

Estimating the Divergence Time of Molecular Sequences using Bayesian Techniques

Submitted by
Shubhra Gupta
Computational Biosciences
Email: shubhra.gupta@asu.edu

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

ARIZONA STATE UNIVERSITY

August 2004

Estimating the Divergence Time of Molecular Sequences using Bayesian Techniques

**by
Shubhra Gupta**

APPROVE:

Supervisory Committee

Chair, Dr. Sudhir Kumar
Associate Professor
School of Life Sciences

Dr. Rosemary Renaut, Professor
Department of Mathematics and Statistics

Dr. Martin Wojciehowski
Assistant Professor
School of Life Sciences

ACCEPTED:

Department Chair

Table of Contents

Abstract	4
1. Introduction	
1.1 A Timeline for Molecular Clock	5
1.2 Background on Molecular Clock	5
1.3 Background and Development of the Bayesian Method	7
1.4 Advantages and Disadvantages of Bayesian Approaches	10
2. Bayesian Method in Molecular Clock Analysis	10
3. Supported Studies	17
4. Multidivtime Software	23
4.1 Perl Script to Automate Multi-gene Analyses using MULTIDIVTIME	24
4.2 Use of the Perl Script to Conduct Some Analyses	26
5. Challenges and Criticisms in the Bayesian Analysis	28
6. Significance of the Project	30
7. Conclusions	30
8. Acknowledgements	31
9. Appendix	31
10. References	34

Abstract

The ability to estimate the times of divergence between lineages using molecular data provides significant opportunities to answer many important questions in evolutionary biology. Over long evolutionary periods, rates of molecular evolution may vary over time and among lineages. Several methods, including Bayesian approaches, have been developed to estimate divergence times to accommodate this variation. In this internship, the focus was on understanding the theoretical, computational, and practical aspects of these Bayesian methods in constructing timescales of organismal evolution using molecular data. The study of theoretical aspects involved a primary literature survey, computational aspects dealt with understanding the inner-workings of the *multidivtime* software (popularly used to conduct Bayesian analyses), and practical work involved writing a *perl* script to automate the use of *multidivtime* when it needs to be used for a large number of datasets. Through these efforts I have learned how fundamental research is conducted, in addition to understanding how mathematical and statistical methods are allowing scientists to answer longstanding questions in evolutionary biology.

1. Introduction

1.1 A Timeline for Molecular Clock

For most of the 20th century, the fossil record served as the primary source for reconstructing human origins and the field remained virtually the sole province of anthropologists. Around 1940's two researchers Baldwin (1937) and Florkin (1944) were working on proteins and nucleic acids, but were not in the position to give unique values for the molecular evolution of proteins and nucleic acids.

Zuckerkandle et al. (1960) used the fingerprinting technique used for hemoglobin to show that human, gorillas and chimpanzees are closer to each other than either is to orangutan. But, still they were not aware of the molecular clock until 1965, when Zuckerkandle and Pauling (1965) observed that the rates of amino acid substitution in some genes were similar across lineages. This led to the proposal of constant rate of amino acid substitutions at the molecular level. In 1967, Sarich and Wilson (1967) used an albumin molecular clock to estimate the divergence times between primate species and showed for the first time that human and chimpanzee diverged from a common ancestor only 5 million years (Myr) ago in the Pliocene era, rather than the contemporary assumption of 25 Myr divergences between these species.

1.2 Background on Molecular Clock

The molecular clock refers to approximate constancy of accumulation of nucleotide or protein substitutions over time, or in other words, is an evolutionary hypothesis based on the assumption that mutations occur in a regular manner. It has been proposed that, given a calibration date and a molecular clock, the amount of sequence

divergence can be used to calculate the time that has elapsed since two molecules diverged. The molecular clock technique is an important tool in molecular systematics and in determining the correct scientific classification of organisms (see review in Hedges and Kumar 2003).

Due to improved sequencing technology, molecular sequence data are becoming easy to collect and molecular clocks are being increasingly used to estimate the species divergence times to illuminate the evolutionary history of life. It is clear that a large numbers of genes are needed to improve the precision and reduce any bias of the time estimated (Hedges and Kumar 2003, 2004).

The incompleteness of the fossil record has made DNA and protein sequences the main source of information for some evolutionary events (*reviewed in* Hedges and Kumar 2003). This information is typically extracted by assuming that DNA and protein sequences change at a constant rate i.e. molecular clock exists. After estimating the rate, which is assumed to be common between all lineages, the observed amount of sequence divergence can be converted into time.

Rates of molecular evolution may vary over time and among lineages. The rate at which sequences change could depend on a large number of factors. Different genes evolve at different rates due to differences in the intensity of natural selection. Rates among species may vary for the same gene, because of fluctuations in evolutionary constraints in different lineages and changes in genomic mutation rates (e.g., Takezaki, Rzhetsky & Nei 1995; Nei and Kumar 2000). In fact, Kishino et al. (2001) think that differences in natural selection, population size, generation time, and mutation rate may lead to gradual changes in evolutionary rates among lineages. However, it is not easy to

prove that this is indeed the case, especially when a large number of genes are used. Still, for many genes the molecular clock may not apply, especially when long term evolutionary patterns are considered. In such cases, many sequences of a given gene are often available and the use of the Bayesian method provides opportunities to examine whether substantially more accurate estimate of times are obtained when it is used instead of methods that reject non-clock like sequences (e.g., Takezaki et al. 1995 and Kumar and Hedges 1998).

1.3 Background and Development of Bayesian Inference

A variety of statistical methods have been proposed to check the consistency of a particular data set with molecular clock as a null hypothesis (see review in Nei and Kumar 2000). Use of these tests leads to rejection of null hypothesis for many genes and evolutionary lineages (Hedges et al. 1996; Kumar and Hedges 1998; Thorne et al. 1998). Instead of removing those genes and sequences from the dataset, it is possible to model the rate variation among lineages using a sophisticated model. Bayesian approaches provide one way to use a complex model and avoid computational difficulties.

The Bayesian technique was developed by Thomas Bayes (a Presbyterian minister who lived from 1702 to 1761). Bayes worked on the problem of computing a distribution for the parameter of a binomial distribution. His key paper was published posthumously by his friend Richard Price as “Bayes”. The term "Bayesian" actually came into use only around 1950. Laplace independently proved a more general version of Bayes' theorem and put it to good use in solving problems in mechanics, medical statistics and in accounts. The frequentist interpretation of probability was preferred by some of the most

influential figures in statistics during the first half of the twentieth century, including R.A. Fisher, Egon Pearson, and Jerzy Neyman (**Bayesian probability**).

Bayesian inference is a widely used statistical inference in which probabilities are interpreted as degrees of belief. Methods of Bayesian inference are a formalization of the scientific method involving collecting evidence which points towards or away from a given hypothesis. As more evidence accumulates, the degree of belief in a hypothesis will usually become very high (almost 1) or very low (near 0). Bayes theorem is a means of quantifying uncertainty. Based on probability theory, the theorem defines a rule for refining a hypothesis by factoring in additional evidence and background information, and leads to a number representing the degree of probability that the hypothesis is true. Bayes theorem states that, given some data X and a model (or hypothesis) H that depends on a set of parameters θ , the posterior probability of the parameters is

$$p(\theta | X, H) = \frac{p(X | \theta, H)p(\theta | H)}{p(X | H)}$$

Here $P(\theta|X,H)$ is called the *posterior probability* of the parameters when the data X and the model H are given, $P(X|\theta, H)$ is called the *likelihood or conditional probability* of the data when the model H and its parameters are given, $P(\theta|H)$ is the *prior probability* of the parameters before looking at the data X and the model H , $P(X|H)$ is called the *evidence* of the model H (Berger 1985), e.g. searching of the US nuclear submarine Scorpion by applying Bayesian technique which was failed to arrive as expected at her home port of Norfolk, Virginia. The US Navy's deep water expert, John Craven, believed that it was in south west of the Azores based on a controversial approximate triangulation by hydrophones. He took advice from a firm of consultant mathematicians. The sea area was

divided up into grid squares and a probability assigned to each square. The probability attached to each square was then the probability that the wreck was in that square. A second grid was constructed with probabilities that represented the probability of successfully finding the wreck if that square were to be searched and the wreck were to be actually there. This was a known function of water depth. The result of combining this grid with the previous grid is a grid which gives the probability of finding the wreck in each grid square of the sea if it were to be searched. This sea grid was systematically searched in a manner which started with the high probability regions first and worked down to the low probability regions last. Each time a grid square was searched and found to be empty its probability was reassessed using Bayes' theorem. This then forced the probabilities of all the other grid squares to be reassessed (upwards), also by Bayes' theorem. The use of this approach was a major computational challenge for the time but it was eventually successful and the Scorpion was found in six months.

Two general approaches may be used to generate the posterior distribution when unknown parameters occur in the prior density: (1) Empirical Bayesian analysis and (2) hierarchical Bayesian analysis. The empirical Bayesian analysis replaces the unknown parameters with estimates, whereas the hierarchical Bayesian analysis assigns second level priors as densities for the unknown parameters of the prior. Integration is performed over the second-level priors to obtain a new prior that is completely specified (Berger 1985).

1.4 Advantages and Disadvantages of Bayesian Approaches

Bayesian approaches allow computation of probabilities associated with different theories or models in the light of the data. This is considered an advantage, because the standard approaches, in contrast, seek the inverse of that probability (i.e., the probability of data given a theory). Furthermore, Bayesian methods allow for the incorporation of extraneous (but relevant, e.g., results from past and/or other researchers' studies) information through the formulation of the priors. Such a process can be repeated as many times as desired. However, the formalization of prior information into a prior probability density is not always an easy task and often leads to subjectivity. In addition, the mathematics of obtaining the posterior probabilities is often quite involved, requiring computer-intensive numerical methods. It also requires knowledge of the prior distribution, namely what was known before the data was collected. This issue of prior distribution specification arises frequently in Bayesian applications.

2. Bayesian Methods in Molecular Clock Analysis

The development of Bayesian approaches for estimating divergence times is closely tied to maximum likelihood (ML) methods for comparative sequence analysis. Maximum likelihood is a method of inferring phylogenetic relationships using a pre-specified (often user-specified) model of sequence evolution. Given a tree (a particular topology, with branch lengths), the ML process asks the question "What is the likelihood that this tree would have given rise to the observed data matrix, given the pre-specified model of sequence evolution?".

Felsenstein (1981) implemented the first workable algorithm for calculating maximum likelihood estimates for DNA sequence data. This method has gained significant popularity due to the improvements in computing power, availability of easy-to-use computer programs, and its ability to handle sophisticated models of molecular sequence evolution. The ML framework provides a powerful and flexible framework for estimating model parameters and testing interesting biological hypotheses (Felsenstein 2003; Nei and Kumar 2000). It also allows the testing of hypotheses about the constancy of evolutionary rates by likelihood ratio tests (Muse and Weir 1992; Schierup and Hein 2000). In early 1980s, Hasegawa et al. (1985) developed a new statistical method for estimating divergence dates of species from DNA sequence data by a molecular clock approach. This method takes into account effectively the information contained in a set of DNA sequence data using a maximum likelihood approach.

The assumption of global molecular clock was relaxed by a number of authors (Hasegawa et al. 1989, 2003; Kishino and Hasegawa 1990; Uyenoyama 1995; Takezaki et al. 1995) in their distance-based approach in which different rate parameters could be specified for different parts of the tree or times were estimated in a lineage-specific manner. In the approach proposed by Hasegawa and colleagues, the maximum likelihood method was used to estimate different rate parameters and the branching dates. These methods assumed that there are locally constant rates in parts of a tree, despite rate variation at a larger scale.

In contrast, Sanderson (1997) suggested a model in which the substitution rates were allowed to evolve over time; this method is considered to be superior because it does not require specification of lineages in which a local clock needs to be applied.

However, it is still unclear whether the mutation or substitution rates evolve over time in any consistent fashion. Sanderson (1997) also placed constraints on individual nodes in the phylogenetic tree, based on the existing fossil record, rather than using the minimum divergence times to calibrate local clocks. Sanderson also suggested nonparametric rate smoothing (NPRS) approaches for estimating times in order to avoid making assumptions. Simulations suggested that NPRS performs well when sequence lengths are sufficiently long and evolutionary rates are truly non-clocklike.

In 1997, Yang and Rannala presented an improved version of their earlier Bayesian method (1996), which was an alternative to the classical maximum likelihood method (Felsenstein 1981) for inferring phylogenetic trees using DNA sequences. The Felsenstein (1981) method differs from the conventional maximum likelihood parameter estimation in two ways: the functional form of the likelihood depends on the tree topology (Nei 1987), and the regularity conditions required for the asymptotic properties of maximum likelihood estimators are not satisfied (Yang 1996). As a result, it was unclear whether this method of topology estimation shares all the asymptotic properties (especially efficiency) of maximum likelihood estimators of parameters. Another difficulty was the lack of a reliable method for evaluating the significance of the estimated tree. Therefore, Yang and Rannala (1996) took a different approach, following the earlier work of Edwards (1970) on the problem of estimating phylogeny using gene frequency data from human populations. They used a birth-death process to specify the prior distribution of phylogenetic trees and ancestral speciation times and also employed a Markovian process to model nucleotide substitution.

Usually, a birth-death process is used to model populations of entities in a system. A birth death process can also be used to model most Markov chain process. An arrival is considered a “birth” and a service completion is considered a “death”. Birth and death are independent. A birth increases the state from j to $j+1$, λ_j gives birth rate for state $j > 0$ whereas death decreases the state from j to $j - 1$, μ_j gives death rate for state j and $\mu_0 = 0$. Births and deaths occur at a constant rate (like a Poisson model). In the Poisson process, $\mu_j = 0$ and $\lambda_j = \lambda$. In the Yule process $\mu_j = 0$ and $\lambda_j = j\lambda$ (the extinction rate parameter μ_j is set to zero, the birth-and-death process reduces to the Yule process). In this process, all entities are assumed to act independently.

The parameters of the birth-death process and the substitution model were estimated using a maximum likelihood approach in Yang and Rannala (1996). This method was computationally not feasible for more than five species. Therefore, they suggested the use of Monte Carlo integration to evaluate integral efficiently (Yang and Rannala, 1997). This was necessary, because their first attempt included summation over all tree topologies, which increases very quickly with the number of species. Furthermore, the model for the prior distributions of trees and speciation times were also improved by considering species sampling (which reduced the number of internal branch lengths and results in a more realistic prior distribution of trees) and by treating the birth-and-death rates of the prior distribution as random variables (and eliminated by integration [known as hierarchical Bayesian analysis]) to make the posterior probabilities more robust. Yang and Rannala (1997) used continuous time Markov process to model nucleotide substitution with different equilibrium nucleotide frequencies and hierarchical Bayesian analysis approach with Markov Chain Monte Carlo (MCMC) method to

estimate and evaluate the posterior distribution of phylogenetic trees, under a molecular clock assumption to simplify the calculations.

In their paper (1997), the conditional probability of observing the sequence data, given the labeled history τ and the node times t , is a product over nucleotides sites $f(X|\tau, t; m, \kappa) = \prod_{n=1}^N f(x_n|\tau, t; m, \kappa)$; where $f(x_n|\tau, t; m, \kappa)$ is the conditional probability of observing the nucleotides at the n^{th} site. Substitutions are assumed to occur independently at different nucleotide sites. The posterior probability of the labeled history τ , conditional on the observed sequence data has been given by $f(\tau|X) = \frac{f(X|\tau)f(\tau)}{f(X)}$. The conditional probability is specified by the nucleotide substitution model.

Thorne et al. (1998) developed a maximum likelihood based Bayesian methods to estimate divergence times in which the substitution rates evolved over time and constraints were placed on phylogenetic nodes, following Sanderson (1997). In their method, rate of evolution is constant on any particular branch (which is different from Sanderson [1997]), but rates are allowed to differ among branches and be correlated with each other over time. While this concept uses the molecular clock concept, it does not require the assumption of a global clock or need pre-specification of areas to which local clocks should be applied in a phylogenetic tree. Mathematically in their paper (1998) the posterior distribution for Bayesian analysis is given by

$$p(T, R, v|X) = p(X|T,R)p(R|T,v)p(T)p(v)/p(X) \quad (3)$$

where X is a data set, $R = (R_0, R_1, \dots, R_k)$ is the rates of molecular evolution on the $k+1$ branches of the rooted tree, T is a vector that specifies the internal node times (including

the root) and v is a constant (value of v determines the prior distribution for the rates of molecular evolution on different branches given the internal node times).

In their model, to generate approximately random samples from the posterior distribution, some runs were performed without evaluating $p(X|T, R)$ from (3) and because the denominator $p(X)$ of (3) is more difficult to evaluate due to multiple integration over T , R and v , Metropolis-Hastings algorithm was used, to obtain an approximately random sample from $p(T, R, v|X)$. Metropolis-Hastings algorithm is a MCMC technique that permits construction of a Markov chain on the parameter (T, R, v) . This algorithm was cycled through a series of steps (e.g. for v , Internal node time, Rate & Mixing). In this way the resulting Markov Chain becomes irreducible.

In one step of the cycle i.e. for v , a state (T', R', v') is proposed that differs from the current state (T, R, v) only by the value of v . v' can be generated by randomly sampling a value U from a uniform distribution on the $(0, 1)$ interval and then setting $v' = ve^{H_1(U - 0.5)}$, where H_1 is a constant with a pre-specified value. In next part of the cycle i.e. internal node time, a new time for each internal node has been proposed exactly once and then this part of the cycle is exited. The internal node has parental node p , eldest-child node e and youngest-child node y . The rate on a branch was indexed according to the node at which the branch ends. If node was not the root, its proposed time T_i' should be greater than the time T_p of its parental node p and less than the time T_e of its eldest-child node e and proposed a new time T_o' for the root node of the in-group, it must be ensured that the proposed time of this root node should be less than the time T_e of its eldest-child node. New time can be calculated by $T_o' = T_e - (T_e - T_o)e^{H_2(U - 0.5)}$; where U is

a uniform random variable on the interval (0, 1) and H_2 is the value of a pre-specified constant.

For rate cycle, each of the $k+1$ branches on the in-group, suggests a new rate $R = (R_0, R_1, \dots, R_k)$ from the current state R_i . This was sampled by a value U from a uniform distribution on (0, 1) and using a pre-specified constant H_3 to get $R_i' = R_i e^{H_3(U - 0.5)}$. Mixing step is given to improve convergence of the Markov chain. In this step, all proposed node times differ from the current node times by a factor of M . The value of M can be obtained by $M = e^{H_4(U - 0.5)}$; where U is a uniform random variable on the interval (0, 1), and H_4 is a pre-specified constant. All node times T_i , can be determined by $T_i' = MT_i$ and rate for all branches can be determined by $R_i' = \frac{1}{M} R_i$.

Rambaut et al. (1998) also developed a maximum likelihood approach to estimate divergence times that deals explicitly with the problem of rate variation. In their method rate constancy test were included (excluding the data for which rate heterogeneity is detected, following Takezaki 1995 and Kumar and Hedges 1998). They advocated the use of multiple calibration points, even though they may be severe underestimates. The use of one (or a few) good calibration points (those that are expected to be minimal underestimates of divergence times) versus a large number of calibration points that are severe underestimates of time is a major point of controversy today (*see* Hedges and Kumar 2004).

In year 2001, Kishino et al. (2001) extend the Bayesian techniques for estimating divergence times from their 1998 method and explored their behavior via simulation. In this study mean of the rates at the two nodes is simply approximation of the average rate on a branch. Rates were assigned to branches of a rooted tree rather than to nodes of the

tree. Their implementation sets the mean of the normal distribution for the logarithm of the ending rate by forcing the expected ending rate to be equal to the beginning rate. In earlier work, the prior distribution for divergence times was a Yule process, but a generalization of the Dirichlet distribution to rooted tree structures is given in 2001. This prior was selected because it is statistically simple and provides flexibility. The term “branch length” is representing an expected amount of sequence change on a branch rather than the time duration of the branch. Their method also required specification of many more priors. Thorne et al. (2002) improved their method by specifying the prior distribution for the rate of molecular evolution at the root node with a gamma distribution and proposed Bayesian techniques for estimating divergence times to analyze multiple gene sequences for each taxon of interest.

3. Supported Studies

Yang and Rannala (1997) used birth-death process to specify the prior distribution of phylogenetic trees and ancestral speciation times. The method was applied on two datasets of DNA sequences consisting of a segment of the mitochondrial genomes of human, chimpanzee, gorilla, orangutan, gibbon, macaque, squirrel monkey, tarsier, and lemur. Both empirical and hierarchical Bayesian analyses were performed. Bayesian method generated the same best trees as were obtained by maximum-likelihood analyses, but the posterior probabilities for these trees were quite different and suggested that adding a second level prior for the birth-death rates did not change the posterior probabilities.

Thorne et al. (1998) rooted the tree with an out-group and obtained maximum-likelihood estimates of branch lengths for the unrooted topology consisting of the out-group and the in-group. To prove the efficiency of their model they used 31 amino acids sequences rather than DNA sequences, from the *rbcL* chloroplast gene and set “Marchantia” sequence as outgroup. A model of amino acid replacement used to incorporate the impact of protein secondary structure. They used PAML (Yang 1997) package to analyze the 30 *rbcL* in-group sequences using JTT (Jones, Taylor, and Thornton 1992) model under a hypothesis of molecular clock, with estimated depth of the in-group root 6.03 amino acid replacements per 100 sites. Fossil evidence was not used to calibrate the molecular clock. In their model, the expected number of amino acid replacements between root and tip was varying among tips. The MCMC algorithm completed 100,000 initial cycles before the state of the Markov chain was sampled. Thereafter, the Markov chain was sampled every 1,000 cycles until a total of 1,000 samples were collected.

The resulted posterior means of the normalized times to the root from their program (divtime) was closer but greater than the root depth of 6.03 replacements per 100 sites which was estimated by PAML, because PAML ignored alignment columns with gaps, whereas their method treated gaps as missing data and the model of amino acid replacement assumed in the Bayesian analysis allows rate heterogeneity among sites. Their analysis found that the ages of conifer and angiosperm clades to be more similar than they probably were. Later Kishino et al. (2001) did this experiment with constraints on node times and without constraints on node times.

Case 1: Simulation with constraints on node times. One hundred simulated data sets were generated according to a tree topology and node times; each dataset contained one out-group and 16 in-group sequences. Every simulation began by randomly sampling the rate of evolution at the in-group root node from a gamma distribution with mean 1 and standard deviation 0.5. In addition, the rate at the tip of the out-group node was randomly sampled based on the rate at the in-group root node and according to their model. The simulated rates, together with the node times determine the branch lengths of the tree. The resulting branch lengths allowed simulated evolution of DNA sequences along the tree. All sequences were of length 1,000 and were generated according to the Jukes-Cantor (JC, 1969) model. In all MCMC analyses, 10,000 cycles were used for “burn in” and 10,000 numbers of samples had been collected and constraints, that forced a node age to exceed a specific value or that restrict a node age to less than specific values, were allowed.

Case 2: Simulation without constraints on node times. In this case, data were simulated according to a perfect molecular clock. Therefore, rates at all nodes on the tree were identical to the rate at the in-group root node. Simulated results they have presented in the form of trees in their paper; three trees for without constraints on node time and five trees for with constraint on node time.

In 2002, Thorne et al. explored the effect on divergence time estimates of the number of genes in the data set by simulation. Simulation was based on the two tree topologies, 64 genes, and 16 in-group and one out-group taxa. True in-group root time for both trees was 0.5 time units. The total time represented on the path from the in-group root back to the common ancestor of all sequences and then forward to the outgroup

taxon was 0.625 time units for both trees. For each gene, rate at the in-group root node was obtained from a gamma distribution with a mean of 1.0 and standard deviation (SD) of 0.5. All genes were evolved according to JC model and analyzed with the nucleotide substitution model. Branch lengths were generated for simulation according to the JC model. They considered two cases for analysis (1) constant rate of evolution over time, (2) evolutionary rates change over time and performed three varieties of MCMC analysis

- Assumed rates were constant throughout evolution.
- Assumed all genes shared some common value and
- Assumed that each gene had a separate value.

First tree analyzed with both cases but second tree was analyzed with only the first case. They found that the first tree produced a better result than the second because the time interval between nodes on the first tree was evenly spaced, which was not the case in the second tree. As expected, authors correctly state that without the knowledge of the prior (for more accurate estimation of the in-group root time), no amount of sequence data will allow perfect separation of rates and times.

Their study also shows results from multi-gene analysis using the same dataset of 64 genes. They performed multiple MCMC runs from different initial states on the data and determined different runs yield similar approximations of the posterior distribution (long MCMC runs is necessary to achieve convergence, computational tractability). Performance of the multigene divergence time was relatively good when constancy of rates was assumed; posterior means of the divergence time estimates for the in-group root was closer to their true value. These analyses showed that divergence time estimates based on single genes revealed that the constant rate assumption often produced poor

estimates of divergence times (Kishino et al. 2001), which is a well known fact (Kumar and Hedges 1998, Hedges and Kumar 2003).

When rate variation over time existed, the assumption of a constant rate of evolution does not yield good estimates of divergence times (as expected) even when multiple genes were employed to estimate divergence times (Kishino et al. 2001). Overall, the divergence time estimated from single gene analyses, when rates were allowed to vary, was similar for most nodes. The same pattern was observed when the divergence time estimated from the multi-gene analysis. As expected the credibility intervals were narrower for the multi-gene analysis than for single gene analysis.

Case 3: Empirical analysis of data. Springer et al. (2003) studied the data set of placental mammal using their Bayesian approach to avoid making the molecular clock assumption and to be able to use multiple constraints from the fossil record. This is different from previous studies, in which one reliable fossil record was used to calibrate the molecular clock and the molecular clock assumption was used for genes that passed the molecular clock test (Hedges et al. 1996; Kumar and Hedges 1998). Springer et al. (2003) also had much larger number of species from a larger number of mammalian orders than the previous authors, but a much fewer number of genes.

Springer et al. data contained a total of 16,397 aligned nucleotide positions for 42 placental mammals. They investigated the competing hypotheses for the timing of the placental mammal and focus on whether extant placental orders originated and diversified before or after the Cretaceous-Tertiary (K/T) boundary (65 Myr). Branch lengths were estimated with the “ESTBRANCHES” program of Thorne et al. (1998) for the complete dataset. They used Felsenstein’s (1984) model of sequence evolution and

allowed a gamma distribution of rates among sites. The transition/transversion parameter and rate categories of the gamma distribution were calculated with PAUP 4.0 (1998) for data set. They placed opossum as an out-group. Program “DIVTIME” was used to estimate divergence times. They set 105 million years for the mean of the prior distribution of the root of the ingroup tree. In the results, the deepest split among placentals was in between Afrotheria and other taxa at 107 million years, which is remarkably similar to an estimate of 105 Myr for Elephants and Humans reported by Kumar and Hedges (1998) without using the Bayesian approach and using an external calibration point (Bird-Mammal split of 310 Myr). The split between Euarchontoglires and Laurasiatheria was estimated at 94 million years, which is close to the 90-112 Myr date for primates and rodent splits (primates-lagomorpha=90 Myr, primate-scuirognathi = 112 Myr) in Kumar and Hedges 1998.

Interordinal splits within Euarchontoglires were in the range of 82 - 87 million years and those within Laurasiatheria were in the range of 77 - 85 million years. The oldest split within Primates was at 77 million years (compare with 865 Myr for human-Scandandia in Kumar-Hedges 1998). The earliest divergence within Rodentia was estimated at 74 million years, and the rat-mouse split was at 16 million years. Their analysis also supported the diversification of placental mammals before the K/T boundary, a hypothesis put forth by Hedges et al. 1996. The splitting of rat-mouse was dated at 16-23 million years ago instead of 30-41 million years reported by many others (e.g., Kumar and Hedges 1998; Nei et al. 2000; Hedges and Kumar 2004). However, Springer et al. (2003) clearly mention that they would also obtain an estimate of 38 Myr for mouse-rat split if the rodents were placed as a basal group to primates. In addition,

paleontological evidence supported that *Mus* and *Rattus* diverged at 10 - 14 Myr (Jacobs and Downs 1994) whereas many molecular estimates have departed significantly from fossil evidence as 46 Myr (Kumar and Hedges 1998; Adkins, Gelke, Rowe & Honeycutt 2001). But Springer et al. (2003) ML topology suggested dates 14-24 Myr, which is more concordant with the fossil record. When Placentalia rooted at the base of myomorph rodents Springer et al. estimated 35 Myr for the rat–mouse split. Therefore, Bayesian analyses can be severely affected by the phylogeny used.

4. Multidivtime Software

The Multidivtime software developed by Thorne et al. (Thorne, Kishino and Painter 1998) is intended to compute divergence times using the Bayesian approach. It consists of two main programs (1) ESTBRANCHES and (2) MULTIDIVTIME. The ESTBRANCHES program is used for estimating branch lengths of the evolutionary tree given the node relationship, whereas the MULTIDIVTIME generates the divergence times. In order to get output from ESTBRANCHES, the user needs to provide sequences of either nucleotide or proteins in the TESTSEQ file and the reference of model (e.g. JTT, JC) file and tree topology file in its control file HMMCNTL.DAT. Then it generates one output in a file and other on the computer screen, output written in the file used as an input for MULTIDIVTIME.

MULTIDIVTIME program takes output file of ESTBRANCHES program as an input file. MULTIDIVTIME also has its control file MULTICNTL.DAT in which user

can change the parameters according to available information about organism and sequence. Some of the important features of MULTICNTRL.DAT are

- Rttm that is a prior expected number of time units between tip and root (in-group depth) of the given tree.
- Rttmsd is a standard deviation of prior for time between tip and root.
- 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).
- Rratesd is a standard deviation of prior for rate at root node.
- Zero value of “brownmean” and “brownstd” means divergence time estimated assuming molecular clock, other than zero value shows non-molecular clock exist.
- Bigtime – a number higher than time units between tip and root could be in our wildest imagination.

4.1 Perl Script to Automate Multigene Analyses using MULTIDIVTIME.

In order to learn about the computational biology, and to assist my host laboratory in their research on Bayesian analyses, I wrote a perl script to automatically execute the software multidivtime and provide final output in a file.

The steps in this algorithm are given below (see Appendix also).

Input: type perl command with script file name.

Output: print actual divergence times among lineages in a file.

Step 1: open gene0 file and copy it to “testseq” file.

Step 2: execute ESTBRANCHES program and write computed output in the file oest.gene0 (the user can give other output file name according to convenience).

Step 3: repeat step1 and step2 until all genes estimated by the program ESTBRANCHES.

Step 4: parse tree with branch length from file oest.gene0.

Step 5: print output in the file estout.tree0.

Step 6: repeat step4 and step5 until all oest.gene files parsed.

Step 7: start a.out program to compute distance from every tip to root and give the average of those distances.

Step 8: Print output in a rate file.

Step 9: repeat step7 and step8 until distances computed from all tree files.

Step 10: parse average distance from each tree file and store into array.

Step 11: take the median of those array values.

Step 12: print that median value in the f_rates file.

Step 13: Divide the value of f_rates file by rttm.

Step 14: print that value in the file final_rate.

Step 15: open the file multidivtime1, which is exactly the same as control file of MULTIDIVTIME.

Step 16: make changes in multidivtime1 whatever is needed.

Step 17: copy multidivtime1 file to MULTICNTRL.DAT.

Step 18: execute MULTIDIVTIME program.

Step 19: print output in the output.gene file.

Step 20: parse actual divergence times from the output.gene file.

Step 21: print divergence times in the file output.txt.

The flow chart for this PERL script is given below.

4.2 Use of the Perl Script to Conduct Some Analyses.

The objective of the present internship was to learn about the mathematical, statistical, and computational aspects of Bayesian analyses for estimating divergence times in a 3 month time period. So the extensive analysis of biological data or design of experiments to conduct biological research was beyond the scope of this work. However, the development of the PERL script was motivated by the need to automate execution of multidivtime program and parsing its output for use in the ongoing research in Dr. Kumar's laboratory, which is examining the effectiveness of the Bayesian approaches by means of computer simulations. I was provided by some of those datasets by Dr. Kumar for testing my scripts and to make some general observations so I could learn how evolutionary bioinformatics research is carried out. In the following, I summarize a few observations that the Kumar laboratory has made using these scripts to explore how different priors effect posteriors (time estimates).

Analysis: In the analyses of 50 combined genes, I had run the ESTBRANCHES program 50 times for 50 single genes. After getting all the 50 results from ESTBRANCHES, I combined resultant output files of ESTBRANCHES in the control file (MULTICNTRL.DAT) of multidivtime (takes as a input) and did the experiment with MULTIDIVTIME program. Divergence times were estimated among species by MULTIDIVTIME and the final output was saved in a file. In order to execute the programs automatically for large amount of files (about 1000 files each containing 50 genes), I have written a perl script. In this analysis different priors were used to get good

estimated times (closer to true time) among species based on fossil record and other organism information.

Time taken by the MULTIDIVTIME program: MULTIDIVTIME program took smaller time when the molecular clock method was assumed. In my datasets containing 20 (lamper, shark, fish, frog, snake, bird, alligator, possum, pig, cow, horse, dog, cat, rabbit, rat, mouse, macaque, gorilla, human, chimpanzee) species, the molecular clock case took half (approx 10 hours) as much time as the non-molecular clock (approx 20 hours) for 50 genes.

Effect of bigtime: This prior does not seem to have a major effect on estimated times except for very old divergences, where it can lead to significant overestimates, when the fossil record from in-group species divergences is used [e.g. Bird - Mammal time 310 Myr used as a constraint so lineages far away such as human – chimp (5 Myr) and shark (528 Myr) showed significant overestimate or underestimate].

Effect of some other priors: *r_{ttm}* (expected number of time units between tip and root) and *r_{ttmsd}* (standard deviation of prior for time units between tip and root) do not seem to have much effect on estimated time but we need an accurate value for *r_{ttm}* and *r_{ttmsd}*. *r_{trate}* (mean of prior distribution for rate at root node) and *r_{ratesd}* (standard deviation of prior for rate at root node) do have an effect on estimated times. Bigger values of these priors give good estimates of times.

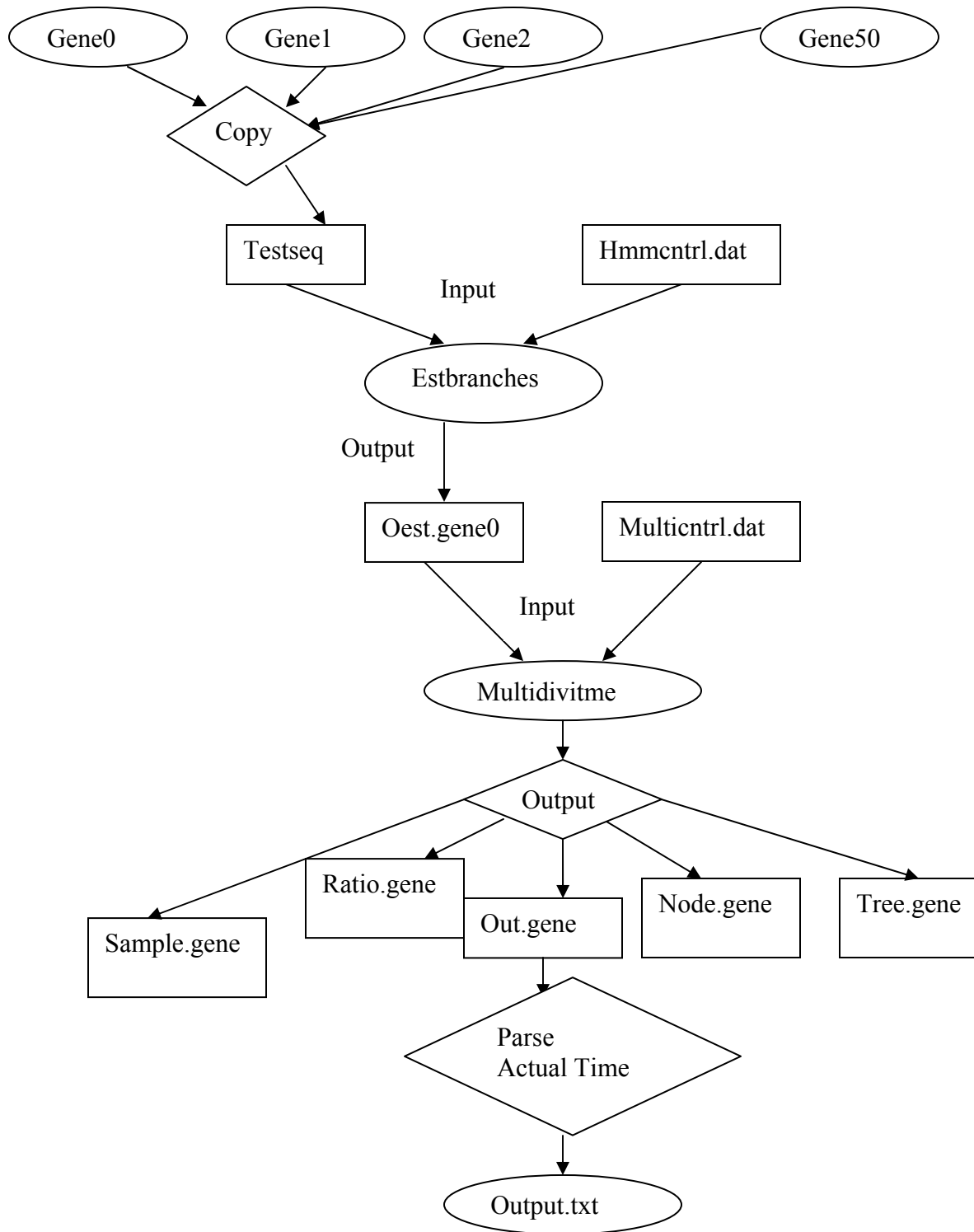
Effect of upper and lower-bounds: When wider ranges for upper and lower bounds were used, the confidence intervals changed proportionally. Therefore, when one or a few calibration points are used, the results are not invariant to the selection of fossil

record calibration. In particular, it is clear that good fossil record is still important to obtain reliable estimates of time.

Effect of using multiple genes: Analysis of combined genes showed that estimated times were closer to the true times, as compared to the individual (single gene) time estimates.

5. Challenges and Criticisms in the Bayesian Analysis

There are some challenges in the Bayesian analysis such as defining what we know before the data are collected that is the issue of prior distribution specification arises frequently in Bayesian applications and deciding how complex our models should be. It is clear from some of the results mentioned above, good prior parameters are needed to get good posterior. This is true even if we use an accurate model that fit the data well. Bayesian analysis basically compromise is between the fit of the model and how highly dimensional and diffuse is the prior.



Flow chart for the Perl Script process

6. Significance of the Project

I learned the computational aspect of software multidivtime, application of Bayesian techniques in molecular biology and theoretical aspect of Bayesian Method. Using the software multidivtime I was estimating the times of divergence between species by applying the Bayesian method. I observed the effect of different prior estimates on posteriors in order to know which one giving better estimation of time. I have written a perl script which can automatically execute the software multidivtime and provides final output in a file.

7. Conclusions

Bayesian techniques are of interest to biologists because they can work well when rates are varying among lineages, which is actually the case. Divergence times among lineages are calculated using the MCMC method. In the report I have described some papers which are based on Bayesian methods. There are some strengths and weakness in these papers. Thorne and Kishino (2002) improved their method from single to multiple-gene studies but have not mentioned that how rate for root node for multiple-genes can be selected. On the other hand Yang and Rannala (1997) improved birth-death process using MCMC such as to remove the limitation of their 1996 paper which was not able to handle more than five species. Some of our preliminary results indicate that the use of a larger number of genes will yield better estimates than a single gene and that some priors have a larger effect than others on the final time estimates. While working with Bayesian techniques, it became clear to me that the robustness of Bayesian approaches for

estimating divergence times needs better characterization and should be the work of a future study.

8. Acknowledgements

I would like to thank Dr. Sudhir Kumar for suggesting the project, providing guidance, and allowing me to use the data from an ongoing research project and also Dr. Alan Filipksi for helpful discussion. Dr. Sudhir Kumar also thoroughly edited most of this report to make it scientifically correct and even rewrote many sections to make them easy to understand. I am also thankful to other members of the lab for their help specially Shanker and Vinod.

9. Appendix

Perl Script:

```
#!/usr/local/bin/perl
use warnings;
for ($i = 0; $i<=49; $i++)
{
    $#arr= -1;
    open(AT, "gene$i");          # opens gene file for reading
    open(OUT, ">testseq");      # opens testseq file for overwriting

    @arr = <AT>;
    print OUT "@arr";          # print array in OUT file
    # execute the "estbranches"
    # program and output from estbranches is going to print in
    # oest.gene and screen output in out. oest.gene.
    system ("./estbranches oest.gene$i > out.oest.gene$i");
    close(AT);                 # close the gene file
    close(OUT);                 # close the testseq file
}

for ($i = 0; $i<=49; $i++)
{
    $#arr= -1;
    open(AT, "oest.gene$i");    # opens oest.gene file for reading
```

```

open(OUT, ">estout.tree$i"); # opens estout.tree file for overwriting

while (<AT>)
{
    $line = $_;          # store text in line
    if ($line =~m /\(/g) # parse text from (
    {
        print OUT "$line\n"; # write in the file estout.tree
    }
}
close(AT);          # close the oest.gene file
close(OUT);        # close the estout.tree file
}

for ($i = 0; $i<=49; $i++)
{
    system (". /a.out estout.tree$i > rate$i"); # execute the "a.out"
                                                # program
    # output from a.out is going to print in estout.gene and
    # screen output in rate.
}

for ($i = 0; $i<=49; $i++)
{
    open(AT, "rate$i")||die "Couldn't open file\n"; # opens rate file for
                                                    # reading
    open(OUT, ">f_rates"); # opens f_rate file for overwriting
    $#arr = -1;
    while (<AT>)
    {
        $line = $_; # store text in line
        chomp $line; # chomp the end character of a specified string
        if ($line =~m /mean/g) # parse text from "mean"
        {
            @arr = split("=", $line); # split from "=" and store in a array
            push (@arr2, $arr[1]);
        }
    }
    close(AT); # close the rate file
}
@sorted = sort {$a <=> $b} @arr2; # sort an array
print "@sorted\n";
$A = $sorted[24] + $sorted[25]; # sorted value in a variable
print OUT "$A\n"; # print array in OUT file
close (OUT); # close the file f_rate

$variable = 495.00;
$var = 2;
open(AT, "f_rates")||die "Couldn't open file\n"; # opens f_rate file
                                                    # for reading
open(OUT, ">final_rate"); # opens final_rate file for overwriting

while (<AT>)
{
    $line = $_;

```

```

chomp $line;
if ($line =~m /\^\s *$/g) # parse text
{
    next;
}
$R = $line/$variable; # value store in $line is divided by
                        # $variable and store in another variable
$V = $R/$var; # value store in $R is divided by $var and store in $V
}
close (AT); # close the file f_rate
printf OUT "%.3f\n" , $V; # print value up to 3 decimal places in the
                        # file OUT
close (OUT); # close the file final_rate
open(FR, "final_rate"); # opens final_rate file for reading
open(AT, "multidivtime1"); # opens multidivtime1 file for reading
open(OUT, ">multicntrl.dat"); # opens multicntrl.dat file for
                        # overwriting
while (<FR>)
{
    $var = $_;
    chomp($var)
}

while (<AT>)
{
    $line = $_;
    chomp($line);
    if ($line =~m /rtrate/g) # parse text from rtrate
    {
        @arr = split (/\.\\.\\.\/); # split text and store in a array
        print OUT "$var ...$arr[1]";
        next;
    }
    print OUT "$line\n";
}
close (FR); # close the file final_rate
close (AT) # close the file multidivtime1
close (OUT); # close the file multicntrl.dat

{
    system (". /multidivtime gene > out.gene"); # execute the
        # "multidivtime" program and output from a.out is going
        # to print in estout.gene and screen output in rate.
}

open(AT, "out.gene"); # opens out.gene file for reading
open(OUT, ">output.txt"); # opens output.txt file for overwriting

while (<AT>)
{
    $line = $_;
    if ($line =~m /Actual/g && $line == /0\.00000/) # parse text from
        # Actual

    {
        @arr = split("=", $line); # split text from = and store in a array
        @arr1 = split(/\(/, $arr[1]); # split text and store in a array
    }
}

```

```

    print OUT "$arr1[0]\n";          # print array in the file OUT
  }
}
close (AT); # close the file out.gene
close(OUT); # close the file output.txt

```

10. References

1. **Adkins, R. M., Gelke, E. L., Rowe, D. & Honeycutt, R. L.**, Molecular phylogeny and divergence time estimates of major rodent groups: evidence from multiple genes, 2001, *Mol. Biol. Evol.* **18**, 777–791.
2. **Baldwin, E.**, An introduction to comparative biochemistry, 1937, Cambridge University Press. Cambridge, England.
3. **Berger, J. O.**, Statistical decision theory and Bayesian analysis, 1985, Springer - Verlag, New York.
4. **Felsenstein, J.**, Evolutionary trees from DNA sequences: a maximum likelihood approach, 1981, *J. Mol. Evol.* **17**, 368 – 376.
5. **Felsenstein, J.**, Distance methods for inferring phylogenies: a justification, 1984 *Evolution* **38**,16-24.
6. **Felsenstein, J.**, *Inferring Phylogenies*, 2003, Sinauer Associates, Sunderland, MA.
7. **Florkin, M.**, *L'evolution biochimique*, 1944, Masson, Paris.
8. **Hasegawa M., Kishino, H. and Yano, T.**, Estimation of branching dates among primates by molecular clocks of nuclear DNA which slowed down in Hominoidea, 1989, *J. Human Evol.* **18**, 461-476.
9. **Hasegawa, M., Kishino, H. and Yano, T.**, Dating of the human-ape splitting by a molecular clock of mitochondrial DNA, 1985, *J. Mol. Evol.* **22**, 160-174.

- 10. Hasegawa, M., Thorne, J.L. and Kishino, H.,** Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution, 2003, *Genes Genet. Syst.*, **78**, 267 – 283.
- 11. Hedges, S. B., Parker, P. H., Sibley, C. G., and Kumar, S.,** Continental breakup and the ordinal diversification of birds and mammals, 1996, *Nature* **381**, 226-229.
- 12. Hedges, S. B. and Kumar, S.,** Genomic clock and evolutionary timescales, 2003, *Trends in Genetics* **19**, 200 – 206.
- 13. Hedges, S. B. and Kumar, S.,** Precision of molecular time estimates, 2004, *Trends in Genetics* **20(5)**, 242- 247.
- 14. Jacobs, L. L., and Downs, W. R.,** in *Rodent and Lagomorph Families of Asian Origins and Diversification*, eds, 1994, Tomida, Y., Li, C.-K. & Setoguchi, T. (National Science Museum Monographs, Tokyo), 149–156.
- 15. Jones, D.T., Taylor, W.R. and Thornton, J.M.,** The rapid generation of mutation data matrices from protein sequences, 1992, *Comput. Applic. Biosci.* **8**, 275-282.
- 16. Jukes, T. H., and Cantor C. R.,** Evolution of protein molecules, 1969, 21–32 *in* H. N. MUNRO, ed. *Mammalian protein metabolism*. Academic Press, New York.
- 17. Kishino, H., Thorne, J.L. and Bruno, W.J.,** Performance of a Divergence Time Estimation Method under a probabilistic Model of Rate Evolution, 2001, *Mol. Biol. Evol.* **18**, 352-361.
- 18. Kishino, H. and Hasegawa, M.,** Converting distance to time: an application to human evolution, 1990, *Methods in Enzymology*, **183**, 550-570.
- 19. Kumar, S. and Hedges, S. B.,** A molecular timescale for vertebrate evolution, 1998, *Nature* **392**, 917 – 920.

- 20. Multidivtime Software:** <http://statgen.ncsu.edu/thorne/multidivtime.html>
- 21. Muse, S. and Weir, B.,** Testing for equality of evolutionary rates, 1992, Genetics **132**, 269 – 276.
- 22. Nei, M.,** Phylogenetic trees. In: Molecular evolutionary genetics., 1987, 287-326, Columbia University Press, New York.
- 23. Nei, M. and Kumar, S.,** Molecular Evolution and Phylogenetics, 2000, Oxford University Press, New York.
- 24. Rambaut, A. and Bromham, L.,** Estimating Divergence Dates from Molecular Sequences, 1998, Mol. Biol. Evol. **15**, 442 – 448.
- 25. Sanderson, M. J.,** A nonparametric approach to estimating divergence times in the absence of rate constancy, 1997, Mol. Biol. Evol. **14**, 1218 – 1232.
- 26. Sarich, V. M. and Wilson, A. C.,** Immunological Time Scale for Hominid Evolution, 1967, Science **158**, 1200 – 1203.
- 27. Schierup, M. H. and Hein, J.,** Recombination and the Molecular Clock, 2000, Mol. Biol. Evol. **17**, 1578 – 1579.
- 28. Springer, M.S., Murphy, W.J., Eizirik, E. and O'Brien, S.J.,** Placental mammal diversification and the Cretaceous-Tertiary boundary, 2003, PANS **100**, 1056-1061.
- 29. Takezaki, N., Rzhetsky, A. and Nei, M.,** Phylogenetic test of the molecular clock and linearized trees, 1995, Mol. Biol. Evol. **12**, 823 – 833.
- 30. Thorne, J. L., Kishino, H. and Painter, I. S.,** Estimating the Rate of Evolution of the Rate of Molecular Evolution, 1998, Mol. Biol. Evol. **15**, 1647 – 1657.
- 31. Thorne, J. L. and Kishino, H.,** Divergence Time and Evolutionary Rate Estimation with Multilocus Data, 2002, Syst.Biol. **51**, 689 – 702.

- 32. Uyenoyama, M.,** A generalized least squares estimate of the origin of sporophytic self-incompatibility, 1995, *Genetics* **139**, 975–992.
- 33. Yang, Z.,** Maximum likelihood models for combined analyses of multiple sequence data, 1996, *J. Mol. Evol.* **42**, 587-596.
- 34. Yang, Z.,** PAML: a program package for phylogenetic analysis by maximum likelihood, 1997, *Comput. Appl. Biosci.* **15**, 555-556.
- 35. Yang, Z. and Rannala, B.,** Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference, 1996, *J. Mol. Evol.* **43**, 304-311.
- 36. Yang, Z. and Rannala, B.,** Bayesian Phylogenetic Inference using DNA Sequences: A Markov Chain Monte Carlo Method, 1997, *Mol. Biol. Evol.* **14**, 717 – 724.
- 37. Zuckerkandl, E. and Pauling, L.,** Evolutionary divergence and convergence in proteins, 1965, In: Bryson V. and Vogel H. (symposium) *Evolving Genes and Proteins*, New York and London Academic press, 97 – 166.
- 38. Zuckerkandl, E., Jones, R. T. and Pauling, L.,** A comparison of animal hemoglobins by tryptic peptide pattern analysis, 1960, *Proc. Natl. Acad. Sci. USA*, **46**, 1349 – 1360.
- 39. Bayesian probability:** http://en.wikipedia.org/wiki/Bayesian_probability
- 40. Phylogenetic reconstruction:** <http://abacus.gene.ucl.ac.uk/C337/lecture9.pdf>