coli	5.40E-11	
LUCAEukaryota from LUCA			**5.40E-11**
LUCAMetazoa from LUCAEukaryota			**5.40E-11**
	C. elegans	2.80E-09	
	D. melanogaster	1.59E-09	
LUCAMammalia from LUCAMetazoa			**2.20E-09**
HUMAN from LUCAMammalia			**2.20E-09**
		Ratio	
LUCA			
	E. coli	5.00	
LUCAEukaryota from LUCA			**5.00**
LUCAMetazoa from LUCAEukaryota			**5.00**
	C. elegans	2.00	
LUCAMammalia from LUCAMetazoa			**2.00**
HUMAN from LUCAMammalia			**2.00**

Table 17: Deletion rates

We also modeled inversions, which has been noted to be a significant mutation mechanism [128-130 132 133]. While the available data for the previous mutations may be accurately described as less than abundant, the data for inversions is truly scant. We use the rate from a single E. coli observation for all of our reference species. From the data in Feuk et al [133], we derived a power law for the distribution of inversions as a function of length, much as was done by Fan et al [126].

The very short inversions (4 bases) seem to occur only in the presence of unique sequences that allow hairpin secondary structures in the DNA strand. In addition, the large inversions (> 1000 bases) will tend to simply duplicate the gene on the opposite strand running in the opposite direction and so have little impact on gene function. Since we are modeling mutations in coding DNA, we must consider relatively commonly occurring mutations with lengths between these two extremes. We somewhat arbitrarily chose 12 (3 * 4, the minimum observed) and model inversions that are this length or longer.

As we did with indels, we normalized our inversion distribution so that our minimum length inversions have the observed probability and longer inversions have lower probability.

Reference species evolution	Relevant species	Rate	Rate mean
LUCA			
	E. coli	2.88E-11	
LUCAEukaryota from LUCA			**2.88E-11**
LUCAMetazoa from LUCAEukaryota			**2.88E-11**
LUCAMammalia from LUCAMetazoa			**2.88E-11**
HUMAN from LUCAMammalia			**2.88E-11**
Length			
minimum length	**12**		
maximum length	**unlimited**		
Power law parameters			
coefficient	**2.1**		
power	**0.295**		

Table 18: Inversion rates

The inversion rates that we used in our model are shown in Table 18.

The remaining mutation types are treated as described in their respective sections. In particular, many (e.g. polyploidy) are treated simply as mechanisms that add random DNA to the genome. Their rate is sufficiently high that they do not affect overall evolution duration.

7 Sequence evolution model

7.1 Related work

Much work over the last several decades has gone into sequence evolution models. Graur and Li [149] give a comprehensive overview of genome evolution, the principles of which are common to any evolution model. Two types of evolution models are relevant to this research. One type is a broad model, which attempts to cover long expanses of evolutionary time but is not specific to sequence evolution per se. Wright [150] did the best-known early work in this area; more recent thinking has come from Fisher [151]. These models are composed of biological trait frequency distributions, including compound distributions and correlations, and path analysis attempting to describe causation.

Another type is probabilistic models, of which Beerenwinkel et al [152] and Komarova and Wodarz [153] are exemplary. These models use Bayesian networks of specific mutations. Bayesian networks [154] are probabilistic models that represent a set of random variables and their dependencies in a directed graph.

Coalescent theory [155 156] creates gene phylogenetic trees, starting with a sample distribution of the sequences of the leaf nodes. It can be used, for example, to determine when most recent common ancestor (the root of the tree) of all of the leaves existed. It is at this root node that all of the gene sequences at the leaves have coalesced. This is primarily a population evolution model, though it uses specifics of sequences to make its predictions.

Ma et al [132] suggests an infinite sites model of sequence evolution. This model provides that genomes are sufficient in length that a reasonable approximation of long-term evolution uses each site in a genome at most once. The results of this model were our first indication that inversions were a significant mutation mechanism.

We described, in chapter 6, Kimura's [16] neutral theory of evolution which states that the rate of evolution in terms of nucleotide substitutions is sufficiently high that most mutations must be nearly neutral with respect to natural selection. Once again, this is primarily a population evolution model that makes little use of sequence specifics. In our sequence evolution model, we made little in the way of assumptions with respect to the neutral theory.

Felsenstein and Churchill [157] describe the use of hidden Markov models to allow mutation rates to vary as a function of location in a genome. These results are used to estimate phylogenetic tree branch lengths. In our model, we take an approach of mutations having rates dependent on reference species but independent of location in a sequence.

Considerable research has been done on sequence alignment algorithms [58 67 69-72 74 76 77 158-168] and we applied that work throughout our sequence evolution model.

7.2 Model description

With the exception of LUCA, each of our reference species is a successor to a previous reference species; we refer to these as successor reference species. As described in chapter 3, we constructed a hidden Markov model for each successor reference species, containing the most probable paths over which it might have evolved from its predecessor. In particular, we constructed a Markov model for each gene in our reference species' genomes. For those genes with a predecessor homolog, we constructed the Markov model between the gene and its homolog with the sequence lengths left unchanged. For genes with no predecessor homolog, we used a sample set of nonhomologous genes from the immediate predecessor reference species and a sample set of random sequences generated through the use of a random number generator.

7.2.1 Basic global alignment algorithm

Note that a path through the Markov model for a given gene corresponds to a path through the edit graph (Figure 5) for the gene, the target, and the gene from which it is being evolved, the source. Thus, to construct the Markov model for two genes, we align their sequences using edit graphs. We made extensive use of prior work in sequence alignment, extending it in some areas.

Our model has some advantages in comparison to others. We can identify the source and the target and are therefore able to distinguish insertions from deletions, a distinction that cannot be made in general. This is important as we find that typically insertions and deletions have different probabilities. We know which mutation mechanisms to apply and their rates; therefore can produce a biologically accurate alignment with scoring corresponding to actual mutation probabilities.

The basis for our sequence evolution model is the two sequence alignment described by Waterman [67] as a global distance alignment. It is a global alignment due to its accounting for all of the bases in both the source and target sequences; this is simply an alignment that begins at the upper left corner of the edit graph and ends at the bottom right.

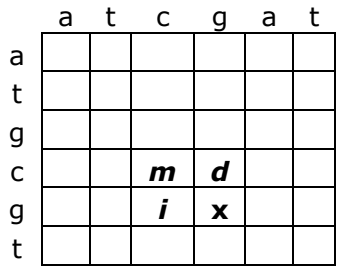

Figure 15: Sequence alignment edit graph

In Figure 15, the source sequence is written down the left side and the target sequence is written across the top. Once any path arrives at a cell, there are three ways to traverse the cell. For example, to traverse cell x:

1. The base to its left, a source base, may be matched to the base above it, a target base. In the case of cell x, this matches the "g" to its left in the source with the "g" above it in the target. Since these bases are identical in this case, this is an identity match; if the bases were different, this would indicate a substitution mutation. This would be a diagonal traversal through the cell. The total score for this cell traversal is the score from

predecessor cell *m* + the score for the match (identity or substitution, the former in this case).

2. The base to its left, a source base, may be deleted. It is not matched to the corresponding base in the target. This would be a vertical traversal through the cell. The total score for this traversal is the score from predecessor cell *d* + the score for an incremental deletion. As we are using affine scores, the score for a deletion of length k ≠ k * the score for a deletion of length 1 and the incremental score must be calculated.
3. The base above it, a target base, may be inserted. It is not matched to the corresponding base in the source. This would be a horizontal traversal through the cell. The total score for this traversal is the score from predecessor cell *i* + the score for an incremental insertion. Again, the affine score for an insertion of length k ≠ k * the score for a insertion of length 1 and the incremental score must be calculated.

To find the optimal path through the edit graph, we start at the upper left cell and proceed left to right row by row, computing the three scores for each cell by computing the three predecessor cells (*m, d, i*) before computing the score at the given cell. At each cell, the best of the three scores becomes the score for that cell, which indicates the optimal path ending at that cell. The score for the global alignment is the score for the lower right cell and the optimal path can be traced backwards from the lower right cell, always choosing the predecessor cell with the best score, until the cell at the upper left is reached.

To create a biologically accurate model, our scores are based on actual mutation probabilities from chapter 6. As these probabilities are very small, our actual scores tended to be very small, often below the limits of conventional computer arithmetic. To avoid this drawback, we used log values in our calculations. As a result, our scores were all negative with the exception of the case of two identical sequences, resulting in a score of 0. This approach yields a single, optimal path with a single, optimal score. For our purposes, it has several drawbacks.

7.2.2 Multiple paths

Evolution will tend to take very probable paths but is not likely to always take the optimal path. Thus, we would like to find multiple, high probability paths in our alignments. Again, there is prior work in algorithm development for finding multiple paths [169-171]. Our algorithm sought the k-best paths through each sequence alignment. At each cell, it combined the traversal score (match, deletion, or insertion) with the k-best scores of the appropriate predecessor cell, yielding a total of 3*k best scores for the cell. The k-best out of all of these scores were then chosen to be the scores for the cell.

There are clearly more paths through longer sequences and we chose our k-best to reflect that fact. The total number of paths through two sequences of length n is approx. $(1+\sqrt{2})^{2n+1} * \sqrt{n}$; two sequences of length 1000 have approximately 10^{767} paths through them [67]. Obviously, we could only examine a tiny fraction of the possible paths. However, the highest scoring paths were the most probable and these were the paths we sought. We chose k-best to be a slowly increasing function of the two sequence lengths. We computed k-best as

$$k = ((length_{source} * length_{target})^{\wedge}.25)/2$$

This is one half of the square root of the geometric mean of the two sequence lengths. We set k-best to have a minimum of 4, applicable only to very short sequences we used for testing purposes.

7.2.3 Diagonal constraint

As we performed a comparatively large number of alignments in total, we aligned sequences of nontrivial length (up to approx. 9000 base pairs), and we computed the k-best paths for each alignment, the total required computation time was a concern as the algorithmic complexity is order of $length_{source}$ * $length_{target}$ * k-best. Computer memory space was a similar concern as a straightforward implementation would also require order of $length_{source}$ * $length_{target}$ * k-best space. An insight of ours, also noted by others such as Chao et al [159], was that the most probable alignments are constrained to lie near the major diagonal of the edit graph.

Consider the edit graph with the shaded major diagonal in Figure 16. Any path that includes cells above and to the right of the shaded diagonal will have had many insertions to get to the right of the diagonal and would then require many deletions to get back down to the lower right corner. Such a path has too many insertions and deletions to be very probable. Similarly, any path that includes cells below and to the left of the diagonal will have had many deletions to get below the diagonal and would then require many insertions to get back right to the lower right corner. Again, such a path has too many insertions and deletions to be very probable.

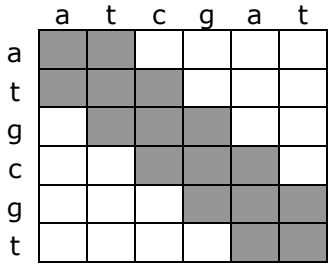

Figure 16: Diagonal in edit graph

Hence, we constrained our search for optimal paths to those that occupied only cells within a distance $d_{diagonal}$ from the major diagonal. To calculate an appropriate diagonal distance, we first computed the diagonal length, as the allowed distance from the diagonal must be greater for longer sequences. We then took the square root of the diagonal length so that the distance would not grow too quickly as sequence length increased. Finally, we took twice that result to be conservative compared to a reported, successful result. Thus, the diagonal distance we used was

$$length_{diagonal} = \sqrt{(length_{source}\text{^}2 + length_{target}\text{^}2)}$$
$$d_{diagonal} = 2 * \sqrt{length_{diagonal}}$$

We considered only paths whose nodes had a perpendicular distance from the diagonal no larger than $d_{diagonal}$. For two sequences of length 1000, $d_{diagonal}$ = 75; thus the diagonal was effectively 150 bases wide. This is significantly larger than the successful result reported by Chao et al [159].

7.2.4 Inversions

As described in chapter 6, inversions are a significant mutation mechanism and we included them in our model. There is a smaller, but still significant, body of work in algorithms to identify inversions [160 164 166 171].

The problem of identifying inversions is related to, distinct from, the global alignment problem. In global alignment, we search for sequences that align each of the sequences from end to end, accounting for each base in each sequence. Every path must start at the upper left of the edit graph and end at the lower right, despite potentially having a very low score. Conversely, inversions may start at any point and end at any point in the edit graph, but we seek only good inversions with comparatively high scores and therefore high probabilities of occurrence – these are known as local alignments. In both cases, sequences are aligned with a combination of matches and indels.

To find inversions, we find good local alignments of a copy of the source sequence, which has been reversed and complemented, and the original target sequence. To find inversions for the sequences in Figure 17a, we create the edit graph shown in Figure 17b. The sequence written down the left side of the graph is the reversed and complemented source sequence. It has been capitalized to indicate that it is an inverted sequence.

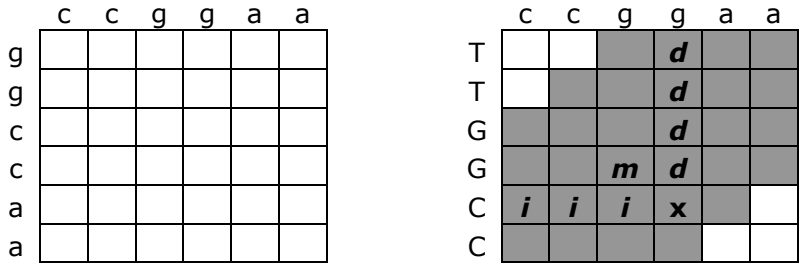

Figure 17a,b: Inversion edit graph

Our goal was to find the best inversion from each cell in the original graph, so we attempted to find a good inversion that ends at each cell in the inversion edit graph; many had no good inversion. Our fundamental local alignment algorithm was taken from Waterman [67].

Once any inversion path arrives at a cell, there are three ways to traverse the cell. For example, to traverse cell x in Figure 17b:

1. The base to its left, a source base, may be matched to the base above it, a target base. In the case of cell x, this matches the "g" to its left in the source with the "g" above it in the target. Since these bases are identical in this case, this is an identity match; if the bases were different, this would indicate a substitution mutation. This would be a diagonal traversal through the cell. The total score for this cell traversal is the score from predecessor cell *m* + the score for the match (identity or substitution, the former in this case). Only inversions that begin with an identity match are optimal and so only those are ultimately considered in our search.
2. Since we seek inversions anywhere earlier than cell x in the edit graph, any number of source bases may be deleted. These are not matched to the corresponding bases in the target. This would be a vertical traversal through the cell. We chose the score from the highest scoring of the cells labeled *d* and refer to that cell as d_{highest} The total score for

this traversal is the score from predecessor cell $d_{highest}$ + the score for a deletion of corresponding length.

3. Similar to deletions, any number of target bases may be inserted. These are not matched to the corresponding bases in the source. This would be a horizontal traversal through the cell. We chose the score from the highest scoring of the cells labeled *i* and refer to that cell as $i_{highest}$ The total score for this traversal is the score from predecessor cell $i_{highest}$ + the score for an insertion of corresponding length.

To find the best inversions in the edit graph, we start at the upper left cell and proceed left to right row by row, computing the three scores for each cell by computing the three predecessor cells (m, $d_{highest}$, $i_{highest}$) before computing the score at the given cell. At each cell, the best of the three scores becomes the score for that cell, which indicates the best inversion ending at that cell.

Waterman [67] describes an efficient algorithm for finding inversions. Part of its efficiency is due to its removing from the edit graph all but the longest, best inversions. Our goal was distinct in that we wished to find the best inversion for each cell in the edit graph. As a result, we used the less efficient local alignment algorithm to identify a complete set of inversions. Our version of the algorithm was order $length_{source}*length_{target}*(length_{source}+length_{target})$ in complexity. As we did with our global alignment algorithm, we constrained our search to high probability inversions that lie in the diagonal. However, since the source sequence is reversed, we constrained our search to the minor diagonal as shown in Figure 17b. Since the diagonal distance is a constraint on the length of an inversion, we used twice the diagonal distance used in our global alignment algorithm.

During the inversion identification process, scoring is done in a unique way as well. During global alignment, we used actual mutation probabilities for scoring. As the logs of these probabilities are always negative, the score was an arbitrarily large negative number. While identifying good inversions, we must end an inversion whose score has fallen too low rather than always extending it all the way back to the upper left corner cell. We used the scoring from Waterman [67]:

1. Match identities scored 10
2. Match substitutions scored -11.
3. Indels of length k scored -15 – 5*k

When the score for an inversion fell to 0 or less, the inversion was ended. We also normalized inversion scores based on total length; if this was not done, long inversions dominated. When the inversions were subsequently used in the context of a global alignment, our biologically accurate scoring was used for the mutations contained within the inversion as well as the inversion mutation itself.

The inversion coordinate space is distinct from the global alignment coordinate space because the source sequence in reversed in the former case. The inversion coordinate space must be transformed into the global alignment coordinate space in order to determine which global alignment cells begin and end inversions. The column coordinates are identical as the target sequence is unchanged for inversions.

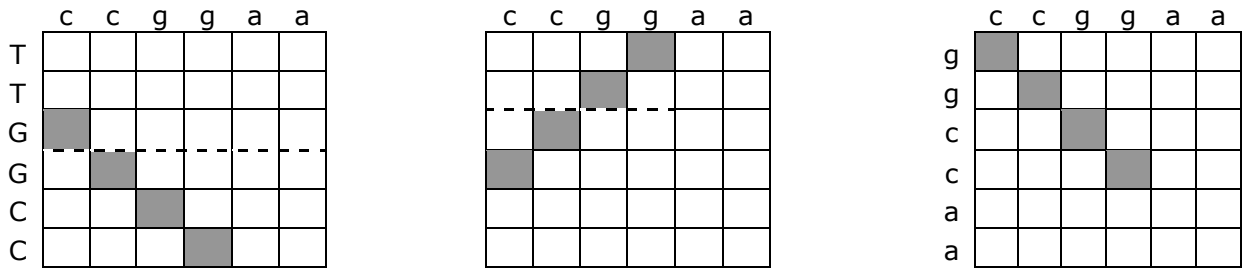

Figure 18a,b,c: Inversion coordinate transformation

To transform the row coordinates, we first reflected the coordinates about the horizontal center of the edit graph as shown in the transformation from Figure 18a to b. This brings the row coordinates into the original location of the inversion before the source sequence was reversed. However, the start and end of the inversion are transposed; a reflection about the horizontal center of the inversion, as shown in the transformation from Figure 18b to c, puts the inversion into the correct orientation.

To integrate inversions into our basic global alignment algorithm, we add a 4th step to those described in 7.2.1:

4. The cell may be traversed by an inversion that ends at this cell if a good such inversion exists. The total score for the inversion is the score of an inversion mutation of length k + the score of the matches and indels contained in the inversion itself. This score is added to the k-best scores from the cell at the beginning of the inversion, resulting in a 4th set of k-best scores for the current cell.

This 4th set of k-best scores is combined with the other three sets and the k-best scores from this combination make up the k-best scores for the cell.

7.2.5 Global alignments performed

As described above, for those genes with a predecessor homolog, we constructed the Markov model between the gene and its homolog with the sequence lengths left unchanged. For genes with no predecessor homolog, we used a sample set of nonhomologous genes from the immediate predecessor reference species and a sample set of random sequences generated through the use of a random number generator.

	LUCAMammalia				
	LUCAMammalia unique	LUCAMetazoa homolog	LUCAEukaryota homolog	LUCA homolog	Random sequences
HUMAN	1	1	1	1	3

	LUCAMetazoa			
	LUCAMetazoa unique	LUCAEukaryota homolog	LUCA homolog	Random sequences
LUCAMammalia	1	1	1	3

	LUCAEukaryota		
	LUCAEukaryota unique	LUCA homolog	Random sequences
LUCAMetazoa	1	1	3

	LUCA	
	LUCA unique	Random sequences
LUCAEukaryota	1	3

Table 19: Nonhomologous alignments

The specific alignments performed are indicated in Table 19. The rows are labeled with the reference species. Each nonhomologous gene from a reference species is aligned with the type and number of genes given in its rows. For example, a nonhomologous LUCAMetazoa gene is aligned with 2 genes from its predecessor, LUCAEukaryota: one gene unique to LUCAEukaryota, one LUCAEukaryota gene that is homologous to a LUCA gene, and three random sequences. We took one sample from each such homology group in order to have a broad range of sequences.

The nonhomologous source genes were cropped (with the center of the sequence retained) or extended (with the original sequence centered between two randomly generated sequences of appropriate length) to match the length of the target gene as terminal insertions or deletions in such cases do not impact our model. Random source sequences were generated identical in length to the target gene.

Reference species	Nonhomologous genes	Nonhomologous alignments	Homologous genes/alignments	Total genes	Total alignments
LUCA	*33*			*33*	
LUCAEukaryota	43	172	33	76	205
LUCAMetazoa	43	215	76	119	291
LUCAMammalia	44	264	119	163	383
HUMAN	39	273	163	202	436

Table 20: Gene and alignment counts

In Table 20, we show total gene and alignment counts. As LUCA is our initial reference species, it has only nonhomologous genes and no alignments are performed on them. For each alignment, the data from multiple paths are calculated and aggregated as described in 7.2.2.

7.2.6 Distance measure

In order to make comparisons between different alignment paths, we defined a positive distance measure to quantify the distance between any given source and target sequence through a specific alignment path. For distance, we used the negative of the score for each path computed during the alignment process.

In our calculations, we aggregated these distances. We aggregated individual path distances up to an alignment pair distance, alignment pair distances up to a gene distance, and gene distances up to a reference species distance. In most calculations to aggregate distances, we calculated a mean; since the scores were logs, these were geometric means rather than arithmetic means. The one potential exception was aggregating path distances up to a gene distance. In this case, each path increases the probability of the overall alignment, rather than simply being a part of the mean.

To aggregate path distances, we note that high scoring paths in a given alignment will typically have the majority of their mutations in common. These mutations in common do not represent independent probabilities and do not increase the overall probability. We must distinguish the common mutations from those mutations of each path that are unique and so represent independent probabilities.

```
1.  cgatcacaatcggaatgacaaggaacaaa
    |   || | ||||||  |    |||||||
    catgcaaagtcggaagaattcagaacaaa


2.  cgat-cacaatcggaatgacaaggaacaaa
    | || || | ||||||  |    |||||||
    c-atgcaaagtcggaagaattcagaacaaa
```

Table 21: Alignment path examples

Consider the two alignment paths through the same sequence pair in Table 21. Path 1 has 11 substitutions and no indels; path 2 has 8 substitutions, 1 insertion of length 1, and 1 deletion of length 1. They share most of the same mutations, i.e. the 8 substitutions.

Our results indicated that our most probable paths shared nearly all of the same mutations. As a result, we concluded that the multiple paths do not represent significant, probabilistic independence. Hence, we ultimately calculated geometric means to aggregate path distances up to a gene distance.

7.3 Model results

Our sequence evolution model generated a plethora of results. To provide a sense of the magnitude, alignment path counts are shown in Table 22.

Reference species	Total genes	Total alignments	Total paths
LUCA	*33*		
LUCAEukaryota	76	205	3992
LUCAMetazoa	119	291	5152
LUCAMammalia	163	383	8272
HUMAN	202	436	5991
Total with inversions			23407
Total without inversions			23407
Grand total			46814

Table 22: Alignment paths

Since we produced nearly 50,000 paths, we summarize our results. We note that though the HUMAN genes we modeled were more numerous than the LUCAMammalia genes, LUCAMammalia had more alignment paths. This is because LUCAMammalia had more large, nonhomologous genes than HUMAN did and more paths are generated for large genes than are generated for smaller ones. We present results from the model that included inversions. For a comparison with the model without inversions, see 7.4.1.

7.3.1 Homologous gene

One of our homologous LUCAMammalia genes is 38089LUCAMammalia, which has a homolog in LUCA. This gene is an ATP synthase [172], which is an enzyme that synthesizes ATP from ADP. It has a homolog in each reference species. The LUCAMammalia gene was aligned with its predecessor homolog, 38089LUCAMetazoa.

```
       Path:  38089LUCAMetazoa to 38089LUCAMammalia
       paths: 1      source bases: 1859   target bases: 1857
       distance:     4.0855E3      per Kb:       2.2000E3
       subs dist:    1.5987E4
       subs:         454     per Kb:       244
       inserts:      9       per Kb:       4
       deletes:      11      per Kb:       5
       inverts:      0       per Kb:       0
       invCount:     0       per Kb:       0
       insertions:   len     count
                     1       9
       deletions:    len     count
                     1       8
                     3       1


.....................................................................................
tcccctgaaggacatcaacgaactgacccaaagcatctacattcccagtggtgttaacactcccgctttgagccg-catg
 ||  ||  ||| |||| |   | ||||||||||||||||| ||||| || || ||     | ||| | ||| |  ||
accttttgtcggatatcagcagtcagacccaaagcatctacatccccagaggagtcaatgtgtctgctcttagcagagat-
```

Figure 19: Homologous alignment example

In Figure 19, we present the model results from one path in this alignment and a portion of the aligned sequences. The source and target sequence lengths are nearly identical, as is often but not always the case with homologs. We show our raw distance measure as well as the distance normalized to distance per kilobase so that alignments of difference length sequences can be compared. The subs dist is the distance that would apply if the only mutations were n

substitutions where n is the length of the target sequence. This approximates a worst-case distance; it equals the worst-case distance when the source and target are the same length. We expected our alignments to produce better distances than this and, in fact, they did except when source and target lengths are significantly mismatched.

There were 454 substitutions in this path. For random sequences of the same length, the expected number of substitutions would be .75 * length = 1393. As 454 is much less than 1393, it is clear that these two sequences are much more similar than two random ones, as expected for homologs. There were relatively few insertions or deletions, and their net result, 2, was the difference in the sequence lengths as expected. No valuable inversions were identified, so both inverts, the number of inverted bases, and inversions, the number of inversion subsequences, are 0. Information on mutation spectra, the quantity and size of the individual insertions, deletions, and inversions, is included.

We also show a portion of the sequence alignment itself. It is depicted in the same way as other alignments with the addition of an initial line indicating the location of any inversions; as there were none in this case, it consists of all '.'s. This sequence portion includes one base insertion and one base deletion, both near the right end of the sequences.

7.3.2 Nonhomologous gene

One of our nonhomologous LUCAMammalia genes is 28898LUCAMammalia, which has no homolog in earlier reference species. 28898LUCAMammalia is a bone morphogenetic protein receptor type-2 precursor [173], which plays a role in bone regeneration as well as embryo development. 28898LUCAMammalia was aligned with 3 nonhomologous genes from LUCAMetazoa as well as with 3 random sequences.

```
      Path:  215350LUCAMetazoa to 28898LUCAMammalia
      paths:       1     source bases:     3123  target bases:     3123
      distance:    1.4765E4    per Kb:     4.7278E3
      subs dist:   2.6886E4
      subs:        1239  per Kb:     396
      inserts:     212   per Kb:     67
      deletes:     212   per Kb:     67
      inverts:     978   per Kb:     313
      invCount:    44    per Kb:     14
      insertions: len   count
                   1     163
                   2     16
                   3     3
                   4     2
      deletions:   len   count
                   1     152
                   2     19
                   3     4
                   5     2
      inversions: len   count
                   12    8
                   13    7
                   14    5
                   15    4
                   16    1
                   19    1
                   21    1
                   22    3
                   24    2
                   26    1
                   28    1
                   29    1
                   30    1
                   33    1
                   34    1
                   37    1
                   42    1
                   43    1
                   45    1
                   57    1
                   87    1

..ctg--gtaacggctt..............................................................
-tGAC--CATTGCCGAAgggagaaggggcctgagtccgtgtcgggaatggcccttcaaattgtgcatccaatgcaagtaa
 ||||  | | || | || | |    |||  | | || || |    |||| || ||    ||  ||  || |||   |
atgacttcctcgctgcagcg-gccctggcgggtg-ccctggctaccatggaccatcctgctggtca-gcactgcggctgc
```

Figure 20: Nonhomologous alignment example

In Figure 20, we show the results of aligning 28898LUCAMammalia with 215350LUCAMetazoa, a nonhomologous gene arbitrarily chosen from among those in LUCAMetazoa; we also show a portion of the sequence alignment. The length of 215350LUCAMetazoa was augmented with random sequence to match the length of 28898LUCAMammalia.

As with the homologous alignment, this alignment distance is less than the substitution distance, $1.4765*10^4$ versus $2.6886*10^4$. Compared to the homologous alignment, the distance in this nonhomologous alignment is higher per kilobase, $4.727*10^3$ versus $2.200*10^3$. Substitutions, insertions, and deletions per kilobase were all higher in this nonhomologous case, 244, 4, and 5 versus 396, 67, and 67 respectively.

Arguably the greatest contrast with the homologous alignment was the inversion result: none in the homologous case versus 14 inversions per kilobase and 313 inverted bases per kilobase in the nonhomologous case. Note that these inversions improve the overall distance of a path even while they cause an increase in distance due to their presence. In the portion of the nonhomologous alignment shown, there is a single, length 15 inversion. The lowercase bases in the top line indicate the bases in the original source with their order reversed; the uppercase bases below are the complemented bases aligned with the target. Note that there were mutations found within the inversion as well, 4 substitutions and an insertion of length 2.

Insertion, deletion, and inversion spectra are shown. While indels become less probable rapidly with length, insertion probability falls much more slowly with length, despite being governed by a power law. As a result, few long indels are evident while longer inversions are more common.

```
       Path:  RANDOMA3123 to 28898LUCAMammalia
       paths: 1      source bases: 3123   target bases: 3123
       distance:     1.4930E4     per Kb:       4.7808E3
       subs dist:    2.6886E4
       subs:         1282   per Kb:       410
       inserts:      204    per Kb:       65
       deletes:      204    per Kb:       65
       inverts:      831    per Kb:       266
       invCount:     39     per Kb:       12
       insertions:   len    count
                     1      172
                     2      13
                     3      2
       deletions:    len    count
                     1      174
                     2      11
                     3      1
                     5      1
       inversions:   len    count
                     12     7
                     13     2
                     14     2
                     15     2
                     16     1
                     17     4
                     18     3
                     19     4
                     20     2
                     24     1
                     25     2
                     26     2
                     31     1
                     32     2
                     36     1
                     47     1
                     49     1
                     56     1
.............................................tacttggttgtatcgtaata.................
agtttttacccacagttac-gatct-ac-aaaactctgg-ttgtATGAACCAACATAGCATTATttagttatttttaat-
|    || | | | |   | |  ||  |     | |||| |   ||| |||| | | || |  |  ||  | |     |
atgacttcctcgctgcagcggccctggcgggtgccctggctaccatggaccatcct-gc-tggt-cag-cactgcggctg
```

Figure 21: Random alignment example

In Figure 21, we show the results of 28898LUCAMammalia aligned with a random sequence of the same length. Of note is that the distance in the random alignment is greater than the distance in nonhomologous gene alignment, $1.4930*10^4$ versus $1.4765*10^4$. This result is generalized and quantitatively analyzed in 7.4.2.

7.3.3 Homologous and nonhomologous results

In Figure 22, we show the model distances for homologous and nonhomologous alignments for the reference species; all values are per kilobase. For all reference species, distances were higher in the nonhomologous alignments compared to the homologous alignments.

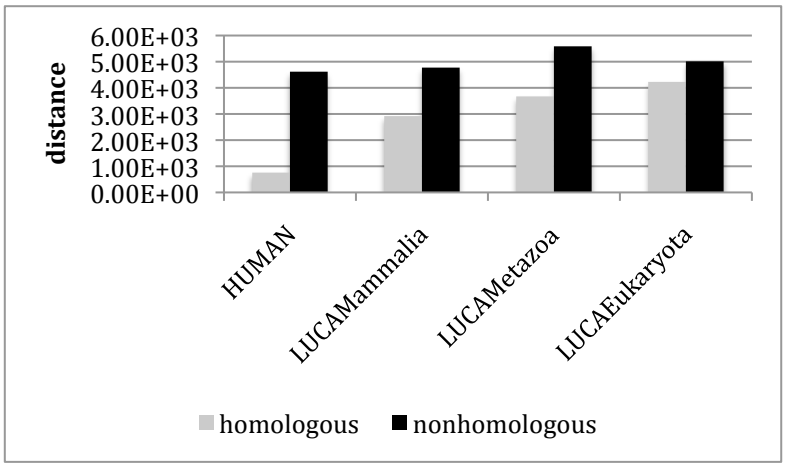

Figure 22: Distance comparison

In Figure 23, we compare a) substitutions, b) inserted bases, c) deleted bases, and d) inverted bases; all values are per kilobase. Note that the inserted bases count always equals deleted bases count in the nonhomologous results. This is due to the source length being adjusted so that it is equal to the target length in such alignments in our model.

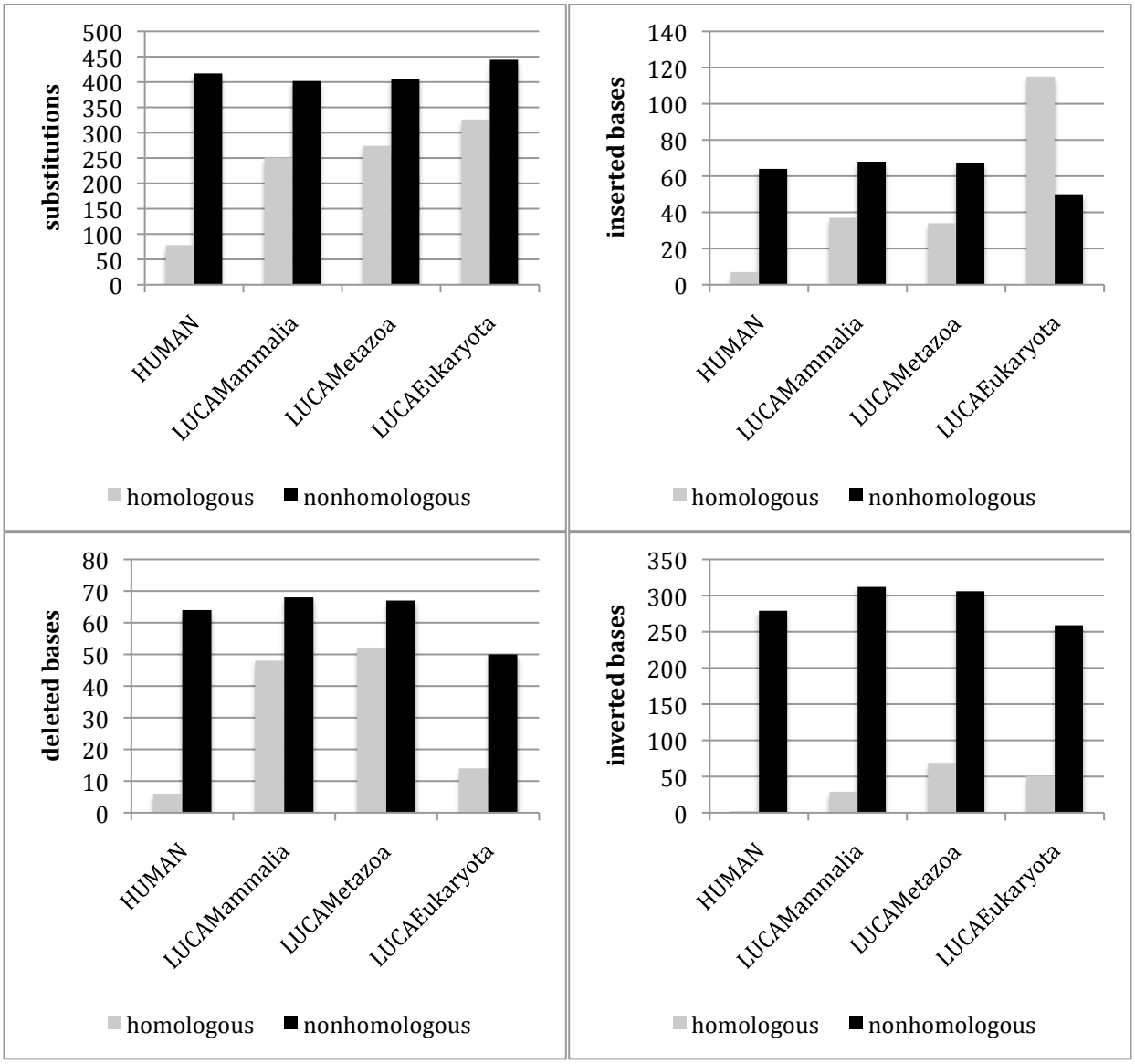

Figure 23a,b,c,d: Nonhomologous mutation comparison

For each reference species, mutation counts are higher in the nonhomologous alignments compared to the homologous alignments with one exception. For LUCAEukaryota evolving from LUCA, the homologous insertion count was higher than the nonhomologous insertion count, 115, versus 50. While nonhomologous alignment pairs were always matched in length, homologous pairs were not. We suggest that this result may be due to the incorporation of introns into the LUCAEukaryotic genome in evolving from the LUCA genome. These introns increased the length of homologous LUCAEukaryotic genes. While our sequences include only exons, the inclusion of introns may have been accompanied by increases in exon length. This provides modest, further evidence that LUCA was a prokaryote, as prokaryotes lack introns.

While we expected distances to be shorter for homologous alignments, the consistency of this result for so many measures was less expected. Most prominent is the significant difference

in inversion results. From this result, it is clear that inversions are a significant factor in the evolution of new function.

7.3.4 Reference species results

In Figure 24, we show a summary of the mutations for each reference species. The vertical axis is bases with the given mutation per kilobase of sequence. The inverts values are inverted bases rather than inversion counts as this make these values comparable to the others. This summary includes both homologous and nonhomologous results.

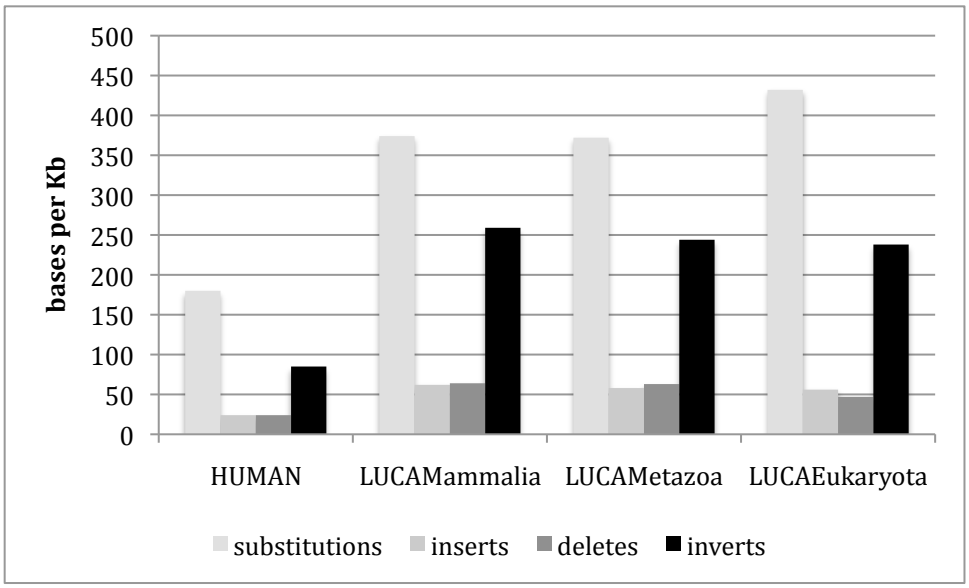

Figure 24: Reference species mutation comparison

Substitutions are clearly the dominant mutation mechanism as expected given their comparatively high probability. The higher value for substitutions in LUCAEukaryota is likely due to its comparatively long distance from its predecessor reference species, LUCA. A comparatively large number of bases are inverted in each reference species. However, since the probability of an inversion falls slowly as a function of length, the inverted bases may almost as likely be contained in a small number of large inversions as they are to be contained in a large number of small inversions. The mutation spectra can determine which of these two explanations is accurate.

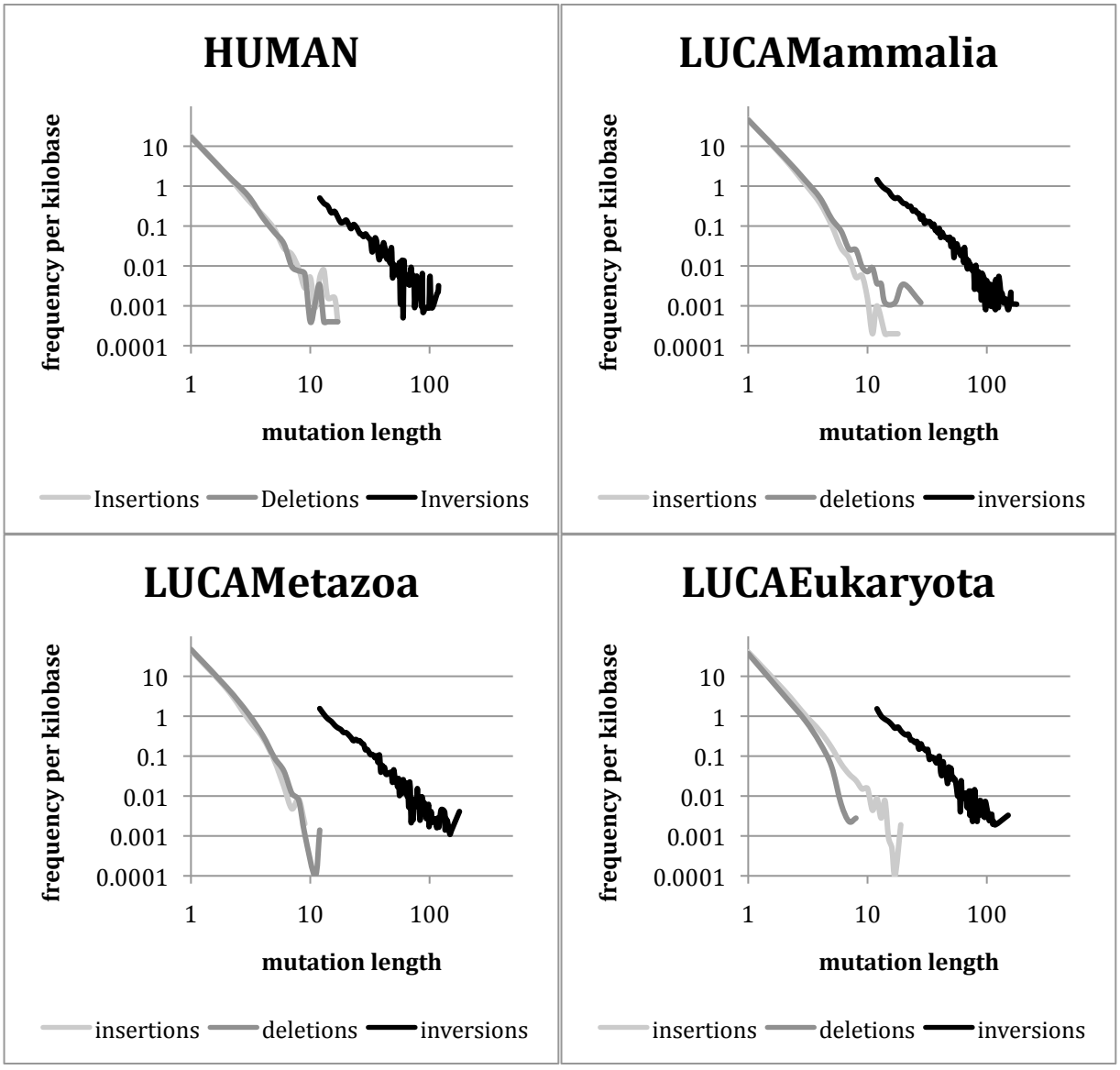

Figure 25: Mutation spectra

In Figure 25, we show mutation spectra for each reference species. These show the frequency of each length of each mutation type per kilobase of sequence; the graphs have a log-log scale. As expected insertion and deletion frequencies fall rapidly with increasing length; this is much less the case with inversions. Using the inversion spectra results and normalizing by the number of bases in each inversion length, we conclude that most inverted bases are in short length inversions.

7.4 Model analysis

From our sequence evolution model results, we were able to investigate several aspects of sequence evolution, described below.

7.4.1 Inversion mutations

Inversions as a significant mutation mechanism was suggested by Ma et al [132] who found them in large-scale studies of whole chromosomes. Smaller inversions, referred to as microinversions, were found significant in a study by Chaisson et al [131]. Our inversion identification algorithm, described in 7.2.4, found a substantial number of inversions; we even found at least one, unexpected case of contiguous inversions.

As part of our sequence evolution model analysis, we assessed inversion significance. We first performed each set of gene alignments with and without inversions. The inversions alignment scores include the scores for the inversions themselves as well as any substitutions and indels within the inverted subsequences.

Our null hypothesis was that the number of inversion alignments with shorter distances would be equal to the number with the same distance as the corresponding alignments without inversions. Because our model produces the highest scoring, most probable paths, adding an insignificant mutation type to the model would result in no change alignment distances; hence, we treated an equal distance as the no-difference result in our assessment.

Since there are two possible outcomes for each independent trial (shorter distance or equal distance), we had a binomial distribution [174]. With our 50% probability of success and our relatively large sample sizes, the normal distribution applies.

	HUMAN	LUCAMammalia	LUCAMetazoa	LUCAEukaryota
shorter inversion distances	283	331	269	199
alignments	436	381	291	205
mean	218	190.5	145.5	102.5
standard deviation	10.44	9.76	8.53	7.16
difference from mean	65.0	140.5	123.5	96.5
sd's from mean	6.23	14.40	14.48	13.48
diff / no inversion distance	0.03159	0.04235	0.04134	0.03806

Table 23: Inversions in alignments

In Table 23 we show the inversion analysis results. The first row shows the number of successful trials, i.e. those inversion alignments with shorter distances out of the total number of reference species alignments shown in the second row. The next two rows show the mean and standard deviation of the corresponding normal distribution. Following these are a row indicating the difference between the number of successful trials and the mean, and a row showing how many standard deviations above the mean our result lies. Since all of the reference species' results are more than 3 standard deviations above the mean, we reject the null hypothesis and conclude that inversions are a significant mutation mechanism in our model. Note that these results include both nonhomologous and homologous alignments, with the latter having relatively few inversions. If we had only looked at the nonhomologous results, the effects of inversions would be even more dramatic.

The final row indicates the ratio of the mean difference in distance between alignments with and without inversions to the mean distance without inversions. While these ratios amount to only a few percent, since they are ratios of logs their effect is significant.

7.4.2 Evolution of nonhomologous genes

Each of our successor reference species has a genome with a set of nonhomologous genes. These are the genes in the genome that are unique to the reference species and thus do not a have

a corresponding homolog in the predecessor reference species. Clearly, these nonhomologous genes evolved from a sequence in the predecessor reference species; there are two possibilities for how this may have occurred. The first possibility is that the nonhomologous gene evolved from an unrelated gene in the predecessor species. The second possibility is that the nonhomologous gene evolved from relatively random, noncoding sequence in the predecessor species.

In our model, we investigated both possibilities. In particular, we attempted to answer the question of whether or not arbitrary coding sequence had an advantage in evolving novel function as compared to random sequence. For each nonhomologous gene in each reference species (excluding LUCA as it has no predecessor in our model) we did an alignment with three random sequences and a sample from each group of genes in its predecessor's genome as described above. Note that while each gene was aligned only with gene sequences from its immediate predecessor, many of those genes have homologs in yet earlier species.

We noted that in most cases, the coding sequence distance to the nonhomologous, target gene was shorter than the random sequence distance to the nonhomologous, target gene. To quantify this difference, for each target gene we computed the mean m_r and standard deviation s_r of the $distance_r$ from the random sequences to the target. Using this mean and standard deviation for each gene, we computed the normalized differences $diff_{nonhomo}$ between each nonhomologous gene $distance_{nonhomo}$ and m_r:

$$diff_{nonhomo} = (distance_{nonhomo} - m_r) / s_r$$

We performed the analogous calculation for each random sequence $distance_r$:

$$diff_r = (distance_r - m_r) / s_r$$

Under the null hypothesis of no difference between nonhomologous and random distances, these normalized differences should be comparable in scale for all genes and should have 0 mean and unit variance. We performed a ttest on the two sets of differences using a one-tailed distribution, as the $diff_{nonhomo}$ were typically negative. We used a heteroscedastic ttest [175] as the samples were of unequal variance. We also computed a mean for the $diff_{nonhomo}$. The results are shown Table 24.

	HUMAN	LUCAMammalia	LUCAMetazoa	LUCAEukaryota
ttest	1.134%	1.846%	0.002%	19.513%
mean	-0.422	-1.611	2.560	-0.484

Table 24: Nonhomologous gene ttest

With the exception of LUCAMetazoa, we reject the null hypothesis for each reference species as these results indicate that $distance_{nonhomo}$ and $distance_r$ are somewhat likely (>80%) to have different means. In particular, the mean of $diff_{nonhomo}$ was negative with the exception of LUCAMetazoa. While the overall sample sizes were comparatively small, this analysis provides modest evidence that coding sequence in general has an overall advantage, compared to random sequence, in evolving into an unrelated coding sequence. This may be due, for example, to protein secondary and tertiary structures that are functional in many different contexts in biology.

Since the overall sample sizes were comparatively small and LUCAMetazoa showed a contradictory result, we elected to retain the random sequence results in our model along with the nonhomologous gene results.

7.4.3 Universal source sequences

Noting our conclusion from 7.4.2 that novel sequence is more likely to come from coding sequence than from random sequence, we investigated whether there might be universal source sequences, i.e. sequences that are comparatively more likely to evolve into many other nonhomologous, coding sequences. To shed some light on this question, we selected four genes from LUCAMammalia, one from each of its groups, to be our source genes. We looked at the distance from each of these four genes to each of our nonhomologous, target genes, the 39 genes unique to HUMAN. We normalized the distances to their target sequence lengths so that they could be compared. We performed a one-factor analysis of variance (ANOVA [176]) on the four groups of normalized distances.

Our null hypothesis was that the means of the distances for the four source genes would be the equal. Our analysis of variance result indicated ($p < .035$) that we should reject the null hypothesis and conclude that the means are not equal. This result provides modest evidence for universal source sequences.

The source gene with the lowest mean distance was 215350LUCAMammalia, a gene with homologs going all the way back to LUCA. The function of this gene is described as uncharacterized according to the OMA database [91]. We suggest that a possibility is that its primary function is to act as a universal source sequence.

7.4.4 LUCA to HUMAN estimate

With the results from our sequence evolution model, we are now in a position to estimate the required mutations that must have taken place in the evolution from LUCA to HUMAN. We used our effective HUMAN genome size from chapter 2, $1.3*10^8$ base pairs and the results shown in Figure 24. We used our expected progression from LUCA through our intermediate reference species to HUMAN.

	substitutions	inserts	deletes	inverts
HUMAN	180	24	24	85
LUCAMammalia	374	62	64	259
LUCAMetazoa	372	58	63	244
LUCAEukaryota	432	56	47	238
Total / Kb	1358	200	198	826
Grand total	1.77E+08	2.60E+07	2.57E+07	1.07E+08

Table 25: Total LUCA to HUMAN mutations

The mutations counts are shown in Table 25. The Total/Kb row contains total mutation counts per kilobase. We note that the total substitution count is greater than 1000. This indicates many sites per kilobase changed more than once during the period of evolution from LUCA to HUMAN. The last row indicates total mutation counts needed to produce the entire effective portion of the HUMAN genome. These counts represent our estimate of the total count of mutations between LUCA and HUMAN.

8 Population evolution model

8.1 Related work

Gillespie [177] gives a broad overview of population genetics. Population models in the context of the ecological processes that determine size and structure of a population is covered by Begon et al [38].

In finite populations, random changes in allele frequencies (variations in one or more genome sites that cause a gene to have multiple forms) result from reproduction rate variation from one individual to another and to sexual reproduction. These random changes are known as genetic drift [30 178]. Genetic drift can reduce the genetic variation of a population by causing one of several alleles to fix (become ubiquitous throughout the population). It also affects the probability of survival of new mutations.

A significant component of population evolution is natural selection that acts on fitness variations caused by differing phenotypes, which are in turn caused by differing genotypes. Natural selection reduces the prevalence of the less fit genotypes. Futuyma [30] states that many geneticists believe that genetic drift explains most sequence differences observed unless there is evidence of natural selection or other factors. However, there is significant evidence that selection pressure affects population evolution. One such example is the differing mutation rates observed for synonymous versus nonsynonymous codons [179]. If selection pressure were absent, these rates would always be identical. A comprehensive population evolution model must therefore provide a means to incorporate selection effects.

Coalescent theory [155 156 178] defines the lineage of a set of alleles. The alleles in a sample are traced backwards in time, joining at common ancestors, often until a single common ancestor of all of the alleles is reached. While our population model produces the genomic future from a past, population starting point, coalescent theory produces the genomic past from a current, population starting point. This allows the likelihood analysis of the sample genomes under various mutation models and could provide a useful complement to our approach.

The concept of carrying capacity [180] describes how a population is limited by its environment. According to Malthus [181], an unconstrained population will have a rate of population increase $d\mathit{population}/dt = g*\mathit{population}$ where g is the growth rate of the population. The value of g is approximately *birthRate* - *deathRate*; immigration and emigration also affect g. This exponential growth rate assumes unlimited resources to support the population. Recognizing that resources are ultimately always limited, Verhulst [182] developed a logistic model of population growth, with $d\mathit{population}/dt = g*\mathit{population}*(1-(\mathit{population}/K))$, where K is known as the carrying capacity of the environment for the population. Note that as *population* approaches K, $d\mathit{population}/dt$ approaches 0. It follows that in an actual population, g approaches

0 and *birthRate* approaches *deathRate*; these latter rates, however, do not fall to 0 except in the case of extinction.

Research into carrying capacity has that the way *birthRate* and *deathRate* change due to carrying capacity limitations varies depending on the specific situation. Some research indicates that *birthRate* is constant and *deathRate* rises; other research indicates that *birthRate* falls and *deathRate* is constant [183 184]. As a compromise in our model, *birthRate* and *deathRate* approach each other linearly under the effect of a carrying capacity limitation.

Among our reference species, LUCA was likely purely asexual, LUCAEukaryota likely engaged in sexual reproduction part of the time (a proxy for LUCAEukaryota, the yeast Saccharomyces cerevisiae reproduces mostly by budding [185]), and the remaining reference species reproduced primarily or solely sexually. Sexual reproduction has an associated cost compared to parthenogenetic reproduction [186], as each parent has only half of its genome reproduced in its offspring compared to parthenogenetic reproduction (reproduction by cloning). Yet sexual reproduction dominates in Animalia and much of Plantae.

Clearly sexual reproduction provides some advantages [186 187]. One such advantage, genetic segregation, was confirmed by Kirkpatrick and Jenkins [188]. In sexual reproduction, genetic segregation occurs when chromosome pairs separate during the production of gametes. Kirkpatrick and Jenkins assumed that advantageous mutations are relatively rare and that they fix through achieving the homozygous state. Due to genetic segregation, an advantageous mutation can achieve a homozygous state through a single mutation in a lineage, while two mutations are required to achieve a homozygous state with parthenogenetic reproduction.

Another advantage is due to crossing-over and genetic recombination that occurs during meiosis [186]. We assume that there is frequent evolutionary advantage to bringing together certain alleles of different genes on the same chromosome. Recombination will cause this to happen more frequently than it will happen in the absence of recombination, even when the frequencies of the two alleles are comparatively low. In our population model, we confirmed this advantage in a somewhat novel context.

The Hardy-Weinberg law [189] describes the relationship between allele frequencies and genotypic frequencies. It applies at equilibrium in a randomly mating, sexually reproducing population. If allele A_1 has frequency p_1 and allele A_2 has frequency p_2, according to the Hardy-Weinberg law, the genome frequencies are $p_1 * p_1$ for genotype A_1A_1, $2 * p_1 * p_2$ for A_1A_2, and $p_2 * p_2$ for A_2A_2.

However, populations often do not mate randomly [190]. One such case is inbreeding where individuals are more likely to mate with relatives than with other individuals. This results in a higher probability of homozygotes; this has a detrimental effect on the population. Another case of nonrandom mating is when a species occupies a large geographic area or one containing physical barriers such that individuals are not able to have sexual contact with all other individuals. In this case there will be genetic distinctions between the isolated subpopulations and Hardy-Weinberg will not accurately represent the entire species.

An extreme example of nonrandom mating occurs after speciation [191]. Speciation is the division of one species into two, different species; members of one species cannot mate with those of the other species and produce fertile offspring. It is thought to occur when two subpopulations of a species are geographically isolated and genetic variation reaches the point where members of the subpopulations can no longer successfully mate. In our model, we include speciation as well as a novel form of nonrandom mating.

Further limits to random mating arise due to geographic proximity. To model this effect, Kimura [192] devised a stepping-stone model which divided a population into subpopulations. The members of a subpopulation are local to each other and mating accours randomly within a subpopulation. Exchange of individuals occurs only between adjacent subpopulations. Kimura showed that, under some circumstances, the stepping-stone model more accurately reflects natural populations than other models. We used a unique variant of the stepping-stone model in our population evolution model.

Our population evolution model shares some characteristics with a Markov chain [36 37 193]. A Markov chain is a finite set of states with probabilities for the transition from one state to another and the property that the future state of the system depends only on the previous state. Our model also has a finite set of population pools that resemble states; the probability that an offspring will belong in a given pool depends only on the pools of its parent(s) and does not depend on the pools of more distant ancestors.

8.2 Model input data

Much of the input data for our population evolution model was obtained from our sequence evolution model. This information included effective genome length, fraction of the genome that is homologous to predecessor reference species, and mutation types, lengths, and counts for both homologous and nonhomologous genes. The mutation rates that we used are described in chapter 6.

However, our population model required additional data for its operation. For reasons described later in this chapter, we confined our population model efforts to the evolution from LUCAMammalia to Homo sapiens. Hence our population model data needs were limited to these two reference species.

There are a number of other model parameters, described below, that we varied for different model runs. For comparison purposes, we defined a standard model that uses a standard value for each of these parameters. Results from variations in these parameters were compared to standard model results.

Our population model required population growth rates separated into birth and death rates. For Homo sapiens, we used the values described in chapter 2. For LUCAMammalia, we used Mus musculus as a proxy and relied on information from the Louisiana Veterinary Medical Association [194] and the International Science and Technology Center [195]. Modeling the reproduction using this information yielded a birth rate of 1630/year and a death rate of .67/year. These are our standard values for birth and death rates.

We wanted our model to be able to simulate a range of reproduction mechanisms. Different species engage in various degrees of sexual reproduction and asexual reproduction. E. coli reproduce only asexually by binary fission, Homo sapiens reproduces only sexually, and the yeast Saccharomyces cerevisiae reproduces both sexually and asexually by budding. We defined a model parameter, sexual reproduction fraction, to capture these variations. Both Homo sapiens and LUCAMammalia reproduced solely sexually and so the sexual reproduction fraction parameter has a standard value of 1.0, meaning 100% sexual reproduction and 0% asexual reproduction.

Our population model also required carrying capacity estimates for prehistoric times; few specific numbers can be found in the literature for any time period. Yue et al [196] estimated the Homo sapiens carrying capacity of China under several scenarios in the post-industrial age to be on the order of 10^9. In prehistoric times, populations were limited by physical geographic

boundaries but not by political ones; this would yield a higher total figure than that in Yue et al. However, industrialization has increased carrying capacity; preindustrial carrying capacity was substantially lower. Based on this information, we estimated a Homo sapiens carrying capacity to be 10^7. We also needed a carrying capacity for LUCAMammalia. No specific numbers were found in the literature for Mus Musculus, a suitable proxy for LUCAMammalia. Mus musculus has a mass three orders of magnitude lower than Homo sapiens, a longevity approximately one order of magnitude lower, and a comparatively high birth rate. Considering these factors, we estimated the LUCAMammalia carrying capacity to be 10^9. These are our standard values for carrying capacity.

8.3 Model description

Our simple evolution model, described in chapter 2, consisted of the creation of More Recent Ancestors (MRAs) through transition mutations. Included in the simple model were several significant constraints not found in biological evolution processes:

1. All bases must change.
2. One specific transition mutation had to happen at any given time.
3. Only substitutions occurred.
4. No transition mutation was lost by a subsequent mutation that nullified the earlier transition mutation.

Our population evolution model obviates these constraints while containing additional components that more accurately simulate biology.

8.3.1 Model fundamentals

In our population model, we define two types of mutations:

1. A mutation$_+$: a plus mutation that creates an MRA; we have also referred to these as transition mutations.
2. A mutation$_-$: a minus mutation that nullifies a mutation$_+$ and causes the organism to revert to a previous MRA compared to its parent(s).

Our population model consists of a set of pools with dynamically varying populations, shown in Figure 26. Each pool is numbered according to the net number of mutation$_+$s contained in the individuals making up the pool. An individual's net number of mutation$_+$s is equal to total mutation$_+$s that have occurred in the evolutionary history of that individual's genome minus total mutation$_-$s that have occurred. Pools are distinguished by net mutation$_+$ counts and not other characteristics such as geographic location; individuals in different pools will typically be physically intermingled. We describe two pools as similar if their numbers are close in value.

While each individual in a given pool has the same net number of mutation$_+$s, they do not all have the same mutation$_+$s; in particular, any pair of individuals in a given pool will typically have some mutation$_+$s in common and some unique to each of them. In addition, each individual in each pool will have mutations unrelated to the evolution to the successor reference species; these are not tracked in our model.

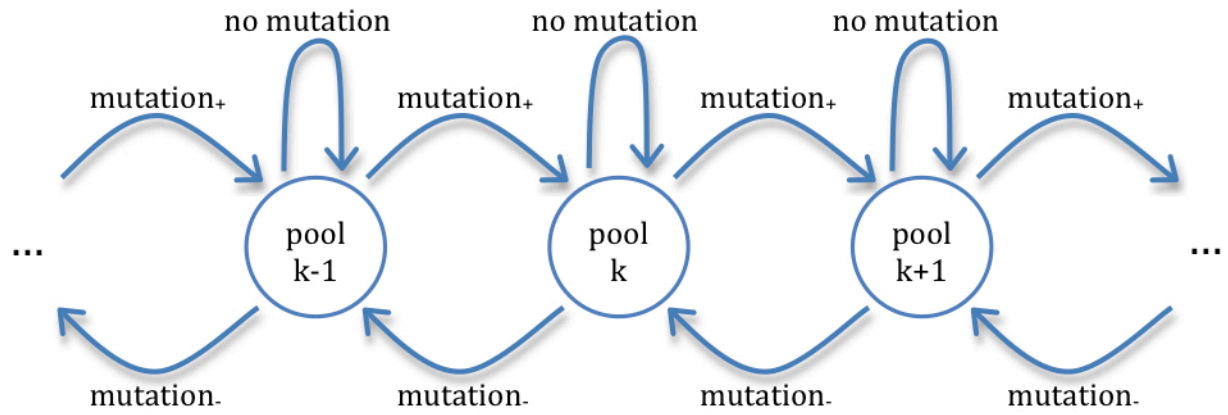

Figure 26: Fundamental population model

We began each model run with a single individual in pool 0. When a new individual was produced in a pool *k*, we assumed at most a single, relevant mutation occurred due the overall low mutation probability. If a new individual had no mutation, it remained in the same pool as its parent(s). If it had a mutation$_+$, it was promoted to pool *k*+1; if it had a mutation$_-$, it was demoted to pool *k*-1. Each of these possibilities has a probability, and each of the three probabilities is distinct for each pool.

At any point in a model run, there was a window of active pools. Active pools were defined as those pools with a population ≥ 1. When a pool's population reached 1, it was added to the window; when a pool's population fell below 1, it was deleted from the window. Mutation, reproduction, and other model actions occurred only within the window. Individuals could still be demoted below the lowest pool in the window, but they were no longer included in the model. They could also be promoted above the highest pool in the window, where they would lie dormant until the pool joined the window.

The effective sequence length for Homo sapiens and the net number of mutation$_+$s required to produce it from the LUCAMammalia genome were derived in chapter 2 and remained constant throughout our population modeling. These numbers were too large to model with acceptable compute time due to the algorithmic complexity of our model. Hence, each model run used a smaller model sequence length and corresponding specific number of net mutation$_+$s, *n*. Results were scaled up from these parameters as needed. Our standard model sequence length was 200.

Given the model sequence, we calculated the number of mutation$_+$s *n* required to evolve it from the source reference species (LUCAMammalia) to the target one (Homo sapiens):

$$n = (HomologousFraction * \sum_{i \in MutationTypes} HomologousCount_i)$$

$$+((1 - HomologousFraction) * \sum_{i \in MutationTypes} NonhomologousCount_i)$$

This value is approximately 255 mutation$_+$s per kilobase for LUCAMammalia to Homo sapiens evolution.

8.3.2 Mutation$_+$ probability

For our population model mutations, we used the data produced in our sequence models. In particular, for each reference species we defined a mutation$_+$ probability P_+ that is a weighted mean of the probabilities of the required substitutions, indels, and inversion mutations for the homologous and nonhomologous genes for that species. The weighting is based on the number of mutations for homologous and nonhomologous genes and the fraction of genes that are homologous.

For substitutions, we adjusted our counts for codon redundancy. While we have been modeling sequences of bases, they are in fact coding sequences and are composed of codons (triplets of bases). The 64 possible codons translate into 20 amino acids plus stop, which ends transcription (and so translation), for a total of 21 different translation results; hence the translation mapping is redundant. Without redundancy, there would be an expected value of three substitutions for a mutation$_+$ as there is a 1/3 chance of an incorrect base mutating to the correct base. A single substitution in any location in a codon can create any of 64-1=63 different codons and there are 21-1=20 different translation results. Thus, only 20/63 of those codons produce a different translation result. Two different codons produce the same translation result with a 1-(20/63) = 0.683 probability. The expected number of substitutions for a mutation$_+$ is then 3*0.683 = 2.05.

For indels, we used length 1 probabilities as they dominate the indel spectra.

For inversions, minimum length mutations do not dominate the spectra as they do with indels. We calculated a mean inversion length by dividing the number of inverted bases by the number of inversions. We then used the probability of an inversion of mean length in our mutation$_+$ calculation.

For each mutation type (substitution, insertion, deletion, and inversion) we took the reciprocal of its probability to obtain the expected population required to produce that mutation type, $Pop_{type} = 1/P_{type}$. We then calculated

$$Pop_{mean} = \frac{\sum_{i \in MutationTypes} Pop_i * Count_i}{\sum_{i \in MutationTypes} Count_i}$$

where Pop_{mean} is the mean population for the required mutation$_+$s and $Count_i$ is the count of mutations required for mutation type i. Finally, we set $P_+ = 1/Pop_{mean}$.

8.3.3 Mutation$_-$ probability

We also defined the probability of a mutation$_-$ undoing a mutation$_+$s, P_-. Mutation$_-$s have a different probability than mutation$_+$s that was calculated as follows.

Without redundancy, one substitution would be required for a mutation$_-$. A single substitution in any location in a codon can create any of 64-1=63 different codons. However, only 21-1=20 of those codons produce a different translation result. A single substitution will result in a different translation result (and so will be a mutation$_-$) with a probability of 20/63 = .317. The expected number of substitutions required for a mutation$_-$ is the reciprocal of this probability 1/0.317 = 3.15. We note that the expected number of substitutions for a mutation$_-$ is then higher than the expected number of substitutions for a mutation$_+$.

For indels, we used length 1 probabilities, as those are most probable and thus most common.

For inversions, we used length 12 probabilities, as those minimum length inversions are most probable. Such an inversion affects multiple sites. A length 12 inversion is the equivalent of replacing 12 bases with 12 others. However, only an expected ¾*12 = 9 will result a substitution; an expected 12-9=3 bases will not change. We treated an inversion like a substitution that has 9 times the probability of an inversion.

Since the different types of mutations are independent, we simply summed them to calculate P_-.

$$P_- = \sum_{i \in MutationTypes} P_i$$

8.3.4 Pool dynamics

As new individuals are produced in each pool due to reproduction, the populations of the pools change depending on the growth rate g, P_+, and P_-. We define:

$P_{no+} = 1 - P_+$	probability of no mutation$_+$
P_{no+}^{n-k}	probability of no mutation$_+$ out of (n-k) possible mutation$_+$s
$1 - P_{no+}^{n-k}$	probability of at least 1 mutation$_+$ out of (n-k) possible mutation$_+$s

and

$P_{no-} = 1 - P_-$	probability of no mutation$_-$
P_{no-}^{k}	probability of no mutation$_-$ undoing any of k mutation$_+$s
$1 - P_{no-}^{k}$	probability of at least 1 mutation$_-$ undoing any of k mutation$_+$s.

New individuals leave pool k with probabilities:

$P_{promoFromK}$	probability a new individual is promoted from pool k into pool k+1 $= P_{no-}^{k} * (1 - P_{no+}^{n-k})$

and

$P_{demoFromK}$	probability a new individual is demoted from pool k into pool k-1 $= P_{no+}^{n-k} * (1 - P_{no-}^{k})$.

New individuals stay in pool k with probability:

$P_{staysInK}$	probability a new individual stays in pool k $= 1 - (P_{promoFromK} + P_{demoFromK})$.

8.3.5 Population size and growth

Wright [150] defined effective population size as the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration. It is the effective population size that should be used in population modeling. LUCAMammalia and Homo sapiens are dioecious species [178] (a given individual can produce only eggs or sperm) with males and females produced in equal numbers. Given this situation and the fact that we mostly have large populations, our effective population size approximately equals overall population size.

Population size changes were based on a growth rate g where $g = birthRate - deathRate$ and $birthRate \geq deathRate$. We defined $pop_k(t)$ to be the population of pool k at time t. Our standard value for t was 1 year. We then have:

$$
\begin{aligned}
pop_0(0) &= 1 \\
pop_1(0) &= 1 \\
pop_i(0) &= 0 \text{ for } i > 1 \\
pop_k(t) &= pop_k(t\text{-}1) \\
&\quad + g * (pop_k(t\text{-}1) * P_{staysInK} \\
&\quad + pop_{k-1}(t\text{-}1) * P_{promoToK} \\
&\quad + pop_{k+1}(t\text{-}1) * P_{demoToK}).
\end{aligned}
$$

We began each model run with a single individual in each of the two lowest pools and the remaining pools empty. The population of a pool at time t is its population at time t-1 plus its growth times the fraction of offspring that stay in the pool (have the same net number of mutation$_{+}$s) plus offspring promoted from the previous pool plus the offspring demoted from the next pool.

8.3.6 Sexual reproduction

The majority of Eukaryotes reproduce sexually – this was a significant component of our model. It was implemented as an overlay on the fundamental model.

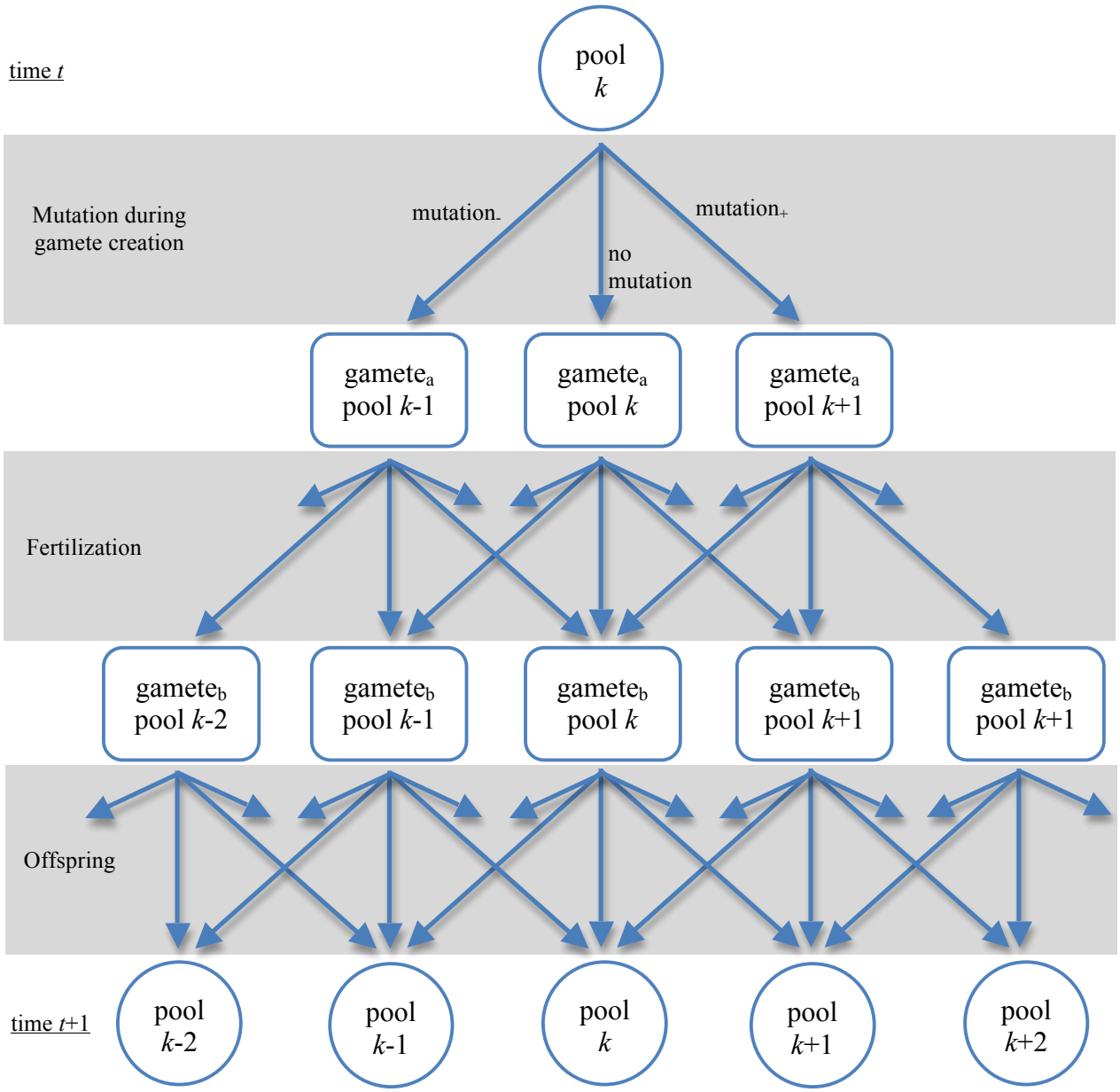

Figure 27: Sexual reproduction in population model

Figure 27 depicts the sexual reproduction component of our model. We first determined the mutation$_+$s and mutation$_-$s present in the gametes produced during a time period. These mutations occur during the process of creating the gametes, from the time of fertilization producing the first complete cell of a new individual until the time that individual creates gametes. Each gamete produced was placed into the appropriate gamete population pool depending on the net number of mutation$_+$s it contained that in turn depended on the mutation probabilities for that pool. Each gamete$_a$ in a pool then fertilized a gamete$_b$, from the same pool or a different pool. This was probabilistically determined based on pool populations – the probability of mating was proportional to the population of the pool containing gamete$_b$.

For a gamete from population pool k and a gamete from population pool l, the probability that they have a specific mutation$_+$ in common is $(k/n)*(l/n)$, where n is the total number of

mutation$_+$s. They have an expected number of mutation$_+$s in common = $n*(k/n)*(l/n) = (k*l)/n$. Their expected number of distinct mutation$_+$s = (k+l) – (2 * mutation$_+$s in common). When one gamete fertilizes another, the resulting offspring are very likely to inherit all of the mutation$_+$s that their parents have in common. The offspring will show a binomial distribution for the number of mutation$_+$s that are distinct between their parents.

For example if n=10, a gamete from pool 8 and one from pool 9 have approximately 7 mutation$_+$s in common and 3 that are distinct. When one fertilizes the other, 12.5% of the offspring would go into pool 7+0=7, 37.5% would go into pool 7+1=8, 37.5% would go into pool 7+2=9, and 12.5% would go into pool 7+3=10. Note that the range of offspring pools is larger than the range of gamete pools due to the range of the distinct mutation$_+$s.

In our model, we used this binomial distribution to determine the mutation$_+$ count of the offspring produced; these were then placed into the appropriate pools. Thus, the new population of each pool is the superposition of these binomial distributions.

Our model allowed a reference species to reproduce both sexually and asexually. We captured this ability in a sexual reproduction fraction parameter. This parameter is 0.0 for LUCA, 0.1 for LUCAEukaryota (using the yeast Saccharomyces cerevisiae as a proxy) and 1.0 for LUCAMammalia and Homo sapiens.

8.3.7 Environmental carrying capacity

As noted earlier, under carrying capacity limitations, d*population*/dt = g*$population$*(1-($population$/K)), where K is the carrying capacity of the environment for the population. As *population* approaches K, d*population*/dt approaches 0 and so g must approach 0. For g to approach 0, *birthrate* must approach *deathRate*. Since empirical evidence indicates variation about how these rates change, in our model, *birthrate* and *deathRate* approach each other linearly under the effect of a carrying capacity limitation.

Note that births in general don't fall to 0; only growth falls to 0. The *birthrate* and *deathRate* approach each other as the total population nears the carrying capacity. When the total population reaches carrying capacity *birthRate* = *deathRate*. However, *birthRate* cannot exceed 1 when carrying capacity is reached, as the total population cannot grow beyond the carrying capacity. These calculations are done for each pool using their relative proportion of the carrying capacity as the latter applies to the total model population.

8.3.8 Fitness

Futuyma [197] defines the relative fitness of the most fit genotype to be 1.0. A less fit genotype i has a relative fitness of 1-s_i where s_i is the selection coefficient and measures the relative selection against genotype i compared to the most fit genotype. Note that a higher selection coefficient results in a lower fitness.

In our model, we defined the fittest pool$_{fittest}$ as that pool with the highest number of mutation$_+$s that has a population ≥ 1.0. A pool$_i$ with a smaller number of mutation$_+$s has its birth rate reduced through multiplying it by a factor of 1-((*fittest*-i)*s), where s is the selection coefficient for the model. However, the birth rate cannot fall below 0.

For our standard model, we needed to choose a moderately low selection coefficient as we expected that a significant fraction of the net mutation$_+$s were likely to be neutral [111]. We used a moderate value of 1% for the standard selection coefficient [198].

8.3.9 Nonrandom mating

Typical population models, e.g. the Hardy-Weinberg law [189], assume individuals mate randomly. However, the random mating assumption is frequently inaccurate [190]. In our population model, we implemented two specific forms of nonrandom mating.

The first form of nonrandom mating is evolution of new species, speciation. It is clear that many speciation events occurred during the evolution from LUCAMammalia to Homo sapiens. It is difficult to estimate the number of speciation events as a single mutation, e.g. a polyploidy mutation, may cause a speciation event in a short time or a significant number of mutations over a long period may be required.

As described above, the model sequence length is less than the effective Homo sapiens sequence length and that a model run begins with minimum population in the two lowest pools and end when the final pool population reaches 1. There is thus an implied speciation event both prior to and subsequent to a model run. One implication of the prior speciation event is that individuals in low-numbered pools are prevented from mating with individuals in the earlier species and so net mutation$_+$s cannot be lost through this mechanism.

The second form of nonrandom mating we implemented was a new concept we refer to as mating radius. The mating radius is defined as the maximum difference in pool numbers that two mates may have. For example, an individual in pool 20 using a mating radius of 5 may mate with individuals in pools 15-25 inclusive and not with individuals in pools outside of that range.

Clearly a speciation event limits mating radius. We claim that geographic proximity is also a cause of limited mating radius. Consider the offspring resulting from the mating of an individual in pool a with an individual in pool b, where $a \leq b$. These offspring will go into pools with a binomial distribution, describe in section 8.3.6, with a maximum in pool $(a+b)/2$. Hence, the offspring will go into pools similar to pools a and b. Since offspring are likely to have small natal dispersal (geographic distance between birthplace and breeding place) [199], individuals on the whole tend to mate with other individuals from similar pools.

While related to other forms of nonrandom mating, the mating radius concept is novel and the available literature provided no guidance in choosing a numerical value. Our value of 5 was somewhat arbitrarily chosen to be a moderate value; in particular, we wished to avoid a mating radius that was too narrow. To better understand the implications of our choice, we examined the sensitivity of the model to different values of the mating radius; these results are described in section 8.4.3.

8.3.10 Operation

Our model operated by iterating through multiple time periods. During each time period, the model iterated through all of the active pools and performed the following:

1. Births and deaths – it would calculate the births and deaths that would take place during this time period. The birth rate was adjusted for fitness (using the selection coefficient and pool number) and carrying capacity limit; the death rate was adjusted for carrying capacity limit.
2. Mutations – the mutations for the newborn population were calculated and newborn population fractions were placed into demotion, promotion, or stays in pool categories depending on net mutation$_+$s.
3. Gamete production – the number of gametes created was calculated based on the number of newborns and the sexual reproduction fraction.

4. Asexual reproduction – the fraction of the newborn population that did not become gametes was calculated.
5. Sexual reproduction – a mating window for the gametes was calculated based on the active pools with at least one gamete and the mating radius. The pools in the mating window are called mate pools. For each mate pool, the fraction of gametes to mate with those in the mate pool was calculated based on the mate pool gamete population size. The expected number of common and distinct mutation$_+$s was also calculated. These were used to create the binomial distribution of distinct mutation$_+$s for the zygotes. From these, the new zygote populations were calculated.
6. Population update – the asexually produced newborns and the zygotes were added to the appropriate pools. The dead were subtracted from the appropriate pools.

When pool *n* reached a population of 1, it meant that a net number of *n* mutation$_+$s were present in that individual and the model run was complete. At the completion of a run, a number of results were collected.

Our model encompassed relatively wide range numeric magnitudes. Some probabilities were below and some populations were above the range of magnitudes that typical computer arithmetic formats could express and that typical arithmetic floating point instructions could operate upon. To successfully proceed with the implementation and operation of our model, we used the Java BigDecimal class [200] for many model variables. The BigDecimal class allows numeric magnitudes well beyond the range we required at the expense of greater storage space and much greater compute time. An additional disadvantage of the BigDecimal class is a dearth of available arithmetic operations beyond the 4 basic operations. In particular, we required exponentiation for our probability calculations and so we produced our own implementation.

We estimated model algorithmic complexity as a function of the model sequence length *n*. There are *p* pools manipulated over a period of *y* years. During sexual reproduction, each pool is mated with 2**matingRadius* other pools and produces offspring that are placed into the *p* pools. The complexity is then order *p***y**((2**matingRadius*)**p*). Since *p* and *y* tended to be of the same order as *n* and *matingRadius* was typically small, the complexity is order n^3. While this order is not an extreme one, the inner loop iterations of our model included a considerable number of BigDecimal arithmetic operations and comparisons; these caused the model to run relatively slowly even with modest values of *n*.

8.4 Model Results

8.4.1 Fundamental results

We began with a simplified initial model and added components to it to assess their effects. Our initial model used our standard parameter set with the exception of the following:

1. No carrying capacity limit
2. No sexual reproduction (and mating radius was therefore irrelevant)
3. No fitness effects

It produced the results shown in Figure 28. Figure 28a shows the time per net mutation$_+$. Note that the net mutation$_+$ count is equivalent to the pool number. Using our standard sequence length parameter of 200, there are 51 net mutation$_+$s required to complete the evolution process

and so our graph ends at 51. The vertical axis indicates the number of years required to produce an additional net mutation$_+$.

For evolution from one reference species to its successor to take place, the time per mutation$_+$ must be relatively low and relatively constant. If the result were too high, the evolutionary process would be too lengthy. If the time per mutation$_+$ began low but grew exponentially as the total number of mutation$_+$s increased, again the evolutionary process would be too lengthy; this latter situation was observed with some combinations of parameter settings.

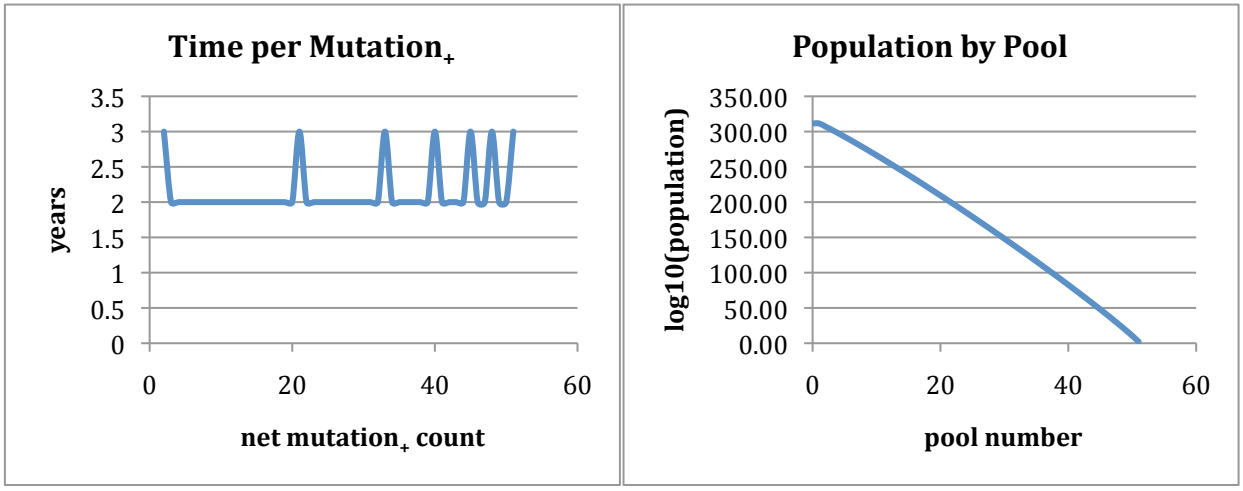

Figure 28a,b: Initial model results

Our initial model produced a time per mutation$_+$ result that was both relatively low and relatively constant. However, the initial model had a result incompatible with any biological process. Figure 28b indicates the population size by pool number at the conclusion of the evolution process. Note that the vertical axis does not indicate population size but it indicates the base 10 logarithm of the population size. The largest pools have populations exceeding 10^{300}; worse yet, all of these populations exist simultaneously. Clearly this result is not consistent with any realizable biological habitat. From our experience with this initial model, we imposed upon all subsequent model versions a carrying capacity parameter, limiting the total population to a physically realizable value.

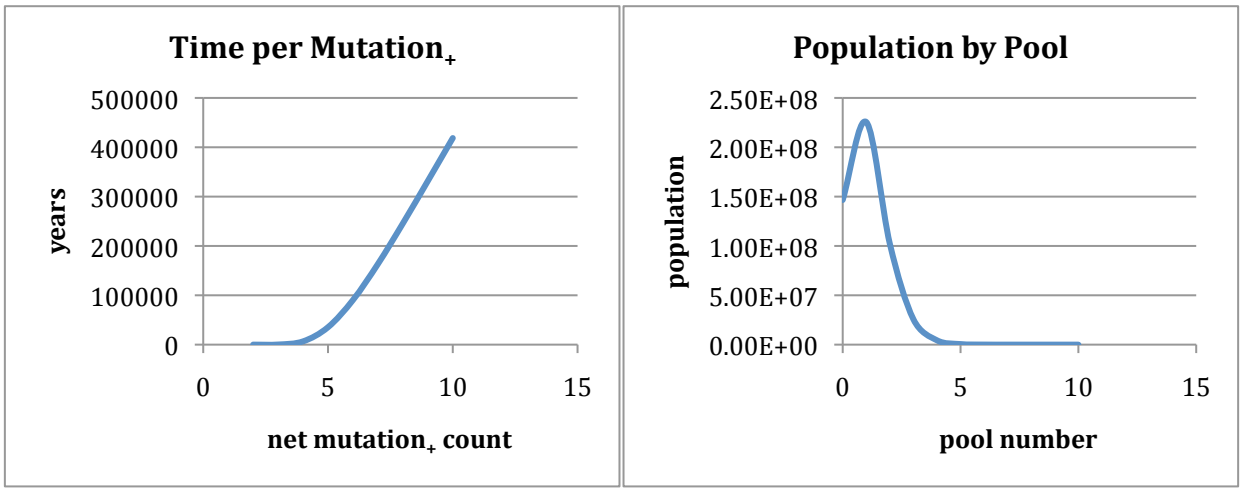

Figure 29a,b: Initial model with carrying capacity limit

Figure 29 presents the results from the initial model with our standard carrying capacity limit imposed. Our final pool populations, shown in Figure 29b, are large but realistic. However, the time per mutation$_+$, shown in Figure 29a, grows too rapidly with mutation$_+$ count. After over 10^6 years duration (at which point the model run was halted), a total net of only 10 mutation$_+$s had occurred. This result was too lengthy to be consistent with the known evolutionary time for Homo sapiens.

We then investigated the effect of enabling sexual reproduction in our initial model. Much has been written about the evolutionary advantages of sexual reproduction [186 187] and both LUCAMammalia and Homo sapiens reproduced sexually.

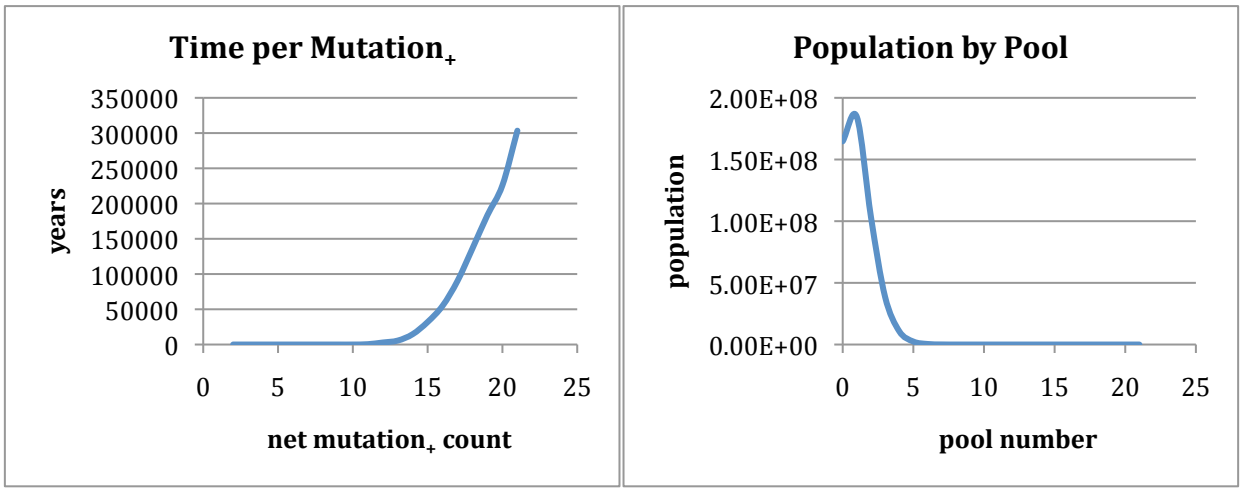

Figure 30a,b: Initial model with carrying capacity and sexual reproduction

Figure 30 presents the results that are improved somewhat from those without sexual reproduction. In Figure 30a, a total of 21 net mutation$_+$s occurred within approximately 10^6 years; this is more than double the result without sexual reproduction. However, since only 21 mutation$_+$s occurred and the time per mutation$_+$ was still rising, this result is also too lengthy to be consistent with known evolutionary time.

The population profile in Figure 30b was similar to that in Figure 29b, with the population being spread among more pools. In both models, it is clear that the smallest

numbered pools, those with the least net mutation$_+$s, dominate the population profile. Even with sexual reproduction, the majority of offspring will tend to go into a pool near their parents' pool(s).

The time-rate of mutation$_+$s would increase if the population of the smallest numbered pools fell over time. If the mutation$_+$s have a positive fitness effect, this would cause the populations of lower numbered pools to be reduced relative to those of higher numbered pools. To investigate this effect, we added a large, selection coefficient of 10%.

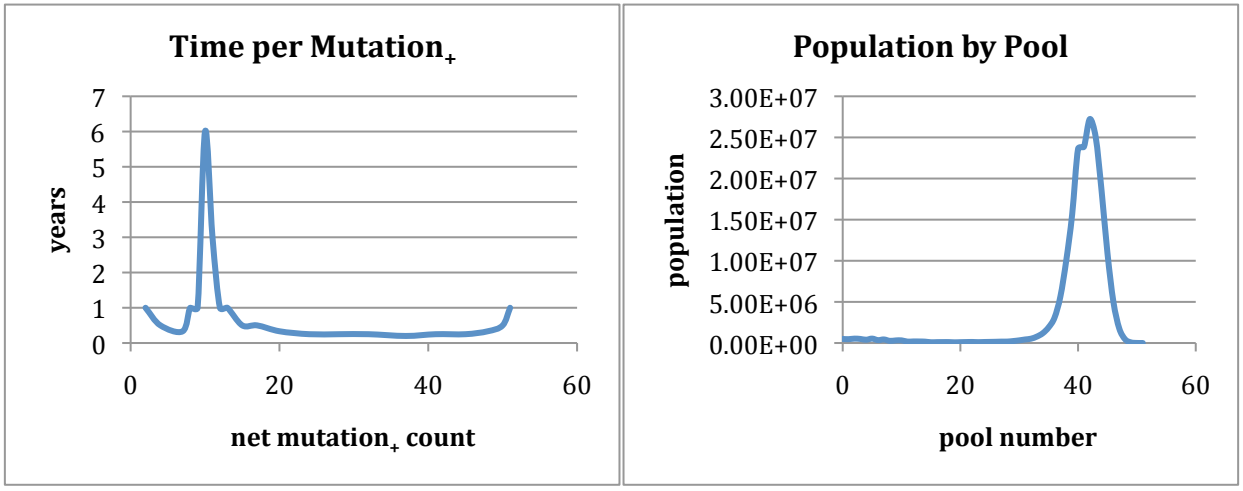

Figure 31a,b: Initial model with carrying capacity, sexual reproduction, and 10% selection coefficient

The results of the 10% selection coefficient are shown in Figure 31 and the changes from Figure 30 are dramatic. With the 10% selection coefficient, the time per mutation$_+$ shown in Figure 31a is both low and relatively constant as required. Figure 31b indicates that the population peak in the pools is near to the final pool as opposed to remaining near the initial pools when the selection coefficient is 0 as in Figure 30b. Observation of the position of this peak as the model runs indicates that it moves from left to right over time as expected. It is this population peak movement that enables comparatively rapid accumulation of mutation$_+$s over time.

A 10% selection coefficient per net mutation$_+$ is too high to be realistic, especially given that a significant fraction of the net mutation$_+$s were likely to be neutral. Using our standard value of 1% for the selection coefficient in the model yielded results inconsistent with known evolutionary time and improved over but not substantially different from the Figure 30 results, for which a selection coefficient of 0 was used. Clearly fitness alone was not the solution to our conundrum.

Retaining the 1% selection coefficient and imposing our standard mating radius of 5 was the next step in our population evolution modeling.

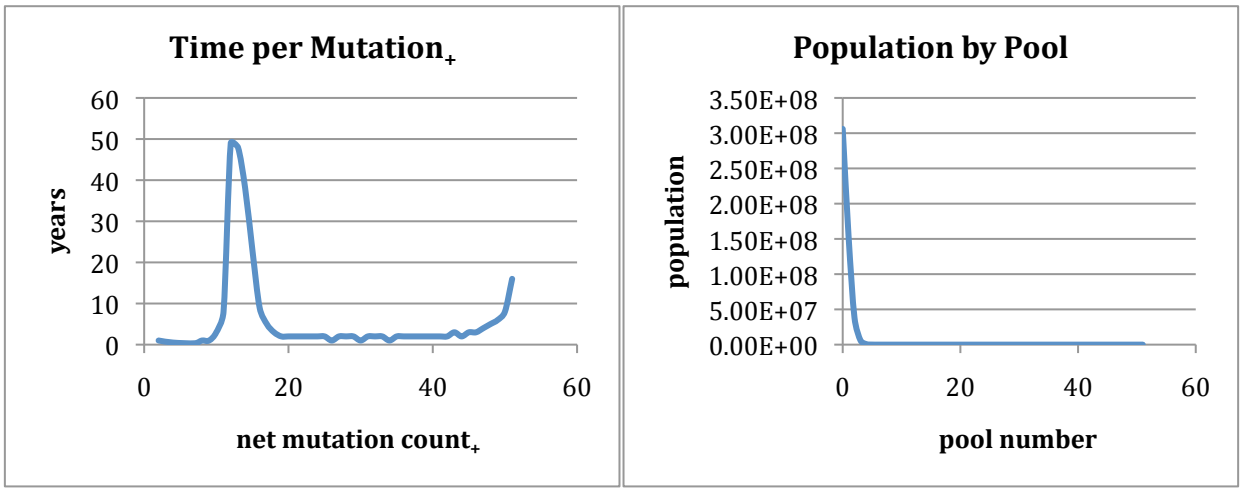

Figure 32a,b: Standard model

The results shown in Figure 32 use parameter values that make up our standard model: carrying capacity limited, sexual reproduction enabled, selection coefficient of 1%, and mating radius of 5. The time per mutation$_+$ shown in Figure 32a, is both low and relatively constant as required by known evolutionary time. We noted 4 epochs in this result:

1. In the first epoch, the time per mutation$_+$ is very low as the early pools grow at their full rate, having not yet reached the carrying capacity limit. In addition, when the highest numbered pool k is low, there are such a large number of possible sites for a mutation$_+$ and so comparatively few for a mutation$_-$ that virtually all offspring are promoted to the next pool. The number of sites for a mutation$_+$ is given by the binomial coefficient $\binom{n-k}{1}$ where n is the total number of mutation$_+$s; this large number results in a high probability of offspring being promoted to a higher pool number.
2. The second epoch begins just as the 7th net mutation$_+$ occurs and pool 7 has population $\geq$ 1. This is the time at which the total population reaches the carrying capacity limit. The time per mutation$_+$ climbs, reaching a peak of 49 at mutation$_+$ 12, as the high populations in the early pools and limited growth throughout don't support rapidly populating increasingly high numbered pools.
3. The third epoch begins approximately when the 18th net mutation$_+$ occurs. At this point, the high numbered pools themselves have increased in population sufficiently to support populating yet higher numbered pools. The time per mutation$_+$ then falls to a low value.
4. The last epoch begins when the maximum net mutation$_+$ count is nearly reached at approximately the 46th net mutation$_+$. Here the time per mutation$_+$ climbs again due to three effects. The first effect is that with a large majority of mutation$_+$s having taken place, the probability of a net mutation$_-$ is much higher than the probability of a net mutation$_+$; a binomial coefficient analogous to that in epoch 1 applies. The second effect is that these high numbered pools have fewer high numbered pools to mate with due to the cutoff at pool 51, the highest pool in this model. The third effect is that offspring that would go into pools higher than pool 51 are lost – there is effectively a speciation event after pool 51. These lost offspring cannot produce offspring to contribute to pools numbered between 46 and 51.

Figure 32b once again indicates that the lowest numbered pools dominate the population profile.

8.4.2 Duration results

The duration of the evolutionary process derived from this model remained of some interest. As stated above, we concentrated our efforts on the LUCAMammalia to Homo sapiens evolutionary process. We calculated a duration for this process for all of our model runs that ran to completion. We present here the result from our standard model.

Our standard model had a sequence length of 200 bases, requiring 51 net mutation$_+$s to evolve, and the duration to evolve that sequence with our model using to LUCAMammalia to Homo sapiens parameters was 286 years. The effective Homo sapiens genome length is $1.3*10^8$ as derived in chapter 2. We then calculated:

286 years * ($1.3*10^8$/200) = 186 million years

This value compares well with the broadly accepted duration of approximately 200 million years [201]. Based on this result, we concluded that our model could produce results that are consistent with the biological process that took place on earth.

However, we do not describe this result as an independent confirmation of the 200 million year value. Depending on its parameter values, our model generated a range of values for this duration from 0.5 million years to values greater than the age of the universe. In theory, it is possible to mitigate this issue by merely confining parameters to accurate values. However, our confidence in parameter value accuracy varied considerably.

Model Parameters
High confidence:
mutations – types, counts, rates
sexual reproduction fraction
effective sequence length
fraction of genes homologous with predecessor
Modest confidence:
birth rate
death rate
Poor confidence:
environmental carrying capacity
fitness
mating radius

Table 26: Confidence level in parameter value accuracy

Table 26 lists the significant parameters for our population model and our confidence that the values we used for them were biologically accurate, i.e. reflected the actual values of those measures during the biological evolution of LUCAMammalia to Homo sapiens. Our comparatively high confidence in the accuracy of some parameter values is due to solid

empirical data and well-founded algorithms to manipulate it. We have only modest confidence in the birth and death rates because, while we have good empirical values for current species, values for theoretical species like LUCAMammalia are speculative and past environmental conditions likely had a significant effect on these rates. We have but poor confidence in some parameters, such as carrying capacity, as there is little empirical data related to them and/or they clearly varied substantially with environmental changes (e.g. changes in climate, competition, and predation) over time; one might justly describe our chosen values for these parameters as little better than wild guesses.

Given the wide range of our confidence in parameter value accuracy, one should not describe our duration result as definitive or as an independent confirmation of other duration measurements. It is simply consistent with other measurements. Nonetheless, we claim that our standard model, with a range of parameter values, provided us with significant insights into the biological evolutionary process.

8.4.3 Parameter sensitivity results

To fully understand the sensitivity of the model to its parameters, we ran it with a range of parameter values and assessed the changes in model results compared to the model results with standard parameter values.

Mutation rate

Experience with the model demonstrated that small changes in mutation rates had little effect. In Figure 33, we show the effects of varying the mutation rates determined in chapter 6 over a range of four orders of magnitude.

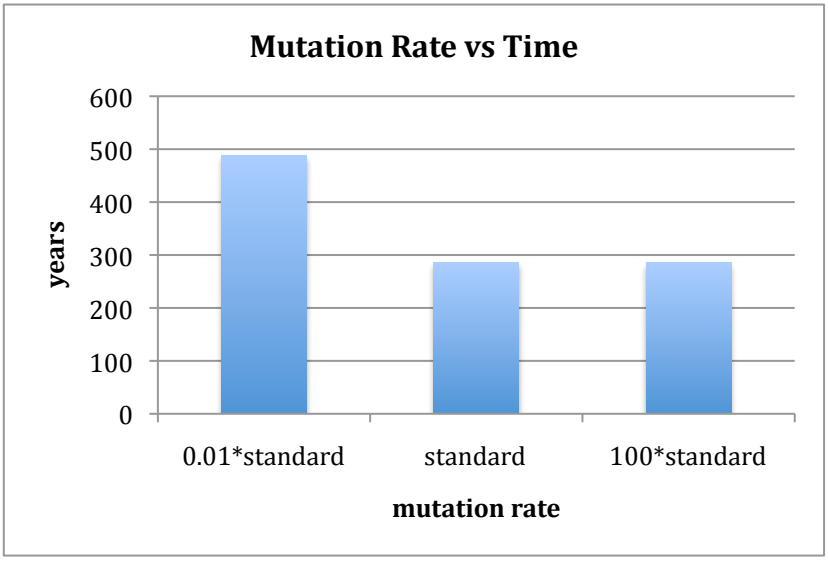

Figure 33: Mutation rate vs time

A hundredfold decrease in the mutation rate did increase the time as expected, from just less than 200 years to just less than 500 years. Increasing the mutation rate a hundredfold had little effect.

Sexual reproduction fraction

We ran the model with no sexual reproduction, a 0.5 fraction of sexual reproduction (and the remainder of the births produced asexually), and with the standard parameter value for LUCAMammalia and Homo sapiens, a 1.0 sexual reproduction fraction.

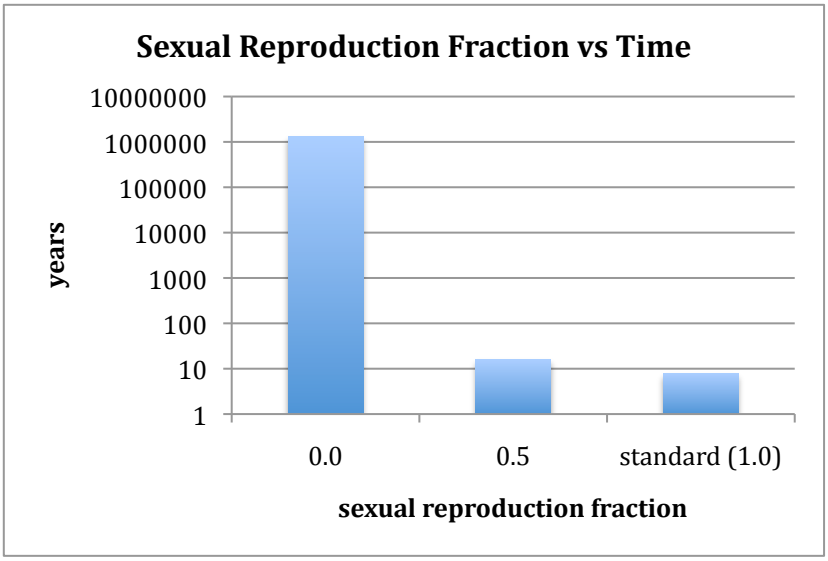

Figure 34: Sexual reproduction fraction vs time

Varying the fraction of sexual reproduction had a dramatic effect on the model results; note the logarithmic time scale in Figure 34. With no sexual reproduction (and therefore complete asexual reproduction), the evolution time was nearly 5 orders of magnitude larger than with full sexual reproduction. Reducing the sexual reproduction fraction to 0.5 had little effect, however.

Mating radius

Mating radius is one of the model parameters for which we have the least confidence in having an accurate value. Hence, its assessment over a range of values was unusually important.

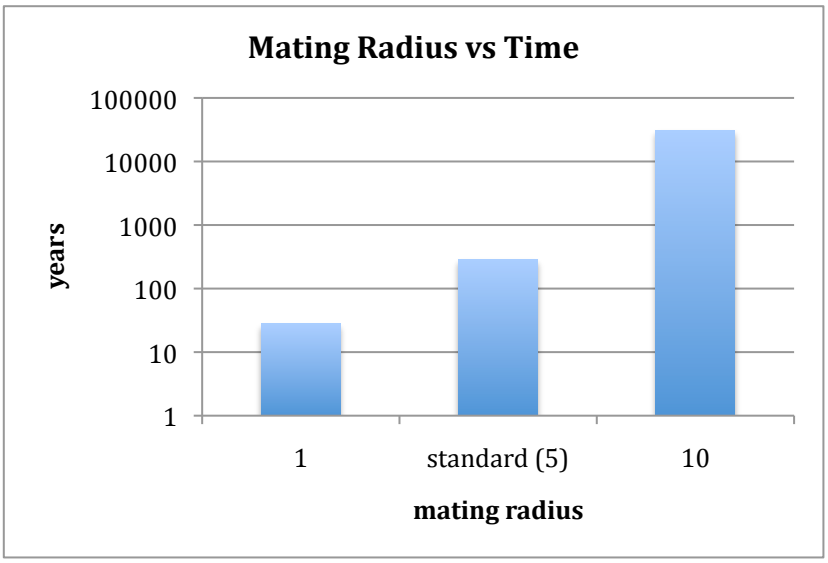

Figure 35: Mating radius vs time

Mating radius variation had a dramatic effect on evolution time; note the logarithmic time scale in Figure 35. Varying the mating radius from 1 to 10 caused an increase in time of approximately 3 orders of magnitude.

Birth and death rates

Experience with the model demonstrated that small variations in birth and death rates had little effect, so we assessed a range for these rates over 4 orders of magnitude in size, centered on the standard rates. The results are shown in Figure 36; the time scale has been expanded to make some subtle differences visible.

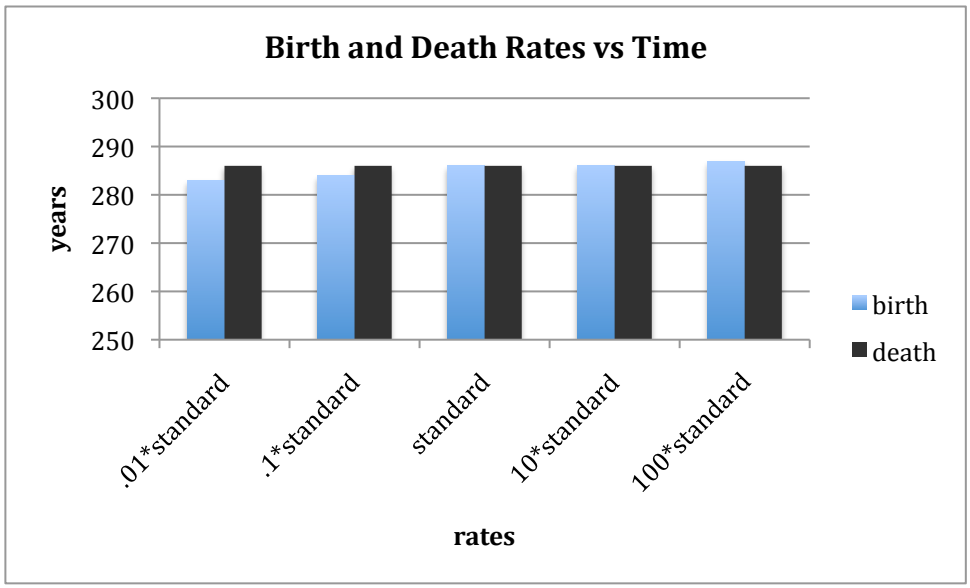

Figure 36: Birth and death rates vs time

On the whole, significant changes in birth or death rate had little effect on evolution time – this is due to the limits on these rates imposed by carrying capacity. We noted one unexpected effect:

a lower birth rate very slightly decreases evolution time. A detailed examination of the model results showed that the difference occurs just before carrying capacity is reached. A lower birth rate reaches carrying capacity more slowly and allows more time for higher numbered pools to be populated before the birth rate is limited by the carrying capacity.

Carrying capacity

Environmental carrying capacity is one of the model parameters for which we have the least confidence in having an accurate value. Hence, its assessment over a range of values was unusually important.

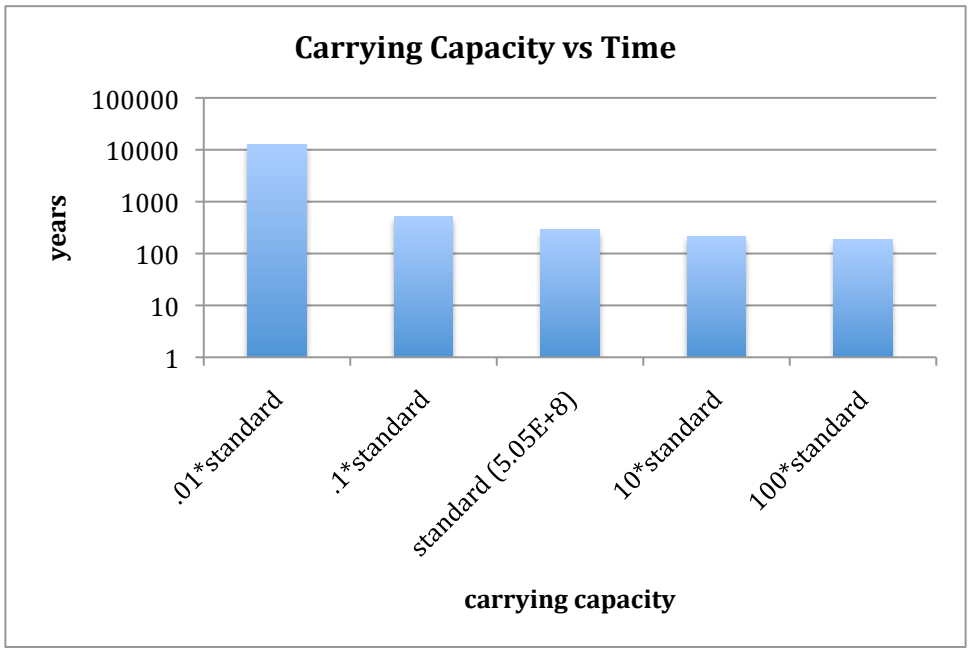

Figure 37: Carrying capacity vs time

Carrying capacity had a significant effect on evolution time; note the logarithmic scale of the time axis in Figure 37. In particular, a hundredfold-reduced carrying capacity caused an increase in time of similar magnitude. A hundredfold increase in carrying capacity caused only a modest reduction in time.

Fitness

To accurately gauge relative fitness, a natural population must be observed over time. As this was not possible, we ran the model with a range of fitness values, using selection coefficients from 0.0 up to 10 times our standard value.

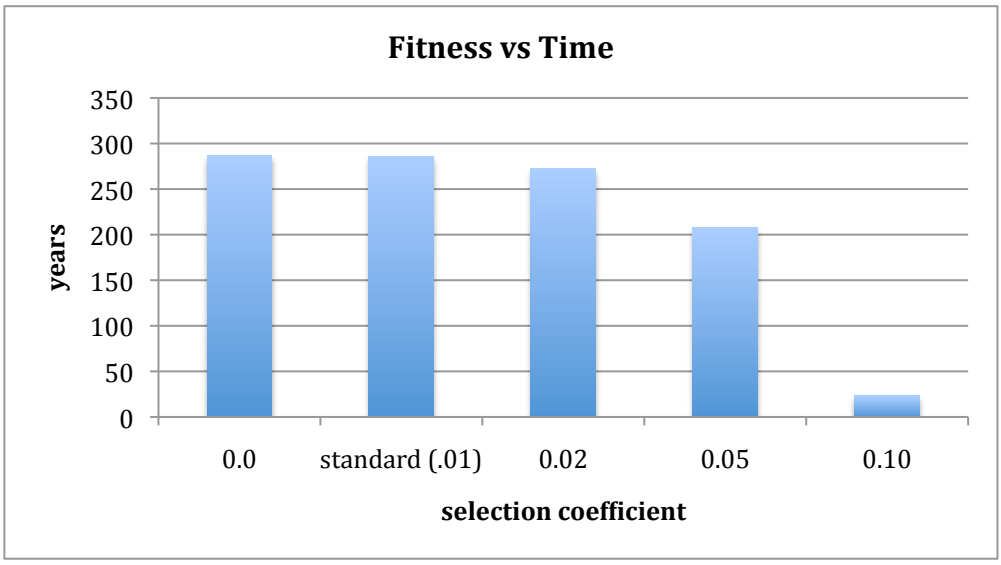

Figure 38: Fitness vs time

As expected, a higher selection coefficient caused a reduction in evolution time as shown in Figure 38. A higher selection coefficient causes the relative population of earlier pools to be reduced as their fitness declines. This reduction in early pool population is accompanied by an increase in later pool populations while subject to the carrying capacity limit.

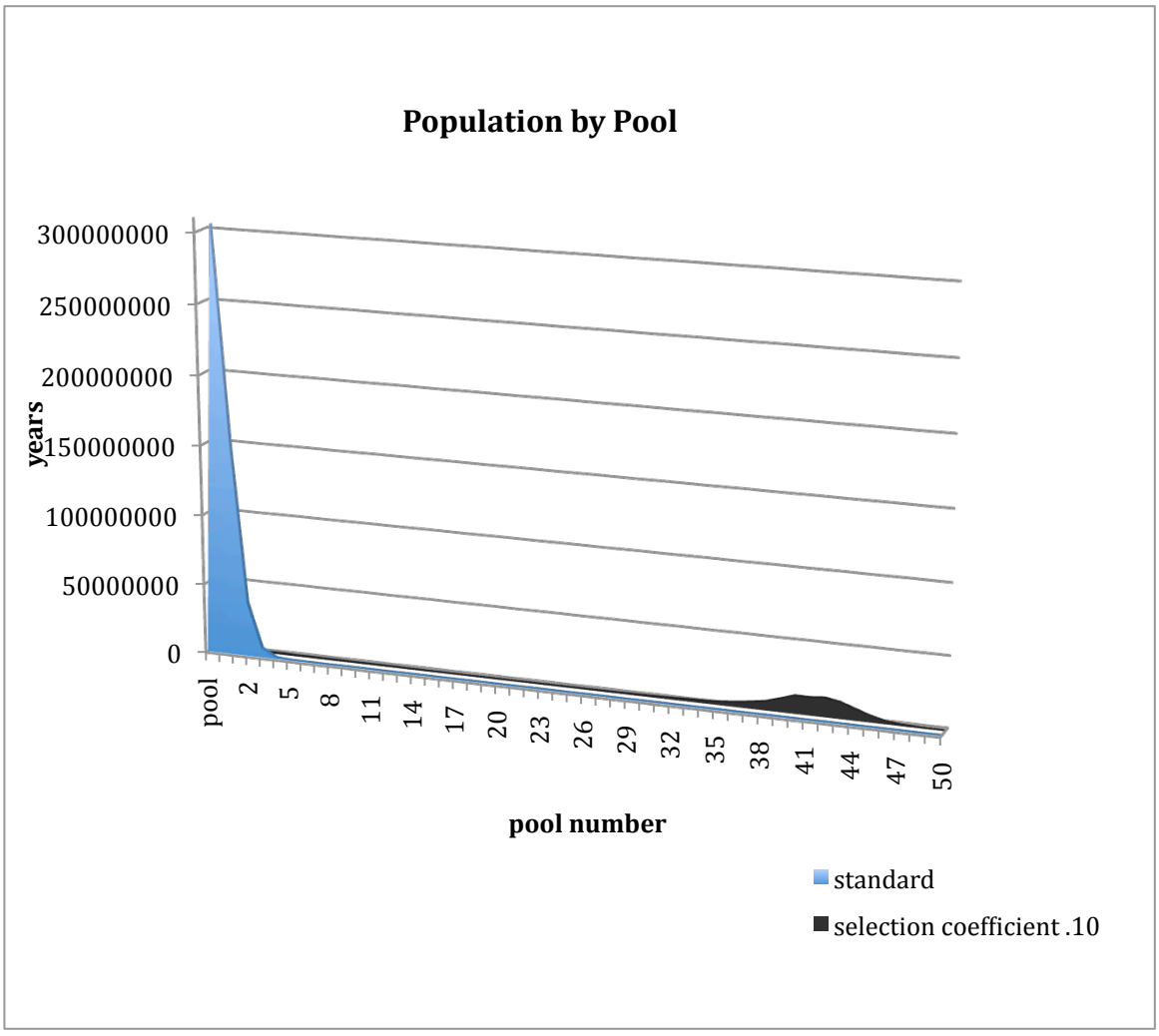

Figure 39: Population by pool for fitness variants

In Figure 39, we show a comparison of pool population profiles for standard fitness and a selection coefficient of .10. The majority of the population resides in the lowest pools in the standard model while the population profile of the .10 fitness model shows a binomial distribution in the high numbered pools centered at pool 42. This distribution is due to the binomial distribution of unique mutation$_{+}$s in offspring from sexual reproduction described in section 8.3.6.

Model sequence length

The effective sequence length for Homo sapiens was derived in chapter 2 to be $1.3*10^8$ bases; this value was used consistently in our population model. However, the computational complexity of our model precluded our using this full length in our model runs. In addition, there were clearly speciation events during the evolution from LUCAMammalia to Homo sapiens; these events acted as mating boundaries between population pools and our model runs were implicitly bracketed by speciation events. Our standard model used a model sequence length of 200 and we investigated model sequence lengths greater than this to determine their effects.

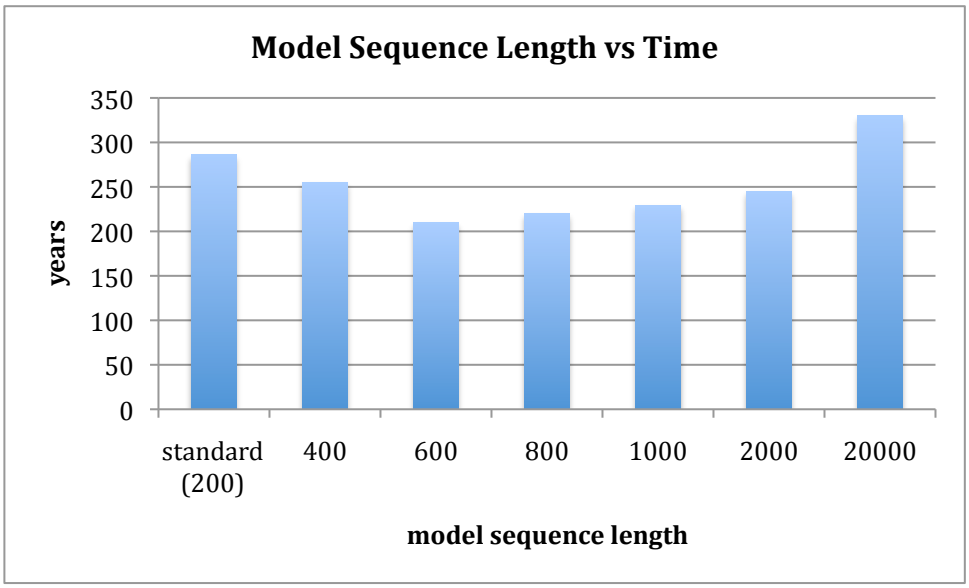

Figure 40: Model sequence length vs time

As shown in Figure 40, total evolution time falls from sequence length 200 to 600 and rises after that. This is drop in time is due to fitness effects reducing the early pools' population increasingly rapidly. However, this is a boundary effect that is reduced as model sequence length increases.

We also looked at evolution time per sequence base as a function of model sequence length. Taking the accepted value for LUCAMammalia to Homo sapiens evolution of $2*10^8$ years [201] and dividing by the effective sequence length for Homo sapiens $1.3*10^8$, we know that the biological time per base was 1.54 years/base. If our model accurately captured the biological process, it would yield a time/base of approximately 1.

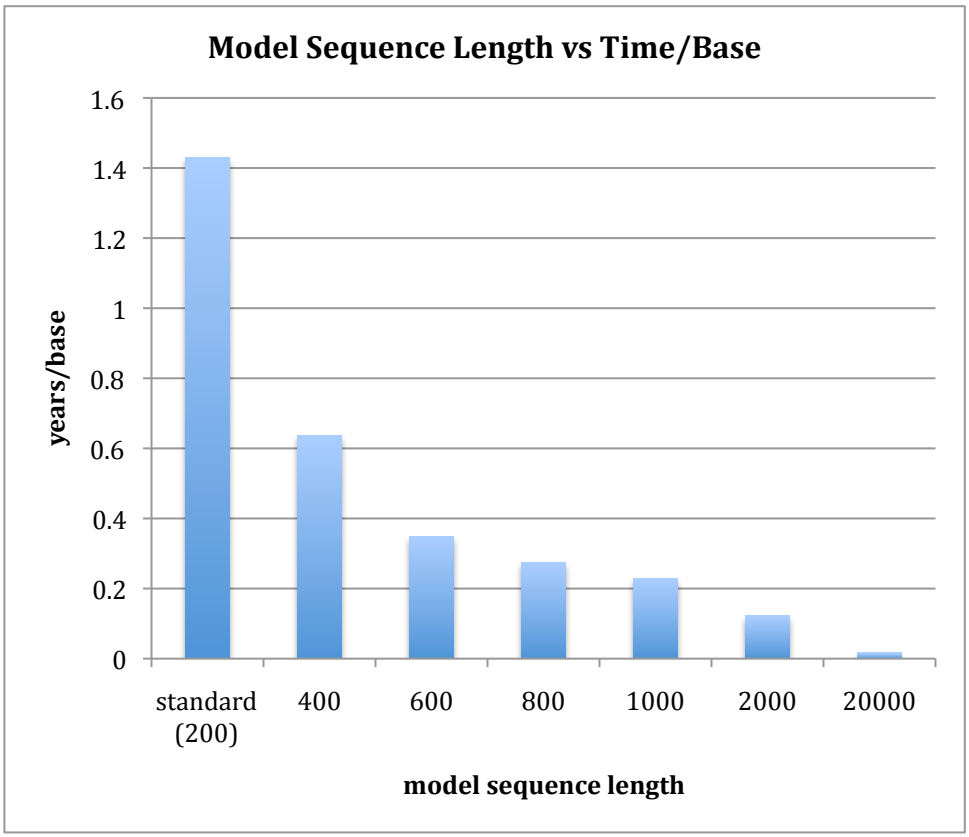

Figure 41: Model sequence length vs time/base

Rather than staying near a value of 1, Figure 41 shows that the time/base falls rapidly as model sequence length increases. It is clear that another parameter must be changing as model sequence length changes in order for the model to accurately reflect biological evolution. As the model sequence length increases, the number of pools correspondingly increases. With a larger number of pools with which to mate, the mating radius (to which model results are very sensitive) must also increase.

To investigate this possibility, we ran the model over a combination of model sequence lengths and mating radii. The results are shown in Figure 42.

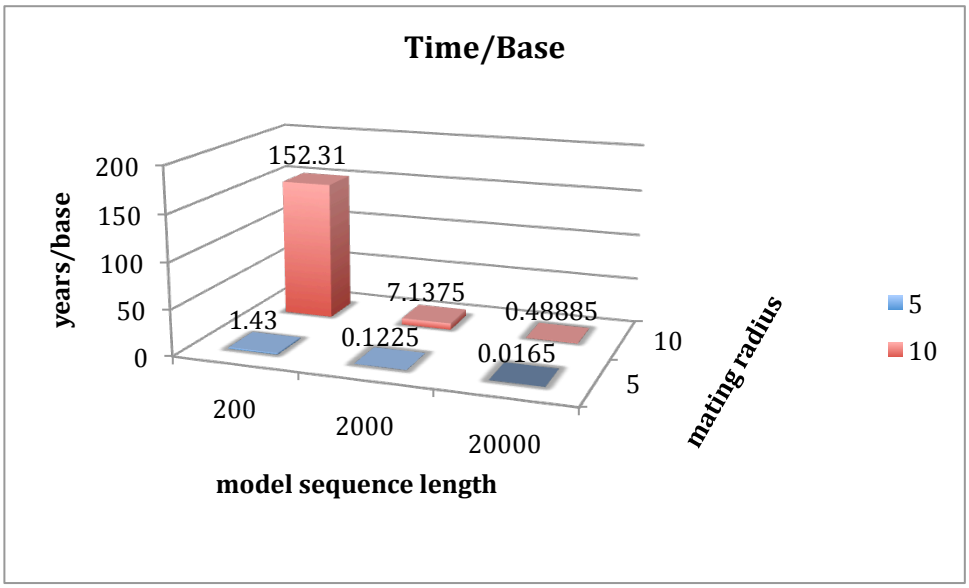

Figure 42: Time/base

When the standard mating radius is used, the time/base decreases from 1.43 at model sequence length 200, a value consistent with biological evolution, to .0165 at length 20,000, a value far too low implying unrealistically fast evolution. Increasing the mating radius to 10, the time/base decreases from 152, implying unrealistically slow evolution, to .48885, a value just below a realistic one. From this model experience, it is clear that mating radius is a monotonically increasing function of sequence length.

8.5 Model Analysis

A central result of our model is that, with reasonable estimates for the relevant parameters, the time/base is approximately 1 year/base and comparatively constant with model sequence length. A significantly time/base that declined with sequence length would imply an evolution time too short to be consistent with the broadly accepted time. Conversely, a time/base that grew with sequence length, as shown in Figure 29, would imply an evolution time too long. This result gives us confidence in the veracity of our population model.

8.5.1 Significant factors in population evolution

From a large number of model parameters, four emerged as very significant factors because model results proved very sensitive to their values:

Sexual reproduction fraction

Our model demonstrated that sexual reproduction is essential to the evolution from LUCAMammalia to Homo sapiens. As shown in Figure 34, evolution time was huge when it was absent. In its absence, the population source for higher numbered pools would be strictly from mutation; since the probability of a net mutation$_+$ falls as the net mutation$_+$ count rises, it very improbable for higher-numbered pools to be populated and so evolution time rises rapidly.

It is clear that both partially and fully asexually reproducing species also evolved successfully; for example, we expect that the evolution from LUCA to LUCAEukaryota took place with little to no sexual reproduction. We note that the 0.5 sexual reproduction fraction

time in Figure 34 is very similar to the 1.0 fraction value; this suggests a comparatively small fraction of sexual reproduction, or its equivalent, is sufficient. For LUCA and its asexually reproducing progeny, this equivalent is Horizontal Gene Transfer (HGT). Treangen and Rocha [202] analyzed a number of prokaryote genomes of varying sizes and concluded that the vast majority of expansions of protein families are due to HGT. Our model is clearly consistent with this result.

Mating radius

Our model indicated that mating radius is essential to evolution time. Doubling the mating radius in our standard model caused the evolution duration to increase exponentially. However, it is also clear that mating radius must increase with model sequence length as shown in Figure 42 or evolution time falls with increasing sequence length. When mating radius increases with model sequence length, the latter (and by implication speciation rate) is not a significant parameter of the model.

We described in section 8.3.9 mating radius is effectively limited by the geographic proximity of similar pools. We speculate that variation in sexual attraction may also limit mating radius. It is known that two individuals must be sufficiently similar (by definition, the same species) in order to mate successfully. However, it is advantageous that they not too similar lest they suffer the adverse fitness effects of inbreeding depression [203]. Studies have shown that mammals have various mechanisms for avoiding mating with individuals that are too similar, e.g. have a similar major histocompatibility complex (MHC) [204].

Mating with an individual from a similar pool may effectively provide that successful combination of similarity (similar number of mutation$_{+}$s) and dissimilarity (different specific mutation$_{+}$s). If this speculation is correct and is a factor in sexual attraction, this would tend to effectively limit mating radius.

Carrying capacity

Our model demonstrated that carrying capacity is an essential factor in evolution time. In its absence, time/base is acceptable almost independent of other model parameters but population sizes grow to unrealistic levels even for short sequence lengths. Our model demonstrated that evolution is significantly increased when carrying capacity is lowered from our standard model value by only modestly affected when the carrying capacity is raised.

This result suggests that as the carrying capacity of an environment for a species falls, the ability of the species to develop successful adaptations similarly falls. We speculate that this result is a partial explanation for the common occurrence of extinction, roughly 3 species/year [205].

Fitness

Our model indicated that fitness could be a significant factor; evolution times are significantly shortened with a very high selection coefficient. However, the time difference between a 0 selection coefficient and our standard model value of .01 was negligible; thus, the model is not sensitive to the selection coefficient value when it is low.

There is significant variation in the fitness effect of a mutation. A frameshift mutation is one that causes the all of the codons following it in a gene to change; an insertion of length 1 typically has this effect. A single mutation, e.g. an early frameshift mutation in a human's beta globin gene (one of the proteins that make up hemoglobin), can increase the selection coefficient

to 1, i.e. it is a fatal mutation. Conversely, a single mutation can confer antibiotic drug resistance to a bacterium and allow survival in an antibiotic-laden environment fatal to bacteria lacking the mutation [206]. In addition, mutations with only marginal (<<1%) effect on fitness are conserved [198]. Based on our model results and those obtained from the literature, we conclude that our mutation+s had a mean fitness effect that was slightly above neutral with significant variation around this mean.

We note that among these four most significant parameters, only one, sexual reproduction fraction, is in our high confidence of accurate value category. The remaining three parameters are in the poor confidence category. Better estimates for these parameters are a clear opportunity for future work.

Other parameters affected the model of course. For example, a significant reduction in mutation rate caused a nontrivial increase in evolution time as shown in Figure 33. However, none of the remaining parameters affected our model to the extent of our four most significant ones.

8.5.2 Generalizations

From our model results and analysis, we make several generalizations about population evolution:

Fundamental population evolution operation

From our model results, we summarize evolution from LUCAMammalia to Homo sapiens as follows:

1. Mutation+s occurred, resulting in gametes whose zygotes would likely be in higher-numbered pools. Mutation-s occurred, resulting in gametes whose zygotes would likely end up in lower-numbered pools.
2. Sexual reproduction created zygotes destined for a range of pools broader than that encompassed by their gametes. A limited mating radius reduced the zygote pool range so that higher-numbered pools were populated despite most of the gametes being in low-numbered pools.
3. Increased fitness coefficients slowed the growth of, and ultimately reduced the population of, lower-numbered pools; this resulted in increased population in higher-numbered pools.
4. By limiting how rapidly population pools could grow, carrying capacity slowed the evolution process to the rates we observe in nature.

Small population characteristics

When the population of a species is comparatively small, i.e. when the population is much lower than the carrying capacity, virtually any sequence can be produced in a time linear with the sequence length. This is true almost independent of other evolution parameters. Such a situation might exist immediately after a speciation event.

While speciation in prokaryotes is not well defined, a useful working definition might be when an individual evolves in such a way that it has a carrying capacity independent of the remainder of the strain. This occurs when an individual evolves antibiotic resistance in an antibiotic-laden environment. Andersson and Levin [206] observed that antibiotic resistance

carries a fitness cost, but that the cost is mitigated by subsequent evolution. We speculate that this subsequent evolution is enabled by small population characteristics.

Progress versus regress

In the evolution of a population, an offspring from an individual may belong to a higher-numbered pool, resulting in progress towards the successor reference species, or it may belong to a lower-numbered pool, resulting in regress. Most of the population evolution mechanisms we have been describing are relatively balanced with respect to progress versus regress. Mutation$_+$s are balanced by mutation$_-$s; the latter are in fact more likely than the former. Sexual reproduction broadens the pool range of offspring, but does so symmetrically; mating radius narrows the offspring range, also symmetrically. Carrying capacity limits the population size of all pools proportionally.

The sole exception to this balance is fitness, which favors higher-numbered pools. Progress can take place in the absence of pool fitness variation – higher-numbered pools will eventually be populated. But fitness variation speeds progress.

The speciation ratchet

When a speciation event occurs, it protects the mutation$_+$s contained in the new species genome from regression due to sexual reproduction. Individuals belonging to the new species cannot mate successfully with members of lower-numbered pools as they are a different species. Thus, speciation acts as a ratchet against dilution by sexual reproduction. It does not, however, protect against regression due to mutation$_-$s.

9 Conclusion and future work

9.1 Summary

The effort in this thesis research was concentrated in four tasks:

1. Creation of and calculation using a simple evolution model to confirm that even such a model offered verifiable predictions.
2. Identification of a set of reference species, starting with LUCA, and partial reconstruction of the reference species genomes.
3. Creation, implementation, and operation of a sequence evolution model, aligning the genomes of adjacent reference species to determine the mutations that likely took place during their evolution.
4. Creation, implementation, and operation of a population model that applied the results of previous tasks to determine the evolution of a population from one reference species genotype to the successor reference species genotype.

Each task produced specific results:

1. The simple model produced an independent time estimate for LUCA's evolution to Homo sapiens consistent with the generally agreed upon value.
2. The reference species identification and genome reconstruction task produced partial genomes for LUCA and its successor reference species.
3. The sequence evolution model produced the mutation types and quantities that took place during the reference species evolution. It also determined that inversions are an important mutation type, that nonhomologous sequences are more likely to evolve from unrelated coding sequence than from random sequence, and that some sequences may be more likely to evolve into unrelated sequences than others.
4. The population model allowed us to observe 4 epochs in population evolution and to discover a new evolution parameter, mating radius. It also determined the 4 most significant parameters in population evolution and produced an independent, long-term evolution time estimate consistent with the generally agreed upon value.

9.2 Future work

This research suggests valuable, future work in each of the 3 realms of computational biology:

1. In vivo: Determine specific values for some of the most significant evolution parameters, in particular carrying capacity, fitness, and mating radius. These determinations should take place in natural environments.
2. In vitro: Measure in the laboratory more mutation rates, especially inversion rates and lengths.
3. In silico:
    a. Complete the reconstruction of the LUCA and other reference species genomes.
    b. Make additional nonhomologous gene evolution comparisons. In particular, further confirm that such evolution is more probable from coding sequence than from random sequence. Also, confirm or refute our universal source sequence hypothesis: that some coding sequences are significantly better, compared to other coding sequences, at evolving into nonhomologous genes.
    c. Implement a parallelized version of the population model and operate it on very large model sequence lengths, simulating long time periods between speciation events.
    d. Determine hetereozygosity effects during population evolution.
    e. Determine mutation locality effects during population evolution. At the outset of evolution from one reference species to its successor, mutation$_{+}$s can occur at essentially location. Eventually, there must be some locality of these mutations in order to make a gene composed of contiguous nucleotides. This need for such locality may partially explain the presence of introns in genes.

10 Appendix

10.1 Software tools

Two significant software packages were created in the course of this research, Blind WatchMaker Path (BWMPath) and DropBox Distributed Processing (DDP). In addition, several existing software tools were extensively used. All of these are described below.

10.1.1 Blind WatchMaker Path (BWMPath)

BWMPath contains the bulk of the software written for this thesis, some 12,000 lines of Java code. It offers both a GUI and command line interface, and is able to function as a node in a DDP distributed processing environment. Its specific functions are described more thoroughly in the thesis; we summarize them here:

1. Manipulation and analysis of genome database information.
2. Creation and operation of a pipeline of existing tools to reconstruct genomes.
3. Sequence evolution modeling and analysis through sequence alignment.
4. Population evolution modeling and analysis using sequence, mutation, and species data.

10.1.2 DropBox Distributed Processing (DDP)

DDP is a distributed processing framework based on a distributed file system. Our computational resources for this research resided in multiple, disparate locations, each connected to the Internet through a strict firewall. The only distributed file system we found that could effectively penetrate these firewalls and provide incremental backup with history was Dropbox [207]. DDP can be based on any distributed file system that allows all computational resources read/write access to all relevant directories and files. We used it universally with Dropbox for the reasons cited above. DDP is written in Java, which provides multiplatform compatibility. A DDP master application is provided and a DDPLibrary is provided for Java-based worker code to include.

DDP assumes a single master node and any number of worker nodes that communicate through a shared directory universally accessible under the auspices of the distributed file system. Workers have unique names based on their Internet host names. Communication is via a formal protocol and each protocol item is instantiated as a file in the shared directory with a name in a well-defined format and contents that are specific to the protocol item type. The file name format is "*type_worker_time*.ddp", where *type* is the protocol item type, *worker* is the name of the worker (source or destination depending on item type) or '*' indicating a destination that is all workers, and *time* is the time in seconds obtained from the Java method

`Calendar.getInstance().getTimeInMillis()`. There is no requirement that the nodes' clocks be synchronized; *time* is used to distinguish otherwise identical protocol items.

type	*worker*	*file contents*
epoch	*	
announce	source	none required; useful to put software name and version
task	typically *	task description
init	source	task description
lockreq	source	name of item for which lock is requested
lock	destination	name of item locked
unlock	source	name of item to unlock
complete	source	time
interrupt	typically *	time
exit	typically *	time
exited	source	time
restart	typically *	time
restarted	source	time

Table 27: DDP protocol types

In Table 27, we list each DDP protocol *type*, whether *worker* is the source of the item or the destination, and the protocol item *file contents*. A task description must be parseable by the worker. We use a task description that consists of a task type and its parameters, all separated by |'s. A time entry in the *file contents* column indicates the time of a protocol type; responses, e.g. exited, use the time of the request, exit.

The `lockreq/lock/unlock` set of *types* is used to implement mutual exclusion – the DDP master ensures that there is at most one lock on any given item at any time. Anything with a unique name can be locked; in our usage, we locked directories in the distributed file system. DDP does nothing to enforce the locks and so assumes cooperative processes.

```
import ddplibrary.LibWorker;

String directoryBase = System.getProperty("user.home") +
"/Dropbox/";
String[] processDirs = {"dir1", "dir2", "dir3"};
String DDPProtocolDir = directoryBase  + "DDPProtocol/";

LibWorker.myLibWorker = new LibWorker(DDPProtocolDir);
LibWorker.myLibWorker.announce("Example code version 1.0);

for(String dir : processDirs) {
     if(LibWorker.myLibWorker.lock(directoryBase + dir)) {
          // lock granted; process the directory
     }
}
```

Figure 43: DDP worker code example

In Figure 43, we show a listing of minimal Java code for a DDP worker. The code imports the required DDP library file, defines a set of directories to process, and identifies the directory where the DDP protocol items will be placed. All of these directories must be part of the distributed file system. The code then makes a new `LibWorker` object and announces it to the DDP master. Finally, the code iterates through its directory list, processing only those directories on which it is granted a lock. In this example, up to three workers can usefully process the directories simultaneously, since there are three directories to process.

Our implementation of a DDP worker was considerably more elaborate than the minimal one described in Figure 43. Using additional `LibWorker` methods, our implementation did the following:

1. Look for tasks to perform.
2. When such a task is found, indicate that a task was initiated.
3. Request locks on task items and process items on which locks were granted.
4. Indicate the task is completed when all task items have been processed.
5. Respond to interrupt, exit, and restart.

Any comprehensive implementation of a DDP worker would need to do these same things.

The DDP master program and its daemon are run before any workers are started. The DDP master program provides a continuously updated display with the status of workers, locks, and tasks. It offers commands to:

1. Start, reset, and stop its daemon process.
2. Interrupt, exit, and restart workers; the latter is useful for updating the software for all workers in parallel.
3. Compose task descriptions and run tasks.
4. Test DDP master/worker interaction.

10.1.3 Multiple Alignment with Fast Fourier Transform (MAFFT)

MAFFT [71] is an efficient, effective sequence alignment program. It is primarily oriented toward aligning amino acid sequences and, by implication, coding DNA sequences such as those used in our research.

Much effort has been expended on the overall problem of sequence alignment. Two successful efforts at alignment software are the commonly used CLUSTALW [208] and T-COFFEE [209], which is known for accuracy. However, neither software tool has made great strides in reducing processor time that, in a typical approach using dynamic programming to optimize the alignment, involves time proportional to N^K, where N is the sequence length and K is the number of sequences.

The frequency of amino acid substitutions, which MAFFT users for alignment, depends on the difference of certain amino acid properties, particularly volume and polarity. Substitutions that involve amino acids, which are similar in these two properties, tend to preserve the structure of the resulting protein and so tend to be neutral with respect to selection pressure. MAFFT converts an amino acid sequence into a sequence of vectors, each of which is a normalized form of the volume or polarity of the amino acid at that site.

The correlation between two sequences is the sum of the volume correlation and the polarity correlation. The volume correlation between two sequences with positional lag k is the product of the normalized volume of the first sequence at site n and the normalized volume of the second sequence at site n+k, summed over the lengths of the two sequences. The polarity correlation is computed analogously.

A Fourier Transform [210] is a mathematical algorithm which characterizes a sequence by its frequency components rather than its amplitude component. The Fast Fourier Transform (FFT) is a computationally efficient version of the algorithm. Using an FFT to compute sequence correlations reduces the $O(N^2)$ computation to one that is O(N log N).

The performance of the MAFFT approach is substantial. The processor time consumed by MAFFT is drastically reduced compared to CLUSTALW with comparable accuracy in alignment results. When the number of sequences to be aligned is large, MAFFT is two orders of magnitude faster than T-COFFEE with comparable accuracy.

10.1.4 Randomized Axelerated Maximum Likelihood (RAxML)

RAxML [211 212] is a program for constructing phylogenetic trees of sequences based on ML inference. While ML methods provide more accurate results than parsimony methods, they are computationally intensive. RAxML begins by creating an initial tree based on parsimony. This is done for two reasons:

1. Parsimony trees are related to ML trees under some evolutionary models. This ensures that the initial tree has a relatively high likelihood.
2. Parsimony algorithms are not computationally demanding. This allows RAxML to perform several iterations, each with a different initial tree, in a reasonable time frame.

The central step in any phylogenetic tree reconstruction algorithm is the tree rearrangement step. This step begins with the best (highest likelihood) tree found up to this point. A range of distances is defined, usually beginning with a minimum distance of 1 and a maximum distance that is dependent on the specific overall algorithm being used. Inefficient algorithms must use a relatively small maximum rearrangement distance in order to reach completion in a reasonable time frame. New trees are constructed by moving all subtrees in the best tree a distance between the minimum and maximum rearrangement distance, inclusive. If one of the newly created trees has a higher likelihood than the current best tree, it replaces the current best tree. The process repeats until no better tree topology is found.

A typical approach to the above step is to optimize and compute the likelihood of the entire tree each time a subtree is rearranged. RAxML improves upon this approach by only optimizing and computing the likelihood of the local branches adjacent to the newly inserted subtree. This fast pre-scoring is used to make a small list of best potential trees that are likely to improve the likelihood of the entire tree compared to the current best tree. After completion of a rearrangement step, RAxML performs global optimization on only those topologies in its best potential tree list. Due to the efficiency of this approach, RAxML is able to analyze significantly more topologies and so allow a higher maximum rearrangement distance. This results in significantly better final trees.

Another variation of the process that RAxML performs takes place during the initial optimization phase, i.e. during the first few rearrangement steps. If during the insertion of a specific subtree in a new location a better likelihood results, this new topology is kept

immediately and all subsequent subtree rearrangements are performed on this improved topology. This results in rapid optimization of early trees.

Overall, the RAxML approach performs better on real data than the other common pyholgenetic tree construction programs PHYML, MrBayes, and PAxML [211]. RAxML is thus a fast, accurate ML program that allows inference of large trees in reasonable timeframes on commodity processor architectures.

10.1.5 Simultaneous Alignment and Tree estimation (SATe)

SATe [162 213] is a program that performs simultaneous alignment of a set of leaf sequences and construction of the phylogenetic tree that relates them using the ML criterion. In the cited study, it showed both improved tree and alignment accuracy compared to best current two-phase methods.

Methods to do alignment only typically estimate a phylogenetic tree for the input sequences and an alignment is then produced, guided by the tree. These methods are very sensitive to the guide tree and often require manual realignment. Methods to do tree construction only from unaligned sequences typically have limitations in either the accuracy of their trees or the number of sequences they can handle.

SATe uses a ML approach and treats gaps in sequences as missing data. SATe begins with a tree and alignment created by RAxML and MAFFT, respectively. It then searches for a tree/alignment pair with a higher ML score by performing hill-climbing searches from the current pair(s). This is done with an iterative, greedy search heuristic to find new pairs with better scores.

In the iterative process, new alignments are proposed by a divide-and-conquer method. A branch in a current tree is selected and the subtrees around the branch are determined. The sequences for each of the subtrees (the selected branch and its neighbors) are realigned by MAFFT. These subtree realignments are used to realign the entire sequence set. RAxML is then used to create a tree based on the new global alignment. Iterations continue until either predefined time limit or iteration limit is reached.

SATe is effective for simultaneous estimation of trees and alignments with a large number of sequences. For rapidly evolving sequences or over long evolutionary time periods, it offers shorter execution times and improved results compared to methods which first estimate sequence alignment and then build the phylogenetic tree.

10.1.6 Phylogenetic Analysis by Maximum Likelihood (PAML)

PAML [87] is a suite of programs to analyze DNA and amino acid sequences using ML. The suite offers a wide variety of analyses; in this description we focus on the three types of analysis that are significantly valuable in our research.

One type of analysis offered by PAML is comparison of a set of phylogenetic trees. It will calculate their likelihood values under a variety of nucleotide and amino acid models. These models estimate parameters of interest, such as species phylogenies, while allowing other parameters, such as substitution rate, to vary at different loci. The use of ML methods allows one locus to use information from other loci in the case of heterogeneous data sets.

Another type of analysis is estimation of species divergence times. PAML uses a likelihood method that assigns rates to different branches on the tree and then estimates both branch rate and divergence times from the sequences. A rate-smoothing procedure helps assigns rates to branches automatically.

Finally, PAML offers reconstruction of ancestral sequences. These are the sequences of extinct ancestors in a phylogenetic tree. It uses an empirical Bayes (EB) method that takes into account differences in branch lengths and in the relative substitution rates between characters, e.g. nucleotides. Both marginal and joint reconstructions are implemented. The marginal reconstruction assigns a character state to a single node, often the root node, in a tree. The joint reconstructions assign a set of character states to all ancestral nodes in a tree; this desirable when counts of changes at each site are needed. The EB approach uses ML estimates of branch lengths and substitution rates. This approach produces accurate character states with suboptimal handling of indels. The ancestral reconstructions may be used in a number of ways, e.g. to estimate selective pressures along lineages and to estimate evolution times between ancestral and existing species.

10.1.7 PRobablistic AligNment Kit (PRANK)

PRANK [158 163] is a program to analyze DNA and amino acid sequences using ML techniques. One challenge in typical, multiple sequence alignment algorithms is the approach to indels that significantly affect sequences of different lengths. With a typical approach, either the indel (also known as a gap) cost is so high that long gaps are not considered or so low that too many gaps are found that fragment the sequences. Separating the gap opening (which initiates a gap) cost and the gap extension (indicating that a gap is increasing in length) cost produces better results.

Progressive algorithms for multiple alignment iterate pairwise alignments from the terminal nodes (leaves) toward the root, guided by a phylogenetic tree that relates the sequences. The cost of an insertion or deletion should only be borne once, at the position in the tree where it actually occurs; PRANK does this properly while other multiple alignment algorithms often do not. The PRANK approach skips over a preexisting gap, with whatever preceded the gap being extended, so the gap cost is paid only once. PRANK also allows long deletions to overlap insertions by keeping track of overlapping gaps in their child sequences.

In a comparison with a traditional multiple alignment tool, PRANK results with respect to variation in sequence length is preferentially explained by single insertion events rather than multiple, independent deletions. The former is significantly more probable than the latter and so is likely to be more accurate.

10.1.8 Dendroscope

Dendroscope [214] is a program to visualize phylogenetic trees. A particular of strength of Dendroscope is the ability to view large trees, those with greater than 100,000 taxa. Such large trees are especially challenging in that their organization and graphical layout are computationally complex and therefore time-consuming. Dendroscope assigns to each subtree in a phylogenetic tree its own bounding box, within which it will be laid out and portrayed. Trees are presented from the root down, and a subtree is drawn only if its bounding box is visible. When information at a selected position is desired, the tree is searched from the root down, and bounding boxes that do not contain the selected position are not considered; this significantly improves search time.

Though such large trees have not thus far resulted from our research, other Dendroscope capabilities, often not shared by other tree viewers, make it an attractive choice. Among them are:

1. Collapsing subtrees.
2. Search function.
3. Comprehensive set of graphic export formats.
4. Comprehensive set of tree view formats.
5. Platform independence.

References

[1] Charles Darwin. (1976). *The Origin of Species*. New York, NY: Random House.

[2] Richard Dawkins. (1987). *The Blind Watchmaker*. New York: W. W. Norton & Company, Inc.

[3] NAIDS/World Health Organization. (2005). World HIV and AIDS Statistics. Retrieved January 23, 2006, from http://www.avert.org/worldstats.htm

[4] European Centre for Disease Prevention and Control. (2006). Avian Influenza Viruses. Retrieved January 23, 2006, from http://www.ecdc.eu.int/avian_influenza/index.php

[5] World Health Organization. (2003). Influenza. Retrieved January 23, 2006, from http://www.who.int/mediacentre/factsheets/2003/fs211/en/

[6] Ricki Lewis. (1995). The Rise of Antibiotic-Resistant Infections. *FDA Consumer Magazine*, *29*, 1.

[7] Douglas Hanahan, & Robert A. Weinberg. (2000). The Hallmarks of Cancer. *Cell*, *100*, 57-70.

[8] Centers for Disease Control and Prevention. FASTSTATS - Leading Causes of Death. Retrieved July 5, 2010, from http://www.cdc.gov/nchs/fastats/lcod.htm

[9] Douglas J. Futuyma. (1998). A Short History of Evolutionary Biology. In *Evolutionary Biology*. (pp. 17-30). Sunderland, MA: Sinauer Associates.

[10] Michael S. Waterman. (2000). Sequence Assembly. In *Introduction to Computational Biology*. (pp. 135-60). Boca Raton, FL: Chapman & Hall / CRC Press.

[11] John L. Hennessey, & David A. Patterson. (2003). Memory Hierarchy Design. In *Computer Architecture -- A Quantitative Approach*. (pp. 390-527). San Francisco, CA: Morgan Kaufmann Publishers.

[12] George Church. (2005). The Personal Genome Project. Retrieved 4 January 2006, from http://www.nature.com/msb/journal/v1/n1/full/msb4100040.html

[13] National Library of Medicine. (2005). Public Collections of DNA and RNA Sequence Reach 100 Gigabases. Retrieved January 8, 2006, from http://www.nlm.nih.gov/news/press_releases/dna_rna_100_gig.html

[14] Douglas J. Futuyma. (1998). Genetics and Development. In *Evolutionary Biology*. (pp. 31-50). Sunderland, MA: Sinauer Associates.

[15] Douglas J. Futuyma. (1998). Evolutionary biology. In *Evolutionary Biology*. (pp. 3-16). Sunderland, MA: Sinauer Associates.

[16] Motoo Kimura. (1994). *Population Genetics, Molecular Evolution, and the Neutral Theory*. Chicago: University of Chicago Press.
[17] Anthony M. Poole. (2002). My Name is LUCA -- The Last Universal Common Ancestor. *ActionBioscience*, *1*, 1-13.
[18] Douglas J. Futuyma. (1998). A History of Life on Earth. In *Evolutionary Biology*. (pp. 165-200). Sunderland, MA: Sinauer Associates.
[19] Douglas J. Futuyma. (1998). The Origin of Genetic Variation. In *Evolutionary Biology*. (pp. 267-96). Sunderland, MA: Sinauer Associates.
[20] Dee R. Denver, Krystalynne Morris, Michael Lynch, & W. Kelley Thomas. (2004). High mutation rate and predominance of insertions in the C. elegans nuclear genome. *Nature*, *430*, 679-82.
[21] Ryan J. Taft, & John S. Mattick. (2003). Increasing biological complexity is positively correlated with the relative genome-wide expansion of non-protein-coding DNA sequences. *Genome Biology*, *5*, 1-22.
[22] Miriam H. Meisler. (2001). Evolutionarily Conserved Noncoding DNA in the Human Genome: How Much and What For? *Genome Research*, *11*, 1617-18.
[23] Population Reference Bureau. (2004). *2004 World Population Sheet*. Washington, DC: Population Reference Bureau.
[24] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, & Peter Walter. (2002). DNA and Chromosomes. In *Molecular Biology of the Cell*. (pp. 191-234). New York, NY: Garland Science.
[25] John W. Drake, Brian Charlesworth, Deborah Charlesworth, & James F. Crow. (1998). Rates of Spontaneous Mutation. *Genetics*, *148*, 1667-86.
[26] Susan M. Rosenberg, & P. J. Hastings. (2004). Adaptive Point Mutation and Adaptive Amplification Pathways in the E. coli Lac System: Stress Responses Producing Genetic Change. *Journal of Bacteriology*, *186*, 4838-43.
[27] Kelly A. Frazer, Laura Elnitski, Deanna M. Church, Inna Dubchak, & Ross C. Hardison. (2003). Cross-Species Sequence Comparisons: A Review of Methods and Available Resources. *Genome Research*, *13*, 1-12.
[28] Ziheng Zang, & Anne D. Yoder. (1999). Estimation of Transition/Transversion Rate Bias and Species Sampling. *Journal of Molecular Evolution*, *48*, 274-83.
[29] B. Hoyle, & N. C. Wickramasinghe. (1999). Astronomical Origins of Life -- Steps Toward Panspermia. *Astrophysics and Space Science*, *268*, 1-398.
[30] Douglas J. Futuyma. (1998). Population Structure and Genetic Drift. In *Evolutionary Biology*. (pp. 297-336). Sunderland, MA: Sinauer Associates.
[31] Douglas J. Futuyma. (1998). Ecology: The Environmental Context of Evolutionary Change. In *Evolutionary Biology*. (pp. 59-84). Sunderland, MA: Sinauer Associates.
[32] Olivier Gascuel. (2005). The minimum evolution distance-based approach to phylogenetic inference. In *Mathematics of Evolution & Phylogeny*. (pp. 1-32). Oxford: Oxford University Press.
[33] Benoit B. Mandelbrot. (1982). *The Fractal Geometry of Nature*. New York, NY: W. H. Freeman.
[34] Olivier Gascuel. (2005). Genome Rearangements with Gene Families. In *Mathematics of Evolution & Phylogeny*. (pp. 291-320). Oxford: Oxford University Press.

[35] Stefan Kurtz, Adam Phillippy, Arthur L. Delcher, Michael Smoot, Martin Shumway, Corina Antonescu, & Steven L. Salzberg. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*, R12.
[36] Ben Noble. (1969). A Markov Chain Example. In *Applied Linear Algebra*. (pp. 46-54). Englewood Cliffs, NJ: Prentice-Hall, Inc.
[37] Richard Durbin, Sean R. Eddy, Anders Krogh, & Graeme Mitchison. (1999). Markov chains and hidden Markov models. In *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. (pp. 46-79). Cambridge, UK: Canbridge University Press.
[38] Michael Begon, Martin Mortimer, & David J. Thompson. (1996). *Population Ecology*. Oxford, UK: Blackwell Science, LTD.
[39] Mark Kot. (2001). *Elements of Mathematical Ecology*. Cambridge, UK: Cambridge University Press.
[40] -. (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research*, *38*, D142-D148.
[41] David R. Maddison, Katja-Sabine Schulz, & Wayne P. Maddison. (2007). The Tree of Life Web Project. *Zootaxa*, *1668*, 19-40.
[42] Matt Ridley. (2000). The search for LUCA. *Natural History*, *11*, 82-85.
[43] Nicolas Glansdorff, Ying Xu, & Bernard Labedan. (2008). The Last Universal Common Ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biology Direct*, *3(29)*, 1-35.
[44] Anthony M. Poole, & Derek T. Logan. (2005). Modern mRAN proofreading and repair: clues that the last universal common ancestor possessed an RNA genome? *Molecular Biology and Evolution*, *22*, 1444-55.
[45] L Delaye, A Beccara, & A Lazcano. (2004). The Nature of the Last Common Ancestor. In *The Genetic Code and the Origin of Life*. New York: Landes Bioscience.
[46] Geoffrey M Cooper. (2000). *The Cell, 2nd edition*. Sunderland, MA: Sinauer Associates.
[47] -. Heterotrophic - Definition and More from the Free Merriam-Webster Dictionary. Retrieved June 27, 2010, from http://www.merriam-webster.com/dictionary/heterotrophic
[48] -. Molecular Expressions Cell Biology: Animal Cell Structure. Retrieved June 27, 2010, from http://micro.magnet.fsu.edu/cells/animalcell.html
[49] S. G. Saupe. Concepts of Biology. Retrieved June 27, 2010, from http://employees.csbsju.edu/SSAUPE/biol116/Zoology/digestion.htm
[50] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, & Peter Walter. (2002). Development of Multicellular Organisms. In *Molecular Biology of the Cell*. (pp. 1157-258). New York, NY: Garland Science.
[51] -. Mammal - Definition and More from the Free Merriam-Webster Dictionary. Retrieved June 27, 2010, from http://www.merriam-webster.com/dictionary/mammal
[52] P. R. Hof, I. I. Glezer, E. A. Nimchinsky, & J. M. Erwin. (2000). Neurochemical and Cellular Specializations in the Mammalian Neocortex Reflect Phylogenetic Relationships: Evidence from Primates, Cetaceans, and Artiodactyls. *Brain, Behavior and Evolution*, *55(6)*, 300-10.
[53] Christopher B. Daniels, & Sandra Orgeig. (2003). Pulmonary Surfactant: The Key to the Evolution of Air Breathing. *News in Physiological Sciences*, *18(4)*, 151-57.
[54] Douglas J. Futuyma. (1998). Human Evolution and Variation. In *Evolutionary Biology*. (pp. 727-51). Sunderland, MA: Sinauer Associates.

[55] Francis S. Collins, Michael Morgan, & Aristides Patrinos. (2003). The Human Genome Project: Lessons from Large-Scale Biology. *Science*, *300*, 286-90.

[56] -. Functional and Comparative Genomics Fact Sheet. Retrieved June 28, 2010, from http://www.ornl.gov/sci/techresources/Human_Genome/faq/compgen.shtml#genomesize

[57] T. Ryan Gregory, James A. Nicol, Heidi Tamm, Bellis Kullman, Kaur Kullman, Ilia J. Leitch, Brian G. Murray, Donald F. Kapraun, Johann Greilhuber, & Michael D. Bennett. (2007). Eukaryotic genome size databases. *Nucleic Acids Research*, *35*, 332-38.

[58] Alexandr Andoni, Constantinos Daskalakis, Avinatan Hassidim, & Sebastien Roch. (2009). Global Alignment of Molecular Sequences via Ancestral State Reconstruction. *1st Symposium on Innovations in Computer Science*, 1-22.

[59] Mathieu Blanchette, Eric D. Green, Webb Miller, & David Haussler. (2004). Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Research*, *14*, 2412-23.

[60] David A. Liberles. (2007). *Ancestral Sequence Reconstruction*. Oxford University Press.

[61] Guillaume Bourque, & Pavel A. Pevzner. (2002). Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species. *Genome Research*, *12*, 26-36.

[62] Jian Ma, Louxin Zhang, Bernard B. Suh, Brian J. Raney, Richard C. Burhans, W. James Kent, Mathieu Blanchette, David Haussler, & Webb Miller. (2006). Reconstructing contiguous regions of an ancestral genome. *Genome Research*, *16*, 1557-65.

[63] Jianzhi Zhang, & Masatoshi Nei. (1997). Accuracies of Ancestral Amino Acid Sequences Inferred by the Parsimony, Likelihood, and Distance Methods. *Journal of Molecular Evolution*, *44*, S139-S146.

[64] Tal Pupko, Itsik Peer, Ron Shamir, & Dan Graur. (2000). A Fast Algorithm for Joint Reconstruction of Ancestral Amino Acid Sequences. *Molecular Biology and Evolution*, *17*, 890-96.

[65] Victor Kunin, & Christos A. Ouzounis. (2003). GeneTRACE–reconstruction of gene content of ancestral species. *Bioinformatics*, *19 (11)*, 1412-16.

[66] Ziheng Yang, Sidhir Kumar, & Masatoshi Nei. (1995). A New Method of Interence of Ancestral Nucleotide and Amino Acid Sequences. *Genetics*, *141*, 1641-50.

[67] Michael S. Waterman. (2000). Dynamic Programming Alignment of Two Sequences. In *Introduction to Computational Biology*. (pp. 183-232). Boca Raton, FL: Chapman & Hall / CRC Press.

[68] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, & D. J. Lipman. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, *215(3)*, 403-10.

[69] Michael Brudno, Alexander Pollakov, Asaf Salamov, Gregory M. Cooper, Arend Sidow, Edward M. Rubin, Victor Solovyev, Serafim Batzoglou, & Inna Dubchak. (2004). Automated Whole-Genome Alignment of Rat, Mouse, and Human. *Genome Research*, *14*, 685-92.

[70] Colin N. Dewey, & Lior Pachter. (2006). Evolution at the nucleotide level: the problem of multiple whole-genome alignment. *Human Molecular Genetics*, *15*, R51-R56.

[71] K. Katoh, K. Misawa, Kuma K., & T. Miyata. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*, 3059-66.

[72] Kevin Liu, Serita Nelesen, Sindhu Raghvan, C. Randal Linder, & Tandy Warnow. (2009). Barking Up the Wrong Treelength: The Impact of Gap Penalty on Alignment and Tree Accuracy. *Computational Biology and Bioinformatics*, *6(1)*, 7-21.

[73] T. Heath Ogen, & Michael S. Rosenberg. (2007). Alignment and Topological Accuracy of the Direct Optimization approach via POY and Traditional Phylogenetics via ClustalW + PAUP. *Systematic Biology*, *56(2)*, 182-93.
[74] Abdullah N. Arslan, Omer Egecioglu, & Pavel A. Pevzner. (2001). A new approach to sequence comparison: normalized sequence alignment. *Bioinformatics*, *17(4)*, 327-37.
[75] Richard Bellman. (1952). On the Theory of Dynamic Programming. *Proceedings of the National Academy of Sciences*, *38*, 716-19.
[76] E. W. Myers, & W. Miller. (1988). Optimal alignment in linear space. *Computer Applications in the Biosciences*, *4(1)*, 1-17.
[77] Ward Wheeler. (1998). Alignment characters, dynamic programming, and heuristic solutions. In *Molecular Approaches to Ecology and Evolution*. (pp. 243-51Birkhauser Verlag Basel.
[78] Michael S. Waterman. (2000). Multiple Sequence Alignment. In *Introduction to Computational Biology*. (pp. 233-52). Boca Raton, FL: Chapman & Hall / CRC Press.
[79] Dan Graur, & Li Wen-Hsiung. (2000). Molecular Phylogenetics. In *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates.
[80] C.R. Linder, & T. Warnow. (2005). Overview of phylogeny reconstruction. In S. Aluru (Ed.), *Handbook of Computational Biology*. Chapman & Hall.
[81] W. M. Fitch. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, *20*, 406-16.
[82] David Swofford. (1991). *Phylogenetic Analysis Using Parsimony, Version 3.1*. Champaign, IL: Illinois Natural History Survey.
[83] M. D. Hendy, & D. Penny. (1982). Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences*, *59*, 277-90.
[84] N. Saitou, & M. Nei. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4(4)*, 406-25.
[85] J. Felsenstein. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, *17*, 368-76.
[86] H. Kishino, T. Miyata, & M. Hasegawa. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*, *31*, 151-80.
[87] Z. Yang. (2007). PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*, 1586-91.
[88] D. L. Swofford, G. J. Olsen, P. J. Waddell, & D. M. Hillis. (1996). Phylogeny reconstruction. In *Molecular Systematics*. (pp. 407-543). Sunderland, MA: Sinauer Assoc.
[89] Mark Holder, & Paul O. Lewis. (2003). Phylogeny Estimation: Traditional and Bayesian Approaches. *Nature Reviews of Genetics*, *4*, 275-84.
[90] Alexander CJ Roth, Gaston H Gonnet, & Christophe Dessimoz. (2008). Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics*, *9*, 518.
[91] A. Schneider, C. Dessimoz, & G. H. Gonne. (2007). OMA Browser - Exploring Orthologous Relations across 352 Complete Genomes. *Bioinformatics*, *23(16)*, 2180-82.
[92] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, & D. L. Wheeler. (2008). GenBank. *Nucleic Acids Research*, *36*, 25-30.
[93] T. J. P. Hubbard, & all. (2009). Ensembl 2009. *Nucleic Acids Research*, *37*, 690-97.

[94] G. H. Gonnet, M. A. Cohen, & S. A. Benner. (1992). Exhaustive matching of the entire protein sequence database. *Science*, *256(5062)*, 1443-45.
[95] Christophe Dessimoz, Brigitte Boeckmann, Alexander C. J. Roth, & Gaston H. Gonnet. (2006). Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Research*, *34(11)*, 3309-16.
[96] Bastien Boussau, Samuel Blanquart, Anamaria Necsulea, Nicolas Lartillot, & Manolo Gouy. (2008). Parallel adaptations to high temperatures in the Archaean eon. *Nature*, *456*, 942-46.
[97] Eugene V. Koonin. (2003). Comparative genomics, minimal gene-sets, and the last universal common ancestor. *Microbiology*, *1*, 127-36.
[98] Patrick Forterre, & Herve Philippe. (1999). The Last Universal Common Ancestor (LUCA), Simple or Complex? *Biological Bulletin*, *196*, 373-77.
[99] Christos A. Ouzounis, Victor Kunin, Nikos Darzentas, & Leon Goldovsky. (2006). A minimal estimate for the gene content of the last universal common ancestor—exobiology from a terrestrial perspective. *Research in Microbiology*, *157*, 57-68.
[100] Joel B. Dacks, & W. Ford Doolittle. (2001). Reconstructing/Deconstructing the Earliest Eukaryotes: How Comparative Genomics Can Help. *Cell*, *107*, 419-25.
[101] Nobuko Arisue, Masami Hasegawa, & Tetsuo Hashimoto. (2005). Root of the Eukaryota Tree as Inferred from Combined Maximum Likelihood Analyses of Multiple Molecular Sequence Data. *Molecular Biology and Evolution*, *22(3)*, 409-20.
[102] J. A. Lake. (1990). Origin of the Metazoa. *PNAS*, *87(2)*, 763-66.
[103] Claire Larroux, Bryony Fahey, Sandie M. Degnan, Marcin Adamski, Daniel S. Rokhsar, & Bernard M. Degnan. (2007). The NK Homeobox Gene Cluster Predates the Origin of Hox Genes. *Current Biology*, *17(8)*, 706-10.
[104] Guillaume Bourque, Pavel A. Pevzner, & Glenn Tesler. (2004). Reconstructing the Genomic Architecture of Ancestral Mammals: Lessons From Human, Mouse, and Rat Genomes. *Genome Research*, *14*, 507-16.
[105] Morris Goodman, Ben F. Koop, John Czelusniak, Mark L. Weiss, & Jerry L. Slightom. (1984). The η-globin gene: Its long evolutionary history in the β-globin gene family of mammals. *Journal of Molecular Biology*, *180(4)*, 803-23.
[106] Jennifer A. Marshall Graves, & Jaclyn M. Watson. (1991). Mammalian sex chromosomes: Evolution of organization and function. *Chromosoma*, *101(2)*, 63-68.
[107] International Mouse Genome Sequencing Consortium. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*, 520-62.
[108] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, & Peter Walter. (2002). Cells and Genomes. In *Molecular Biology of the Cell*. (pp. 3-46). New York, NY: Garland Science.
[109] Jeffrey E. Barrick, Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E. Lenski, & Jihyun F. Kim. (2009). Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, *461*, 1243-47.
[110] Motoo Kimura. (1968). Evolutionary Rate at the Molecular Level. *Nature*, *217*, 624-26.
[111] Motoo Kimura. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature*, *267*, 275-76.
[112] Dan Graur, & Li Wen-Hsiung. (2000). Dynamics of Genes in Populations. In *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates.

[113] John Wakeley. (1994). Substitution-Rate Variation among Sites and the Estimation of Transition Bias. *Molecular Biology and Evolution*, *11(3)*, 436-42.
[114] Irene Keller, Douda Bensasson, & Richard A. Nichols. (2007). Transition-Transversion Bias Is Not Universal: A Counter Example from Grasshopper Pseudogenes. *PLOS Genetics*, *3(2)*, 185-91.
[115] Motoo Kimura. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Biology*, *16(2)*, 111-20.
[116] JF Crow. (2000). The origins, patterns and implications of human spontaneous mutation. *Nature Reviews Genetics*, *1(1)*, 40-47.
[117] A Nakabachi, A Yamashita, H Toh, H Ishikawa, HE Dunbar, NA Moran, & M Hattori. (2006). The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science*, *314(5797)*, 267.
[118] Adam Eyre-Walker, & Peter D. Keightley. (1999). High genomic deleterious mutation rates in hominids. *Nature*, *397*, 344-47.
[119] Michael W. Nachman, & Susan L. Crowell. (2000). Estimate of the Mutation Rate per Nucleotide in Humans. *Genetics*, *156*, 297-304.
[120] JC Roach, G Glusman, AFA Smit, CD Huff, R Hubley, PT Shannon, L Rowen, KP Pant, N Goodman, & M Bamshad. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, *328(5978)*, 636.
[121] Richard Ellis Hudson, Ullfar Bergthorsson, John R. Roth, & Howard Ochman. (2002). Effect of Chromosome Location on Bacterial Mutation Rates. *Molecular Biology and Evolution*, *19*, 85-92.
[122] M Lynch, W Sung, K Morris, N Coffey, CR Landry, EB Dopman, WJ Dickinson, K Okamoto, S Kulkarni, DL Hartl, & WK Thomas. (2008). A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A*, *105(27)*, 9272-77.
[123] C Haag-Liautard, M Dorris, X Maside, S Macaskill, DL Halligan, D Houle, B Charlesworth, & PD Keightley. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila. *Nature*, *445(7123)*, 82-85.
[124] Alec J. Jeffreys, Nicola J. Royle, Victoria Wilson, & Zilla Wong. (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature*, *332*, 278-81.
[125] James L.Weber, & Carmen Wong. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics*, *2(8)*, 1123-28.
[126] Yanhui Fan, Wenjuan Wang, Guoji Ma, Lijing Liang, Qi Shi, & Shiheng Tao. (2007). Patterns of Insertion and Deletion in Mammalian Genomes. *Current Genomics*, *8*, 370-78.
[127] Dan Graur, & Li Wen-Hsiung. (2000). Rates and Patterns of Nucleotide Substitution. In *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates.
[128] JP Rossiter, M Young, ML Kimberland, P Hutter, RP Ketterling, J Gitschier, J Horst, MA Morris, DJ Schaid, & P De Moerloose. (1994). Factor VIII gene inversions causing severe hemophilia A originate almost exclusively in male germ cells. *Human molecular genetics*, *3(7)*, 1035.
[129] Scot A. Kelchner, & J. F. Wendel. (1996). Hairpins create minute inversions in non-coding regions of chloroplast DNA. *Current Genetics*, *30(3)*, 259-62.

[130] SR Downie, & JD Palmer. (1992). Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In DE Solits (Ed.), *Molecular systematics of plants*. (pp. 14-35Chapman and Hall.
[131] M.J. Chaisson, B.J. Raphael, & P.A. Pevzner. (2006). Microinversions in mammalian evolution. *Proceedings of the National Academy of Sciences*, *103(52)*, 19824.
[132] J Ma, A Ratan, BJ Raney, BB Suh, W Miller, & D Haussler. (2008). The infinite sites model of genome evolution. *Proceedings of the National Academy of Sciences*, *105(38)*, 14254.
[133] L Feuk, JR MacDonald, T Tang, AR Carson, M Li, G Rao, R Khaja, & SW Scherer. (2005). Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet*, *1(4)*, 489-98.
[134] M Slatkin. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, *139(1)*, 457.
[135] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, & Peter Walter. (2002). DNA Replication, Repair, and Recombination. In *Molecular Biology of the Cell*. (pp. 235-97). New York, NY: Garland Science.
[136] JW Ijdo, A Baldini, DC Ward, ST Reeders, & RA Wells. (1991). Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proceedings of the National Academy of Sciences of the United States of America*, *88(20)*, 9051.
[137] Philip J. Stephens, Chris D. Greenman, Beiyuan Fu, Fengtang Yang, Graham R. Bignell, Laura J. Mudie, Erin D. Pleasance, King Wai Lau, David Beare, Lucy A. Stebbings, Stuart McLaren, Meng-Lay Lin, David J. McBride, Ignacio Varela, Serena Nik-Zainal, Catherine Leroy, Mingming Jia, Andrew Menzies, Adam P. Butler, Jon W. Teague, Michael A. Quail, John Burton, Harold Swerdlow, Nigel P. Carter, Laura A. Morsberger, Christine Iacobuzio-Donahue, George A. Follows, Anthony R. Green, Adrienne M. Flanagan, Michael R. Stratton, P. Andrew Futreal, & Peter J. Campbell. (2011). Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development. *Cell*, *144(1)*, 27-40.
[138] Ernest B Hook. (1981). Rates of Chromosome Abnormalities at Different Maternal Ages. *Obstetrics & Gynecology*, *58(3)*, 282-85.
[139] Justin Ramsey, & Douglas W. Schemske. (1998). PATHWAYS, MECHANISMS, AND RATES OF POLYPLOID FORMATION IN FLOWERING PLANTS. *Annual Review of Ecology and Systematics*, *29*, 467-501.
[140] KH Wolfe, & DC Shields. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, *387*, 708-13.
[141] Tal Dagan, & William Martin. (2007). Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *PNAS*, *104(3)*, 870-75.
[142] Olga Zhaxybayeva, J. Peter Gogarten, Robert L. Charlebois, W. Ford Doolittle, & R. Thane Papke. (2006). Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. *Genome Research*, *16*, 10999-1108.
[143] In-Geol Choi, & Sung-Hou Kim. (2007). Global extent of horizontal gene transfer. *Proceedings of the National Academy of Sciences*, *104(11)*, 4489-94.
[144] RD Finn, J Mistry, J Tate, P Coggill, A Heger, JE Pollington, OL Gavin, P Gunasekaran, G Ceric, & K Forslund. (2009). The Pfam protein families database. *Nucleic acids research*, D211-D222.

[145] Victor Kunin, Leon Goldovsky, Nikos Darzentas, & Christos A. Ouzounis. (2005). The net of life: Reconstructing the microbial phylogenetic network. *Genome Research*, *15*, 954-59.
[146] Javier Costas. (2002). Characterization of the Intragenomic Spread of the Human Endogenous Retrovirus Family HERV-W. *Molecular Biology and Evolution*, *19*, 526-33.
[147] Z. Yang, N. Goldman, & A. Friday. (1994). Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution*, *11*, 316-24.
[148] Nicolas Galtier, Olivier Gascuel, & Alain Jean-Marie. (2005). Markov Models in Molecular Evolution. In Rasmus Nielsen (Ed.), *Statistical Methods in Molecular Evolution*. (pp. 3-24). Ithaca, NY: Springer.
[149] Dan Graur, & Li Wen-Hsiung. (2000). *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates.
[150] Sewall Wright. (1968). *Evolution and the Genetics of Populations*. Chicago, IL: University of Chicago Press.
[151] Michael M. Desai, & Daniel S. Fisher. (2007). Beneficial Mutation–Selection Balance and the Effect of Linkage on Positive Selection. *Genetics*, *176*, 1759-98.
[152] Niko Beerenwinkel, Jorg Rahnenfuhrer, Martin Daumer, Daniel Hoffman, Rolf Kaiser, Joachim Selbig, & Thomas Langauer. (2005). Learning Multiple Evolutionary Pathways from Cross-Sectional Data. *Journal of Computational Biology*, *12*, 584-98.
[153] N. L. Komarova, & D. Wodarz. (2005). Drug resistance in cancer: principles of emergence and prevention. *Proceedings of the National Academy of Sciences*, *102*, 9714-19.
[154] Irad Ben-Gal. (2008). Bayesian Networks. In *Encyclopedia of Statistics in Quality and Reliability*. Wiley.
[155] J. F. C. Kingman. (1982). The coalescent. *Stochastic Processes and their Applications*, *13(3)*, 235-48.
[156] Magnus Nordberg. (2000). Coalescent Theory. *University of Southern California*, 1-37.
[157] J Felsenstein, & G A Churchill. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution Molecular Biology and Evolution*, *13(1)*, 93-104.
[158] Ari Loytynoja, & Nick Goldman. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *PNAS*, *102(30)*, 10557-62.
[159] KM Chao, WR Pearson, & W Miller. (1992). Aligning two sequences within a specified diagonal band. *Computer applications in the biosciences: CABIOS*, *8(5)*, 481.
[160] ZZ Chen, Y Gao, G Lin, R Niewiadomski, Y Wang, & J Wu. (2004). A space-efficient algorithm for sequence alignment with inversions and reversals. *Theoretical Computer Science*, *325(3)*, 361-72.
[161] Layeb Abdesslem, Meshoul Soham, & Batouche Mohamed (2006). *Multiple Sequence Alignment by Quantum Genetic Algorithm*. Paper presented at the 20th International Parallel and Distributed Processing Symposium (IPDPS 2006).
[162] Kevin Liu, Sindhu Raghavan, Serita Nelesen, C. Randal Linder, & Tandy Warnow. (2009). Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees. *Science*, *324*, 1561-64.
[163] Ari Loytynoja, & Nick Goldman. (2009). Uniting Alignments and Trees. *Science*, *324*, 1528-29.

[164] Michael Schoniger, & Michael Waterman. (1992). A local algorithm for DNA sequence alignment with inversions. *Bulletin of Mathematical Biology*, *54(4)*, 521-36.
[165] Thorne JL, Kishino H, & Felsenstein J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *Journal of Molecular Evolution*, *33(2)*, 114-24.
[166] A Vellozo, C Alves, & A Lago. (2006). Alignment with non-overlapping inversions in O (n 3)-time. *Algorithms in Bioinformatics*, 186-96.
[167] MS Waterman, & M Eggert. (1987). A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *Journal of Molecular Biology*, *197(4)*, 723-28.
[168] Dan Graur, & Li Wen-Hsiung. (2000). Evolutionary Change in Nucleotide Sequences. In *Fundamentals of Molecular Evolution*. Sunderland, MA: Sinauer Associates.
[169] B. Fox. (1973). Calculating Kth shortest paths. *Canada J. Oper. Res. Inform. Process.*, *11*, 66-70.
[170] R. Bellman, & R. Kalaba. (1960). On Kth best policies. *J. SIAM*, *8*, 582-88.
[171] M.S. Waterman, & T.H. Byers. (1985). A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. *Mathematical Biosciences*, *77(1-2)*, 179-88.
[172] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, & Peter Walter. (2002). Proteins. In *Molecular Biology of the Cell*. (pp. 129-90). New York, NY: Garland Science.
[173] R. Ruppert, E. Hoffmann, & W. Sebald. (1996). Human bone morphogenetic protein 2 contains a heparin-binding site which modifies its biological activity. *European Journal of Biochemistry*, *237(1)*, 295-302.
[174] Jutta Buschbom, & Arndt von Haeseler. (2005). Introduction to Applications of the Likeihood Function in Molecular Evolution. In Rasmus Nielsen (Ed.), *Statistical Methods in Molecular Evolution*. (pp. 25-44). Ithaca, NY: Springer.
[175] Richard G. Lomax. (2007). Introduction to Hypothesis Testing: Inferences about a Single Mean. In *An Introduction to Statistical Concepts*. (pp. 92-118). New York: Psychology Press.
[176] Richard G. Lomax. (2007). One-Factor Analysis of Variance -- Fixed-Effects Model. In *An Introduction to Statistical Concepts*. (pp. 196-221). New York: Psychology Press.
[177] John H. Gillespie. (1997). *Population Genetics: A Concise Guide*. Baltimore, MD: The John Hopkins University Press.
[178] John H. Gillespie. (1997). Genetic Drift. In *Population Genetics: A Concise Guide*. (pp. 21-58). Baltimore, MD: The John Hopkins University Press.
[179] Laurent Duret, & Dominique Mouchiroud. (2000). Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Molecular Biology and Evolution*, *17(1)*, 68-74.
[180] A.A. Berryman. (1992). The orgins and evolution of predator-prey theory. *Ecology*, *73(5)*, 1530-35.
[181] T.R. Malthus. (1798). An Essay on the Principle of Population. *Godwin, M. Condorcet, and other writers*, 3-143.
[182] P. F. Verhulst. (1838). Notice sur la loi que la population suite dans son accroissement. *Correspondence Mathematique et Physique*, *10*, 113-21.
[183] L.L. Doust. (1981). Population dynamics and local specialization in a clonal perennial (Ranunculus repens): I. The dynamics of ramets in contrasting habitats. *Journal of Ecology*, *69(3)*, 743-55.

[184] M. Mangel, & C. Tier. (1994). Four facts every conservation biologists should know about persistence. *Ecology*, *75(3)*, 607-14.
[185] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, & Peter Walter. (2002). The Cell Cycle and Programmed Cell Death. In *Molecular Biology of the Cell*. (pp. 983-1126). New York, NY: Garland Science.
[186] John H. Gillespie. (1997). The Evolutionary Advantage of Sex. In *Population Genetics: A Concise Guide*. (pp. 169-84). Baltimore, MD: The John Hopkins University Press.
[187] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, & Peter Walter. (2002). Germ Cells and Fertilization. In *Molecular Biology of the Cell*. (pp. 1127-56). New York, NY: Garland Science.
[188] Mark Kirkpatrick, & Cheryl D. Jenkins. (1989). Genetic segregation and the maintenance of sexual reproduction. *Nature*, *339(6222)*, 300-01.
[189] John H. Gillespie. (1997). Genetic Variation. In *Population Genetics: A Concise Guide*. (pp. 1-20). Baltimore, MD: The John Hopkins University Press.
[190] John H. Gillespie. (1997). Nonrandom Mating. In *Population Genetics: A Concise Guide*. (pp. 119-38). Baltimore, MD: The John Hopkins University Press.
[191] Douglas J. Futuyma. (1998). Evolving Lineages in the Fossil Record. In *Evolutionary Biology*. (pp. 127-64). Sunderland, MA: Sinauer Associates.
[192] Motoo Kimura. (1994). "Stepping Stone" Model of Population. In *Population Genetics, Molecular Evolution, and the Neutral Theory*. (pp. 133-34). Chicago: University of Chicago Press.
[193] S.P. Meyn, & R.L. Tweedie. (1993). *Markov Chains and Stochastic Stability*. London: Springer-Verlag.
[194] Louisiana Veterinary Medical Association. (2007). Biology of the Mouse. Retrieved March 26, 2011, from http://www.lvma.org/mouse.html
[195] International Science and Technology Center. (2009). Mus musculus Linnaeus - House Mouse. Retrieved March 26, 2011, from http://www.agroatlas.ru/en/content/pests/Mus_musculus/
[196] Tian-Xiang Yue, Yong-Zhong Tian, Ji-Yuan Liu, & Ze-Meng Fan. (2008). Surface modeling of human carrying capacity of terrestrial ecosystems in China. *Ecological Modelling*, *214(2-4)*, 168-80.
[197] Douglas J. Futuyma. (1998). The Theory of Natural Selection. In *Evolutionary Biology*. (pp. 365-96). Sunderland, MA: Sinauer Associates.
[198] JW Thatcher, J.M. Shaw, & WJ Dickinson. (1998). Marginal fitness contributions of nonessential genes in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, *95(1)*, 253.
[199] G.D. Sutherland, A.S. Harestad, K. Price, & K.P. Lertzman. (2000). Scaling of natal dispersal distances in terrestrial birds and mammals. *Conservation Ecology*, *4(1)*, 16.
[200] -. java.math Class BigDecimal. Retrieved February 1 2010, from http://java.sun.com/j2se/1.5.0/docs/api/java/math/BigDecimal.html
[201] Northcutt Glenn, R., & Jon H. Kaas. (1995). The emergence and evolution of mammalian neocortex. *Trends in Neurosciences*, *18(9)*, 373-79.
[202] T.J. Treangen, & E.P.C. Rocha. (2011). Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLoS Genetics*, *7(1)*,
[203] J. Merila, & B.C. Sheldon. (1999). Genetic architecture of fitness and nonfitness traits: empirical patterns and development of ideas. *Heredity*, *83(2)*, 103-09.

[204] K. Grammer, B. Fink, & N. Neave. (2005). Human pheromones and sexual attraction. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, *118(2)*, 135-42.
[205] John H. Lawton, & Robert McCredie May. (1995). *Extinction Rates*. Oxford University Press.
[206] D.I. Andersson, & B.R. Levin. (1999). The biological cost of antibiotic resistance. *Current opinion in microbiology*, *2(5)*, 489-93.
[207] -. Dropbox. Retrieved December 30, 2009, from http://www.dropbox.com/
[208] J. D. Thompson, D. G. Higgins, & T. J. Gibson. (1994). CLUSTAL W. *Nucleic Acids Research*, *22*, 4673-80.
[209] C. Notredame, D. G. Higgins, & J. Heringa. (2000). T-Coffee. *Journal of Molecular Biology*, *302*, 205-17.
[210] R. N. Bracewell. (2000). *The Fourier Transform and Its Applications*. Boston: McGraw-Hill.
[211] A. Stamatakis, T. Ludwig, & H. Meier. (2005). RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, *21(4)*, 456-63.
[212] Alexandros Stamatakis. (2006). RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. *Bioinformatics*, *22(21)*, 2688-90.
[213] -. SATe Software, University of Kansas. Retrieved April 8, 2010, from http://phylo.bio.ku.edu/software/sate/sate.html
[214] Daniel H Huson, Daniel C Richter, Christian Rausch, Tobias Dezulian, Markus Franz, & Regula Rupp. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, *8(1)*, 460.