

분기연대 베이스 추정 (Ver.25.01.02)

극지연구소
서태건(seo.taekun@gmail.com)

1 (한국진화학회 2025년도 겨울학교용 자료)

2 분자시계(molecular clock)

3 그림 1(A)와 같은 진화 시나리오로 진화해 온 가상의 염기서열 taxa 5개를 상상해보자. 단위 시간은 1백
4 만년(1MY)이다. 각각 8,6,4,2백만년 전에 종분화가 일어났으며 백만년당(per million years), 사이트당(per
5 site) 염기치환의 속도(=진화속도)는 0.05회이다. 진화속도는 계통수 전체에서 일정하다고 가정한다.¹ 이
6 다섯 개의 염기서열로부터 우리가 추정할 수 있는 계통수는 추정 오차를 무시할 수 있다면² 그림 1(C)와
7 같다. 각 계통수 가지위에 표시된 수치는 가지의 길이(=진화적 거리; 염기치환수/사이트)를 의미한다.

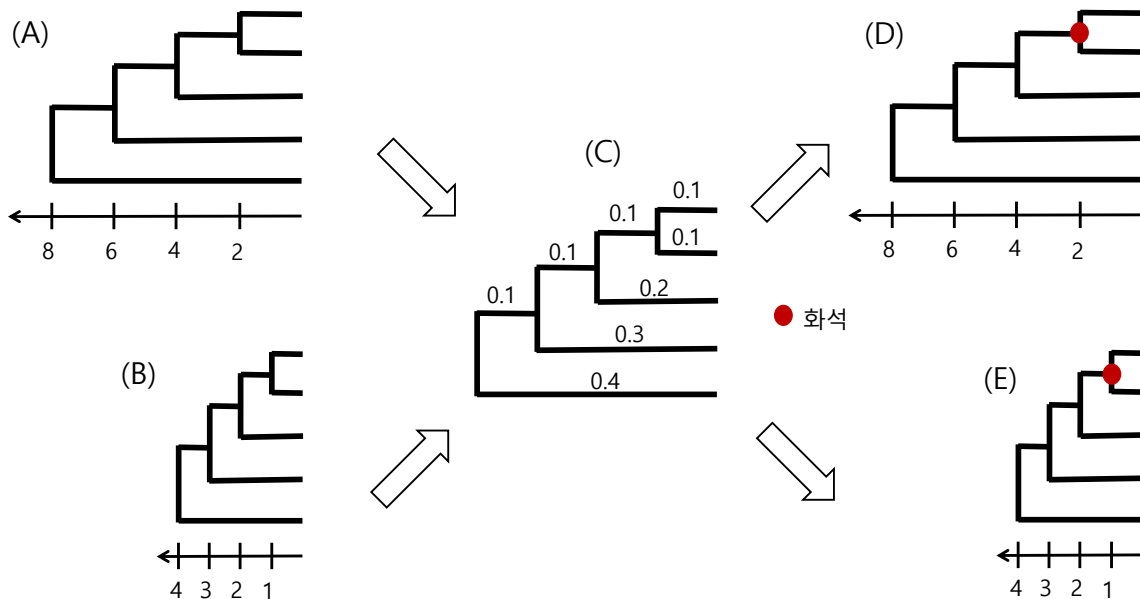


그림 1. Molecular clock과 화석데이터를 이용한 분기연대 추정 예시

8 또 다른 가상의 진화 시나리오를 생각해보자 (그림 1(B)). 이 시나리오에서는 종분화 연대가 각각 그
9 림 1(A)의 반이지만, 염기치환 속도는 두배이다. 따라서 이 진화 시나리오를 따라 진화한 염기서열로부터
10 얻어지는 계통수는 역시 그림 1(C)와 같다.

11 우리에게 다른 외부 정보(화석, 지질연대, 대륙의 분기연대 같은 정보)없이 동일한 시기³에 채취된

¹ 이를 분자시계(molecular clock)이라 한다.

² 사용된 치환모형이 옳고 염기서열의 길이가 무한대이면 오차없이 참값의 branch length를 추정할 수 있다.

³ 동일한 시기에 얻어진 염기서열 데이터를 contemporaneous sequences라고 한다. 데이터 취득에 약간의 시간차가 있게 마련이지만 전체의 타임 스케일에서는 차이가 미미하므로 동일한 시기에 채취된 샘플이라 볼 수 있다.

12 염기서열 데이터만 주어졌을때 우리가 추정할 수 있는 진화시나리오에는 그림 1(C)와 같은 진화적 거리가
13 명시된 계통수이다. 진화적 거리는 ‘단위 시간당 진화 속도’ \times ‘시간’의 관계로 주어지는데 염기서열 데이
14 터로부터 우리가 구할 수 있는 것은 이러한 곱셈의 결과인 진화적 거리이지 개별적으로 분리된 ‘진화속도’
15 와 ‘시간’이 아님에 주의한다. 그림 1(A)와 1(B) 이외에도 그림 1(C)를 생성할 수 있는 진화 시나리오에는
16 분자시계의 가정하에 무수히 많다. 그중에 어떤 시나리오로부터 염기서열 데이터가 생성되었는지는 염
17 기서열 데이터만으로는 알 수 없다.⁴

18 이제 우리에게 화석 데이터가 주어졌다고 생각해보자. 주어진 화석데이터가 2백만년 된 화석이고
19 이것이 가장 최근에 분기된 염기서열의 공동조상(most recent common ancestor; MRCA)인 경우를 상상
20 해보자. 이 화석 데이터는 그림 1(D)와 같이 계통수의 내부 노드(internal node)를 보정(calibration)하는데
21 사용될 수 있고 이를 통해 진화적 거리를 시간(=2MY)과 진화속도(=0.05)로 분리할 수 있다. 진화속도가
22 계통수 전체에 대해서 일정하다고 가정했으므로 다른 내부 노드들에도 같은 진화속도가 적용되어 분기
23 연대의 추정이 가능하게 된다. 발견된 화석데이터가 1백만년된 화석이라면 같은 방식으로 그림 1(E)과
24 같이 분기 연대를 추정할 수 있다.

25 화석 데이터의 적용

26 일반적으로 화석 데이터는 그림 1(D,E)와 같이 한 점으로 주어지지 않고 구간으로 정해지는 경우가 많다.⁵
27 화석 연구에는 주로 형태적 형질(morphological character)이 사용되는데, 예를 들어, 그림 2와 같이 taxon
28 A와 taxon B의 공통형질을 모두 가진 화석이 t_1, t_2 시점에서, taxon A만의 형질을 가진 화석이 t_3, t_4 시점에
29 서, taxon B만의 형질을 가진 화석이 t_5 시점에서 각각 발견되었다고 하자. 그렇다면 taxon A와 taxon B의
30 공동조상⁶은 t_2 이후 그리고 t_3 이전이어야 한다.⁷ 이처럼 화석 데이터를 이용한 보정점(calibration point)
31 은 하나의 점이 아닌 구간으로 주어지는 경우가 많다.

32 화석 데이터를 이용하여 계통수 내부노드 시간의 제한 범위가 정해진 후 분자시계⁸의 가정하에 분
33 기연대를 추정하는 과정을 그림 3에 모식적으로 나타내었다. 만약 화석데이터가 없다면 계통수의 뿌리
34 부분을 무한정 과거 방향으로 잡아당기거나 현재의 방향으로 밀수가 있다. 그 각각의 시나리오에 해당하
35 는 분기연대와 진화속도는 동일한 정도로 진화적 거리와 부합하기에 어느것이 진실인지 알 방도가 없다.
36 하지만, 일부노드에 제한이 걸리면 무한히 과거로 당기거나 현재로 밀수가 없다. 따라서 각각의 노드들
37 이 특정 시간의 범위내에 한정된다. 하단의 시간축에 파란색으로 표시된 root의 분기연대 범위가 주어진
38 화석데이터의 범위(빨간색)보다 넓음에 주목할 필요가 있다. 이처럼 분기연대 추정값은 과거로 갈수록
39 점점 더 불확실성이 증가해 신뢰구간의 폭이 넓어지는 경향이 있다.

⁴이를 identifiability problem이라고 한다.

⁵분기연대의 제한 구간을 정함에 있어 반드시 화석데이터만 사용되는 것은 아니다. 지질학적 정보나 대륙의 분기등의 정보도 구간을 정하는데 활용될 수 있다. 본 문헌에서는 설명의 편의상 이것들을 ‘화석 데이터에 의한 정보’라고 통칭한다.

⁶본 문서에서 가장 최근의 공동조상(most recent common ancestor; MRCA)을 편의상 ‘공동조상’이라고 칭한다. 단순히 공동조상을 의미하는지 MRCA를 의미 하는지는 문맥에 따라 판단하기 바란다.

⁷뒤에서 다룰 mcmctree 프로그램에서는 이를 ' $t_3 < t_2$ ' 혹은 ' $B(t_3, t_2)$ '로 표현하고 계통수 파일에 포함시킨다. 시간의 흐름은 역순으로, 현재시점 0으로부터 과거로 갈수록 시간이 증가하는 형식이다.

⁸Molecular clock 가정은 통상 진화속도가 “일정하다”고 가정하는 것이다. 진화속도의 일정함을 가정하지 않는 경우 ‘relaxed molecular clock’이라는 용어를 흔히 사용한다.

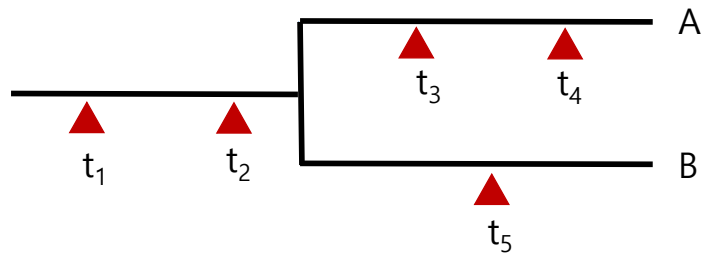


그림 2. 화석 데이터에 의한 MRCA의 제한. A와 B의 MRCA는 t_2 와 t_3 의 사이가 된다.

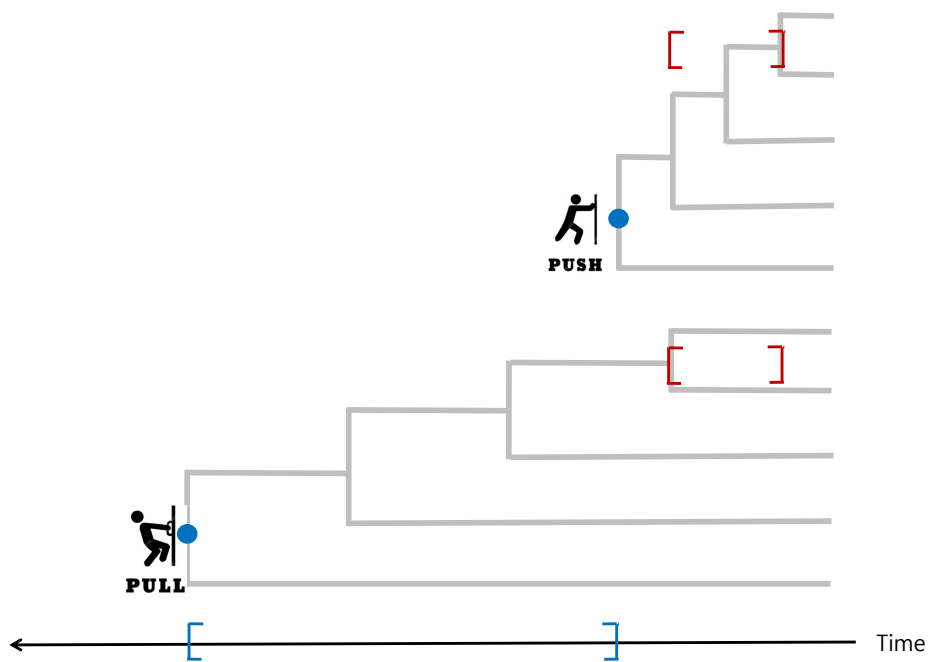


그림 3. 분자시계(Molecular clock)와 화석데이터를 이용한 분기연대 추정 모식도

40 화석데이터에 의해 내부노드 시간에 제한이 가해질 때 주어진 범위를 절대로 벗어날 수 없다는 강한
 41 제한을 hard bound라고 한다. 화석데이터의 연대 추정에는 여러가지 오차가 포함될 수 있기 때문에⁹ 주어진
 42 범위를 벗어날 수 있는 약간의 여지를 허용하기도 한다. 이를 soft bound라고 한다 (Inoue et al. 2010).
 43 흔히 그림 4와 같이 주어진 범위에서 좌우로 각각 2.5% 벗어날 확률을 허용하며 확률밀도는 코시분포
 44 형태로 감소한다.¹⁰

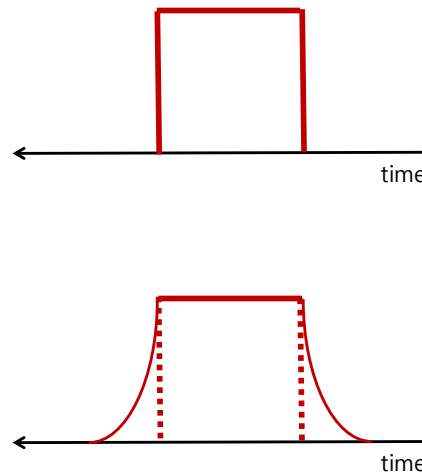


그림 4. Hard bound (위)와 soft bound (아래) 모식도

45 분기연대의 베이지 추정

46 그림 1은 진화속도가 변하지 않고 계통수 전체에 걸쳐 일정한, 즉 분자시계가 성립하는 가상적인 상황
 47 이다. 이 경우 염기서열 데이터만을 이용하여 계통수를 추정하면 그림 1(C)처럼 오른쪽 끝이 가지런히
 48 정렬된 상태를 보이게 된다.¹¹ 하지만, 실제 추정된 계통수의 오른쪽 끝은 상당한 편차를 보이므로 진화
 49 속도의 일정함을 가정하지 않는 분기연대 추정법이 필요하고 이 과정에서 베이지 방법론이 매우 유용한
 50 도구가 된다. 분자 계통수의 베이지 추정에 대해서는 이전 논문 (서태건 2024)에서 논의한 바 있다. 본
 51 문헌에서는 추정된 계통관계를 이용하여 분기연대를 추정하는 과정에 주목한다.

52 화석데이터 **C**와 염기서열 데이터 **X**가 주어졌을 때, 분기연대 **T**, 진화속도 **r** 그리고 각종 모형의 모수
 53 **θ** 의 사후 분포를 베이지 추정법으로 구하는 것이 분석의 목적이다. 베이지 정리를 이용하면 $(\mathbf{T}, \mathbf{r}, \boldsymbol{\theta})$ 의

⁹예컨대 그림 2에서 t_2, t_3 의 결정에 오차가 발생할 수 있다. 또한 t_2 지점의 화석이 A,B의 공동조상이 확실한지, t_3 지점의 화석이 A만의 조상이라고 확신할 수 있는지 있는지, 등등 화석데이터를 이용한 연구에는 여러가지 불확실성에 기인한 오차가 발생할 수 있다.

¹⁰mcmctree 프로그램에서는 디폴트로 soft bound를 상정한다. 따라서 그림 2의 노드 제한 ' $\mathbf{B}(t_3, t_2)$ '은 ' $\mathbf{B}(t_3, t_2, 0.25, 0.25)$ '으로 인식되며 hard bound를 사용하고자 한다면 ' $\mathbf{B}(t_3, t_2, 0, 0)$ '으로 설정할 수 있다.

¹¹실제로 완벽하게 분자시계가 성립할 경우라 해도 염기서열의 길이가 유한하기 때문에 오차가 발생, 오른쪽 끝이 가지런히 정렬되지는 않는다. 그 편차가 분자시계하에서 통상 보이는 정도의 편차인지 아니면 분자시계를 부정할 만한 심각한 편차인지는 가능도비 검정(likelihood ratio test; 서태건 2022)을 이용해 검정할 수 있다. 분자시계가 기각되는 경우가 거의 대부분이므로 가능도비 검정은 요즘은 거의 사용되지 않으며 분자시계를 가정하지 않고 곧바로 베이지 추정으로 분기연대를 추정하는 것이 일반적이다.

54 사후분포는 다음과 같이 구할 수 있다.

$$\begin{aligned}
 p(\mathbf{T}, \mathbf{r}, \boldsymbol{\theta} | \mathbf{X}, \mathbf{C}) &= \frac{p(\mathbf{T}, \mathbf{r}, \boldsymbol{\theta}, \mathbf{X} | \mathbf{C})}{p(\mathbf{X} | \mathbf{C})} \\
 &= \frac{p(\mathbf{X} | \mathbf{T}, \mathbf{r}, \boldsymbol{\theta}, \mathbf{C}) p(\boldsymbol{\theta}) p(\mathbf{r} | \mathbf{T}, \mathbf{C}) p(\mathbf{T} | \mathbf{C})}{p(\mathbf{X} | \mathbf{C})} \\
 &\propto p(\mathbf{X} | \mathbf{T}, \mathbf{r}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\mathbf{r} | \mathbf{T}, \mathbf{C}) p(\mathbf{T} | \mathbf{C})
 \end{aligned} \tag{1}$$

55 여기에서, $p(\mathbf{X} | \mathbf{T}, \mathbf{r}, \boldsymbol{\theta})$ 는 염기서열 정보만을 이용하는 가능도 함수, $p(\boldsymbol{\theta})$ 는 각종 모수의 사전확률 밀도함
 56 수, $p(\mathbf{r} | \mathbf{T}, \mathbf{C})$ 는 분기연대와 화석데이터가 주어졌을때 진화속도의 변화양상을 설명하는 확률밀도 함수,
 57 $p(\mathbf{T} | \mathbf{C})$ 화석데이터에 의해 제한이 가해지는 분기연대의 사전확률밀도 함수를 나타낸다.

58 진화속도 변화 모형

59 진화속도의 변화 모형은 식(1)의 $p(\mathbf{r} | \mathbf{T}, \mathbf{C})$ 를 설명하는 모형이며 여기서는 대표적인 두가지 모형에 대해
 60 살펴본다: Independent rates 모형, correlated rates 모형

61 (1) Independent rates 모형

62 Independent rates 모형은 계통수의 각 가지가 갖는 진화 속도는 어떤 확률분포로부터 독립적으로 얻어
 63 진 랜덤 샘플이라는 가정을 한다(그림 5). 흔히 로그정규분포(lognormal distribution) 혹은 감마분포를
 64 이용하여 모형화 한다.

65 로그정규분포를 이용한 진화속도 변화 모형에서는 진화 속도에 로그를 취한 값이 아래와 같이 평균
 66 μ , 분산 σ^2 인 정규분포를 따른다고 가정한다.

$$\log r \sim N(\mu, \sigma^2) \tag{2}$$

67 이 분포의 확률밀도함수는

$$f(r | \mu, \sigma^2) = \frac{1}{r\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{\log r - \mu}{\sigma} \right)^2 \right\}$$

68 이며 평균과 분산은

$$E[R] = \exp \{ \mu + \sigma^2/2 \} \tag{3}$$

$$\text{Var}[R] = \{ \exp(\sigma^2) - 1 \} E[R]^2 \tag{4}$$

69 가 됨이 알려져 있다(Johnson et al. 1994).

70 감마분포(김우철 2021)를 independent rates 모형화에 사용하기도 한다. 감마분포의 확률밀도함수는

71 모수 α, β 가 주어졌을 때 다음과 같고

$$f(r|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\beta r}$$

72 평균과 분산은 다음과 같다.

$$E[R] = \frac{\alpha}{\beta} \quad (5)$$

$$\text{Var}[R] = \frac{\alpha}{\beta^2} \quad (6)$$

73 (2) Correlated rates 모형

74 Correlated rates 모형¹²은 진화속도가 연속적으로 변하는 상황을 상정한 모형이다. 계통수의 내부노드
75 에 진화속도가 할당되며 이들끼리 상관관계를 유지하며 변화한다고 가정하는 모형이다(그림 6). 분산이
76 시간 간격에 의존하는 상관관계를 표현하기 위해 식 (2)를 변형시켜 아래와 같이 모형화 한다.

$$\log r_{i+1} \sim N(\log r_i, t_i \sigma^2) \quad (7)$$

77 즉, $\{i+1\}$ 번째 노드의 로그-진화속도의 평균은 직전 조상인 i 번째 노드의 로그-진화속도와 같고 분산은
78 시간 간격 t 에 비례한다. Root에서의 로그-진화 속도는 $\log r_0 = \mu$ 이다. 시간 간격 t_i 가 매우 작을 경우 분
79 산이 매우 작아져 $\log r_{i+1}$ 은 $\log r_i$ 과 매우 유사한 값을 갖게 된다. 이는 진화속도의 독립성을 가정하는 식
80 (2)과 대조되는 큰 차이이다. 시간 간격 t_i 가 매우 크면 증가된 분산이 상관관계를 감소시켜 두 노드에서
81 사뭇 다른 진화속도를 가질 수 있다.

82 각종 사전 분포들

83 계통수의 형상에 대한 사전분포는 식 (1)의 $p(\mathbf{T}|\mathbf{C})$ 에 해당하며 birth/death process(Karlin and Taylor 1975)),
84 coalescent theory(Kingman 1982)등을 이용해 정의할 수 있다. 또한 로그정규분포 (2)의 μ 와 σ^2 의 사전
85 분포를 정의할 때 감마분포를 사용할 수 있다. DNA 염기치환 모형의 $\{4 \times 4\}$ 치환율 행렬에 여러 모수를
86 포함하는 경우가 있는데 이에 대한 사전분포(식(1)의 $p(\boldsymbol{\theta})$ 에 해당)도 필요할 경우 적절히 설정해야 한
87 다. 여러 모수의 사전분포를 적절하게 정의하면 식(1)에 의해 사후분포가 정의되고 MCMC알고리즘을
88 이용하면 사후분포를 수치적으로 추정할 수 있다.

89 데이터 분석의 예

90 분자데이터를 이용하여 분기연대를 추정하는 대표적인 프로그램으로 MrBayes(Ronquist 2012), Beast2(Bouckaert
91 et al. 2019), mcmctree(PAML 프로그램 패키지의 일부; Yang 2007)등을 들 수 있다. 여기에서는 비교적 사
92 용법이 단순한 mcmctree프로그램에 대해 논의한다. PAML 프로그램 패키지에서 실행파일 mcmctree.exe,

¹²노드 사이의 거리가 가까우면 서로 진화속도가 비슷하고 거리가 멀면 진화속도가 크게 다를 수 있다는 자기상관관계(auto-correlation)를 강조하여 autocorrelated rates model이라고 부르기도 한다.

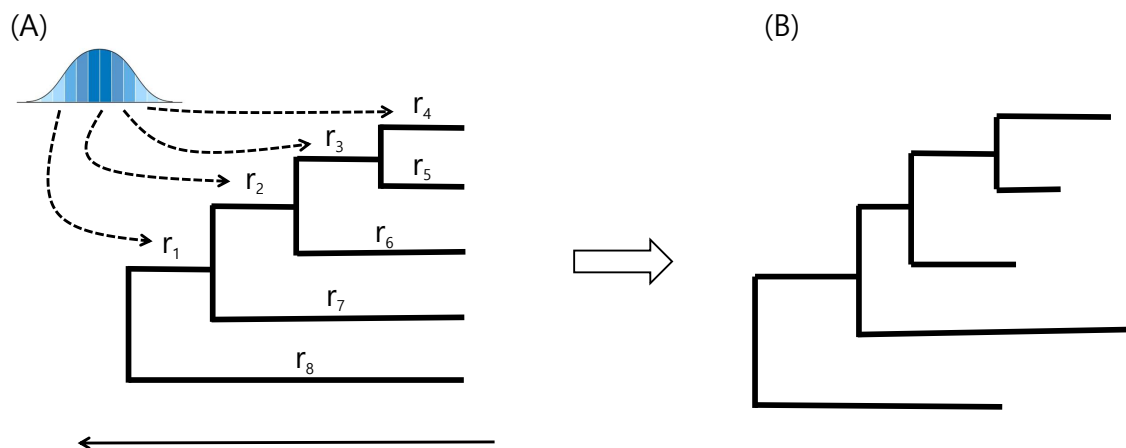


그림 5. Independent rates 모형

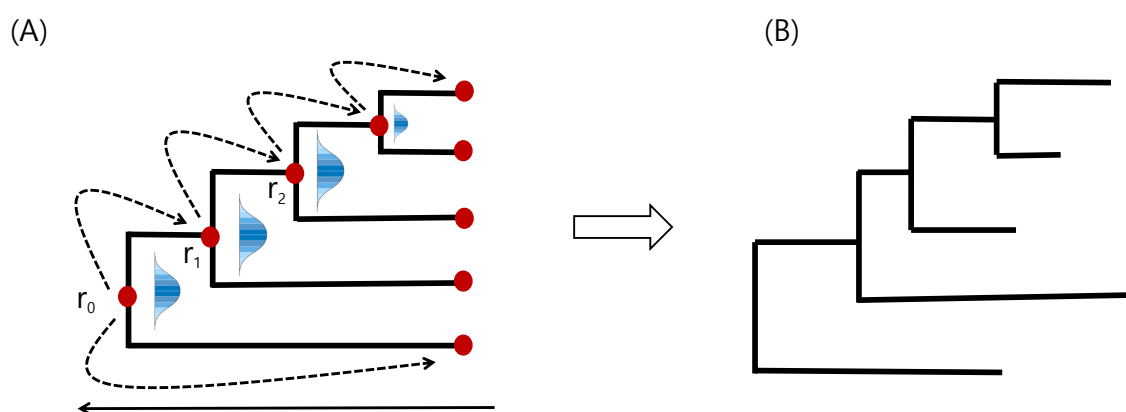


그림 6. Correlated rates 모형

93 baseml.exe과 염기서열 데이터파일 mtCDNApri123.txt, 관련 계통수 파일mtCDNApri.trees을 C:\temp 폴
 94 더 (혹은 적당한 폴더)에 복사한다. 윈도우즈OS에서 cmd.exe 명령어를 실행하여 명령프롬프트 실행한다.
 95 'cd C:\temp' 명령을 실행하여 해당 디렉토리로 이동한다.

96 그림 7은 mcmctree 프로그램이 디폴트로 제공하는 컨트롤 파일 mcmctree.ctl이다. 여기에서 seqfile,
 97 treefile 항목은 각각 염기서열 데이터, 계통수의 Newick포맷이 저장된 파일명을 지정하고, outfile항목은
 98 결과가 저장될 파일명을 지정한다. 본 학습에서는 디폴트로 지정된 염기서열 데이터와 계통수를 이용한다.
 99 다. 계통수 파일의 내부는 그림 8과 같다. Taxa 수와 계통수의 수를 첫줄에 입력하고 그 다음 줄에 Newick
 100 포맷의 계통수와 화석 정보를 기입한다. 입력하는 계통수에는 가지의 길이가 포함되면 안되며 화석정보
 101 에 의한 분기연대의 범위만 입력할 수 있다. 예를 들어 그림 8의 '>0.06<0.08'은 앞에서 언급한 soft bound
 102 로 human, chimpanzee, bonobo의 MRCA의 연대가 6백만년과 8백만년 사이임을 지정하는 것이다. 본
 103 데이터 분석에서 1단위시간은 1억년(=100백만년)으로 0.06은 6백만년에 해당한다.

```

seed = -1
seqfile = examples/DatingSoftBound/mtCDNApri123.txt
treefile = examples/DatingSoftBound/mtCDNApri.trees
outfile = out

ndata = 3
seqtype = 0 * 0: nucleotides; 1:codons; 2:AAs
usedata = 1 * 0: no data; 1:seq like; 2:use in.BV; 3: out.BV
clock = 3 * 1: global clock; 2: independent rates; 3: correlated rates
RootAge = <1.0 * safe constraint on root age, used if no fossil for root.

model = 0 * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
alpha = 0 * alpha for gamma rates at sites
ncatG = 5 * No. categories in discrete gamma

cleandata = 0 * remove sites with ambiguity data (1:yes, 0:no)?

BDparas = 1 1 0 * birth, death, sampling
kappa_gamma = 6 2 * gamma prior for kappa
alpha_gamma = 1 1 * gamma prior for alpha

rgene_gamma = 2 2 * gamma prior for overall rates for genes
sigma2_gamma = 1 10 * gamma prior for sigma^2 (for clock=2 or 3)

finetune = 1: 0.1 0.1 0.1 0.01 .5 * auto (0 or 1) : times, musigma2, rates, mixi

print = 1
burnin = 2000
sampfreq = 2
nsample = 20000

** Note: Make your window wider (100 columns) before running the program.

```

그림 7. mcmctree 프로그램이 제공하는 control file 예.

```

7 1
((((human, (chimpanzee, bonobo)) '>.06<.08', gorilla), (orangutan, sumatran)) '>.12<

```

그림 8. mcmctree treefile 예시

104 대용량의 염기서열 데이터를 mcmctree 프로그램으로 분석할 때 보통 두가지 단계로 진행된다. 첫번째
 105 단계는 가능도함수를 다변량 정규분포로 근사하는 과정이다. 염기서열이 많은 수의 파티션으로 이루어져
 106 그 길이의 총합이 매우 크거나 염기서열의 수가 많을 경우 가능도 함수의 계산에 많은 시간이 소요된다. 다
 107 변량 정규분포의 계산은 가능도 함수의 계산보다 훨씬 빨라 가능도 함수를 다변량 정규분포로 근사하는

108 기법(Thorne et al. 1998)이 널리 사용된다. 두번째 단계는 본격적으로 분기연대와 각종 모수의 사후분포를
109 구하는 과정이다. 이 과정에서 MCMC 알고리즘이 사용된다. 대용량의 염기서열 데이터의 경우 두번째
110 단계에서 많은 계산 시간이 소요된다.

111 첫번째 단계를 실행하기 위해 그림 9과 같이 수정하자. 본 실습은 mcmctree.exe 실행파일과 데이터
112 파일이 같은 폴더에 있는 것을 상정하여 파일의 경로명을 삭제하였다. 첫번째 단계의 작업은 ‘usedata=3’
113 으로 지정한다. 프로그램이 실행되면 out.BV파일이 생성되고 이 파일을 in.BV파일로 복사한 후에 두번째
114 단계 실행을 진행하는 것이다. 염기서열 진화모형은 HKY85, 사이트간 진화속도 모형은 다섯개의 범주를
115 가진 이산형 감마분포를 가정한다.¹³ 그림 9과 같이 수정이 끝났으면 해당파일을 ‘mcmctree.1.ctl.txt’로
116 명명하고 저장하자. 이후 그림 10와 같이 실행한다. 컨트롤 파일을 지정하지 않고 실행하면 디폴트 컨트롤
파일(mcmctree.ctl)이 사용되니 주의한다.

```

seed = 1
seqfile = mtCDNApri123.txt
treefile = mtCDNApri.trees
outfile = out

ndata = 3
segttype = 0 * 0: nucleotides; 1:codons; 2:AAs
usedata = 3 * 0: no data; 1:seq like; 2:use in.BV; 3: out.BV
clock = 3 * 1: global clock; 2: independent rates; 3: correlated rates
RootAge = <1.0 * safe constraint on root age, used if no fossil for root.

model = 4 * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
alpha = 0.5 * alpha for gamma rates at sites
ncatG = 5 * No. categories in discrete gamma

cleandata = 0 * remove sites with ambiguity data (1:yes, 0:no)?

BDparas = 1 1 0 * birth, death, sampling
kappa_gamma = 6 2 * gamma prior for kappa
alpha_gamma = 1 1 * gamma prior for alpha

```

그림 9. mcmctree.1.ctl.txt은 분석의 첫번째 단계를 실행하기 위한 컨트롤 파일이다. 붉은색 상자가 수정부분이고 수정이 필요없는 하단 부분은 표시를 생략하였다. 실행의 결과 out.BV 파일이 생성된다. 이를 복사하여 in.BV 파일을 만들고 두번째 단계의 분석에 사용한다.

117

```
C:\temp>mcmctree mcmctree.1.ctl.txt
```



그림 10. mcmctree 프로그램 실행의 첫번째 단계. out.BV 파일이 생성된다.

118 본 데이터는 3개의 파티션으로 이루어진 데이터이다(ndata=3 옵션에 주목). 1단계 실행이 끝나면 각
119 파티션의 baseml.exe 프로그램 실행 결과가 tmp0001.out, tmp0002.out, tmp0003.out 파일에 저장된다. 또한
120 관련된 baseml 프로그램의 컨트롤 파일(tmp000*.ctl)도 저장되므로 다른 추가적인 데이터 분석을 할 때
121 활용할 수 있다.

122 두번째 단계 실행을 위해서 out.BV파일을 복사하여 in.BV파일을 생성하고 그림 11과 같이 컨트롤
123 파일을 작성한 후 ‘mcmctree.2.ctl.txt’로 명명하고 저장하자. ‘usedata=2’ 설정으로 가능도 함수를 근사한
124 다변량 정규분포의 확률밀도함수(관련 정보가 in.BV에 저장)를 사용할 것을 명시한다. 분기연대 결과
125 조회를 용이하게 하기 위해 결과를 out.txt파일로 저장한다. Mcmctree 프로그램은 root노드의 분기연대

¹³HKY85 모형과 감마 진화속도 모형에 대해서는 이전 논문 (서태진 2022) 참조.

126 제한을 반드시 명시해 주어야 한다(본 데이터의 경우 'RootAge=<1.0' 으로 이미 설정된 것을 그대로 사
 127 용). MCMC 알고리즘의 burn-in step, sampling interval, total generation number는 지정된 설정을 그대로
 128 사용하기로 한다. 이후 그림 12과 같이 실행한다. 편집한 컨트롤 파일을 명시적으로 지정하여 프로그램을
 실행한다.

```

seed = 1
seqfile = mtCDNApri123.txt
treefile = mtCDNApri.trees
outfile = out.txt

ndata = 3
seqtype = 0 * 0: nucleotides; 1:codons; 2:AAs
usedata = 2 * 0: no data; 1:seq like; 2:use in.BV; 3: out.BV
clock = 3 * 1: global clock; 2: independent rates; 3: correlated rates
RootAge = <1.0 * safe constraint on root age, used if no fossil for root.

model = 4 * 0:JC69, 1:K80, 2:F81, 3:F84, 4:HKY85
alpha = 0.5 * alpha for gamma rates at sites
ncatG = 5 * No. categories in discrete gamma

cleandata = 0 * remove sites with ambiguity data (1:yes, 0:no)?

```

그림 11. mcmctree.2.ctl.txt은 분석의 두번째 단계를 실행하기 위한 컨트롤 파일이다. 붉은색 상자가 수정부분이고 수정이 필요없는 하단 부분은 표시를 생략하였다.

129

C:\temp>mcmctree mcmctree.2.ctl.txt

그림 12. mcmctre 프로그램 실행의 두번째 단계. 본격적으로 MCMC 알고리즘을 실행한다.

130 화면에는 그림 13과 같이 MCMC 알고리즘 진행상황이 표시된다. 제일 우측에 현재까지 소요시간,
 131 프로그램 종료까지 남은 시간이 표시된다. 본 데이터는 실습을 위한 간단한 데이터로 빠른 시간내에 종료
 132 된다.

5%	0.24	0.26	0.38	0.36	0.31	0.177	0.157	0.089	0.062	0.023	0.038	-0.292	0.843	-17.7	0:01
10%	0.24	0.26	0.38	0.33	0.31	0.176	0.156	0.088	0.062	0.023	0.038	-0.312	0.842	-17.2	0:01
15%	0.23	0.27	0.38	0.33	0.31	0.178	0.156	0.088	0.062	0.023	0.038	-0.331	0.826	-17.0	0:01
20%	0.23	0.27	0.38	0.32	0.31	0.177	0.156	0.088	0.062	0.023	0.037	-0.332	0.821	-16.9	0:01
25%	0.23	0.27	0.38	0.33	0.31	0.177	0.156	0.088	0.062	0.023	0.037	-0.335	0.831	-16.8	0:01
30%	0.23	0.27	0.38	0.33	0.31	0.177	0.156	0.088	0.062	0.023	0.037	-0.339	0.828	-16.8	0:02
35%	0.23	0.27	0.38	0.34	0.31	0.177	0.156	0.088	0.062	0.023	0.037	-0.350	0.836	-16.7	0:02
40%	0.23	0.26	0.38	0.34	0.31	0.178	0.156	0.088	0.062	0.023	0.038	-0.347	0.841	-16.8	0:02
45%	0.23	0.27	0.37	0.34	0.31	0.178	0.156	0.088	0.062	0.023	0.037	-0.353	0.835	-16.8	0:02
50%	0.23	0.26	0.37	0.34	0.31	0.178	0.156	0.089	0.062	0.023	0.038	-0.348	0.835	-16.8	0:02
55%	0.23	0.26	0.38	0.34	0.31	0.179	0.156	0.088	0.062	0.023	0.037	-0.351	0.832	-16.8	0:03
60%	0.23	0.26	0.37	0.34	0.31	0.179	0.156	0.088	0.062	0.023	0.037	-0.349	0.827	-16.8	0:03
65%	0.23	0.26	0.37	0.34	0.31	0.178	0.156	0.088	0.062	0.023	0.038	-0.347	0.832	-16.8	0:03
70%	0.23	0.26	0.37	0.34	0.31	0.179	0.156	0.088	0.062	0.023	0.038	-0.346	0.831	-16.9	0:03
75%	0.23	0.26	0.37	0.34	0.31	0.179	0.156	0.088	0.062	0.023	0.038	-0.349	0.825	-16.9	0:03
80%	0.23	0.26	0.37	0.34	0.31	0.179	0.156	0.088	0.062	0.023	0.038	-0.346	0.825	-16.9	0:03
85%	0.23	0.26	0.37	0.34	0.31	0.179	0.156	0.088	0.062	0.023	0.037	-0.348	0.826	-16.9	0:04
90%	0.23	0.26	0.37	0.34	0.31	0.180	0.156	0.088	0.062	0.023	0.038	-0.346	0.827	-16.9	0:04
95%	0.23	0.26	0.37	0.34	0.31	0.180	0.156	0.088	0.062	0.023	0.037	-0.349	0.829	-16.9	0:04
100%	0.23	0.26	0.37	0.34	0.31	0.180	0.156	0.088	0.062	0.023	0.037	-0.351	0.827	-16.9	0:04

그림 13. mcmc run screen

133 결과 파일 out.txt 에는 미지의 파라미터에 대한 사후분포 정보가 저장되어 있다. 파일 하단 부분에는
 134 그림 14과 같이 각 노드별 분기연대 사후분포의 평균, 95 % 신뢰구간이 표시된다 (t.n* 부분). mu1 ~ mu3,
 135 sigma2.1 ~ sigma.3은 correlated rates 모형(clock=3 옵션)하에서 세 파티션의 root노드에서의 진화속도의
 136 평균과 분산의 사후분포의 95 % 신뢰구간을 의미한다 (식 (7)).

137 두번째 단계의 분석이 끝나면 mcmc.txt 파일에는 MCMC 알고리즘 실행중 샘플링된 모수들이 세대

```

Posterior mean (95% Equal-tail CI) (95% HPD CI) HPD-CI-width
t_n8      0.1796 ( 0.1566, 0.2140) ( 0.1552, 0.2108) 0.0556 (Jnode 12)
t_n9      0.1562 ( 0.1446, 0.1624) ( 0.1464, 0.1635) 0.0170 (Jnode 11)
t_n10     0.0885 ( 0.0807, 0.0976) ( 0.0803, 0.0970) 0.0167 (Jnode 10)
t_n11     0.0620 ( 0.0588, 0.0679) ( 0.0584, 0.0671) 0.0087 (Jnode 9)
t_n12     0.0232 ( 0.0186, 0.0284) ( 0.0184, 0.0280) 0.0096 (Jnode 8)
t_n13     0.0374 ( 0.0292, 0.0460) ( 0.0294, 0.0461) 0.0167 (Jnode 7)
mu1       0.5149 ( 0.3979, 0.6417) ( 0.4033, 0.6465) 0.2432
mu2       0.1761 ( 0.1295, 0.2257) ( 0.1301, 0.2259) 0.0958
mu3       5.6300 ( 3.9849, 7.5471) ( 3.9429, 7.4926) 3.5497
sigma2_1  0.2410 ( 0.0068, 0.7363) ( 0.0017, 0.6158) 0.6141
sigma2_2  0.3507 ( 0.0300, 0.9282) ( 0.0039, 0.7930) 0.7891
sigma2_3  0.8274 ( 0.2998, 1.6802) ( 0.2646, 1.5698) 1.3052
lnL       -16.8512 (-25.1920, -10.3060) (-24.4020, -9.7220) 14.6800

```

그림 14. 두번째 단계 분석 이후 out.txt 파일 하단에는 내부 노드의 분기연대 사후분포의 평균 및 95 % 신뢰구간이 표시된다. 내부 노드의 위치에 관한 정보는 out.txt 파일내에서 찾을 수 있다.

138 (generation)별로 저장된다(그림 15). 이를 Tracer 프로그램으로 열어보자(메뉴에서 File/import trace file/을
 139 선택한후 mcmc.txt 지정). 그림 16과 같이 각 모수별로 사후분포로부터 얻는 샘플의 변화 양상을 살펴볼
 140 수 있다.

Gen	t_n8	t_n9	t_n10	t_n11	t_n12	t_n13	mu1
1	0.1828834		0.1615101		0.0876872		0.0614
2	0.1828834		0.1526452		0.0854514		0.0614
4	0.1828834		0.1526452		0.0854514		0.0608
6	0.1828834		0.1582246		0.0879473		0.0608
8	0.1847655		0.1518980		0.0924822		0.0637
10	0.1847655		0.1607646		0.0910202		0.0608
12	0.1667355		0.1555069		0.0868780		0.0608
14	0.1840996		0.1555069		0.0936563		0.0608
16	0.1840996		0.1484759		0.0872715		0.0608
18	0.1840996		0.1559565		0.0930122		0.0643
20	0.1702507		0.1504006		0.0930122		0.0660
22	0.1702507		0.1504006		0.0888878		0.0660
24	0.1618147		0.1528821		0.0888878		0.0627
26	0.1618147		0.1528821		0.0888878		0.0654
28	0.1618147		0.1528821		0.0956387		0.0646
30	0.1614548		0.1528821		0.0956387		0.0622
32	0.1739010		0.1604972		0.0956387		0.0622

그림 15. mcmc.txt 파일(디폴트로 설정되어 있음)에는 사후분포로부터 얻는 모수들의 샘플이 저장되어 있다.

141 디폴트로 지정된 figtree.tre 파일에는 분기연대의 사후분포에 대한 정보가 저장된다. 이를 FigTree프
 142 로그로 열어보자(그림 17). 좌측 패널에서 Node Bar 항목에 체크하고 하부의 95%HPD를 선택한다.
 143 또한, Scale Axis 에 체크하고 하부의 Reverse axis 항목을 체크한다. 그외 취향에 따라 폰트의 크기나
 144 계통수 가지의 굵기등을 조절하면 보기 좋은 그림을 얻을 수 있다.

145 고찰

146 화석과 염기서열 데이터의 역할

147 그림 17은 가능도를 정규분포로 근사하고(usedata=2 옵션) MCMC 알고리즘을 적용시켜 얻는 분기연대
 148 결과이다. 염기서열 데이터의 정보가 정규분포의 형태로 근사되어 분석에 반영된 것이다. Mmctree 프로
 149 그램은 염기서열 데이터의 정보를 전혀 반영하지 않고 사전분포만으로 형성되는 분기연대 분포를 살펴
 150 볼 수 있다. 이를 통해 사전분포가 적절한가, 염기서열 데이터가 어떤 차이를 가져 오는가 알아 볼 수 있다.
 151 다른 설정은 모두 동일하게 하고 usedata 설정만 'usedata=0'을 지정하고 두번째 단계의 분석을 실행하자.

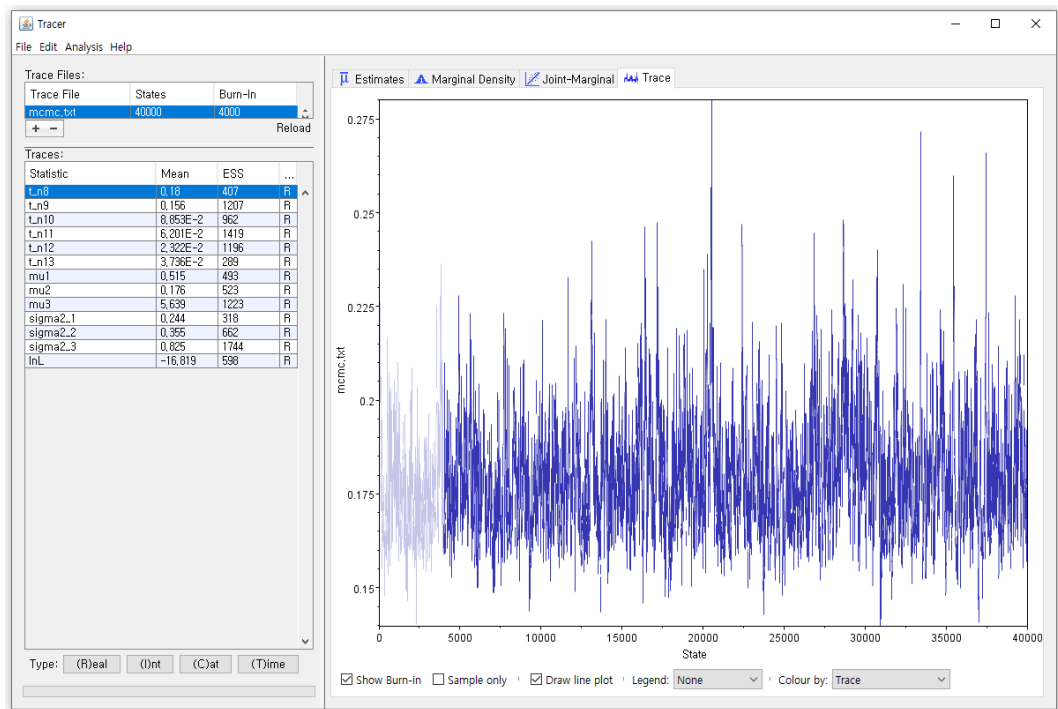


그림 16. mcmc.txt 파일을 Tracer 프로그램으로 읽어들이는 화면.

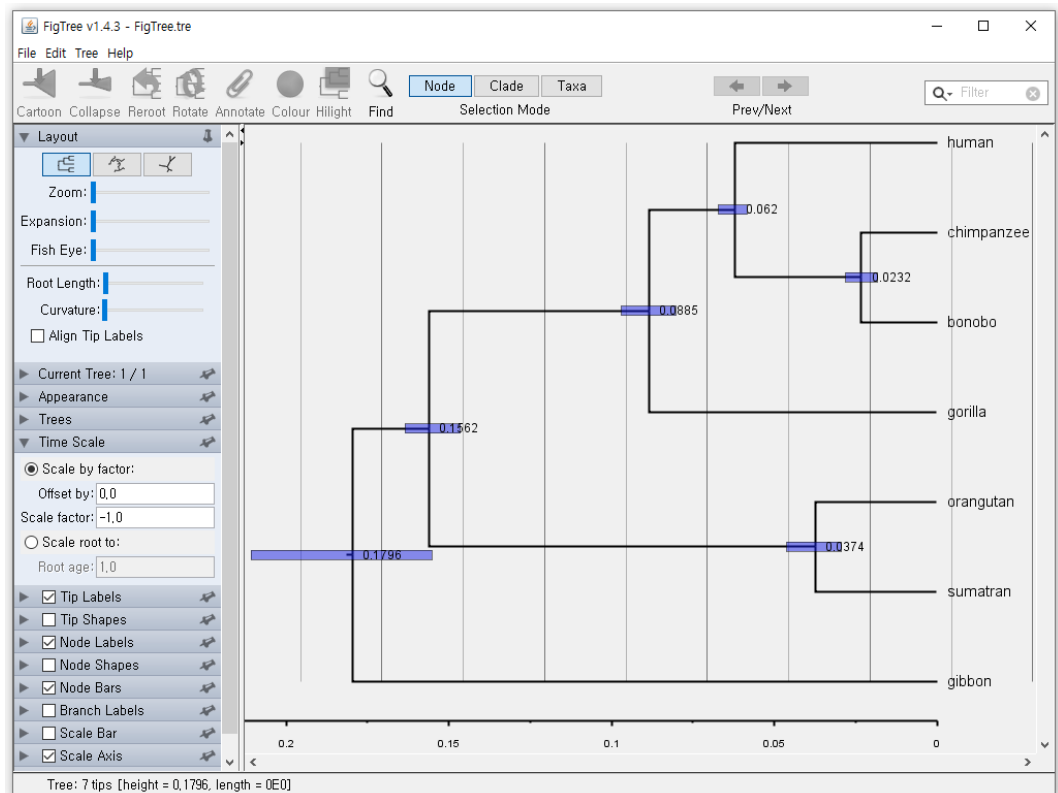


그림 17. 두번째 단계 실행후 생성된 figtree.tre 파일을 FigTree프로그램으로 읽어 들인 화면 모습. 각 내부노드에 분기연대의 사후분포 평균과 95 % 신뢰구간이 표시되어 있다.

152 이는 사전분포와 화석데이터만 정보로서 입력하고 실행한 것이다. 그 결과는 그림 18과 같다. 염기서열
 153 데이터를 포함한 분기연대 추정값의 신뢰구간 폭(그림 17)이 이를 포함하지 않은 추정값의 신뢰구간 폭
 154 (그림 18)에 비해서 좁은 것을 알 수 있다. 특히 뿌리 부분은 신뢰구간의 폭뿐만 아니라 사후분포의 평균
 155 값도 매우 다르다.

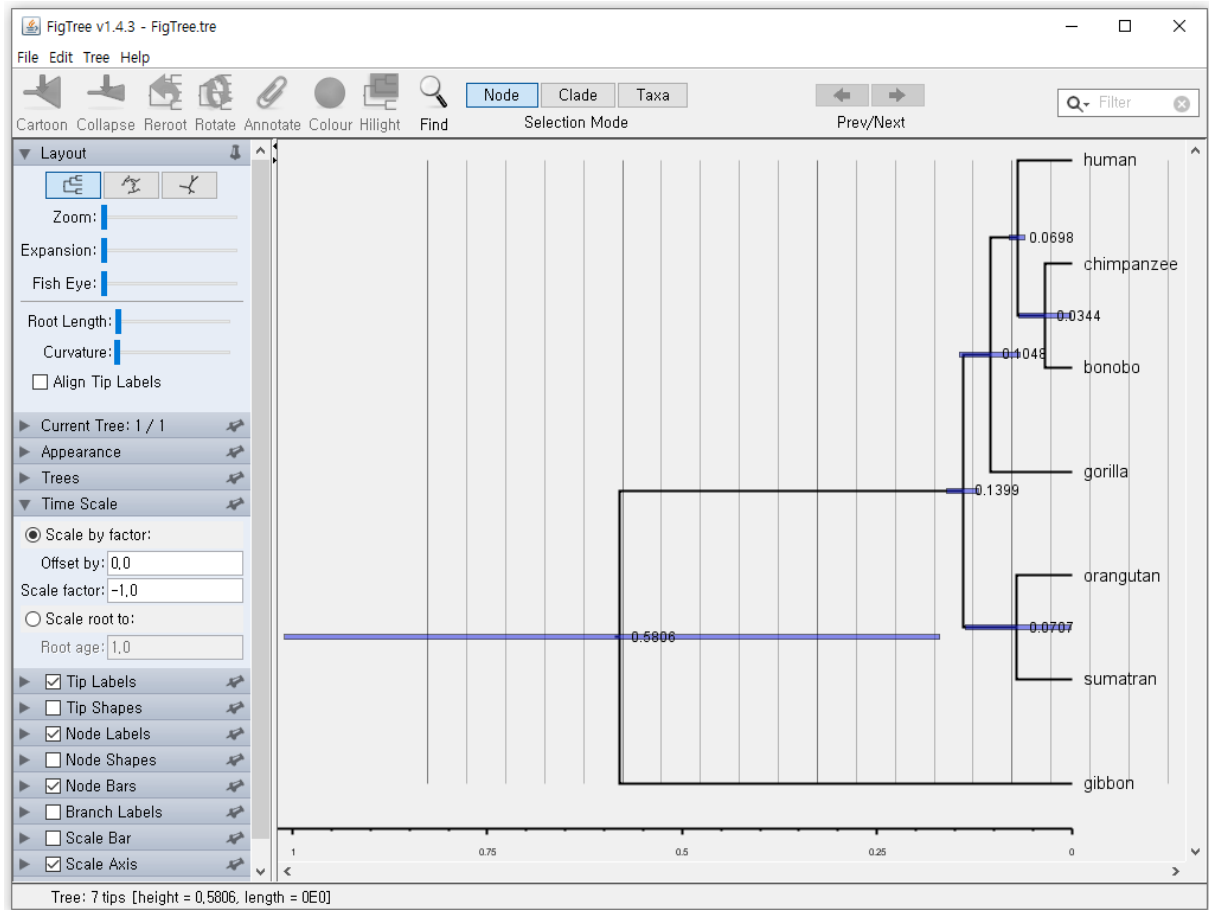


그림 18. 화석정보는 포함하나 염기서열데이터는 포함하지 않은 분기연대 추정값.

156 추가로 분기연대 추정에 있어서 화석의 역할을 살펴보자. 그림 18은 염기서열 정보는 포함하고 있지
 157 않지만, 화석의 정보는 포함하고 있는 결과이다. 화석 정보를 포함하고 있는 mtCDNApri.trees 파일에서
 158 화석 정보를 삭제한¹⁴ 후 ‘usedata=0’을 지정하고 두번째 단계의 분석을 실행하자. 분기연대의 사후 분포
 159 는 그림 19과 같다. 화석정보를 포함한 그림 18의 신뢰구간의 폭이 화석 정보를 포함하지 않는 그림 19
 160 보다 좁다. 본 데이터의 경우 화석데이터는 human-chimpanzee의 MRCA를 600만년~800만년, gibbon을
 161 제외한 모든 taxa의 MRCA를 1200만년~1600만년 사이로 soft bound를 이용하여 제한하고 있다. 따라서
 162 그림 18에서는 root의 MRCA가 과거로 멀리 갈 수 없는 반면, 그림 19에서는 이러한 화석에 의한 제한이
 163 없어 그만큼 더 root의 위치가 과거로 갈 수 있고 다른 MRCA들도 더 과거로 갈 수 있는 것이다. 즉, 계통
 164 수에 포함된 화석정보들은 현재시점에 매우 가까운 것들이고 이것들이 root의 분기연대를 현재 방향으로
 165 당기는 효과를 가져왔다고 볼 수 있다.

¹⁴화석정보를 삭제 할 경우 컨트롤 파일에서 RootAge는 양방향 제한이 필요하므로 임의로 ‘RootAge = <1.0 >0.01’을 설정했다.

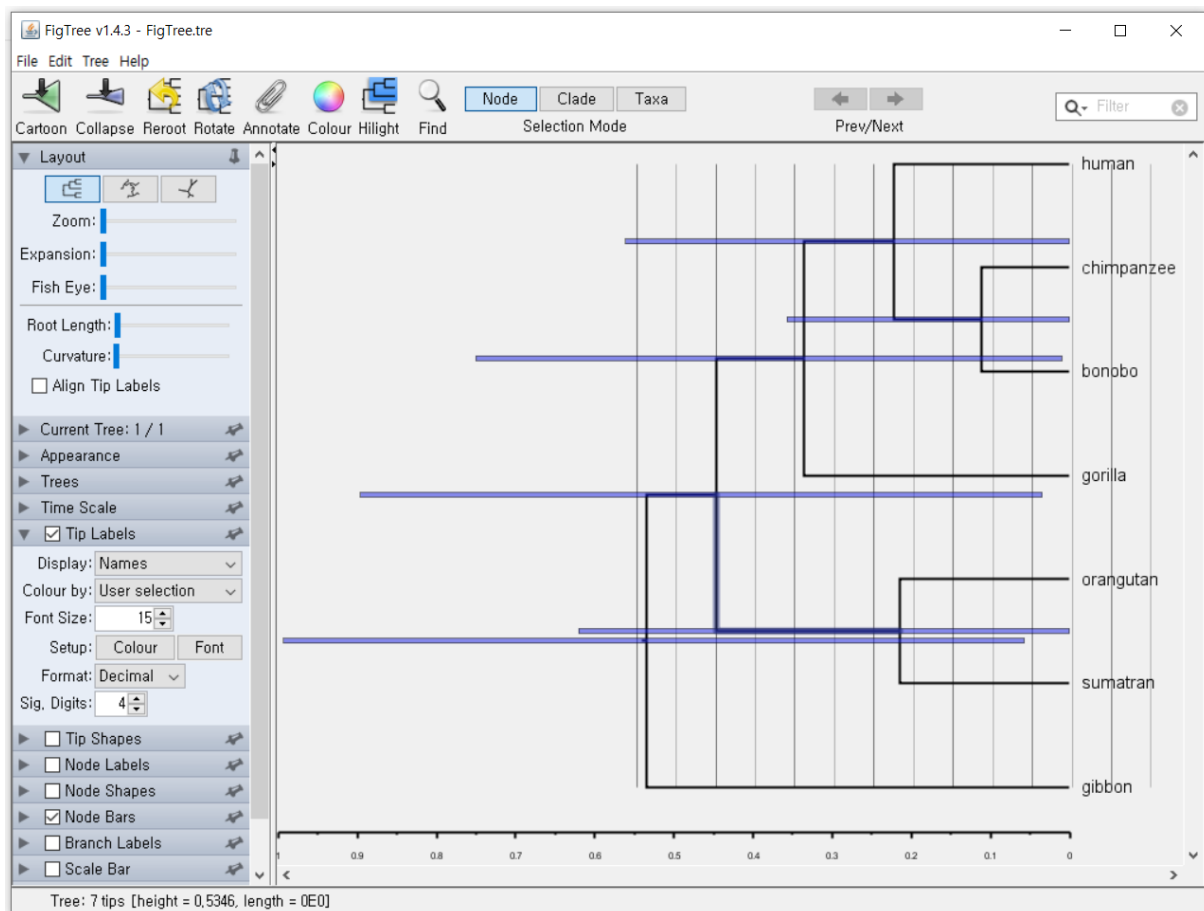


그림 19. 화석정보와 염기서열 정보 모두 포함하지 않았을때의 분기연대 사전분포.

166 **1단위 시간의 변경**

167 위 분석 예시에서 1단위 시간은 1억년(=100백만년)이다. 주어진 7종만을 대상으로 할때는 1단위 시간을
168 1백만년으로 하는 것이 편리할 것이다. 단위 시간을 1/100배로 한다면 컨트롤 파일의 'RootAge=' 항목과
169 계통수내에 포함된 화석정보의 단위에 각각 100을 곱해야 한다. 또한, 식 (2,7)의 (μ, σ^2) 의 사전분포도
170 시간단위의 변화에 따라 적절하게 변경하는 것이 바람직하다.¹⁵ 이때 진화속도 변화 모형에 따라 변경
171 방식에 차이가 있으니 주의가 필요하다.

172 진화속도가 독립적으로 변화한다고 가정하는 모형(independent rate 모형)하에서는 로그노말 분포의
173 분산(식 4에 포함된 σ^2)가 시간 간격에 의존하지 않으므로 그림 20과 같이 μ 에 대한 사전분포 시간단위만
174 변경하면 된다. Mcmctree 프로그램은 식 (5)의 감마분포를 μ 와 σ^2 의 사전분포로 사용하고 있다. μ 에 대한
175 사전분포 시간단위만 변경하면 되므로 'rgene_gamma' 항목의 α, β 모수 중 β 에 해당하는 부분만 100배를
176 해준다. 또한, birth/death process의 관련 모수 (BDparas)도 단위시간이 줄어드는 만큼 1/100배를 해야한다.

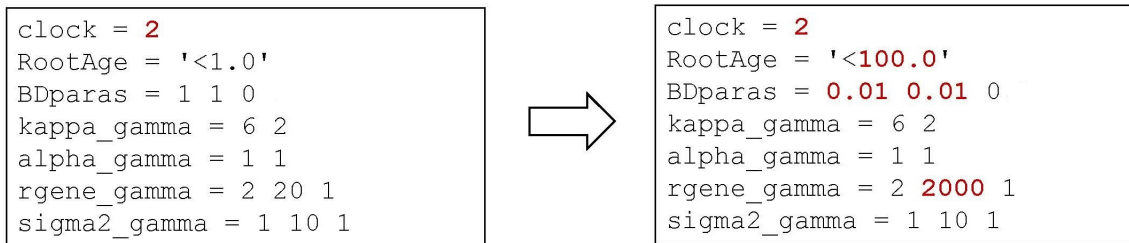


그림 20. independent rate 모형

177 반면, 진화속도가 상관관계를 가지며 변화한다고 가정하는 모형하에서는 식 (4)에 포함된 σ^2 도 시
178 간간격에 의존하므로 σ^2 의 사전분포의 시간단위도 함께 변경해 주어야 한다. 따라서 σ^2 의 사전분포로
179 상정한 감마분포의 β 에 해당하는 부분까지 100배를 해준다(그림 21).

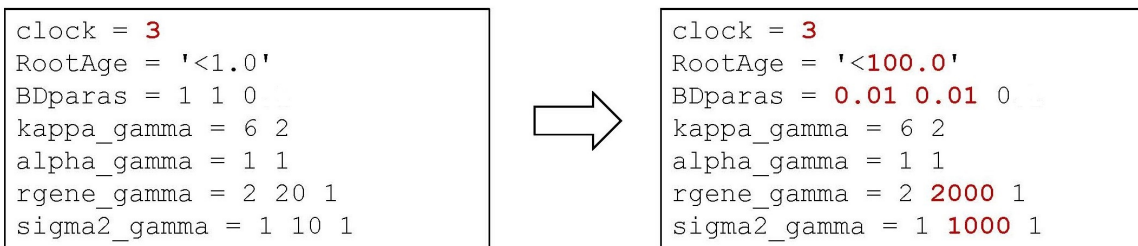


그림 21. Correlated rate 모형

180 **결론**

181 (작성중)

¹⁵원칙적으로 데이터가 충분이 많으면 사후분포는 사전분포에 크게 영향을 받지 않는다(서태진 2024). 하지만 실제 데이터 분석에서는 사전분포에 영향을 받는 경우가 많아 사전분포의 결정에 주의가 요구된다.

참고문헌

- 182
- 183 Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., et al. 2019.
- 184 BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational*
- 185 *biology*, 15(4), e1006650.
- 186 Inoue J, Donoghue PCH, Yang Z. 2010. The impact of the representation of fossil calibrations on Bayesian
- 187 estimation of species divergence times. *Syst Biol* 59:74-89.
- 188 Johnson N.L., Kotz S. Balakrishnan N. 1994. Continuous univariate distribution, Vol 1. Wiley. Chapter 14.
- 189 Karlin S. Taylor H.M. 1975. A first course in stochastic processes. Academic Press Inc.
- 190 Kingman J.F.C. 1982. On the geneology of large populations. *J. Appl. Prob.* 19A:27-43.
- 191 Ronquist F., Teslenko M., van der Mark P., Ayres D., Darling A., Höhna S., Larget B., Liu L., Suchard M.A.,
- 192 Huelsenbeck J.P. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a
- 193 large model space. *Systematic Biology* 61:539–542.
- 194 Thorne J.L., Kishino H., Painter I.S. 1998. Estimating the rate of evolution of the rate of molecular evolution.
- 195 *Mol. Biol. Evol.* 15:1647–1657.
- 196 Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Mol. Biol.*
- 197 *Evol.* 24:1586-1591
- 198 김우철 2021. 수리통계학(개정판). 민영사.
- 199 서태건 2022. DNA 염기치환 모형의 비교. *한국진화학회지* 1:88-104.
- 200 서태건 2023. 아미노산 서열과 코돈 서열의 진화모형. *한국진화학회지* 2(2):41–60.
- 201 서태건 2024. 아미노산 서열과 코돈 서열의 진화모형. *한국진화학회지* 3(2):71–93.