

# Estimating the Divergence Time of Molecular Sequences using Bayesian Techniques

By

**Shubhra Gupta**

**Computational Biosciences**

**Email: [shubhra.gupta@asu.edu](mailto:shubhra.gupta@asu.edu)**

**Supervisor**

**Dr. Sudhir Kumar**

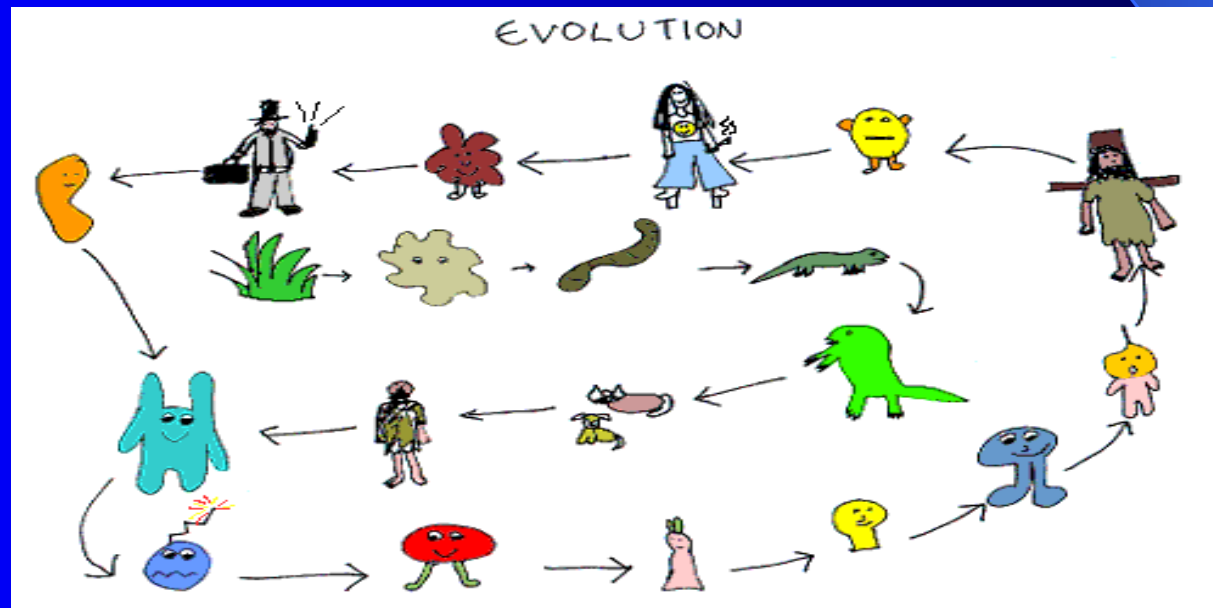
**Email: [S.kumar@asu.edu](mailto:S.kumar@asu.edu)**

**July 2004**



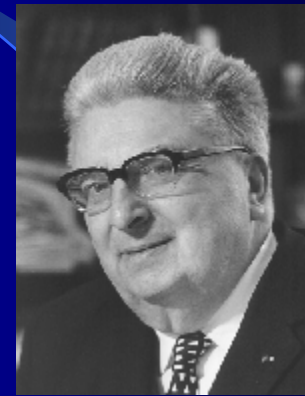
# Problem Statement

- Estimating the divergence times among species using molecular sequences.



# A Timeline for Molecular Clock

- 1940's: Two researchers Baldwin and Florkin were working on the molecular evolution of proteins and nucleic acids.
- 1960s: Zuckerkandl et al. used hemoglobin to show that human, gorillas and chimpanzees are close to each other than to the orangutan.



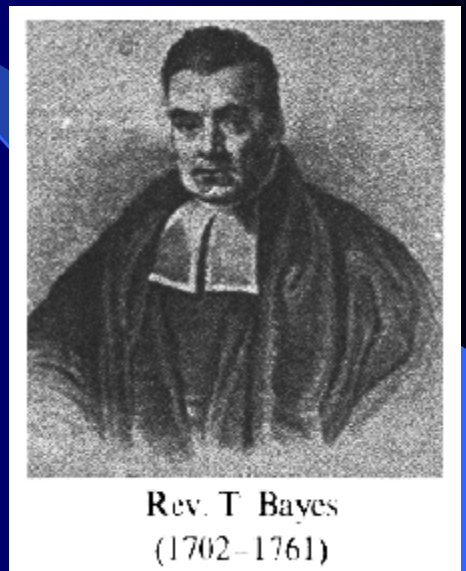
- 1963 and 1965: Margoliash and Zuckerkandle and Pauling appear to have been the first to observe that the rates of amino acid substitution in some genes are the same among lineages.
- 1967: Sarich and Wilson assuming molecular clock showed first time that human and chimp diverged from each other only 5 million years ago, *i.e.* in the Pliocene era.

# Molecular Clock and its Importance

- An evolutionary hypothesis based on the assumption that mutations occur in a regular manner.
- Molecular clocks are used to estimate the time of divergence of genes and species, they have helped to illuminate the temporal history of life.
- Information is typically extracted by assuming that DNA and protein sequences change at a constant rate.

# Background of Bayesian Method

- Bayesian approaches provide one way to use complex model and avoid computational difficulties.
- Bayesian technique was developed by Thomas Bayes. His key paper was published posthumously by his friend Richard Price as “Bayes.”



# Mathematical Basis of Bayesian

- Bayesian probability theory, the probability of an event describes the observer's *degree of belief* on the occurrence of the event.
- Bayesian way of estimating the parameters of a given model focuses around the Bayes theorem (*Probability is that degree of confidence dictated by the evidence through Bayes' theorem – E.T. Jaynes*).

# Bayes' Theorem

- Given some data  $X$  and a model (or hypothesis)  $H$  that depends on a set of parameters  $\theta$ , the posterior probability of the parameters

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

$P(\theta|X,H)$  is called the *posterior probability* of the parameters when the data and the model are given.

$P(X|\theta, H)$  is called the *likelihood* of the data when the model and its parameters are given.

$P(\theta|H)$  is the *prior probability* of the parameters before looking at the data and the model.

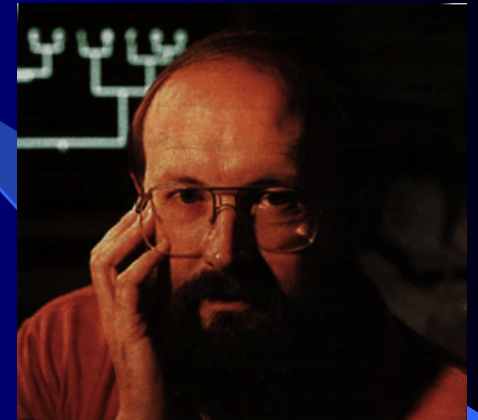
$P(X|H)$  is called the *evidence* of the model  $H$ . E.g. US marine search

# Advantage and Disadvantage of Bayesian Approach

- Allows for the incorporation of extraneous (but relevant, e.g., results from past and/or other researchers' studies) information through the formulation of the priors.
- Allows computation of probabilities associated with different theories or models in the light of the data.
- Gives more accurate forecasts and prediction with proper prior distributions but what we know before the data are collected that is the issue of prior distribution specification arises frequently in the Bayesian applications.

# A Timeline for Bayesian Method

- 1981: Felsenstein implemented the first workable algorithm for ML. The ML framework provides a powerful and flexible framework for estimating model parameters and testing interesting biological hypotheses.
- 1985: Hasegawa, Kishino and Yano gave a new statistical method based on maximum likelihood approach.
- 1989: Hasegawa, Kishino and Yano and Kishino and Hasegawa in 1990 developed a distance-based approach that allowed different rate parameters in different parts of the tree.



- 1997: Sanderson proposed a nonparametric method for estimating divergence times that assumes neither globally (single) nor locally (non-uniform) constant rates of molecular evolution.
- 1997: Yang and Rannala presented an improved version of their 1996 Bayesian method. Improved method by using Monte Carlo integration.
- 1998: Thorne, Kishino and Painter developed a maximum likelihood based Bayesian method to estimate, divergence times that can estimate date of evolutionary event in the absence of the constant rates.



- Thorne et al. and Sanderson method differ in some aspects: Thorne et al. allows different rates on different branches of the tree and assume that rates are constant on individual branches but Sanderson method extracts chronological information by optimizing.
- In their model:
  - Autocorrelation of rates between an ancestral branch and its descendant was given.
  - Distribution of rates for two branches on a bifurcating tree that directly emanate from the root.

- Posterior distribution for a data set.
- Adoption of Metropolis-Hastings algorithm to solve multiple integral over rate, time and constant.
- Allow the Markov chain to reach stationary, proposal of steps for rate, internal time node and constant.
- 1998: Rambaut and Bromham also gave a maximum likelihood approach to estimate divergence times. Rate constancy test were included (excluded that data for which rate heterogeneity is detected).

- 2001: Kishino, Thorne and Bruno extended the Bayesian techniques for estimating divergence times and explored their behavior via simulation.
  - Mean of the rates at the two nodes was 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.
- 2002: Thorne and Kishino proposed Bayesian techniques for estimating divergence times to analyze multiple gene sequences.
- 2003: Springer et al. studied the data set of Placental mammal using the Bayesian approach.

# Supported Studies

- 1997: Yang and Rannala analyzed DNA data set consisting of a segment of the mitochondrial genomes of human, chimpanzee, gorilla, orangutan, gibbon, macaque, squirrel monkey, tarsier, and lemur.
- They compared the results of empirical with hierarchical analysis and suggested that adding a second level prior for the birth-death rates did not change the posterior probabilities.
- 1998: Thorne, Kishino and Painter used data set of 31 amino acids sequences rather than DNA sequences from the *rbcL* chloroplast gene and set “Marchantia” sequence as a out-group.

- PAML package and JTT (Jones-Taylor-Thornton) model used under a hypothesis of molecular clock, fossil evidence was not used.
- The resulted posterior means of the normalized times to the root from their program (divtime) was closer but greater than the root depth.
- 2002: Thorne and Kishino took two tree topologies with 16 in-group and 1 outgroups taxa of 64 genes. They considered two cases
  - Constant rate of evolution over time.
  - Evolutionary rates changing over time.Constrained time put on one particular node.

- First tree analyzed with both cases but second analyzed with only first case. First tree gave a better result than the second one because the time interval between nodes on the first tree was evenly spaced.
- Same study shows the analysis of multigene using the same dataset of 64 genes.
- Performance of the multigene divergence time was relatively good when constancy of rates was assumed.
- The posterior means of the divergence time estimates for the in-group root was closer to their true value.
- The 95% credibility interval for the in-group root time when rates actually varied over time but was narrower when rates were forced to be constant.

- 2003: Springer et al. used data set of total 16,397 aligned nucleotide positions for 42 placental mammals and placed opossum as an out-group.
  - Timing of the placental mammal.
  - Extant placental orders originated and diversified before or after the Cretaceous-Tertiary (K/T) boundary.
- Branch lengths were estimated with the “estbranches” program and “divtime” estimated divergence times.
- Set 105 Myr for the mean of the prior distribution of the root of the ingroup tree.
- Their analysis supported the diversification of placental mammals before the K/T boundary.
- The oldest split within Primates was at 77 million years and splitting of rat-mouse at 16-23 million years ago.

# Multidivtime

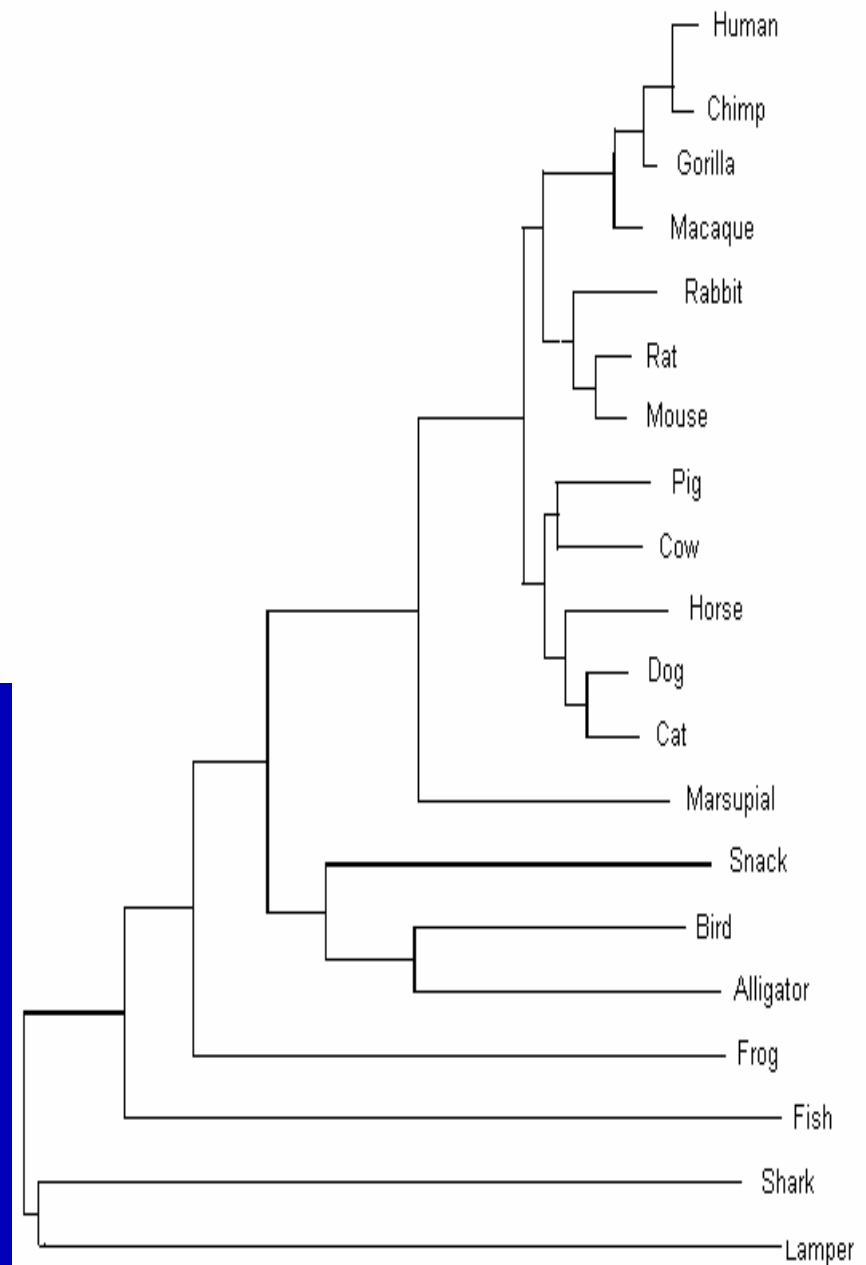
- Multidivtime software computes divergence times using the Bayesian approach.
- It has two main programs
  - ESTBRANCHES.
  - MULTIDIVTIME.
- ESTBRANCHES computes branch length of the given tree.
- MULTIDIVTIME computes estimated divergence times among lineages.

- HMMCNTL.DAT - control file for estbranches.
- MULTICNTL.DAT – control file for multidivtime.
- The user needs to provide in HMMCNTL.DAT
  - Name of the model file.
  - Name of the tree topology file.

- Some of the Important Features of MULTICNTRL.DAT

- Rttm
- Rttmsd
- Rtrate
- Rtratesd
- Brownmean
- Brownsd
- Bigtime

Lineages	True_time
Human - chimp	0.05
(Human, chimp) - gorilla	0.10
(Human, chimp, gorilla) - macaque	0.23
Rat - mouse	0.20
(Rat, mouse) - rabbit	0.85
Primate/Rodent	0.90
Dog - cat	0.46
(Dog, cat) - horse	0.74
Pig - cow	0.65
(Dog, cat, horse) - (Pig, cow)	0.81
Ferungulata	0.92
Marsupial/Placental	1.73
Chicken - alligator	2.22
(Chicken - alligator) - bird	2.76
Bird/Mammal	3.17
Amphibian/Amniote	3.60
Fish/Tetrapod	4.50
Shark	5.28



**It is showing actual time among species in million years and tree topology**

# 50 Combined Genes

- Executed the ESTBRANCHES program 50 times for 50 single genes.
- Output from ESTBRANCHES used in the MULTICNTRL.DAT (control file of multidivtime).
- Executed the MULTIDIVTIME to compute estimated divergence time.

# Effect of Different Priors

- *Time taken by the MULTIDIVTIME program:* Molecular clock case took half (approx 10 hrs) as much time as the non-molecular clock (approx 20 hrs).
- *Effect of bigtime:* Does not seem to have a major effect on estimated times except for very old divergences, where it can lead to significant overestimates.
- *Effect of upper and lower-bounds:* When wider ranges for upper and lower bounds were used, the confidence intervals changed proportionally.

- *Effect of some other priors:*
  - rttm (expected number of time units between tip and root) and rttmsd (standard deviation of prior for time units between tip and root) does not seem to have much effect on estimated time but we need an accurate value for rttm and rttmsd.
  - rtrate (mean of prior distribution for rate at root node) and rtratesd (standard deviation of prior for rate at root node) does have effect on estimated times bigger values of these priors give good estimate of time.
- *Effect of using multiple genes:* Combined genes showed that estimated times were closer to the true times, as compared to the individual time estimates.

# Perl Script

- There were about 1000 files of 50 genes in order to handle all those files. I have written a perl script, which can automatically execute the both programs ESTBRANCHES and MULTIDIVTIME and the final output was saved in a file.

# Conclusions

- Bayesian techniques are becoming interesting among biologist since it can work well when rates are varying among lineages.
- Bayesian technique computes divergence times among lineages using MCMC method.
- Some 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 in the final time estimates.
- The robustness of Bayesian approaches for divergence times needs to characterize better.

# 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 Dr. Alan Filipski for helpful discussion.



# Some References

- **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.
- **Hedges S. B. and Kumar S.**, Genomic clock and evolutionary timescales, 2003, review *Trends in Genetics* **19(4)**, 200 – 206.
- **Multidivtime Software:**  
<http://statgen.ncsu.edu/thorne/multidivtime.html>
- **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(12)**, 1647 – 1657.

