

Molecular Clock based timing of the origin of species: The Human-Chimpanzee divergence

An Internship Report Presented in Partial Fulfillment
of the Requirements for the Degree
of Master of Science

Submitted by

Vinod Swarna

Computational Biosciences,
Arizona State University,
Tempe, AZ 85287-1604 USA

Computational Biosciences

Email: vswarna@asu.edu

Advisor: Dr. Sudhir Kumar

Director, Center for Evolutionary functional Genomics,
The Biodesign Institute.

NOT CONFIDENTIAL

Report No: 05-15

ARIZONA STATE UNIVERSITY

August 2005

Molecular Clock based timing of the origin of species: The Human-Chimpanzee divergence

APPROVED By:
Supervisory Committee:

Dr. Sudhir Kumar,
Associate Professor, School of Life Sciences (SoLS)
Director, Center for Evolutionary functional Genomics,
The Biodesign Institute.

Dr. Jeffrey Touchman, *Associate Director*
Investigator, Translational Genomics Research Institute (TGen) Assistant Professor, School of
Life Sciences (SoLS)

Dr. Michael S. Rosenberg,
Assistant Professor,
School of Life Sciences (SoLS)

ACCEPTED:

Department Chair:

Rosemary Renaut,
Professor, Department of Mathematics and Statistics
Affiliated Professor, Department of Computer Science and Engineering
Director, Computational Biosciences Program
Arizona State University

TABLE OF CONTENTS

I. ACKNOWLEDGEMENTS	5
1. ABSTRACT	6
2. LITERATURE REVIEW	7
(2 A). MOLECULAR CLOCKS	7
(2 B). HUMAN-CHIMPANZEE DIVERGENCE	8
3. INTRODUCTION	9
(3 A). MOLECULAR CLOCK	9
(3 B). BAYESIAN METHODS	11
(3 C). MULTIDIVTIME SOFTWARE PACKAGE	12
4. MATERIALS/DATA AND METHODS	13
(4 A). SEQUENCE COLLECTION	13
(4 B). BAYESIAN TIMING METHOD	17
(4 C). DIVERGENCE TIME ESTIMATION	17
(4 D). MBR CONFIDENCE INTERVAL THROUGH MULTIDIVTIME	17
5. RESULTS	18
6. DISCUSSION	20
7. CONCLUSION	21
8. LITERATURE CITED	21
9. APPENDICES	24
MULTIDIVTIME	24
INPUT TO THE ESTBRANCHES THROUGH HMMCTRL.DAT FILE	24
INPUT TO THE MULTIDIVTIME THROUGH MULTICTRL.DAT FILE	24
PERL PROGRAM	25

TABLE OF FIGURES and TABLES

<i>Figure 1. Phylogenetic relationships of four species.</i>	14
<i>Figure 2. Histograms showing the distribution of the numbers of amino acids [open bars] and fourfold-degenerate non-CpG sites [closed bars]</i>	15
<i>Figure 3. The average evolutionary rates at amino acid (open bars) and 3rd codon positions (solid bars) for 167 protein-coding genes analyzed in this study</i>	16
<i>Figure 4. The percent of simulation replicates in which the 95% confidence interval generated by the MBR approach includes the simulated true value using the Bayesian approaches</i>	19
<i>Table 1. Comparison of divergence estimates for Human and Chimpanzee obtained through Bayesian and ML method.</i>	18

I. Acknowledgements

My sincere thanks to my supervisor, Dr. Sudhir Kumar, Director of Center for Evolutionary Functional Genomics, for helping me in providing this Internship opportunity and for his support, encouragement, and his eminent guidance that always helped me to do my best.

I am very grateful to Graziela Valente for her help in initial tree building and data verification. Special thanks are extended to Alan Filipski and Sankar Subramanian for their advice and guidance.

I would like to thank Dr. Michel Rosenberg and Dr. Jeffery Touchman for being my graduate committee members. I would also like to acknowledge my appreciation for the support and help from Dr. Rosemary Renaut all through the CBS program and for this internship.

I would like to thank Dr. Kumar and Alan Filipski for editing the report and their comments. My internship work was supported by grants from the National Institutes of Health (to S.K.) and by the funds from center for evolutionary functional genomics.

This internship work was aimed at understanding the theoretical and practical limits of the molecular clocks and their application towards estimating the divergence time estimates between humans and chimpanzees. As an Intern at EFG I was involved in this project to understand the functioning of Multidivtime S/W and was also involved in working on several different software tools like MEGA and Paml. Through this internship I learnt how to identify, formulate and solve the research problems besides learning the nuances in the field of molecular evolution.

1. Abstract

Due to the long-standing controversy about the hominoid fossil records and the times of divergence, considerable attention has been focused on molecular clocks during the last three decades. In recent years a large number of authors have investigated the divergence of the lineage that eventually led to modern humans (*Homo sapiens*) from the lineage of our nearest living cousin, the chimpanzees (*Pan troglodytes*), using molecular data (Easteal et al 1997, Kumar et al 1998, Hedges et al 2004, Pickford et al 2005, Thorne et al 2002, Hedges 2003, Chen et al (2001), Stauffer, R. L. et al 2002). A literature study reveals an estimate of divergence time ranging from 3.6 Myr's to 13 Myr's making this issue more controversial. In order to find the real confidence interval and the estimate, we did a thorough study of the divergence times between these species taking all the possible genes available. With a range of 23.8 to 33 Myr's for the Ape-Old World Monkey divergence, we get an estimate ranging from 4.7 – 6.6 Myr's for the Human-Chimpanzee divergence.

This Internship report describes the use of Molecular Clocks to estimate the divergence time between Human and Chimpanzee. The report begins by providing an overview of Molecular Clocks and a review of the basic concepts in determining the divergence times. It then provides a summary of the data acquisition and implementation of Bayesian Methods for this project in estimating the divergence times.

For this project an intern at EFG, I was involved in Data collection through Perl programming and running some phylogenetic software like Mega, Paml, Multidivtime and Estbranches programs.

2. Literature Review

(2 A). Molecular Clocks

The notion of a "molecular clock" was first attributed to Emile Zuckerkandl and Linus Pauling who, in 1962, noticed that the quantity of amino acid differences in hemoglobin between lineages roughly matched the estimated geological time since these species had a common ancestor. They generalized this observation to assert that the rate of evolutionary change of any specified protein was approximately constant over time and over different lineages. It has been applied to DNA sequence evolution also, particularly neutral evolution.

Later Motoo Kimura (1968) observed and formalized that rare spontaneous errors in DNA replication cause the mutations that drive molecular evolution, and that the accumulation of evolutionarily "neutral" differences between two sequences could be used to measure time. One method of calibrating the time was to use as references pairs of groups of living species whose date of speciation was already known from the fossil record. Zuckerkandl and Pauling (1962) wrote a landmark paper in which they demonstrated that indeed in mammalian globin genes substitutions accumulated as a linear function of time. Zuckerkandl and Pauling termed their discovery the molecular clock. Later Doolittle and Bombaek Supported the molecular clock hypothesis by putting forward the correlated times of species divergence with the protein sequence difference. Margoliash in 1963, gave the first direct statement of molecular clock hypothesis. Assuming molecular clock Sarich and Wilson (1967) showed first time that human and chimp diverged from each other only 5 million years ago, i.e. in the Pliocene era.

Later in 1968 Kimura constructed a solid population–genetic basis for the molecular clock, under the assumption that most mutations are neutral or nearly neutral. Neutral mutations appear at a rate proportional to the product of population size N and neutral mutation rate μ , and go to fixation with probability $1/N$. Thus, the rate of neutral evolution is given simply by $N\mu/N = \mu$. The neutral mutation rate μ is itself a product of the organism's overall mutation rate and the fraction of mutations that are neutral. The latter quantity can vary from gene to gene or from species to species. Therefore, according to Kimura's theory, the speed of the molecular clock can vary across species.

(2 B). Human-Chimpanzee Divergence

Sarich and Wilson (1967) assuming molecular clock showed for the first time that human and chimpanzee diverged from each other only 5 million years ago, i.e. in the Pliocene era. Later Chen, F. C (1992) obtain an estimate of 4.6 to 6.2 million years for the Homo-Pan divergence Taking the orangutan speciation date as 12 to 16 million years. Using mitochondrial DNA sequences Horai et al.'s (1992) study revealed that the divergence between human and chimpanzee occurred 4.7 ± 0.5 million years ago. Later in 1996 one other mitochondrial DNA analysis by Arnason U et al reveals an estimate of 13 Myr's for the Human Chimpanzee split. They examined and dated primate divergences by applying the evolutionary separation between artiodactyls and cetaceans anchored at 60 million years before present (MYBP). Easteal, S. & Herbert, G. (1997) showed that the DNA distances between a range of mammalian taxa shows an inconsistent molecular clock with many assumed divergence times irrespective of the assumed substitution rate. They imply a divergence time of humans and chimpanzees of 3.6 - 4.0 million years ago. A Vertebrate timescale was produced with a large scale protein

sequence analysis by Kumar et al in 1998. The calibration was performed based upon sound fossil evidence that mammals and birds diverged 310 million years ago. The result is a evolutionary time line of all the mammals, where the humans have diverged from Chimpanzees 5.5 million years ago. Using the largest number of genes (36 nuclear genes) and divergence of Old World monkeys and hominoids at the Oligocene-Miocene boundary (approximately 23 million years ago), Stauffer RL (2001) provides the best primate calibration point and yields a time and 95% confidence interval of 5.4 +/- 1.1 million years ago for the human-chimpanzee divergence.

3. Introduction

(3 A). Molecular Clock

The hypothesis of the molecular evolutionary clock asserts that informational macromolecules (i.e., proteins and nucleic acids) evolve at rates that are constant through time and for different lineages (Zuckerandl, E. & Pauling 1962). The clock hypothesis has been extremely powerful for determining evolutionary events of the remote past for which the fossil and other evidence is lacking or insufficient.

Motoo Kimura's abiding legacy "The neutral theory of molecular evolution" (Kimura, M. & Ohta, T 1971) provides a basis for conceptual organization in much of molecular evolutionary genetic analysis, and the success of the field of molecular evolution owes much to Kimura's insight in formulating the theory. Although the empirical investigation of molecular evolution (King, J. L. & Jukes, T 1969) started before the neutral theory was developed, it was not until after the theory was proposed that results, particularly of Molecular evolutionary analysis, could be clearly interpreted

Estimating the divergence times from molecular sequence comparisons depends on the existence of a molecular clock, which is one of the most important predictions of the neutral theory. The prediction is based on Kimura's (1968) demonstration that for neutral mutations the rate of substitution between species is equal to the mutation rate per gamete. It only holds, if the mutation rate in different species or lineages remains the same. If the neutral mutation rate varies then the substitution rate will also vary.

The basic approach for estimating molecular dates is to take a measure of the genetic distance between species, then use a calibration rate (the number of genetic changes expected per unit time) to convert the genetic distance to time. There are many available methods, ranging from a simple division of genetic distance by a calibration rate to more sophisticated Maximum likelihood or Bayesian approaches, which estimate molecular dates along with other parameters of models of the DNA substitution process. The reliability of all molecular clock methods depends on the accuracy with which genetic distance is estimated, and on the appropriateness of the calibration Date.

we can estimate the number of amino acid replacements between the two sequences as:

$$D = -\ln(1 - p/L)$$

where p is the number of amino acid differences between the aligned sequences and L is the length of the ungapped alignment. The rate of replacement is:

$$r = d/2T$$

where T ; the time of divergence between the two sequences, is usually inferred from pale ontological data. Under the assumption that all lineages

in a study evolve at the same rate, and assuming that we know the divergence time between two taxa ($T_{\text{cal}} = \text{calibration time}$), we can use the number of amino acid replacements between two sequences from these two taxa (d_{cal}) to calculate a universal rate as:

$$r_{\text{cons}} = d_{\text{cal}} / 2T_{\text{cal}}$$

we can, then, take any pair of sequences from any two taxa, estimate d and calculate the time of divergence as:

$$T = d / 2r_{\text{const}}$$

(3 B). Bayesian Methods

Molecular clocks can be implemented through a number of statistical approaches including maximum likelihood techniques and Bayesian modeling. Bayesian inference is a statistical inference in which probabilities are interpreted not as frequencies or proportions, but rather as degrees of belief.

It uses an estimate of the degree of belief in a hypothesis before the advent of some evidence to give a numerical value to the degree of belief in the hypothesis after the advent of the evidence. Bayes theorem also provides a method for adjusting degrees of belief in the light of new information. Bayes theorem is

$$P(H_0|E) = \frac{P(E|H_0) P(H_0)}{P(E)}.$$

For our purposes, H_0 can be taken to be a hypothesis which may have been developed ab initio or induced from some preceding set of observations, but before the new observation or evidence E .

- The term $P(H_0)$ is called the prior probability of H_0 .

- The term $P(E | H_0)$ is the conditional probability of seeing the observation E given that the hypothesis H_0 is true; as a function of H_0 given E , it is called the likelihood function.
- The term $P(E)$ is called the marginal probability of E ; it is a normalizing constant and can be calculated as the sum of all mutually exclusive hypotheses $\sum P(E|H_i)P(H_i)$.
- The term $P(H_0 | E)$ is called the posterior probability of H_0 given E .

In a Bayes method, a stochastic model of evolutionary rate change is used to specify the prior distribution of rates, and the Bayes theorem is used to derive the posterior distributions of rates and times.

(3 C). Multidivtime Software Package

The first program in the package (Thorne et al. 1998), Estbranches, produces ML estimates of branch lengths for the in-group rooted tree and their approximate variance-covariance matrix. For this purpose, a substitution model should be specified and information about parameters in the substitution model should be provided. We used the Baseml program in the PAML package to obtain estimates of the transition/transversion rate ratio (Kishino and Hasegawa 1990) and the rates for site classes under the discrete-gamma model of rates among sites under the F84G model. Those estimates were used as input to the Estbranches program. The Bayes program also requires an outgroup clade to locate the root in the ingroup tree, for which we used Mouse. The second program in Thorne's packages Multidivtime conducts the Bayes analysis to approximate the posterior distributions of substitution rates and divergence times.

The Multidivtime program requires specification of two gamma prior distributions, for the age of the root and for the rate at the root, which is also the overall prior rate on the phylogeny.

Input Data to Estbranches through Hmncntrl file:

1. Model file name.
2. Tree file name.

Input Data to Multidivtime through Multicntrl file:

- i. Rttm that is a prior expected number of time units between tip and root (in-group depth) of the given tree.
- ii. Rttmsd is a standard deviation of prior for time between tip and root.
- iii. Rtrate is a mean of prior distribution for rate at root node (i.e. calculate the branch length for every tip to root then take the average and divide it by rttm value).
- iv. Rratesd is a standard deviation of prior for rate at root node.
- v. Zero value of "brownmean" and "brownsd" means divergence time estimated assuming molecular clock, other than zero value shows non-molecular clock exist.
- vi. Bigtime – a number higher than time units between tip and root could be in our wildest imagination.

4. Materials/Data and Methods

(4 A). sequence collection

In an effort to estimate the divergence time and the confidence interval between Humans and Chimpanzees we have analyzed 167 nuclear protein-coding genes. Our dataset consisted of all available orthologous nuclear

genes (cDNA and protein sequences) for four species: human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), macaque (*Macaca mulatta*) and mouse (*Mus musculus*) with the phylogenetic relationships shown below.

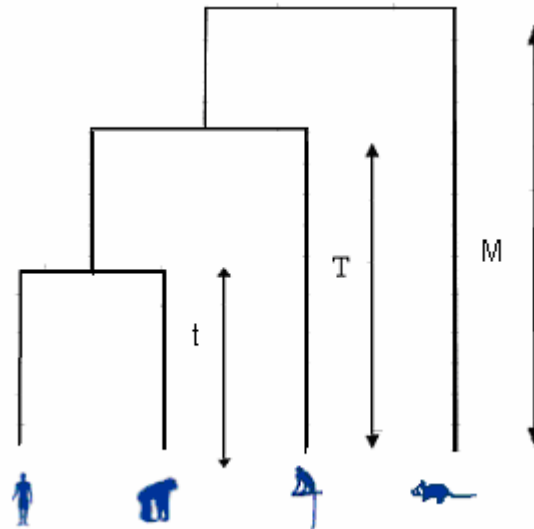


Figure 1. Phylogenetic relationships of four species.

We did not collect any mitochondrial sequences, as mitochondrial rates have been shown to vary among lineages more than the rates of nuclear genes (Glazko & Nei, Gissi et al, 2000). Initially all the available nuclear genes for Macaca were identified and extracted from the Protein sequence database of www.eolproject.org. As the Data on the www.eolproject.org is not maintained on a FTP server the sequence data had to be extracted from the contents of the website with a CGI perl script. There were a total of 1050 Macaca Protein sequences in the database from which 663 distinct sequences were retained after removing multiple sequences of the same gene. Using Macaca as reference we collected all the homologous protein sequences with an E-value greater than 10^{-10} by performing a blast search on GenBank (www.ncbi.nih.gov). Protein Homologues for chimpanzee (*Pan troglodytes*) were obtained by performing a local blast search on the

chimpanzee protein sequences, collected from <http://www.ensembl.org/Download/>. The Protein sequence alignments were carried out with CLUSTAL-W using the default settings (Thompson et al 2000). we took a stringent approach in finding the orthologous by constructing the phylogenetic trees for each thus formed protein pairs by neighbor-joining method using MEGA3 (Kumar S et al., 2004). Short (<40 aminoacid) sequences and sequences with no putative orthologous of Human, Chimpanzee and Mouse were omitted. Protein Sequences that do not have their corresponding coding sequences were purged as were putative genes that failed to support the accepted phylogeny.

In all, 167 nuclear protein sequences met these criteria and could be used to estimate the divergence times. Histogram containing the distributions of amino acid sequence length is given below.

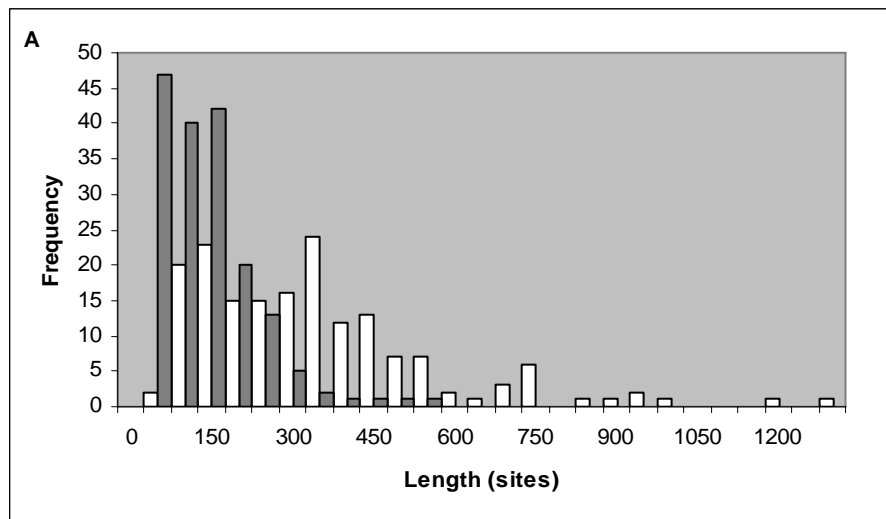


Figure 2. Histograms showing the distribution of the numbers of amino acids [open bars] and fourfold-degenerate non-CpG sites [closed bars]

The corresponding coding DNA sequences for these orthologous Protein sequences of *H.sapiens*, *M.musculus* and *M.mulatta* were collected from <http://www.ncbi.nlm.nih.gov/> with a Perl program. The Chimpanzee

coding DNA sequences were obtained from <http://www.ensembl.org/> through a Batch sequence retrieval system ENSMART. The thus obtained coding DNA sequences were aligned taking the amino acid sequences as guides (for codon boundaries).

We then identified the 3rd codon positions and the fourfold-degenerate sites. The figure below shows the evolutionary rates at 3rd codon positions.

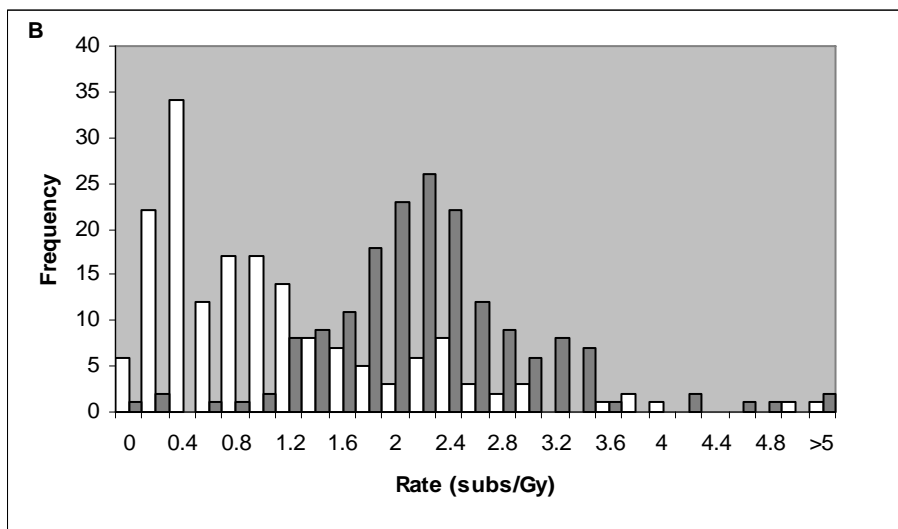


Figure 3. The average evolutionary rates at amino acid (open bars) and 3rd codon positions (solid bars) for 167 protein-coding genes analyzed in this study

The fourfold-degenerate sites were separated by selecting only those sites that have remained fourfold-degenerate throughout the four species under study. As the CpG dinucleotides mutate 7-10 times (Subramanian, S. & Kumar, S 2000) faster than other dinucleotides, the fourfold degenerate sites were separated into those that were involved with CpG dinucleotides and those that were not. We extracted the CpG sites only if there is a Cytosine followed by Guanine or Guanine preceded by Cytosine in the 3rd positions of the codons in the fourfold degenerate sites of any one of the

four species, else the 3rd position of the codon was considered as a non CpG site.

(4 B). Bayesian Timing method

All the 167 genes were concatenated into a supergene and the branch lengths were estimated using the Estbranches program using the topology given in the figure 1. We used the most sophisticated models allowed by Multidivtime (Thorne et al. 1998) software, Felsenstein 84 model for DNA sequence evolution and JTT for amino acid sequences and an allowance for a Γ distribution of rates with five discrete rate categories. The trasversion and transition parameters and estimates of the rate categories of the gamma distribution were calculated with PAML (Yang 2004).

(4 C). Divergence time Estimation

Divergence times were estimated using the program Multidivtime. The multidivtime program requires an input value for the mean of the prior distributions of the root of the ingroup tree. For this purpose we used the boundry of the Oligocene and Miocene, 23.8 Ma as the mean of the prior distributions. The different parameters used in the HMM and Multi control files are given in the appendix.

(4 D). MBR Confidence interval through Multidivtime

To validate the MBR approach proposed by Dr.Kumar, the confidence intervals were calculated through Multidivtime and ML methods using a large number of genes generated through computer simulations. All the data for the computer simulations were generated using SeqGen Software (Rambaut, A. and Grassly, NC 1996) and analyzed using Bayesian procedures in the same way as the empirical data.

5. Results

It is necessary to take account explicitly of the rate heterogeneity (Takezaki et al., 1995; Yoder and Yang, 2000) among lineages in estimations of branching dates. The Bayesian method of Thorne et al. (1998) is useful for this purpose, as shown by Nikaido et al. (2000) and Cao et al. (2000), who applied this method to mammalian proteins.

The table below shows the molecular time estimates (both Bayesian and ML) obtained for the 53,008 bp of 3rd position and 19311bp of the four fold non CpG datasets with the Mouse outgroup and the 23.8 mya prior of HC-OWM for the Root of tree.

	Hsa-Ptr estimate Bayesian	Hsa-Ptr estimate ML
3 rd position	4.98	4.74
4F nonCPG	5.12	4.75

Table 1. Comparison of divergence estimates for Human and Chimpanzee obtained through Bayesian and ML method.

The Bayesian analysis show an 8% higher estimate than the ML estimates but the actual difference becomes smaller when we consider the ratio of the times of HC and HC-OWM divergence events

i.e. $24.00/4.98 = 4.82$ and $24.34/5.12 = 4.75$.

As the results obtained from the 3rd position and the 4fold non-cpg were almost identical and as the size of the dataset for the 3rd position was larger than the four-fold non CpG dataset we consider only 3rd position in all further DNA sequence analysis.

Amino acids gave a 43% larger estimate than the 3rd and 4fold non-CPG sites. Sequencing errors and within species polymorphism can be attributed for this larger estimate. We also calculated the percent of simulation replicates in which the 95% confidence interval generated by the MBR approach includes the simulated true value using the Bayesian approaches. For the 100 sets of 167 gene equivalent simulated datasets we analyzed the three types of rate variation schemes, (equal, random, and correlated rate) for each parameter set independently through Bayesian analysis.

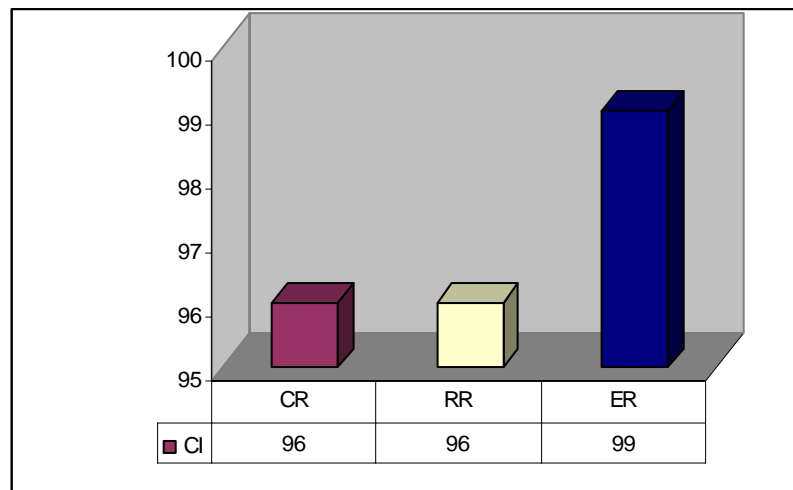


Figure 4. The percent of simulation replicates in which the 95% confidence interval generated by the MBR approach includes the simulated true value using the Bayesian approaches

Our Computer simulation show that the MBR confidence intervals for the Constant Rate, Random Rate and the correlated cases contain the true value with a frequency of 95% in the Bayesian approach so we present the CI generated using the MBR approach.

6. Discussion

A weakness in the Molecular time estimation is the use of different data sets (Mitochondrial, Protein and non coding DNA etc), different methods and different calibration times thus resulting in different time estimates.

As mitochondrial rates have been shown to vary among lineages more than the rates of nuclear genes, we use nuclear data which is supported by Glazko & Nei, Gissi et al, 2000. Most of the other studies produce only point estimates due to the lack of sufficient number of genes. In our study due to the use of a bigger dataset this question has been tackled by a statistical technique called bootstrapping producing the confidence intervals for the point estimates.

Calibration points have a greater influence on the precision of the molecular time estimates. For example, if hominoid or cercopithecoid fossils are found at 30 million years ago, the molecular time estimates would be pushed back. The calibration error takes on greater importance; because fossil calibrations represent minimum time estimates for the divergence of two lineages, the use of poorly constrained calibration points may yield a calibration that is a significant underestimate. The use of 23.8 MYRS for the Ape/OWM split as the calibration point is supported by enough fossil evidences.

Our results are in concordance with many molecular time estimates and recent fossil evidences. Our confidence intervals incorporate the molecular time estimates of Stauffer et al who used the largest number of genes and divergence of Old World monkeys and hominoids at the Oligocene-Miocene boundary (approximately 23 million years ago). In a recent study (Chen and Li 2001) using approximately 24 kb of non-coding sequence, the time

ratio the resulting human-chimpanzee divergence time with those non-coding data (4.9 million years ago) still is within the 95% confidence limit of our estimate. New hominoid fossils named *Orrorin tugenensis* from the approximately 5.4 million-year-old deposits of the Lukeino Formation of Kenya (Pickford and Senut 2001; Senut et al. 2001) are said to be the earliest hominids, and this date is included in our 95% confidence interval for the chimpanzee-human split.

7. Conclusion

The estimation of Molecular time estimates has become a regular step in the analysis of new gene sequences. Bayesian approaches have revolutionized the time estimation in general and it is already clear that these approaches are extending the field by answering previously intractable questions. These new techniques seem poised to teach us a great deal about the tree of life and molecular evolutionary genetics. This study of the divergence time is important for assessing the evolutionary time line and assessing the morphological and molecular changes in humans from the time they diverged from their common ancestors.

8. Literature Cited

1. Kumar S et al 2005 Placing confidence limits on the molecular age of the human-chimpanzee divergence. (Unpublished)
2. Kimura, M. 1968-. Evolutionary rate at the molecular level. *Nature* 217:624-626.
 - . 1968. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet. Res.* 11:247-269.
 - . 1974. Gene pool of higher organisms as a product of evolution. *Cold Spring Harbor Symp. Quant. Biol.* 38:515-524.
 - . 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267:275-276.
 - 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111-120.
 - . 1981. Estimation of evolutionary distances between homologous nucleotide

- sequences.
 Proc. Natl. Acad. Sci. USA 78:454-458. Rare Variant Alleles 93
 -. 1983. The neutral theory of molecular evolution. Cambridge University Press,
 Cambridge. Kimura M (1968). Nature 217: 624–626.
3. Zuckerkandl E, Pauling L (1965). Evolutionary divergence and convergence in proteins.
 4. Kumar, S. (2005) Molecular Clocks: four decades of evolution *Nature Reviews Genetics In press.*
 5. Sarich V. M., A. C. Wilson, 1967 Immunological time scale for hominoid evolution Science 158:1200-1203
 6. Easteal, S., and G. Herbert. 1997. Molecular evidence from the nuclear genome for the time frame of human evolution. J. Mol. Evol. 44(Suppl. 1):S121–S132
 7. Kumar, S., and S. B. Hedges. 1998. A molecular timescale for vertebrate evolution. Nature 392:917 920.
 8. Hedges, S. B., and S. Kumar. 2004. Precision of molecular time estimates. Trends Genet. 20:242–247.
 9. Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51:689-702
 10. Hedges, S.B. and Kumar, S. 2003. Genomic clocks and evolutionary timescales. Trends Genet. 19, 200-6.
 11. Chen, F. C., E. J. Vallender, H. Wang, C. S. Tzeng, and W. H. Li. 2001. Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. J. Hered. 92:481-489.
 12. Stauffer, R. L., A. Walker, O. A. Ryder, M. Lyons-Weiler, and S. B. Hedges. 2001. Human and ape molecular clocks and constraints on paleontological hypotheses. J. Hered. 92:469-474.
 13. Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425
 14. Kumar, S., Tamura, K. & Nei, M. MEGA3: molecular evolutionary genetics analysis (2004) Brief Bioinform 5, 150-63
 15. Kumar, S., and S. Subramanian. 2002. Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. USA 99:803-808
 16. Hasegawa, M., J. L. Thorne, and H. Kishino. 2003. Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. Genes Genet. Syst. 78:267–283
 17. Glazko, G. V. & Nei, M 2003. Estimation of Divergence Times for Major Lineages of Primate Species. Mol. Biol. Evol. 20, 424-34.
 18. Nei, M. & Kumar, S. (2000) Molecular Evolution and Phylogenetics (Oxford University Press, New York).
 19. Thorne, J. L., and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51:689–702.
 20. Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.
 21. Whelan, S., P. Li`O , and N. Goldman. 2001. Molecular phylogenetics: State of the art methods for looking into the past. Trends Genet. 17:262–272.
 22. Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with

23. variable rates over sites: Approximate methods. *J. Mol. Evol.* 39:306–314.
24. Kumar, S. 1996. Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates. *Genetics* 143:537–548.
25. Grassly, N. C., Adachi, J. & Rambaut, A. (1997) *Comput. Appl. Biosci.* 13, 559-60.
26. Rambaut, A. & Grassly, N. C. (1997) *Comput. Appl. Biosci.* 13, 235-8.

9. Appendices

Multidivtime

Input to the Estbranches through Hmmctrl.dat file

```
/* Which Model to use? */
modelinf.f84
L /* How much output? Options: L = Loud mode (prints more output, the default), Q = Quiet mode (prints less output -
use with parametric bootstrap) */
D /* Predict Secondary Structure? Options: P= predict, D = do not predict (the default option) */
N /* Does user tree specify names (N) or specify order (O) of sequences in sequence data file? */
  /* The topology is in the file listed below*/
gene.tree
  /* End of hmmcntrl.dat */
```

Input to the multidivtime through multictrl.dat file

```
/* the following lines are all needed in multicntrl.dat ...
do not add or delete lines but change entry on left of each line as you see fit ... */
gene.tree
1 ... number of genes ... FOLLOWING LINES CONTAIN ONLY NAMES OF DATA FILES
oest.gene1
10000 ... numsamps: How many times should the Markov chain be sampled?
100 ... sampfreq: How many cycles between samples of the Markov chain?
100000 ... burnin: How many cycles before the first sample of Markov chain?
23.8 ... rttm: a priori expected number of time units between tip and root
23.8 ... rttmsd: standard deviation of prior for time between tip and root
**** ... rtrate: mean of prior distribution for rate at root node
**** ... rtratesd: standard deviation of prior for rate at root node
0.04 ... brownmean: mean of prior for brownian motion constant "nu"
```

```

0.04 ... brownsd: std. deviation of prior for brownian motion constant "nu"
/* the following lines are all needed (i.e., do not delete them) but you may
   not want to alter entries unless you are familiar with the computer code */
1.0 ... minab: parameter for beta prior on proportional node depth
0.1 ... newk: parameter in Markov chain proposal step
0.5 ... othk: parameter in Markov chain proposal step
0.5 ... thek: parameter in Markov chain proposal step
110 ... bigtime: number higher than time units between tip and root could be in your wildest imagination
/* the program will expect the entry below to be the number of constraints and then the specified number of constraints
   should follow on subsequent lines */
0 ... number of constraints on node times
0 ... number of tips which are not collected at time 0
0 ... nodata: 1 means approximate prior, 0 means approximate posterior
0 ...commonbrown: 1 if all genes have same tendency to change rate, 0 otherwise

```

**** -> computed from the program

Perl program

```

#!/usr/bin/perl
#use warnings;
use Switch;
print "\n\nMake a choice\n\n 1. Single Gene Analysis\n 2. Multigene analysis\n\n";
$single_multi = <STDIN>;
chomp($single_multi);
switch ($single_multi) {
case 1{
    print "\n\nIs the sequence file\n 1. DNA file or\n 2. AminoAcid file\n(File should be in Phylip format with an empty
line at the end of the file)\n\n";
    $AA_DNA = <STDIN>; #DNA or AA
    chomp($AA_DNA);
    print "\n\nGive the Name of Sequence file\n(File should be in the same folder of the program being executed)\n\n";
    $seqfile = <STDIN>;

```

```

chomp($seqfile);
print "\nGive the Name of the tree file\n(File should be in the same folder of the program being executed and should be
in Phylip format with #Taxa)\n\n";
$Tree = <STDIN>;
chomp($Tree);
open(IQ,"$Tree") or die "can't open filename:$Tree";
while(<IQ>)
{
$lane = $_;
@arr = split(" ",$lane);
last;
}
$firstConstraint = $arr[0]+3;
print "\nEnter the RTTM Value (rttm is the mean of the prior distribution for the time separating the ingroup root
from the present)\n\n";
$RTTM = <STDIN>;
chomp($RTTM);
print "\nEnter the Big Time Value (bigtime is a number that is absolutely positively way bigger than the age of any
node in the data set)\n\n";
$BIGTIME = <STDIN>;
chomp($BIGTIME);
print "\nPlease wait while multidivtime gives the node numbers for the given tree structure For imposing the
calibration points.....\n\n";

#file path
switch ($AA_DNA) { #create a model file
case 1{
open(OUT,">baseml.ctl")or die "can't open filename:baseml.ctl"; #Create Ctl File for Baseml
print OUT "seqfile = $seqfile\n";
print OUT "outfile = Gene.out\n";
print OUT "treefile = $Tree\n";
print OUT "noisy = 2\n";
print OUT "verbose = 1\n";

```

```

print OUT "runmode = 0\n";
  print OUT "model = 3\n";
  print OUT "Mgene = 0\n";
print OUT "fix_kappa = 0\n";
  print OUT "kappa = 2\n";
print OUT "fix_alpha = 0\n";
  print OUT "alpha = 0.01\n";
  print OUT "Malpha = 0\n";
  print OUT "ncatG = 5\n";
print OUT "fix_rho = 1\n";
  print OUT "rho = 0. \n";
  print OUT "nparK = 0 \n";
  print OUT "clock = 0 \n";
  print OUT "nhomo = 1 \n";
  print OUT "getSE = 1 \n";
print OUT "RateAncestor = 0 \n";
  print OUT "cleandata = 0 \n";
close(OUT);
system "./baseml > basemlconsoleoutput";
system "./paml2modelinf Gene.out modelinf.f84";
open(HMM,">hmmcntrl.dat")or die "can't open filename:hmmcntrl.dat";
print HMM " /* Which Model to use? *\n";
print HMM "modelinf.f84\n";
print HMM "L /* How much output? Options: L = Loud mode (prints more output, the\n";
print HMM "   default), Q = Quiet mode (prints less output - use with parametric\n";
print HMM "   bootstrap) *\n";
print HMM "D /* Predict Secondary Structure? Options: P= predict, D = do not predict\n";
print HMM "   (the default option) *\n";
print HMM "N /* Does user tree specify names (N) or specify order (O) of sequences\n";
print HMM "   in sequence data file? *\n";
print HMM " /* The topology is in the file listed below*\n";
print HMM "$Tree\n";
print HMM " /* End of hmmcntrl.dat *\n";

```

```

close(HMM);
$#arr= -1;
open(AT, "$seqfile")or die "can't open filename:$seqfile";
open(OUT, ">testseq")or die "can't open filename:testseq";
  @arr = <AT>;
print OUT "@arr";
system ("./estbranchesdna oest.gene1 > estbranchesconsoleoutput");
close(AT);
close(OUT);
}

case 2{
open(OUT,">codemlctl")or die "can't open filename:codemlctl"; #Create Ctl File for codeml
print OUT "seqfile = $seqfile\n";
print OUT "treefile = $Tree\n";
print OUT "outfile = Gene.out\n";
print OUT "noisy = 9\n";
print OUT "verbose = 1\n";
print OUT "seqtype = 2\n";
print OUT "aaRatefile = dayhoff.dat\n";
print OUT "model = 0\n";
print OUT "fix_alpha = 0\n";
print OUT "alpha = 0.01\n";
print OUT "ncatG = 4\n";
print OUT "clock = 1\n";
print OUT "getSE = 0\n";
print OUT "RateAncestor = 0\n";
print OUT "Small_Diff = 1e-6\n";
close(OUT);
system "./codeml > Codemlconsoleoutput";

open(INGENE,"Gene.out")or die "can't open filename:Gene.out";
while(<INGENE>)

```

```

{
$lineGene = $_;
chomp($lineGene);
if($lineGene =~m /rates:/g)
{
@argene = split (":",$lineGene);
@argene1 = split (" ",$argene[1]);
}
if($lineGene =~m /freqs:/g)
{
@argene11 = split(":",$lineGene);
@argene111 = split(" ",$argene11[1]);
}
}
open(MO,"modelhedges.jtt")or die "can't open filename:modelhedges.jtt";
open(MOUT,">modelinf.jtt")or die "can't open filename:modelinf.jtt";
while(<MO>)
{
$model = $_;
chomp($model);
print MOUT "$model\n";
if($model =~m /Along-Sequence Transition Probabilities follow/g)
{
for($Del = 0; $Del <= $#argene111;$Del++)
{
if($Del eq 0)
{
}
else
{
print MOUT"\n";
}
}
for($Del1 = 0; $Del1 <= $#argene111;$Del1++)

```

```

{
  print MOUT "$argene111[0] ";
}
}
print MOUT"\n";
}
if($model =~m /Structure Frequencies/g)
{
for($Del2 = 0; $Del2 <= $#argene111;$Del2++)
{
  print MOUT "$argene111[0] ";
}
  print MOUT"\n";
}
}
open(HMM,">hmmcntrl.dat")or die "can't open filename:hmmcntrl.dat";
print HMM " /* Which Model to use? *\n";
print HMM "modelinf.jtt\n";
print HMM "L /* How much output? Options: L = Loud mode (prints more output, the\n";
print HMM "  default), Q = Quiet mode (prints less output - use with parametric\n";
print HMM "  bootstrap) *\n";
print HMM "D /* Predict Secondary Structure? Options: P= predict, D = do not predict\n";
print HMM "  (the default option) *\n";
print HMM "N /* Does user tree specify names (N) or specify order (O) of sequences\n";
print HMM "  in sequence data file? *\n";
print HMM " /* The topology is in the file listed below*\n";
print HMM "$Tree\n";
print HMM " /* End of hmmcntrl.dat *\n";
close(HMM);

for($eval = 0;$eval <=3;$eval++)
{
open (EVA,">evalfl.jttc$eval")or die "can't open filename:evalfl.jttc$eval";

```

```

print EVA "G\t".      0.999933957**$argene1[$eval];
print EVA "\nA\t".    0.999609307**$argene1[$eval];
print EVA "\nV\t".    0.999873221**$argene1[$eval];
print EVA "\nL\t".    0.999927603**$argene1[$eval];
print EVA "\nI\t".    0.999819976**$argene1[$eval];
print EVA "\nM\t".    0.999745526**$argene1[$eval];
print EVA "\nF\t".    0.999904554**$argene1[$eval];
print EVA "\nP\t".    0.999865005**$argene1[$eval];
print EVA "\nS\t".    0.999625373**$argene1[$eval];
print EVA "\nT\t".    0.999666541**$argene1[$eval];
print EVA "\nC\t".    0.999885137**$argene1[$eval];
print EVA "\nN\t".    0.999681414**$argene1[$eval];
print EVA "\nQ\t".    0.999792792**$argene1[$eval];
print EVA "\nY\t".    0.999839336**$argene1[$eval];
print EVA "\nW\t".    1**$argene1[$eval];
print EVA "\nD\t".    0.99975707**$argene1[$eval];
print EVA "\nE\t".    0.999778787**$argene1[$eval];
print EVA "\nH\t".    0.999707444**$argene1[$eval];
print EVA "\nK\t".    0.99985355**$argene1[$eval];
print EVA "\nR\t".    0.999727132**$argene1[$eval];
print EVA "\n";
close(EVA);
}

```

```

$#arr= -1;
open(AT, "$seqfile")or die "can't open filename:$seqfile";
open(OUT, ">testseq")or die "can't open filename:testseq";
  @arr = <AT>;
print OUT "@arr";
system("./estbranchesaa oest.gene1 > estbranchesconsoleoutput");
close(AT);
close(OUT);

```

```

}
else {
print "You made a wrong choice\n Make a choice\n 1. Single Gene Analysis 2. Multigene analysis\n" }
}

```

```

$#arr= -1;
open(AT, "oest.gene1")or die "can't open filename:oest.gene1";
open(OUT, ">estout.tree1")or die "can't open filename:estout.tree1";
open(OUT1, ">>outputsingbase.txt")or die "can't open filename:outputsinglebase.txt";
while (<AT>)
{
$line = $_;
chomp($line);
if ($line =~m /^(/g)
{
print OUT "$line\n";
print OUT1 "\n \t $line";
}
}
close(AT);
close(OUT);
close(OUT1);
system("./a.out estout.tree1 > rate");

```

```

open(AT, "rate")||die "Couldn't open file\n";
open(OUT, ">f_rates")or die "can't open filename:f_rates";
$#arr = -1;

```

```

while (<AT>)
{

```

```

$line = $_;
chomp $line;
if ($line =~m /mean/g)
{
  @arr = split("=", $line);
  print OUT "$arr[1]";
}
}
close(AT);
close (OUT);
open(AT, "f_rates")||die "Couldn't open file\n";
open(OUT, ">final_rate")or die "can't open filename:finalrate";

while (<AT>)
{
  $line = $_;
  chomp $line;
  if ($line =~m /^\s *$/g)
  {
    next;
  }

  $R = $line/$RTTM;

}
close (AT);
printf OUT "%.6f\n" ,$R;
close (OUT);

open(OUT, ">multicntrl.dat")or die "can't open filename:multicntrl.dat";
open(FR, "final_rate")or die "can't open filename:final_rate";
while (<FR>)
{

```

```

$var = $_;
chomp($var);
}

print OUT "/* the following lines are all needed in multicntrl.dat ...\n";
print OUT "do not add or delete lines but change entry on left of each\n";
print OUT "line as you see fit ... *\n";
print OUT "$Tree\n";
print OUT "1 ... number of genes ... FOLLOWING LINES CONTAIN ONLY NAMES OF DATA FILES\n";
print OUT "oest.gene1\n";
print OUT "10000 ... numamps: How many times should the Markov chain be sampled?\n";
print OUT "100 ... sampfreq: How many cycles between samples of the Markov chain?\n";
print OUT "100000 ... burnin: How many cycles before the first sample of Markov chain?\n";
print OUT "$RTTM ... rttm: a priori expected number of time units between tip and root\n";
print OUT "$RTTM ... rttmsd: standard deviation of prior for time between tip and root\n";
print OUT "$var ... rtrate: mean of prior distribution for rate at root node\n";
print OUT "$var ... rratesd: standard deviation of prior for rate at root node\n";
print OUT "0.04 ... brownmean: mean of prior for brownian motion constant \"nu\"\n";
print OUT "0.04 ... brownsd: std. deviation of prior for brownian motion constant \"nu\"\n";
print OUT "/* the following lines are all needed (i.e., do not delete them) but you may \n";
print OUT " not want to alter entries unless you are familiar with the computer code *\n";
print OUT "1.0 ... minab: parameter for beta prior on proportional node depth\n";
print OUT "0.1 ... newk: parameter in Markov chain proposal step\n";
print OUT "0.5 ... othk: parameter in Markov chain proposal step\n";
print OUT "0.5 ... thek: parameter in Markov chain proposal step\n";
print OUT "$BIGTIME ... bigtime: number higher than time units between tip and root could\n";
print OUT " be in your wildest imagination\n";
print OUT "/* the program will expect the entry below to be the number of constraints\n";
print OUT " and then the specified number of constraints should follow on\n";
print OUT " subsequent lines *\n";
print OUT "1 ... number of constraints on node times\n";
print OUT "L $firstConstraint 20\n";
print OUT "0 ... number of tips which are not collected at time 0\n";

```

```

print OUT "0 ... nodata: 1 means approximate prior, 0 means approximate posterior\n";
print OUT "0 ...commonbrown: 1 if all genes have same tendency to change rate, 0 otherwise\n";
  close (FR);
  close (OUT);
{
  system (". /multidivtime gene > out.gene");
}
$first = 0;
open(AT, "out.gene")or die "can't open filename:out.gene";
open(OUT, ">>outputsingbase.txt")or die "can't open outputsingbase.txt";
while (<AT>)
{
  $line = $_;
  chomp($line);
  if($first eq 0)
  {
    if ($line =~m /Here are node numbers of other internal nodes on master tree/)
    {
      $first = 1;
      next;
    }
  }
  if($first eq 1)
  {
    print "$line\n\n";
    print "Enter the number of constraints you want to levy\n\n";
    $constraintsNumber = <STDIN>;
    chomp($constraintsNumber);
    for($maka = 1;$maka <= $constraintsNumber;$maka++)
    {
      print "\nEnter the node number\n\n";
      $nodenumber = <STDIN>;
    }
  }
}

```

```

chomp($nodenumber);
push(@arrnodenumber,$nodenumber);
print "\nEnter Whether the constraint is UPPER or LOWER Example: U or L\n\n";
$UL = <STDIN>;
chomp($UL);
push(@arrUL,$UL);
print "\nEnter the Calibration time for the entered node number\n\n";
$constime = <STDIN>;
chomp($consttime);
push(@arrconsttime,$consttime);
}
last;
}
}
close(AT);
close(OUT);
print "\n\nPlease wait while the Multidivtime prepares the results\n\n";
open(OUT, ">multicntrl.dat")or die "can't open filename:multicntrl.dat";
open(FR, "final_rate")or die "can't open filename:final_rate";
while (<FR>)
{
    $var = $_;
    chomp($var);
}
print OUT "/* the following lines are all needed in multicntrl.dat ...\n\n";
print OUT "do not add or delete lines but change entry on left of each\n\n";
print OUT "line as you see fit ... */\n\n";
print OUT "$Tree\n\n";
print OUT "1 ... number of genes ... FOLLOWING LINES CONTAIN ONLY NAMES OF DATA FILES\n\n";
print OUT "oest.gene1\n\n";
print OUT "10000 ... numsamps: How many times should the Markov chain be sampled?\n\n";
print OUT "100 ... sampfreq: How many cycles between samples of the Markov chain?\n\n";
print OUT "100000 ... burnin: How many cycles before the first sample of Markov chain?\n\n";

```

```

print OUT "$RTTM ... rttm: a priori expected number of time units between tip and root\n";
print OUT "$RTTM ... rttmsd: standard deviation of prior for time between tip and root\n";
print OUT "$var ... rtrate: mean of prior distribution for rate at root node\n";
print OUT "$var ... rratesd: standard deviation of prior for rate at root node\n";
print OUT "0.04 ... brownmean: mean of prior for brownian motion constant \"nu\"\n";
print OUT "0.04 ... brownsd: std. deviation of prior for brownian motion constant \"nu\"\n";
print OUT "/* the following lines are all needed (i.e., do not delete them) but you may \n";
  print OUT "  not want to alter entries unless you are familiar with the computer code *\n";
print OUT "1.0 ... minab: parameter for beta prior on proportional node depth\n";
print OUT "0.1 ... newk: parameter in Markov chain proposal step\n";
print OUT "0.5 ... othk: parameter in Markov chain proposal step\n";
print OUT "0.5 ... thek: parameter in Markov chain proposal step\n";
print OUT "$BIGTIME ... bigtime: number higher than time units between tip and root could\n";
print OUT "          be in your wildest imagination\n";
print OUT "/* the program will expect the entry below to be the number of constraints\n";
print OUT "  and then the specified number of constraints should follow on\n";
print OUT "  subsequent lines *\n";
print OUT "$constraintsNumber ... number of constraints on node times\n";
for($loping = 0;$loping <= $constraintsNumber-1;$loping++)
{
print OUT "@arrUL[$loping] $arrnodenumber[$loping] $arrconsttime[$loping]\n";
}
print OUT "0 ... number of tips which are not collected at time 0\n";
print OUT "0 ... nodata: 1 means approximate prior, 0 means approximate posterior\n";
print OUT "0 ...commonbrown: 1 if all genes have same tendency to change rate, 0 otherwise\n";
  close (FR);
  close (OUT);
{
  system ("./multidivtime gene > out.gene");
}
print "output:\n";
print " Input Sequence File: $seqfile\n";
print " Input Tree File   : $Tree\n";

```

```

print " RTTM          : $RTTM\n";
print " BigTime       : $BIGTIME\n";
print " Rrate         : $var\n";
print " Constraints:\n";
for($loping = 0;$loping <= $constraintsNumber-1;$loping++)
{
print "\t\t@arrUL[$loping] $arrnodenumber[$loping] $arrconsttime[$loping]\n";
}

open(AT, "out.gene")or die "can't open filename:out.gene";
open(OUT, ">>output.txt")or die "can't open filename:output.txt";
while (<AT>)
{
$line = $_;
chomp($line);
if ($line =~m /Actual/g && $line == /0\.\00000/)
{
print "\n$line\n";
}
}
close(AT);
close(OUT);
print "\n\nMultidivtime Output file out.txt can also be vewied for the above results\n\n\n";
}#DNA
case 2 {
print "hello 2\n";
for($mu = 1;$mu < 19;$mu++)
{
use warnings;
open(NA,"names.txt")or die "can't open filename:names.txt";
while(<NA>)
{
$ba = $_;

```

```

chomp($ba);
$#arr= -1;
open(AT, "$ba")or die "can't open filename:$ba";
open(OUT, ">testseq")or die "can't open filename:testseq";
  @arr = <AT>;
print OUT "@arr";
system("./estbranchesdna oest.gene1 ");
system("rm testseq");
close(AT);
close(OUT);
$#arr= -1;
open(AT, "oest.gene1")or die "can't open filename:oest.gene1";
open(OUT, ">estout.tree1")or die "can't open filename:estout.tree1";
open(OUT1, ">>output.txt")or die "can't open filename:output.txt";
while (<AT>)
{
  $line = $_;
  chomp($line);
  if ($line =~m /^(/g)
  {
    print OUT "$line\n";
    #print OUT1 "\n \t $line";
  }
}
close(AT);
close(OUT);
close(OUT1);
  system("./a.out estout.tree1 > rate");
open(AT, "rate")||die "Couldn't open file\n";
open(OUT, ">f_rates");
$#arr = -1;
while (<AT>)
{

```

```

$line = $_;
chomp $line;
if ($line =~ m /mean/g)
{
  @arr = split("=", $line);
  print OUT "$arr[1]";
}
}
close(AT);
close (OUT);
$variable = 23.00;
open(AT, "f_rates")||die "Couldn't open file\n";
open(OUT, ">final_rate");
while (<AT>)
{
  $line = $_;
  chomp $line;
  if ($line =~ m /\s *$/g)
  {
    next;
  }
  $R = $line/$variable;
}
close (AT);
printf OUT "%.6f\n" , $R;
close (OUT);
open(FR, "final_rate");
open(AT, "multidivtime$mu");
open(OUT, ">multicntrl.dat");
while (<FR>)
{
  $var = $_;
  chomp($var);
}

```

```

    }
while (<AT>)
{
    $line = $_;
    chomp($line);
    if ($line =~m /trate/g)
    {
        @arr = split (\.\\.\/);
        print OUT "$var ...$arr[1]";
        next;
    }
    print OUT "$line\n";
}
close (FR);
close (AT);
close (OUT);
{
system ("./multidivtime gene > out.gene");
}
$first = 0;
open(AT, "out.gene");
open(OUT, ">>output.txt");
while (<AT>)
{
    $line = $_;
    chomp($line);
    if ($line =~m /Actual/g && $line == /0\.00000/)
    {
        print OUT "$line\n";
    }
}
close(AT);
close(OUT);

```

```
}  
print OUT "\n\n";  
}  
} #second case ends  
else { print "You made a wrong choice\n Make a choice\n 1. Single Gene Analysis 2. Multigene analysis\n" }  
}
```