

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

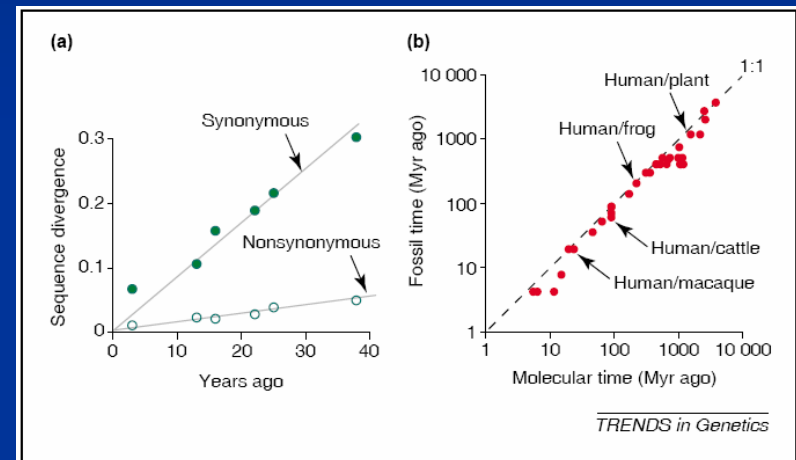
Placing confidence limits on the molecular age of the human-chimpanzee divergence

Sudhir Kumar ^{1,2}, Alan Filipski ^{1,2}, Vinod Swarna ¹, Alan Walker ³, & S. Blair Hedges ^{4,*}

1 Center for Evolutionary Functional Genomics, The Biodesign Institute, 2 School of Life Arizona State University, Tempe AZ, 85287-5301 USA; 3 Department of Anthropology, 4 Department of Biology, Pennsylvania State University, University Park, PA 16802 USA

The molecular clock idea

- First proposed by Zuckerkandl and Pauling (1965) based on haemoglobin data
- Sequences accumulate changes at a constant rate
- There is a linear relationship between sequence divergence (corrected for multiple hits) and time since divergence



S. Blair Hedges and Sudhir Kumar 2003

The Neutral Theory of molecular evolution

States that most mutations are either selectively neutral or nearly so.

Consider population of size N with a neutral mutation rate at a locus of λ mutations per gamete per generation

No. of new mutations = $\lambda \times 2N$

Probability of fixation by genetic drift = frequency, $p = 1/2N$

Number of new mutations per generation that are likely to become fixed by genetic drift = no. of mutations \times probability of fixation = λ

Molecular clocks

The rate of fixation of neutral mutations is equal to the mutation rate

Thus, the sequences diverge at a constant rate

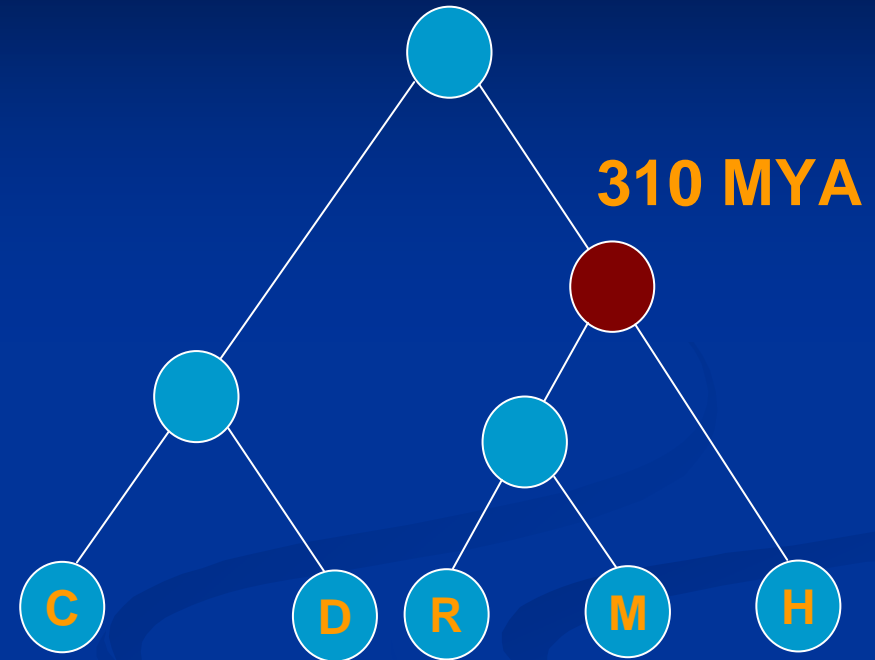
The divergence between two sequences can be used to say when the two organisms diverged from each other

But remember

- Not all mutations are neutral
- Not all loci change at the same rate
- Transitions are more common than transversions
- Rates are strictly based on generations (not years), and reproductive rates vary between species

■ Given

- a phylogenetic tree
- branch lengths (rt)
- a time estimate for one (or more) node



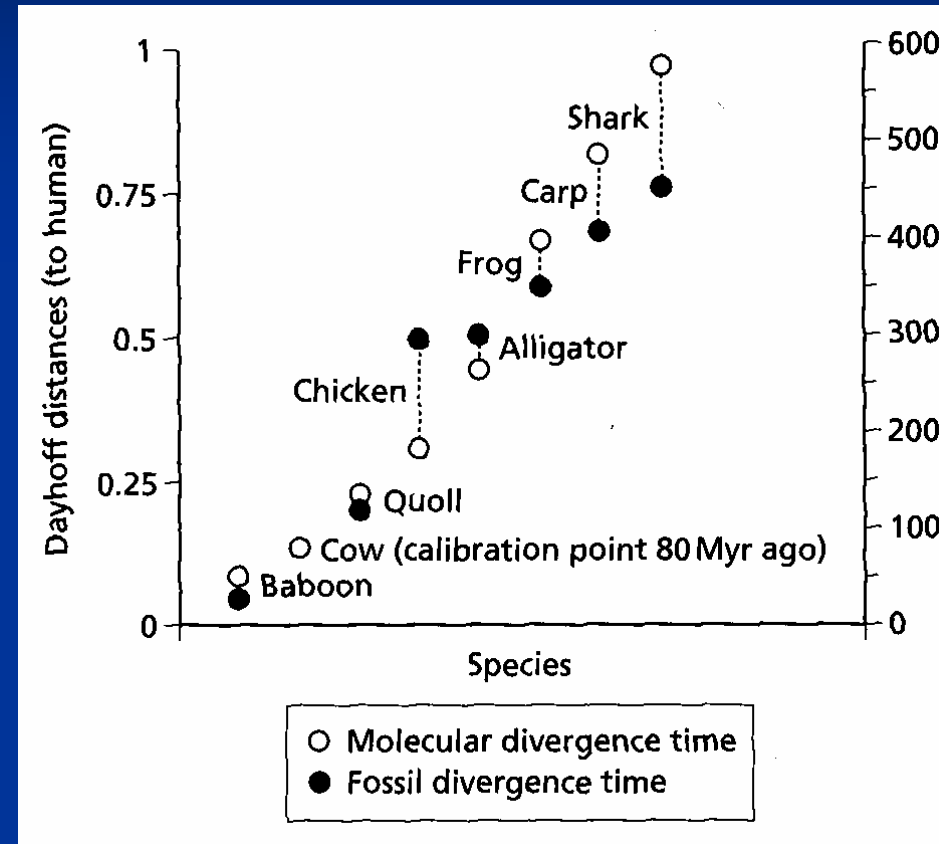
- Can we date other nodes in the tree?

Yes... if the rate of molecular change is *constant* across all branches

Rate Constancy in Hemoglobin gene

Amount of genetic difference between sequences is a function of time since separation.

Rate of molecular change is constant (enough) to predict times of divergence



Overview

- Methods for estimating time under a molecular clock
 - Estimating genetic distance
 - Determining and using calibration points
- Rate heterogeneity
 - reasons for variation
 - how its taken into account when estimating times

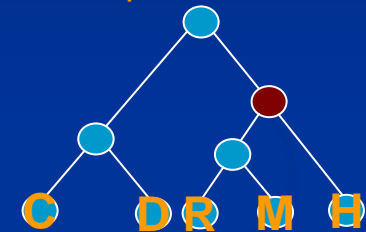
Time Estimation through Molecular clocks

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

$$d_{cal} = -\ln(1 - n/L)$$

Where n is the number of amino acid differences between the aligned sequences and L is the length of the ungapped alignment.

sequences and L



2. Rate of replacement is:

$$r = d/2T$$

Where T ; the time of divergence between the two sequences,

3. 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_{cal} = 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:

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

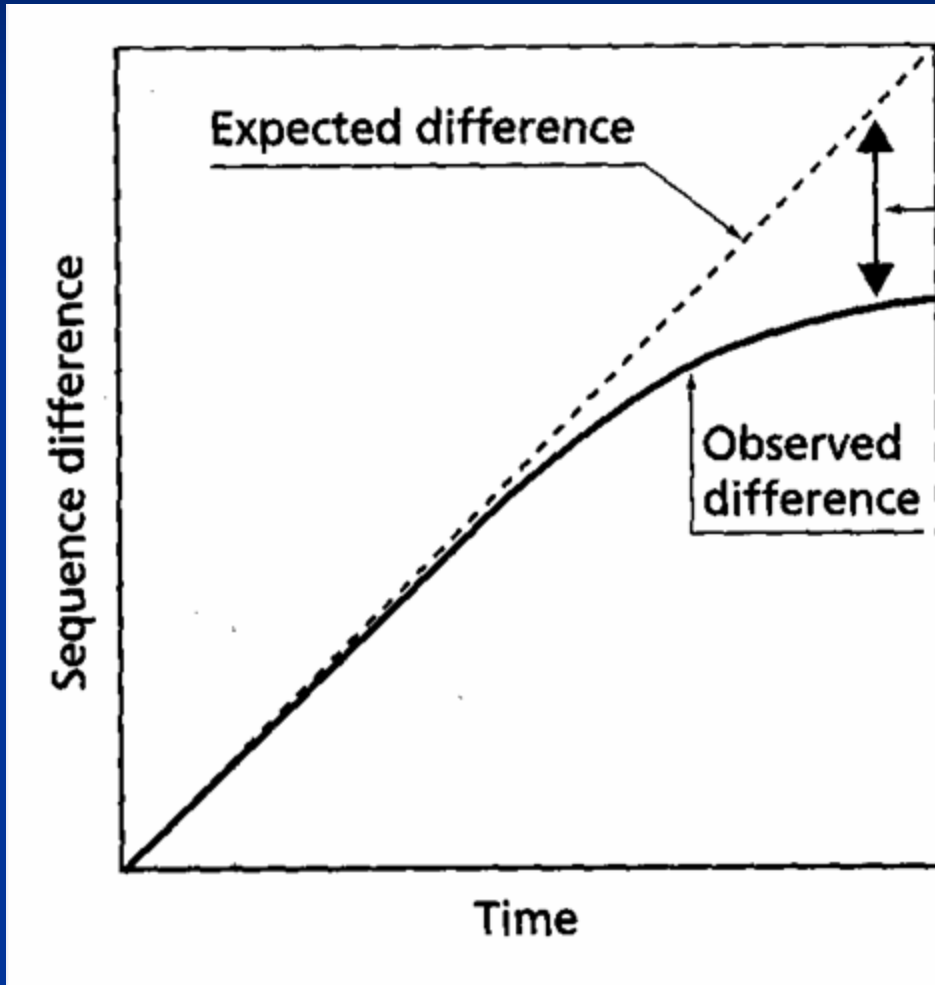
4. We can, then, take any pair of sequences from any two taxa, estimate d ; and calculate the time of divergence as:

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

Overview

- Methods for estimating time under a molecular clock
 - Estimating genetic distance
 - Determining and using calibration points
- Rate heterogeneity
 - reasons for variation
 - how its taken into account when estimating times

Estimating Genetic Differences



Simplest: p-distance

Simply counting differences underestimates distances

Fails to count for multiple hits

(Kumar & Nei p19)

Estimating Genetic Distance with a Substitution Model

- accounts for relative frequency of different types of substitutions
- allows variation in substitution rates between sites
- given learned parameter values
 - nucleotide frequencies
 - transition/transversion bias
 - alpha parameter of gamma distribution
- can infer branch length from differences

Distances from Gamma-Distributed Rates

- rate variation among sites
 - “fast/variable” sites
 - 3rd codon positions
 - codons on surface of globular protein
 - “slow/invariant” sites
 - Tryptophan (1 codon) structurally required
 - 1st or 2nd codon position when di-sulfide bond needed
- alpha parameter of gamma distribution describes degree of variation of rates across positions
- modeling rate variation changes branch length/
sequence differences curve

Overview

- Methods for estimating time under a molecular clock
 - Estimating genetic distance
 - Determining and using calibration points
- Rate heterogeneity
 - reasons for variation
 - how its taken into account when estimating times

Calibration Complexities

- Cannot date fossils perfectly
- Fossils usually not direct ancestors
 - branched off tree before (after?) splitting event.
- Impossible to pinpoint the age of last common ancestor of a group of living species

Overview

- Methods for estimating time under a molecular clock
 - Estimating genetic distance
 - Determining and using calibration points

- Rate heterogeneity
 - reasons for variation
 - how its taken into account when estimating times

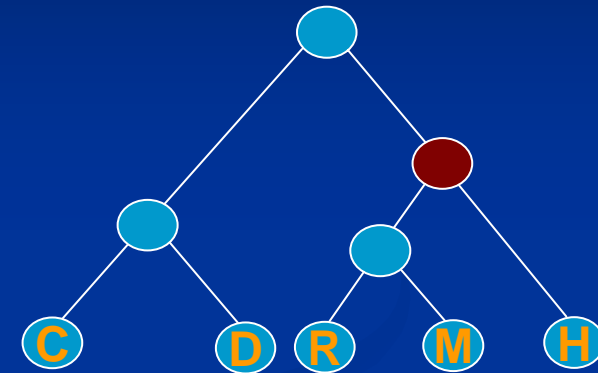
Rate Heterogeneity among Lineages

Cause	Reason
Repair equipment	e.g. RNA viruses have error-prone polymerases
Metabolic rate	More free radicals
Generation time	Copies DNA more frequently
Population size	Effects mutation fixation rate

Search for Genes with Uniform Rate across Taxa

Many 'clock' tests:

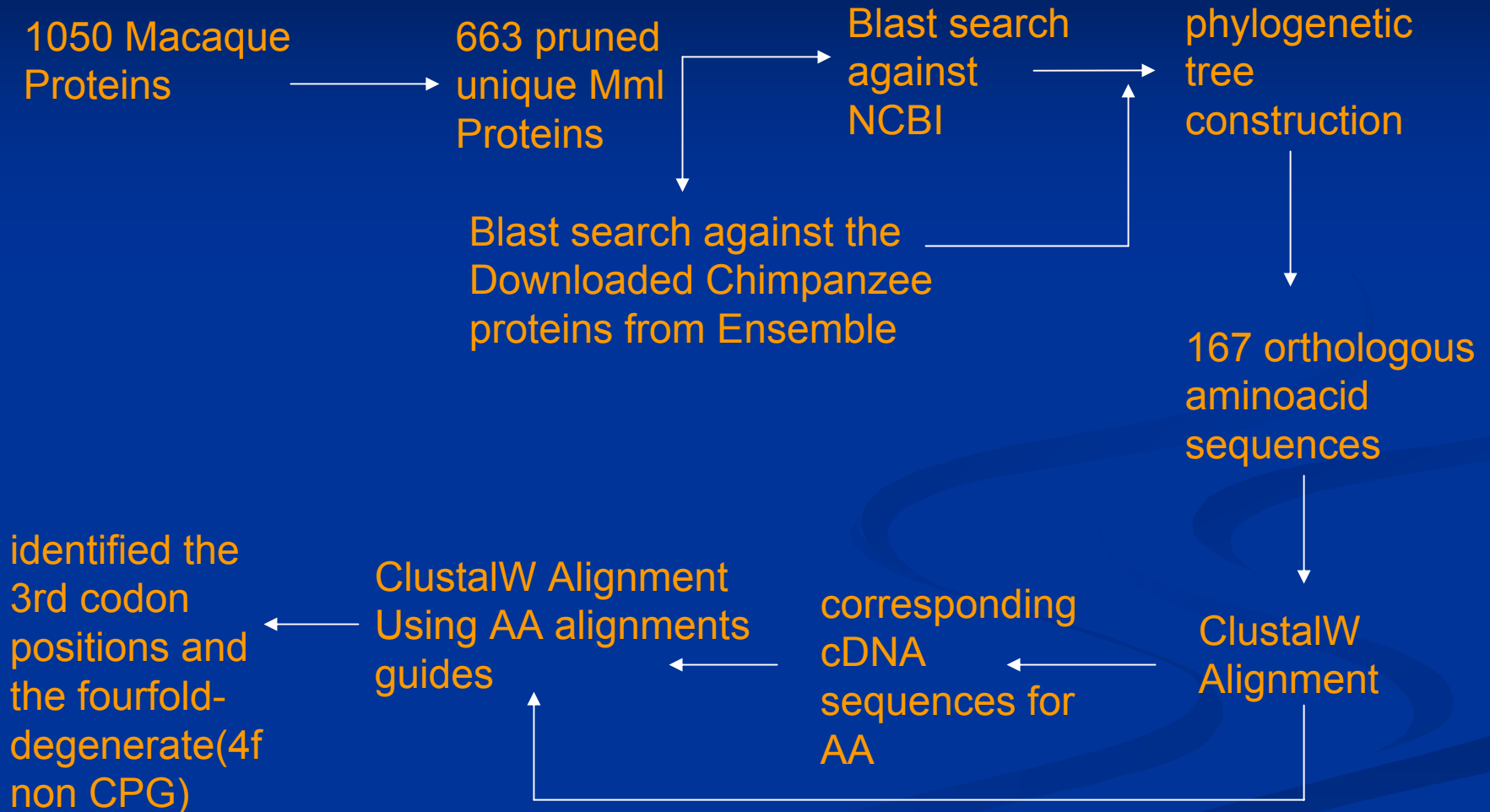
- Relative rates tests
 - compares rates of sister nodes using an outgroup
- Tajima test
 - Number of sites in which character shared by outgroup and only one of two ingroups should be equal for both ingroups



Molecular Dating Sources of Error

1. substitution model could be incorrect
2. tree could be incorrect
3. Lack of rate constancy (due to lineage, population size or selection effects)
4. Errors in orthology assignment
5. Stochastic variability
6. Imprecision of calibration points
7. Human sloppiness in analysis

Data Acquisition Flow Chart



Bayesian Method

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

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

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

$P(X|\theta, H)$ 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

Multidivtime Software

The Multidivtime software developed by Thorne et al. (1998)

- (1) ESTBRANCHES and
- (2) MULTIDIVTIME.

ESTBRANCHES

TESTSEQ

Model

HMMCNTRL.DAT

MULTIDIVTIME

MULTICNTRL.DAT

HMMCTRL.DAT

- `/* 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 hmmctrl.dat */`

MULTICNTRL.DAT

- gene.tree
- 1 ... number of genes ... FOLLOWING LINES CONTAIN ONLY NAMES OF DATA FILES
- oest.gene1
- 10000 ... numsamp: 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
- 0.000429 ... rtrate: mean of prior distribution for rate at root node
- 0.000429 ... rratesd: 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 */

MULTICNTRL.DAT

- 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 */
- 1 ... number of constraints on node times
- L 7 20
- 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

Make a choice

1. Single Gene Analysis
2. Multigene analysis

1

Is the sequence file

1. DNA file or
2. AminoAcid file

(File should be in Phylip format with an empty line at the end of the file)

1

Give the Name of Sequence file

(File should be in the same folder of the program being executed)

gene1

Give the Name of the tree file

(File should be in the same folder of the program being executed and should be in Phylip format with #Taxa)

gene.tree

Enter the RTTM Value (rttm is the mean of the prior distribution for the time separating the ingroup root from the present)

23.8

Enter the Big Time Value (bigtime is a number that is absolutely positively way bigger than the age of any node in the data set)

110

Please wait while multidivtime gives the node numbers for the given tree structure For imposing the calibration points.....

Please wait while multidivtime gives the node numbers for the given tree structure For imposing the calibration points.....

((Human:0,Chimp:1):3,Macac:2);

Enter the number of constraints you want to levy

1

Enter the node number

4

Enter Whether the constraint is UPPER or LOWER Example: U or L

L

Enter the Calibration time for the entered node number

23.8

Please wait while the Multidivtime prepares the results

output:

Input Sequence File: genel
Input Tree File : gene.tree
RTM : 23.8
BigTime : 110
Rrate : 0.000429
Constraints:
L 4 23.8

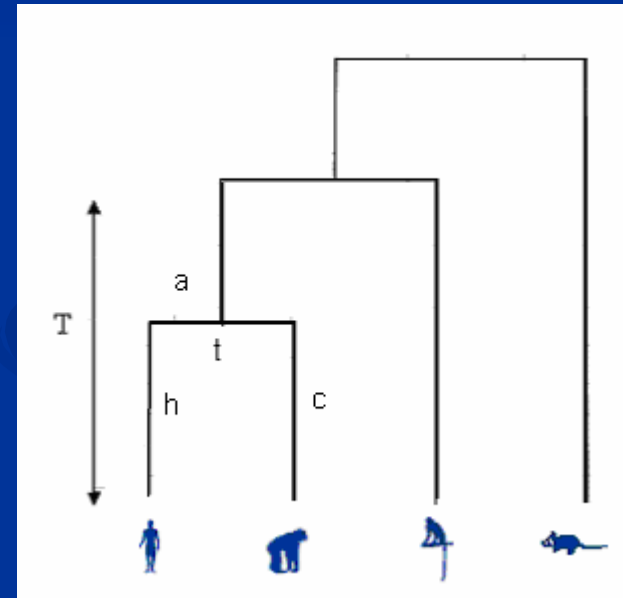
Actual time node 3 = 7.51731 (S.D. = 3.59639) (3.76770, 17.07026)

Actual time node 4 = 40.00025 (S.D. = 15.45772) (24.23958, 82.83113)

Multidivtime Output file out.txt can also be vewied for the above results

Estimation of divergence times

- (a) Phylogenetic relationships of four species.
- (b) The human-chimpanzee divergence time is given by the fraction $((h+c)/2)/(a+[h+c]/2)$ of the time assumed for GA-OWM divergence.
- (c) Ape-OWM 23.8 Calibration point



ML analysis

- ML distance method:
- Calibration of 23.8 Ape-OWM
- GTR + Γ model for DNA, JTT+ Γ model for AA
3rd codon position(53,008)
- 4.74 point estimate 95% CI 3.39 – 5.06
4fold non-CPG
- 4.75 Point estimate
AA
40 % higher point estimate. 95 % CI 4.78 - 8.93

Bayesian Analysis

- RTTM 23.8 MYR's
- F84+ Γ model for DNA, JTT+ Γ model for AA
3rd codon position(53,008)
- 4.98 point estimate Ratio 4.82 (24.00/4.98)
4fold non-CPG
- 5.17 Point estimate Ratio of 4.71(24.37/5.17)

Comparision of ML and Bayesian Results

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

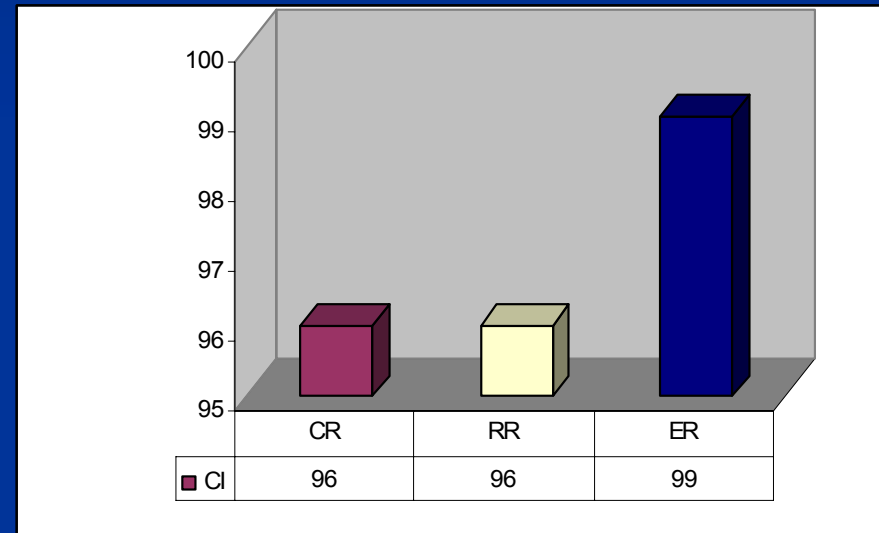
$$24.00/4.98 = 4.82 \text{ and } 24.34/5.12 = 4.75.$$

Multifactor Bootstrap Resampling (MBR) (Kumar et al)

1. Gene Resampling with replacement.
2. Sites Resampling
3. Random selection of lineage of time estimation
4. Random selection of the Calibration time from probability distribution.

Validation of MBR CI's (Kumar et al)

- Computer simulation
- Equal rate, Random rate and Correlated rates.
- MBR contained the true value $>95\%$ of the times.



Conclusion (Kumar et al)

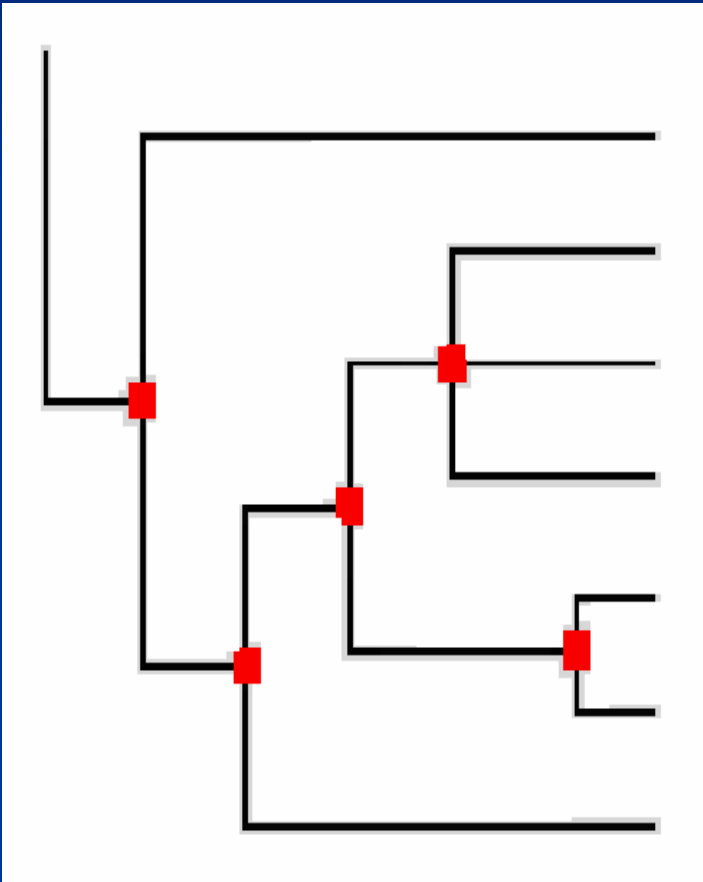
- Min Estimates of 4.74 - 4.98 < other studies

Reasons

- Type of data used,
- Calibration points,
- Number of genes.

Future Work

EFG Evolutionary tree
and divergence times



Future work

Divergence time estimation of other primates.

Phylogenetic Position of Artyodactyles

Tree building.

Divergence time estimation.

Bayesian analysis.

Acknowledgements



Dr. Rosemary Renaut



Questions ??

References

1. Kumar et al (2005) submitted PNAS.
2. Simons, E. L. & Pilbeam, D. R. (1965) *Folia Primat* 3, 81-152.
3. Sarich, V. M. & Wilson, A. C. (1967) *Science* 158, 1200-3.
4. Easteal, S. & Herbert, G. (1997) *J. Mol. Evol.* 44 Suppl 1, S121-32.
5. Kumar, S. & Hedges, S. B. (1998) *Nature* 392, 917-20.
6. Rose hoberman (<http://www-2.cs.cmu.edu/~roseh/Slides/durand03-molclock.ppt>)
7. Chen, F. C., Vallender, E. J., Wang, H., Tzeng, C. S. & Li, W. H. (2001) *J. Hered.* 92, 481-9.
8. Stauffer, R. L., Walker, A., Ryder, O. A., Lyons-Weiler, M. & Hedges, S. B. (2001) *J. Hered.* 92, 469-74.

Data Acquisition

- 1050 Macaque Genes from EOL Project. The largest collection of the Macaque Genes was present here.
- 663 unique 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 lesser than 10^{-10} by performing a blast search on GenBank for Human and mouse orthologous sequences.
- 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/>.
- 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).

- 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/>
- The thus obtained 167 orthologous aminoacid sequences were aligned using the default settings of clustalW. 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.
- As the CpG dinucleotides mutate 7-10 times (Subramanian, S. & Kumar, S 2000) faster than other dinucleotieds, the fourfold degenerate sites were separated into those that were involved with CpG dinucleotides and those that were not