분자계통수의 베이즈 추정

(분자계통수의 베이즈 추정(서태건 2024)논문의 요약 및 추가 설명입니다 ; ver.2026.01.27)

사전분포(prior distribution), 사전확률(prior probability)

아직은 동전을 던지는 실험을 하는 전 단계이다. 즉, '동전을 던지는 실험'을 어떤 '사건(event)'라고 간주하고 사건이 일어나기 전('사전事前'- 이라는 접두어가 붙는다) 확률 대해 생각하는 단계인 것이다. 아무런 추가 언급없이 "동전을 던져서 앞면이 나올 확률은 얼마인가?"라고 물으면 베이즈 통계학을 접한적이 없는 사람들 대다수는 1/2이라고 대답하려는 경향이 있다. 과연 합리적인 대답일까?

앞면이 나올 확률(이를 모수 θ로 표현하자)은 0과 1 사이의 값을 가져야 하고, 구간 $(0,1)$에서의 불확실성을 나타내는 대표적인 확률분포로는 beta 분포가 있다. Beta 분포는 α_1, α_2 두개의 모수로 그 형태가 결정되며 이를 $\text{beta}(\alpha_1, \alpha_2)$로 나타낸다. 그 확률밀도함수는 다음과 같다(김우철 2021, p. 148).

$$p(\theta) = \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1} \quad (0<\theta<1) \tag{1}$$

여기에서 $\Gamma(\cdot)$는 감마함수를 의미한다. $\text{Beta}(\alpha_1, \alpha_2)$분포의 평균($\mu$)과 분산($\sigma^2$)은 다음과 같다.

$$\begin{aligned} \mu &= \frac{\alpha_1}{\alpha_1+\alpha_2} \\ \sigma^2 &= \frac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)} \end{aligned} \tag{2}$$

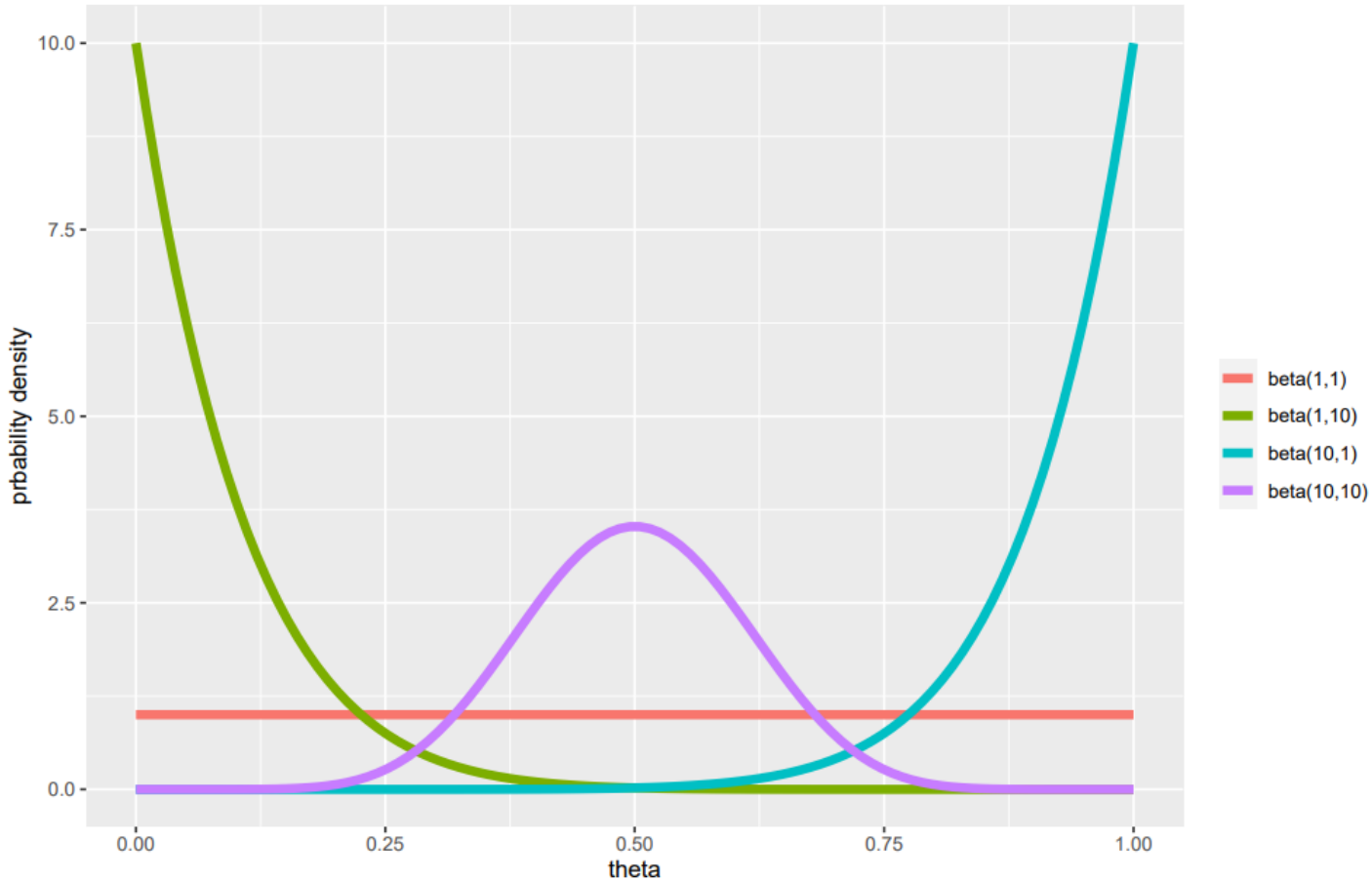


그림 1. Beta 분포 확률밀도함수의 여러 형태

가능도(likelihood)

이제 앞면이 나올 확률을 추정하기 위해 본격적으로 실험을 해보자. 실험은 간단하다. 동전을 n번 던져 앞면이 나온 횟수 k를 세면 된다. 앞면이 나올 확률이 θ일때 n번 동전을 던져 k번 앞면이 나오는 확률은 이항분포(binomial distribution)를 가정하면 다음과 같이 표현할 수 있다.

$$p(k|\theta) = \frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k} \quad (0<\theta<1) \tag{3}$$

estimator)과 n이 충분히 클때 95% 신뢰구간(Confidence Interval; CI)은 다음과 같음을 기초통계학 교재등을 통해 쉽게 확인할 수 있다(김우철 et al. 1994, p.134).

$$\begin{aligned} \widehat{\theta} &= \frac{k}{n} \\ \theta\text{의 95 \% CI(정규근사)} &= \left(\widehat{\theta} - 1.96\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}, \widehat{\theta} + 1.96\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}\right) \end{aligned} \tag{4}$$

가령 문제의 동전을 100번 던져 앞면이 60번 나왔다고 하자. 식 (4)를 적용하면 최대가능도 추정량은 60/100, 95% 신뢰구간은 (0.504, 0.696)과 같이 구해진다. 이 결과는 앞에서 논의한 사전확률을 전혀 고려하지 않은 비-베이지안 추정 결과이다.

빈도주의 방식과 베이지안 방식의 큰 차이 중 하나는 추정된 모수의 불확실함을 해석하는 방식이다. 빈도주의에서는 흔히 신뢰구간(confidence interval; CI)으로 추정된 모수의 불확실함을 표현한다. 위의 동전 던지기를 예로 들면, 식 (4)에 의해 얻은 θ의 95% 신뢰구간은 $(0.504, 0.696)$이다. 이를 직관적으로 "미지의 모수가 구간 $(0.504, 0.696)$ 안에 위치할 확률이 95%이다"라고 해석하는 경우가 종종 있는데 이는 신뢰구간의 의미를 잘못 이해하는 전형적인 예이다.

빈도주의 방식에서 '95% 신뢰구간'의 올바른 해석은 "동전을 100번 던져 앞면이 나오는 횟수를 측정하고 식(4)에 의해 신뢰구간을 구하는 작업을 무한히 반복하면 대략 100번중 95번의 비율로 신뢰구간이 참값을 포함한다. $(0.504, 0.696)$은 이러한 무한히 많은 신뢰구간 중 하나이고 이 구간이 참값을 포함하는지 아닌지는 알 수 없다."라는 것이 올바른 해석이다. 이는 직관적으로 잘 와닿지 않고 이해하기 힘든 해석이다. 그림 3은 θ의 참값이 1/2일때 동전을 100번 던져 식 (4)에 의해 신뢰구간을 구하는 과정을 100번 반복한 모의실험(simulation) 결과이다. 참값 $\theta = 1/2$를 신뢰구간이 포함하면 검은색, 포함하지 않으면 붉은색으로 표시하였다. 이 모의실험에서는 100회중 검은색 신뢰구간이 93회 관찰되었으나 모의실험의 횟수를 늘이면 점근적으로 95%에 근사하게 된다.

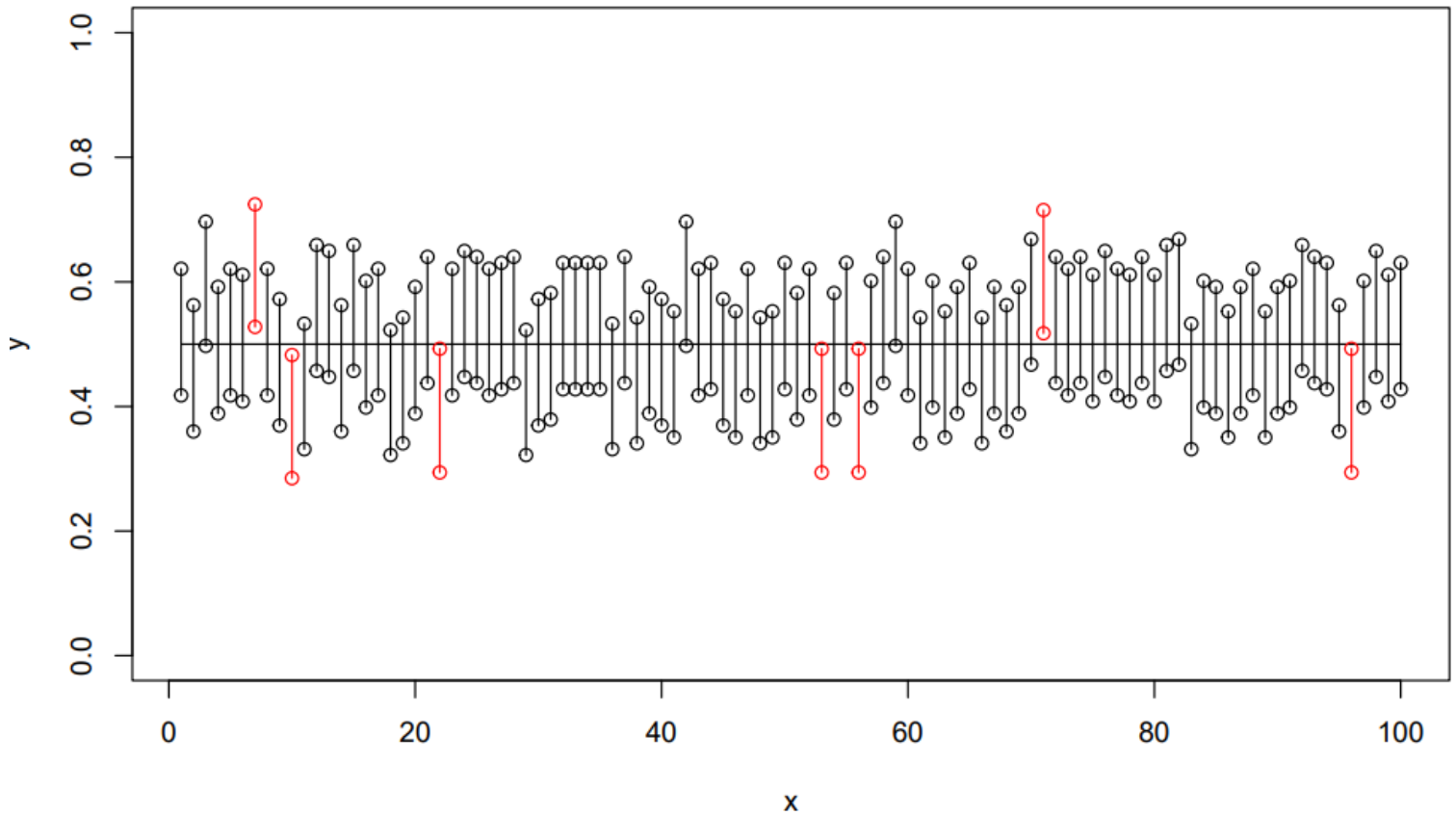


그림 3. 95% 신뢰구간(Confidence Interval)의 의미를 알아보기 위한 모의실험.

$$\text{사후분포} \propto \text{가능도} \times \text{사전분포} \tag{5}$$

동전을 던져 앞면이 나올 확률 θ에 대한 사후 분포를 위의 식을 이용하여 표현하고, 사후분포는 확률분포이므로 θ가 가질 수 있는 범위에서 확률분포의 적분값이 1이 되도록 정규화하면 사후확률분포는 다음과 같이 beta 분포로 주어짐을 확인할 수 있다(김우철 2021, p. 452).

$$p(\theta|k) \propto p(k|\theta)p(\theta) \tag{6}$$

$$= \frac{\Gamma(\alpha_1+\alpha_2+n)}{\Gamma(\alpha_1+k)\Gamma(\alpha_2+n-k)}\theta^{k+\alpha_1-1}(1-\theta)^{n-k+\alpha_2-1} \quad (0<\theta<1) \tag{7}$$

즉, 요약하면 $\text{beta}(\alpha_1,\alpha_2)$를 사전분포로 가정하고 동전 던지기로 (n,k) 결과를 얻었을때 사후분포는 $\text{beta}(\alpha_1+k, n-k+\alpha_2)$로 주어진다.[9] 식 (2)를 이용하여 구한 사후분포의 평균은 $(\alpha_1+k)/(\alpha_1+\alpha_2+n)$이 된다. 평균이 사전분포의 $\alpha_1/(\alpha_1+\alpha_2)$로부터 데이터가 제공하는 정보에 영향을 받아 업데이트된 것이다.

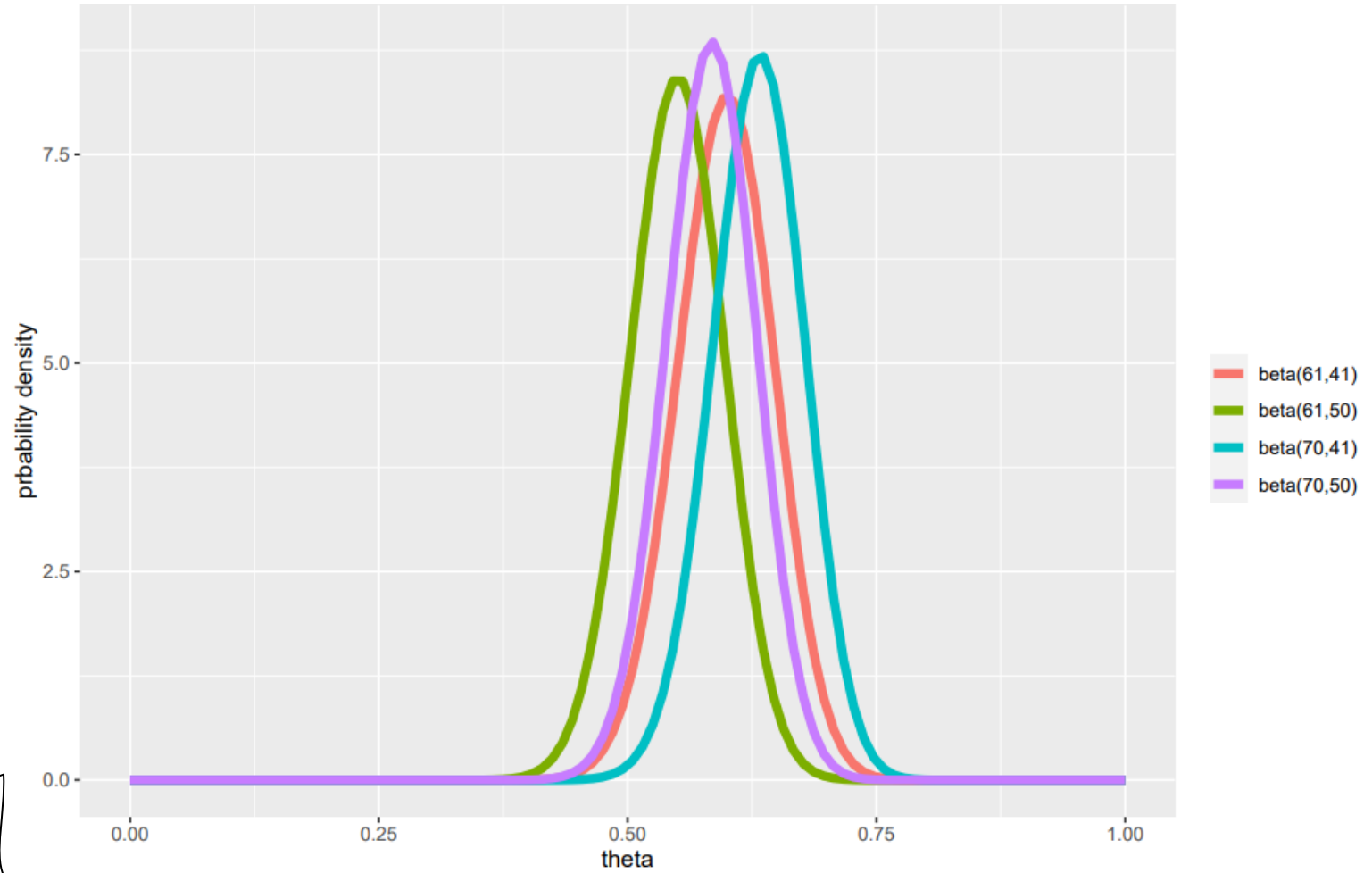


그림 2. 사후 Beta 분포 확률밀도함수의 여러 형태

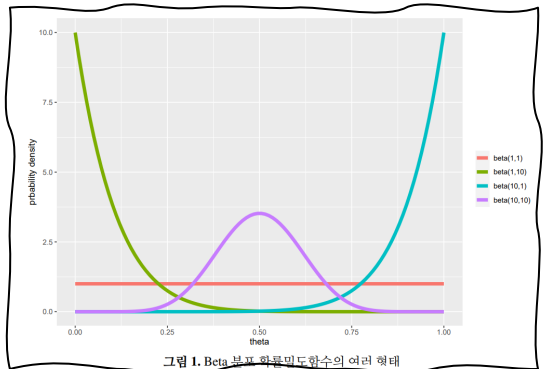
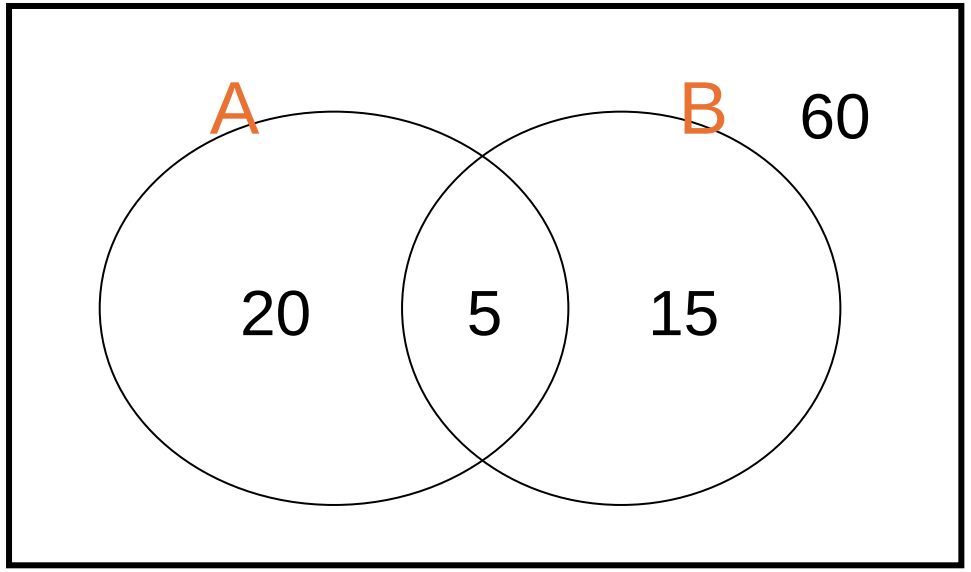


그림 1. Beta 분포 확률밀도함수의 여러 형태

조건부 확률



$$P(A) = (20+5)/100 = 1/4$$

$$P(B) = (15+5)/100 = 1/5$$

$$P(A \cap B) = 5/100 = 1/20$$

$$P(B \mid A) = 5/(20+5) = 1/5$$

$$P(A \mid B) = 5/(15+5) = 1/4$$

$$P(A \cap B) = P(A \mid B)P(B)$$

$$P(A \cap B) = P(B \mid A)P(A)$$

$$\begin{cases} P(B \mid A) = P(B) = 1/5 \\ P(A \mid B) = P(A) = 1/4 \end{cases}$$

A와B는 독립 ⇒ $P(A \cap B) = P(A)P(B)$

Bayes 정리

D: data θ: parameter

$$P(\theta \mid D) = \frac{P(\theta, D)}{P(D)}$$

$$= \frac{P(D \mid \theta)P(\theta)}{P(D)}$$

$$= \frac{P(D \mid \theta)P(\theta)}{\sum_{\theta} P(D \mid \theta)P(\theta)} \quad \text{or} \quad \frac{P(D \mid \theta)P(\theta)}{\int_{\theta} P(D \mid \theta)P(\theta)}$$

데이타가 주어진 상태에서 파라미터의 확률밀도 함수

Tomas Bayes (1703-1761)

P(D|θ) : Non-Bayesian(Frequentist, classical) framework에서는 θ:고정、D:랜덤

P(θ|D) : Bayesian framework 에서는 D:고정、θ:랜덤 → 미지의 파라미터에 대해서 확률분포를 생각하는것이 가능

분자 계통수의 베이지안 추정 모식도

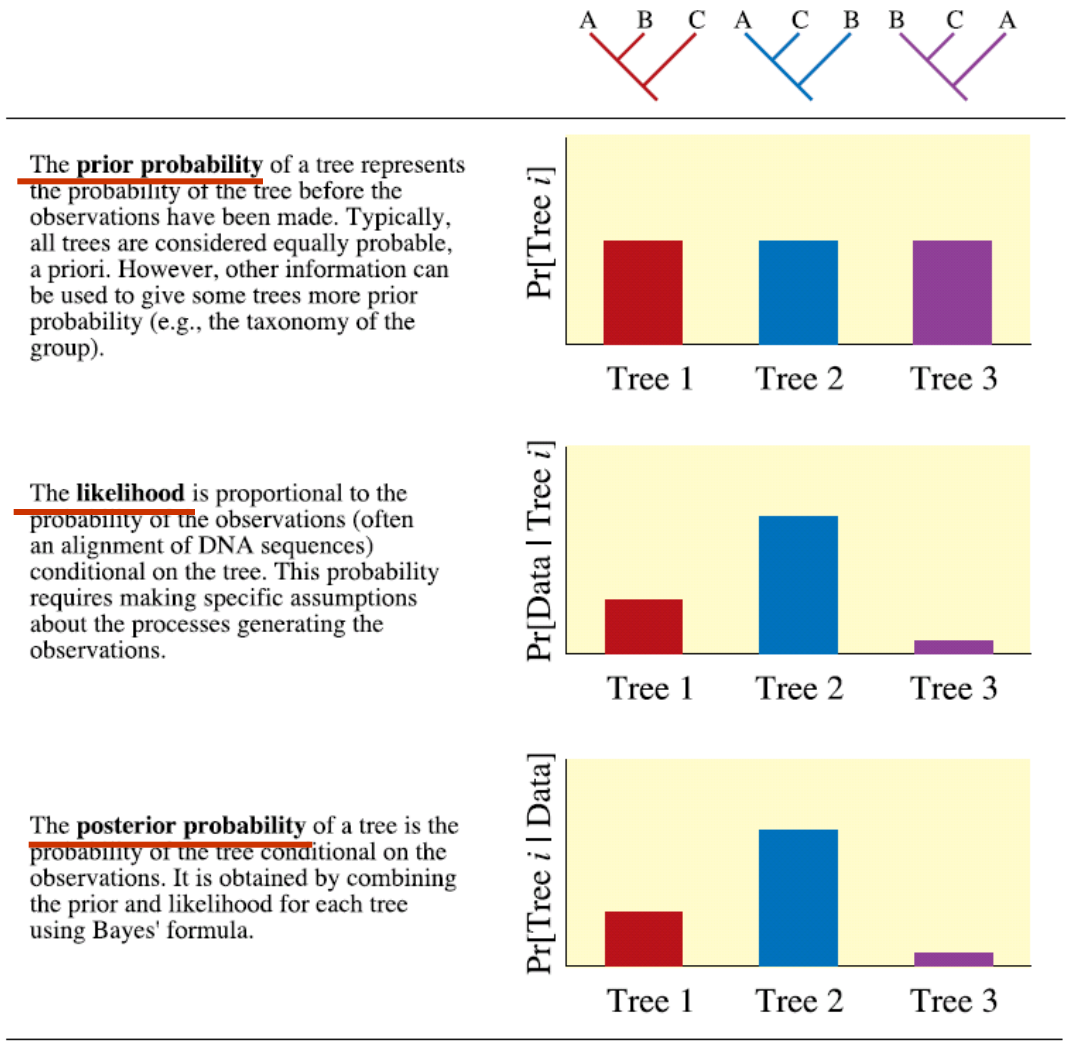


Fig. 1. The main components of a Bayesian analysis.

(1) 사전확률 (**事前確率**) : 테이타를 관측하기전에 사전에 가지고 있던 정보에 의한 확률(정보가 없는 경우 균등한 확률을 생각하는 것이 무난함)

(2) likelihood: 테이타가 계통수에 얼마만큼 잘 적합한가 나타내는 수치

(3) 사후확률(**事後確率**): 사전확률과 likelihood를 곱해서 얻은 확률

Huelsenbeck et al. (2001, Science 294(5550):2310-4)

$$③\ P(Tree \mid Data) = \frac{②\ P(Data \mid Tree)\,①\ P(Tree)}{P(Data)}$$

계통수의 사후확률 계산의 간단한 예*

** branch length, 모델 파라미터 등의 사후분포도 생각해야 하므로, 실제의 계산은 이보다 훨씬 더 복잡함*

Tree1, Tree2, Tree3의 계통수에 대하여 계산한 log-likelihood score가 각각 −10.0, −11.0, −12.0 일 경우, 베이즈 정리를 이용하여 사후확률을 계산해보자

각 계통수의 사전확률을 1/3이라고 가정하면 Tree1의 사후확률은…

$$P(T_1 \mid D) = \frac{P(T_1, D)}{P(D)} = \frac{P(D \mid T_1)P(T_1)}{P(D)}$$

$$= \frac{P(D \mid T_1)P(T_1)}{P(D \mid T_1)P(T_1) + P(D \mid T_2)P(T_2) + P(D \mid T_3)P(T_3)}$$

$$= \frac{e^{-10.0} \times 1/3}{e^{-10.0} \times 1/3 + e^{-11.0} \times 1/3 + e^{-12.0} \times 1/3}$$

$$\approx 0.6652$$

(참고) 사전확률이 각각 1/2, 1/4 , 1/4 일 경우
Tree1의 사후확률을 계산하면 ?

계통수의 사후확률 계산의 간단한 예(2) *

* *branch length, 모델 파라미터 등의 사후분포도 생각해야 하므로, 실제의 계산은 이보다 훨씬 더 복잡함*

Tree1, Tree2, Tree3의 계통수에 대하여 계산한 log-likelihood score가 각각 −30.0, −33.0, −36.0 일 경우, 베이즈 정리를 이용하여 사후확률을 계산해보자 (5페이지의 예에 비하여 염기배열의 길이가 3배)

각 계통수의 사전확률을 1/3이라고 가정하면 Tree1의 사후확률은…

$$P(T_1 \mid D) = \frac{P(T_1, D)}{P(D)} = \frac{P(D \mid T_1)P(T_1)}{P(D)}$$

$$= \frac{P(D \mid T_1)P(T_1)}{P(D \mid T_1)P(T_1) + P(D \mid T_2)P(T_2) + P(D \mid T_3)P(T_3)}$$

$$= \frac{e^{-30.0} \times 1/3}{e^{-30.0} \times 1/3 + e^{-33.0} \times 1/3 + e^{-36.0} \times 1/3}$$

$$\approx 0.95033$$

중요포인트: 데이타의 수가 많으면 사전확률은 사후 확률에 거의 영향을 미치지 않음. 사전확률선택의 주관성에 대한 걱정 불필요.

(참고) 사전확률이 각각 1/2, 1/4 , 1/4 일 경우 Tree1의 사후확률을 계산하면 ?

MCMC (Markov chain Monte Carlo)

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}$$ (D: data, θ: parameter)

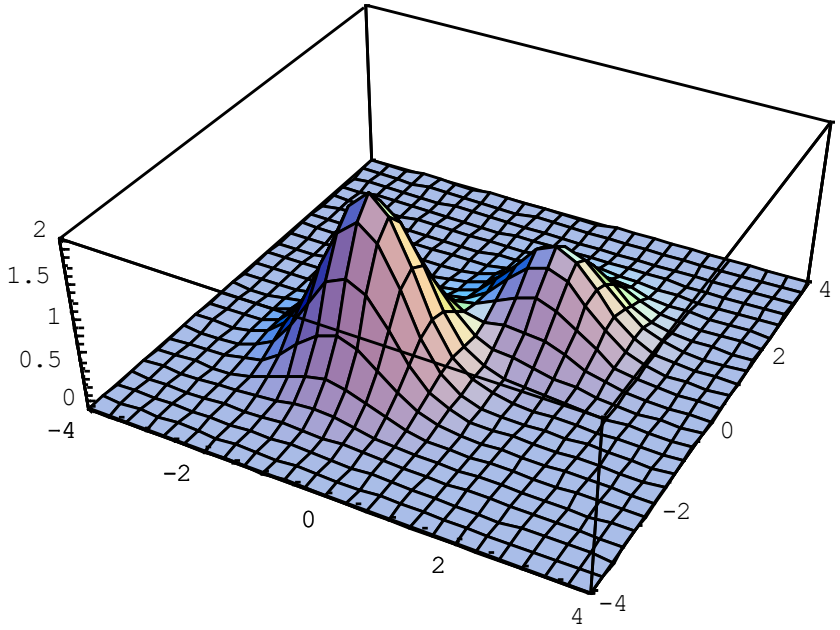
$$= \frac{P(D \mid \theta)P(\theta)}{\sum_{\theta} P(D \mid \theta)P(\theta)} \quad \text{or} \quad \frac{P(D \mid \theta)P(\theta)}{\int_{\theta} P(D \mid \theta)P(\theta)}$$

(1) P(D)를 직접 계산하는것은 어려움 (고차원의 파라미터의 경우 다중적분에 상당한 계산시간이 소요됨)

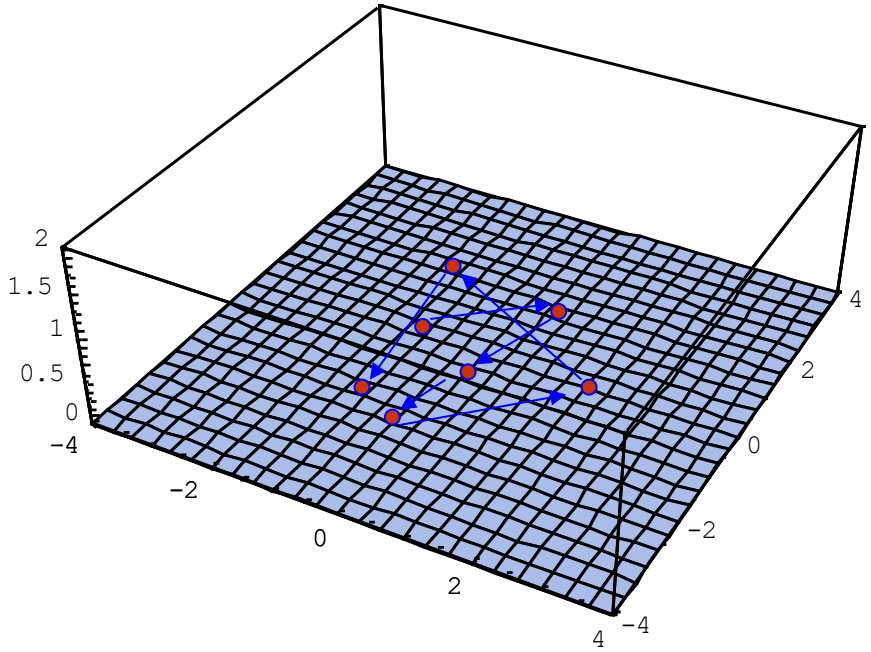
(2) P(D)는 파라미터 θ 에 의존하지 않는 확률이므로 직접 계산하지 않아도 사후확률을 구할수 있음 -> MCMC 이용

MCMC = Metropolis–Hastings algorithm

MCMC (Markov chain Monte Carlo)의 원리



(1)이러한 형태의 사후분포로 부터 샘플링을 하려고 할때...

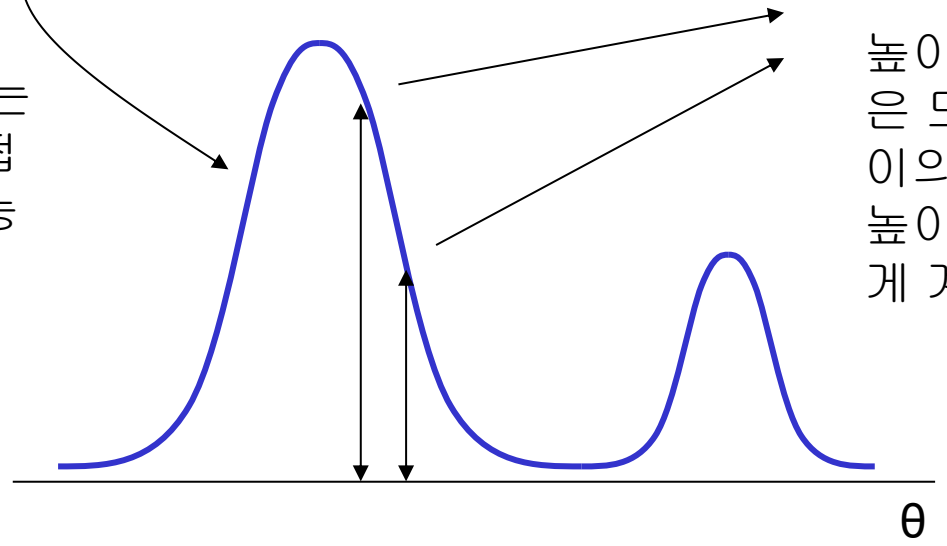


(2)샘플링될 데이타 포인트를 랜덤하게 움직인다 (점의 존재 가능성이 사후분포의 높이에 비례하도록...). 랜덤하게 움직이면서 정기적으로 샘플링함.

MCMC(Markov chain Monte Carlo)의 원리

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} \propto P(D \mid \theta)P(\theta)$$

P(D) 를 모르는 상태에서 직접 계산은 불가능

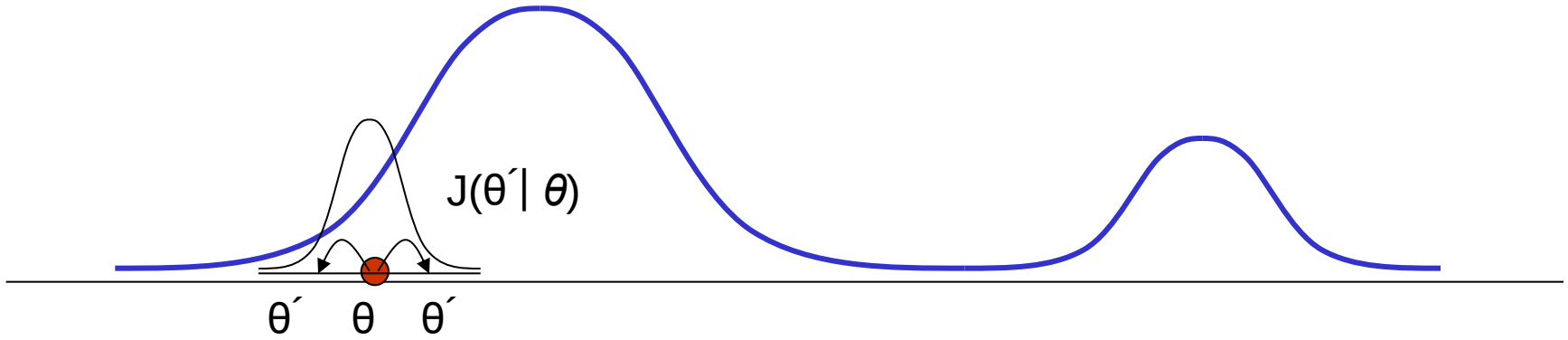


높이의 정확한 값은 모르지만 두 높이의 비 (상대적인 높이)는 비교적 쉽게 계산 할수 있음

D: data θ: parameter

MCMC(Markov chain Monte Carlo)의 원리

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} \propto P(D \mid \theta)P(\theta)$$



(1) Proposal step （현재의 파라미터의 다음 위치를 결정）

(2) Acceptance/rejection step （다음 위치를 채택/기각 할지 결정）

반복하면서 θ 를 움직임

Accept proposed state with probability $\min\left(1, \frac{P(\theta' \mid D)J(\theta \mid \theta')}{P(\theta \mid D)J(\theta' \mid \theta)}\right)$

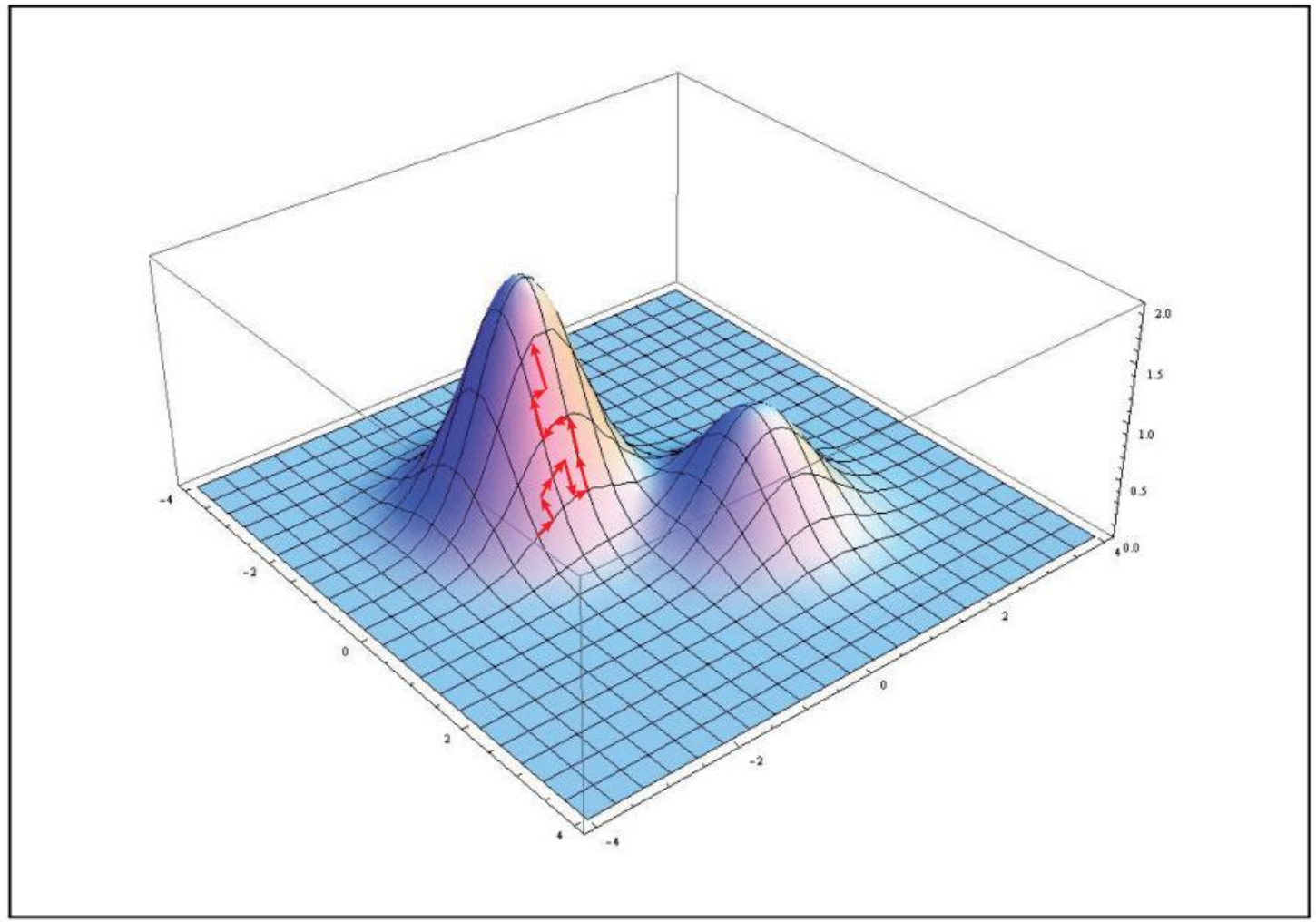
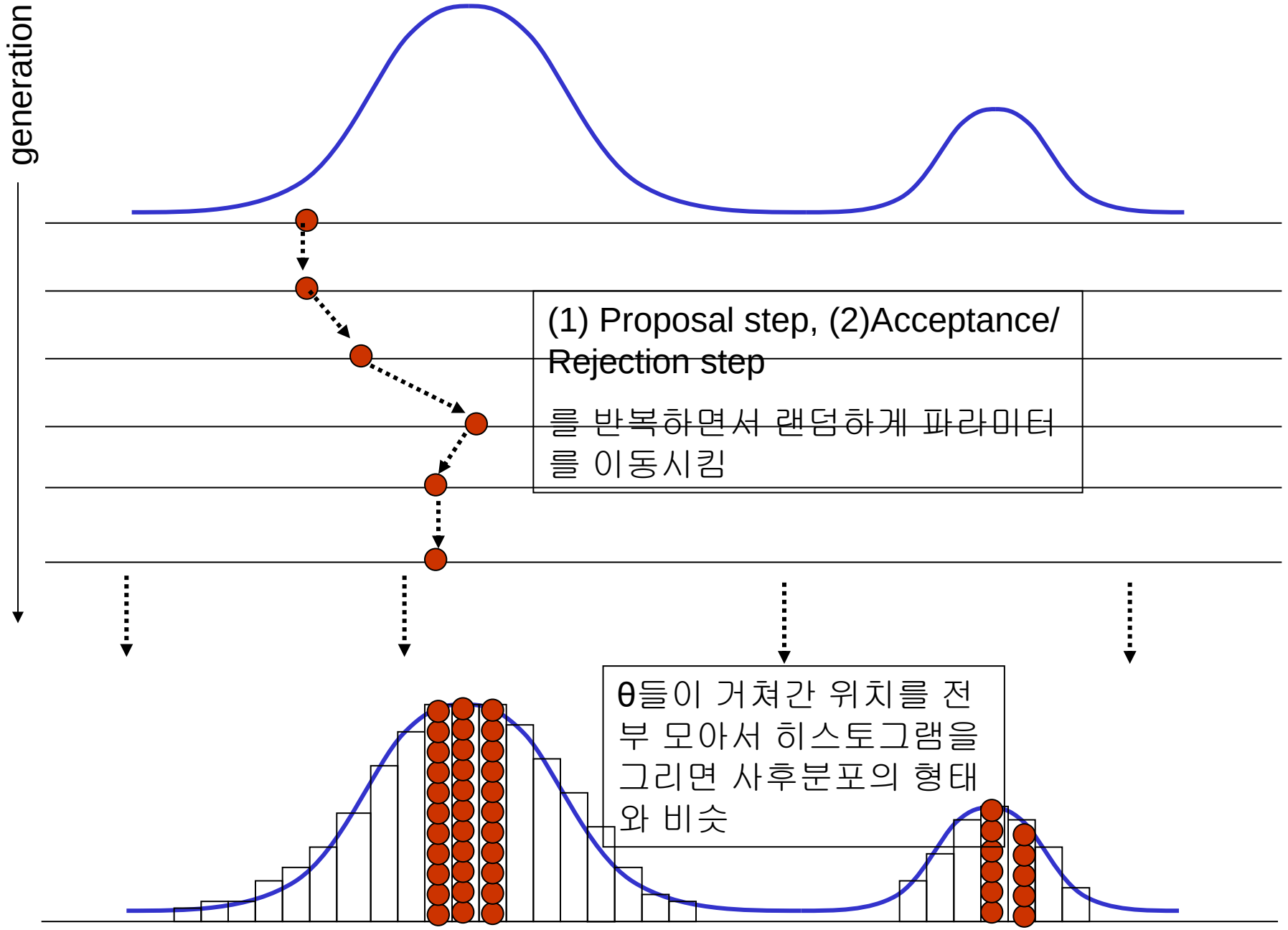
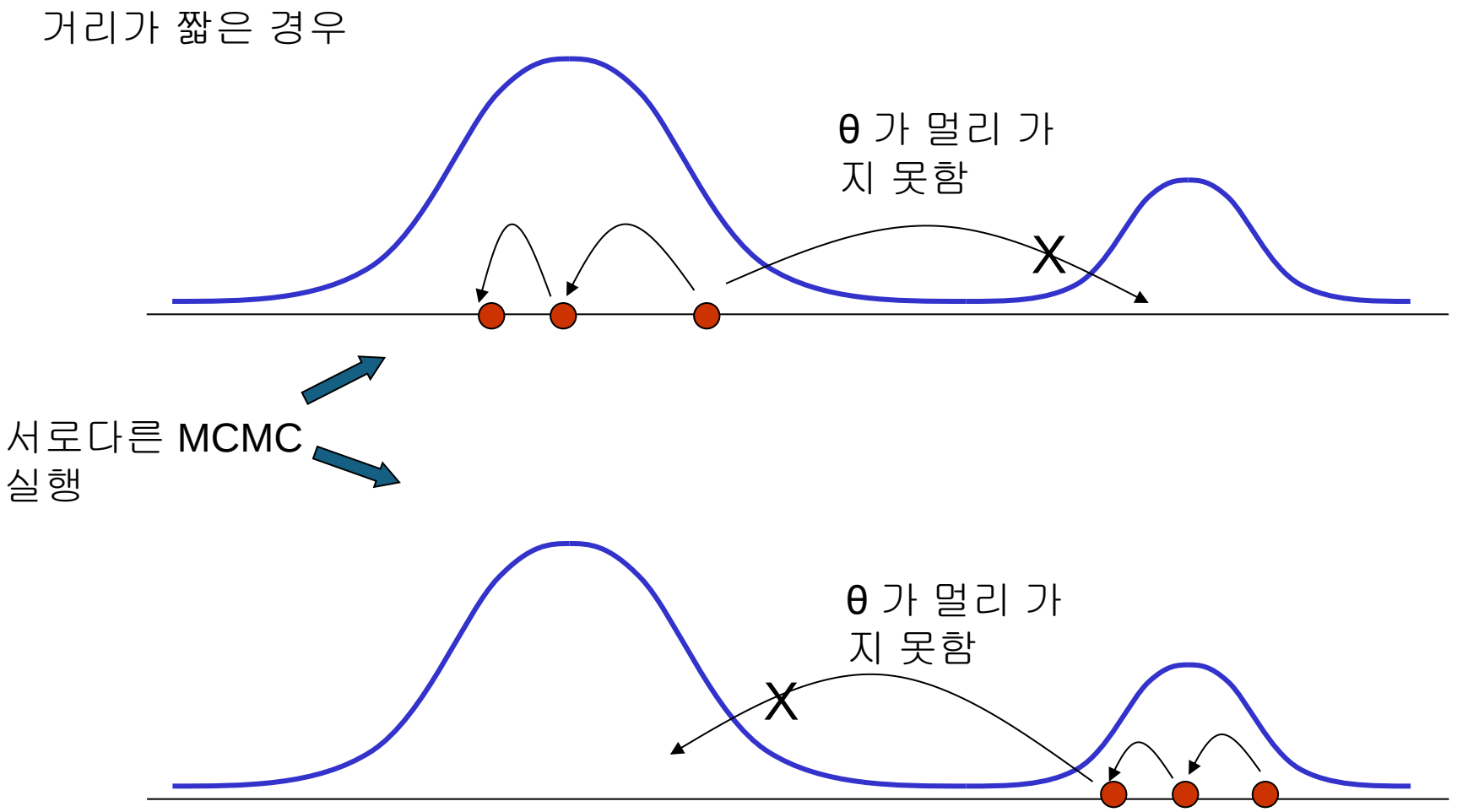


그림 5. 사고실험2의 등산 경로 모식도(붉은색 화살표). 고도가 낮은 곳으로 이동하기도 하지만 전반적으로 높은 곳으로 가려는 경향성이 강하다. 따라서 고도가 높은 곳은 그만큼 자주 방문하게 되어 방문 지점을 2차원 히스토그램으로 나타내면 산의 모양과 대략 비슷하게 된다.


generation
(1) Proposal step, (2)Acceptance/ Rejection step
를 반복하면서 랜덤하게 파라미터를 이동시킴
θ들이 거쳐간 위치를 전부 모아서 히스토그램을 그리면 사후분포의 형태와 비슷

Proposal step에서 θ´와 θ 사이의 거리가 중요



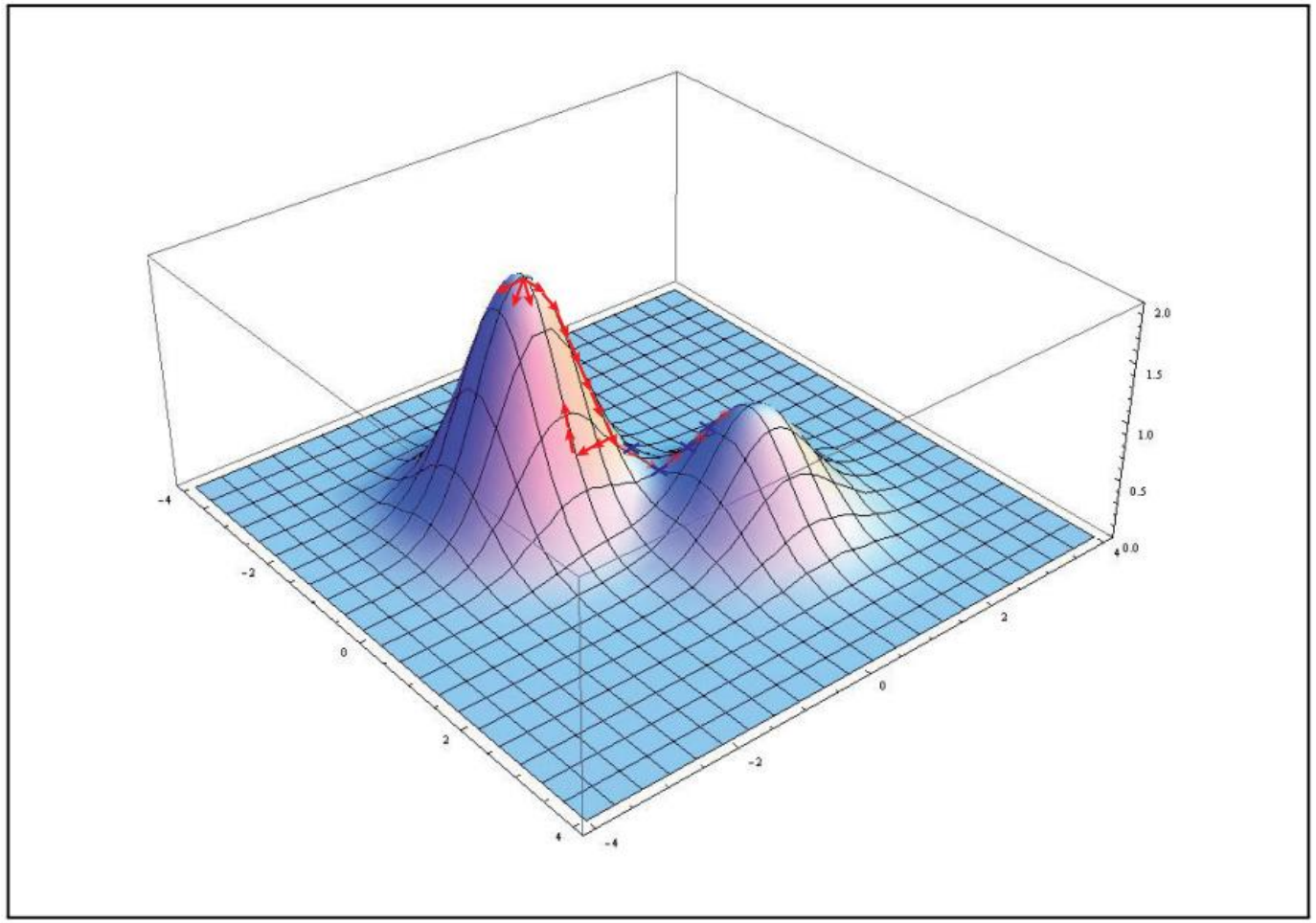
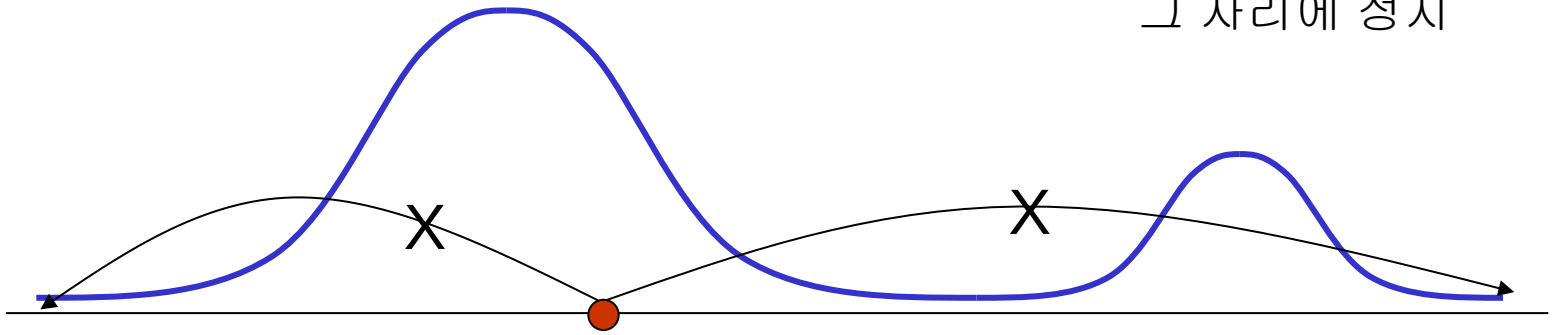


그림 6. $J(\theta'|\theta)$ 함수를 정의할 때 θ와 θ'간의 거리가 너무 가까우면 깊은 골짜기를 지나쳐 다른 산봉우리로 이동하는 것을 하지 못하고 하나의 산봉우리 주위에만 맴돌게 된다. 이 경우 초기 난수(그림 9 참조) 설정을 달리해 MCMC 알고리즘을 실행하면 실행할 때마다 다른 결과가 얻어지곤 한다.

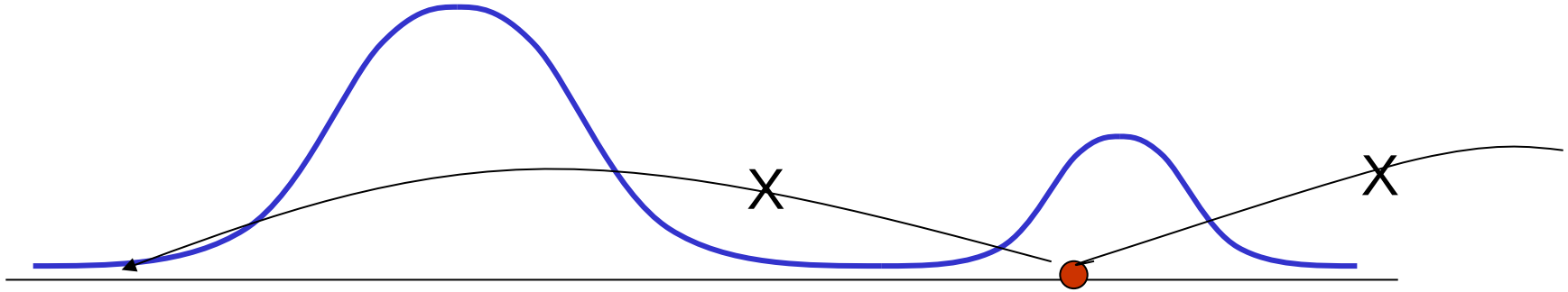
Proposal step에서 θ´와 θ 사이의 거리가 중요

Acceptance rate이 낮다.
→ θ 가 움직이지 않고
그 자리에 정지

거리가 먼 경우



서로다른 MCMC
실행



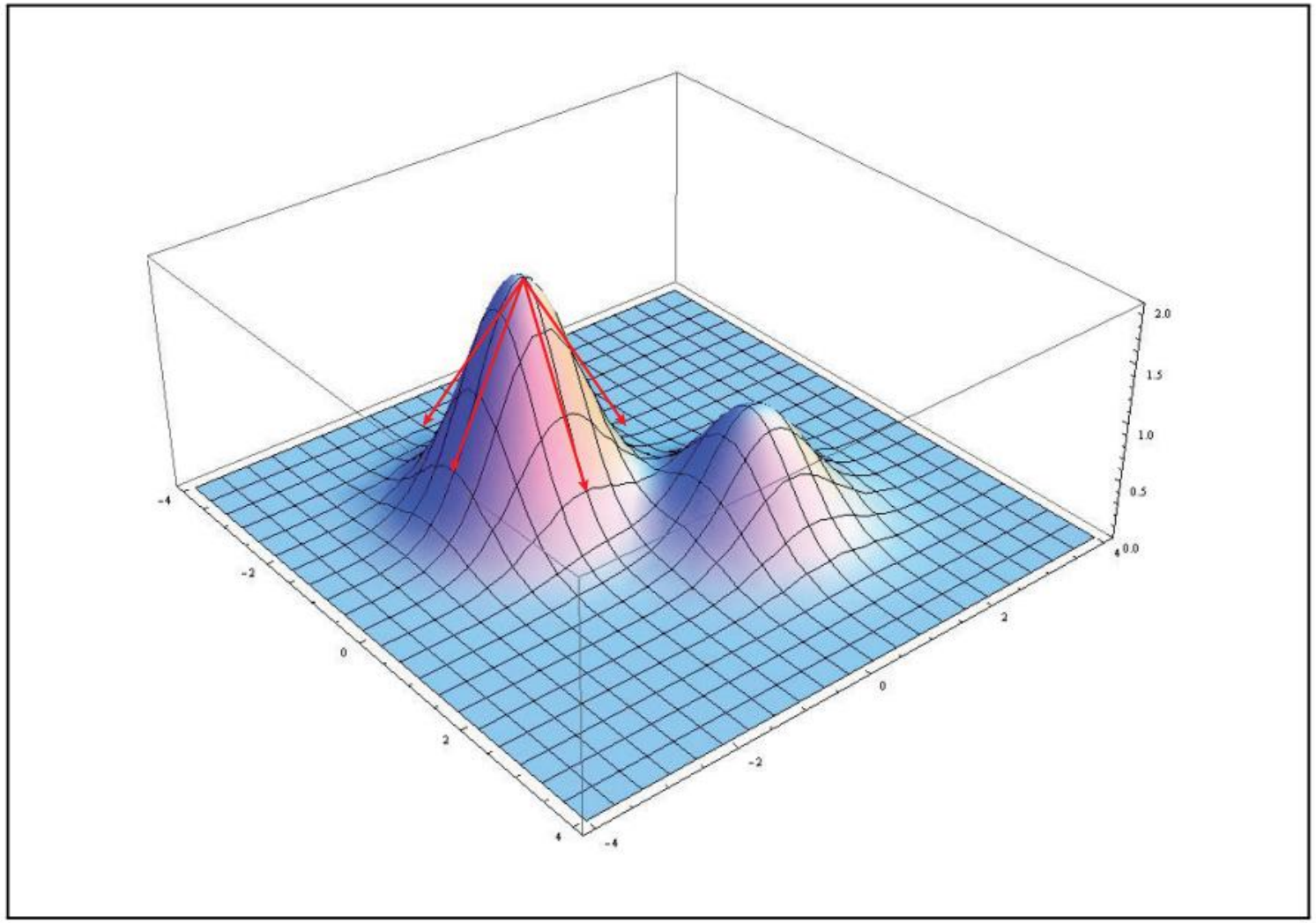
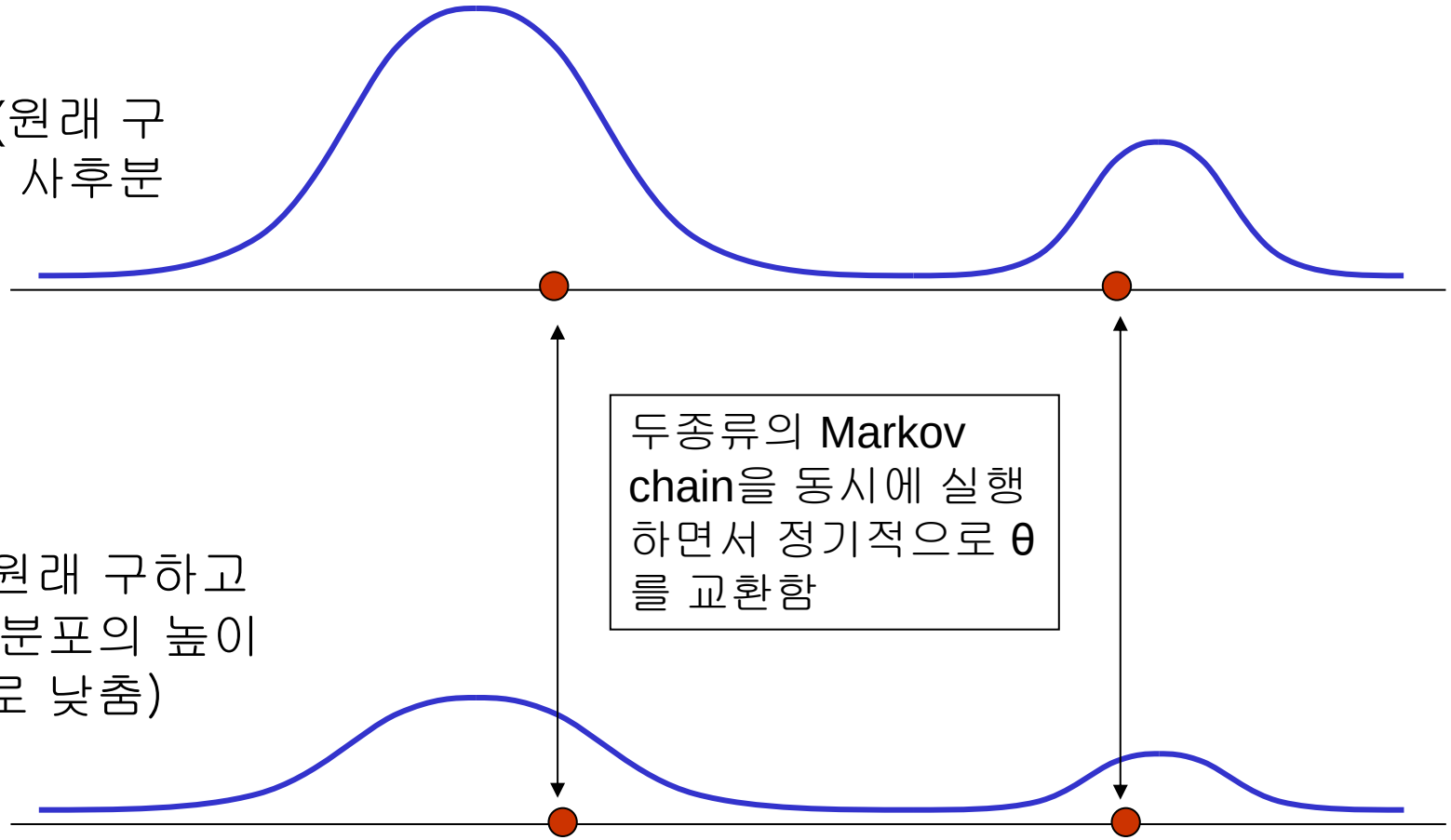


그림 7. $J(\theta'|\theta)$ 함수를 정의할 때 θ와 θ'간의 거리가 너무 멀면 θ의 이동이 잘 이루어지지 않아("MCMC chain is not mixing well.") θ값들의 모임이 사후분포의 형태를 잘 대변해 주지 못하게 된다.

Cold chain & Hot chain

(θ′와θ사이의 거리를 정하기 힘들때 유용한 수단)

Cold chain (원래 구하고자 하는 사후분포)



Hot chain (원래 구하고자 하는 사후분포의 높이를 전체적으로 낮춤)

Hot chain : 열을 가해서 녹아내린 모습을 연상하면 이해하기 쉬움. θ′와θ의 거리에 비교적 영향을 덜 받고 θ가 잘 움직임

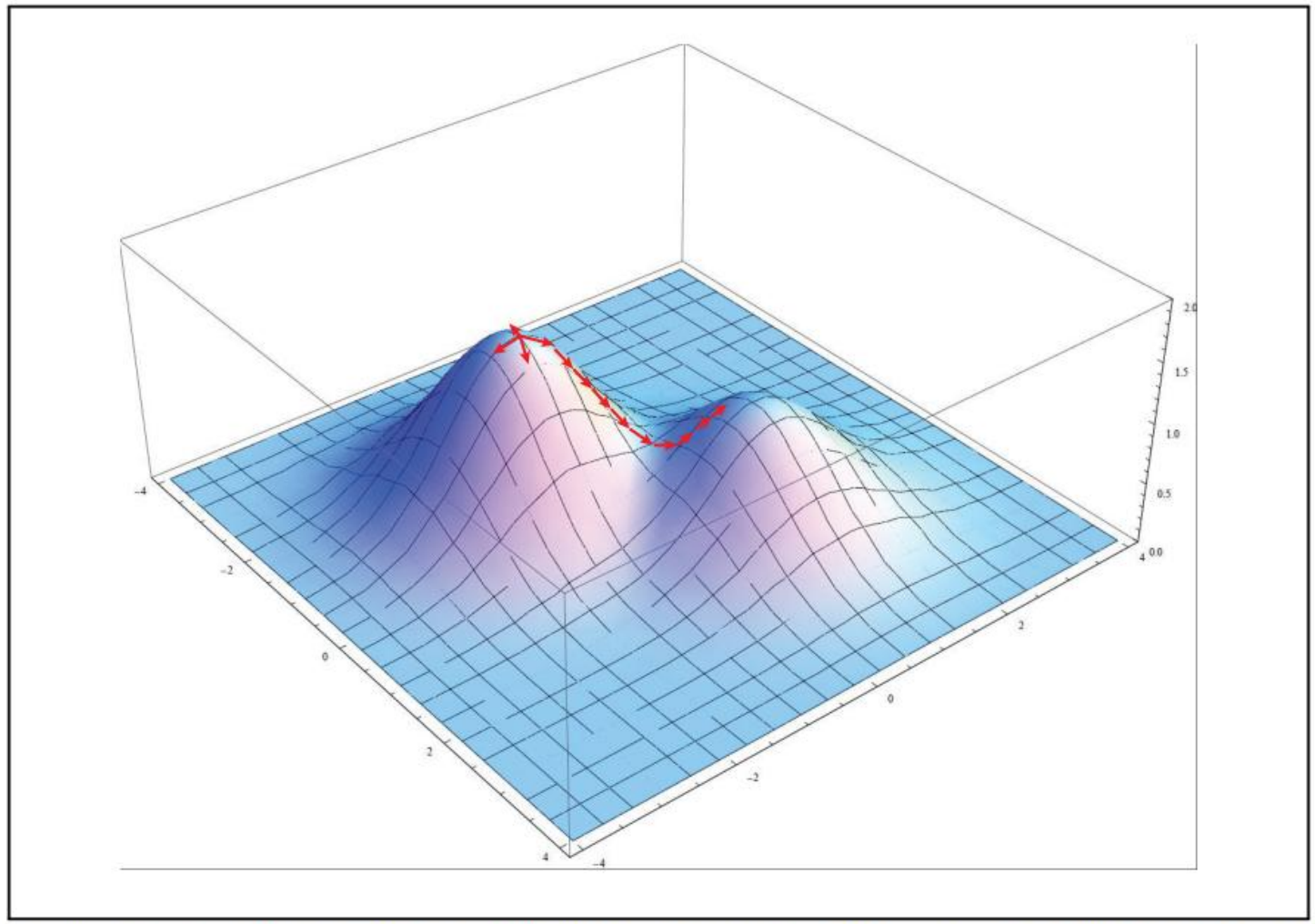
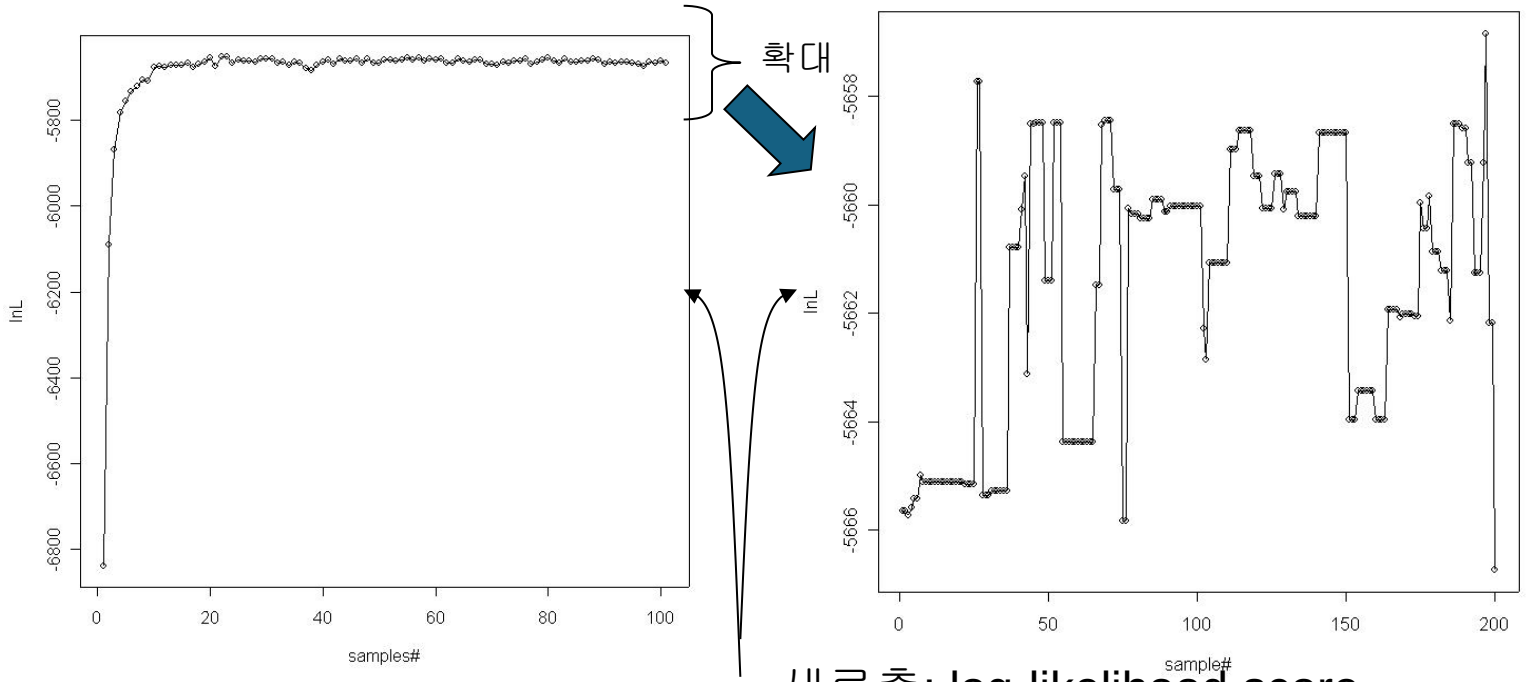


그림 8. 사후분포에 열을 가해 산봉우리가 녹아내려 산정상과 골짜기의 높이차가 줄어든 모식도. 이 경우 θ의 움직임을 hot chain이라고 한다. θ와 θ'간의 거리에 크게 영향을 받지 않고 θ가 양 봉우리를 쉽게 오갈 수 있다.

Number of burn-in : MCMC의 결과가 초기치의 영향을 받지 않게 하기위해 초기의 샘플을 무시한다. 무시하는 generation의 수

Number of interval : MCMC로 부터 샘플링되는 점들 사이의 상관관계를 줄이기 위해서 일정 generation간격으로 점들을 샘플링. 그 간격의 수



$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)} \propto P(D \mid \theta)P(\theta)$$

실습

MrBayes 실행파일과 데이터 파일(mb.3.2.7-win64.exe 와 cynmix.nex)을 적당한 폴더(예를들어 C:\temp)에 복사한다. 명령프롬프트를 실행시켜(cmd.exe 명령어 실행) C:\temp로 이동한 후 MrBayes 프로그램을 실행하면 다음과 같이 MrBayes 프로그램 내부의 프롬프트가 표시된다.

```
MrBayes>
```

이후 모든 명령어는 MrBayes 프롬프트에서 실행한다. 필수적인 것은 아니지만 seed와 swapseed를 MrBayes 실행 직후에 그림 9과 같이 설정하면 분석결과를 완벽하게 재현할 수 있는 장점이 있다("1"대신 다른 양의 정수 설정 가능). 명령어 exec를 이용하여 데이터 파일 cynmix.nex를 읽어들인다.

```
MrBayes>  set seed=1 swapseed =1;  ↩
MrBayes>  exec cynmix.nex  ↩
```

그림 9. 초기 난수 설정과 데이터 읽어들이기

```
#NEXUS

[ Data from: Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL. 2004. Bayesian

Begin data;
   Dimensions ntax=32 nchar=3246;
   Format datatype=mixed(Standard:1-166,DNA:167-3246) interleave=yes gap=- missing=?;
   Matrix
Ibalia          00000000000002-0000000000??000000000000100{01}0100001-00100000-000
Synergus        1-1-1000000002021102010110101101000000001012100112010101010000000011
Periclistus     1-1-1000000002021102010111101101000000001010010100010101001100000000011
Ceroptres       1-1-1000100002021002010111101001000000001110??101010100100000000001
Synophromorpha  1-1-00001000021-1001010011111101000000001010010100010001001000000000001
Xestophanes     1-1-0000100000-10-1100001011010100000000001011010100010001001010000000
```

그림 10. cynmix.nex 파일의 앞부분.

```
begin mrbayes;

   [This block defines several different character sets that could be used
   and then defines and enforces a partition called favored.]

   charset morphology = 1-166;
   charset COI = 167-1244;
   charset EF1a = 1245-1611;
   charset LWRh = 1612-2092;
   charset 28S = 2093-3246;
   partition favored = 5: morphology, COI, EF1a, LWRh, 28S;

   [The following lines set up a particular model (the one discussed in the
```

그림 11. cynmix.nex 파일의 뒷부분.

```
MrBayes>  set partition = favored;
MrBayes>  showmodel
MrBayes>  lset applyto=(1) rates=gamma;
MrBayes>  lset applyto=(2,3,4,5) rates=invgamma nst=6;
MrBayes>  unlink revmat=(all) pinvar=(all) shape=(all)
statefreq=(all);
MrBayes>  showmodel
```

그림 12. 파티션별로 모형 설정하는 예시. 명령어 'lset'은 가능도 모형(likelihood model)을 설정하는 키워드이다.

Active parameters:

Parameters	Partition(s) 1	2	3	4	5
Statefreq	1	2	2	2	2
Ratemultiplier	3	3	3	3	3
Topology	4	4	4	4	4
Brlens	5	5	5	5	5

(a) 그림 12의 첫번째 showmodel 명령어 실행 결과 일부

Active parameters:

Parameters	Partition(s) 1	2	3	4	5
Revmat	.	1	2	3	4
Statefreq	5	6	7	8	9
Shape	10	11	12	13	14
Pinvar	.	15	16	17	18
Ratemultiplier	19	19	19	19	19
Topology	20	20	20	20	20
Brlens	21	21	21	21	21

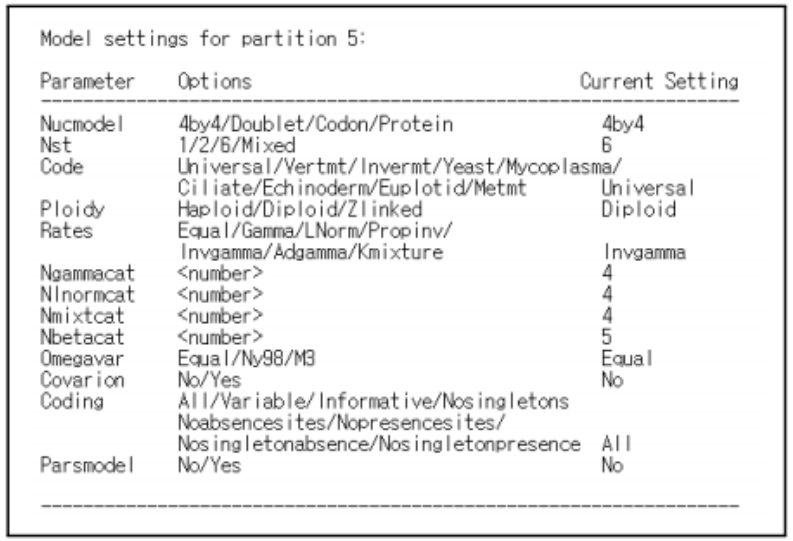
(b) 그림 12의 두번째 showmodel 명령어 실행 결과 일부

그림 13. showmodel 명령어로 파티션별로 모수의 설정을 확인할 수 있다.

하고, 그림 14의 세번째 줄의 'prset ratepr=variable;' 명령을 실행하여 사전분포를 변경한다. 이후 그림 14의 네번째 줄의 'help prset' 명령을 실행하여 사전분포가 바뀐 것을 확인한다(그림 15(b)). 이는 그림 13(b)의 19번 모수와 관련된 내용으로 각 파티션들이 서로 다른 치환속도를 갖게 하는 설정이다.

```
MrBayes>  help lset
MrBayes>  help prset
MrBayes>  prset ratepr=variable;
MrBayes>  help prset
```

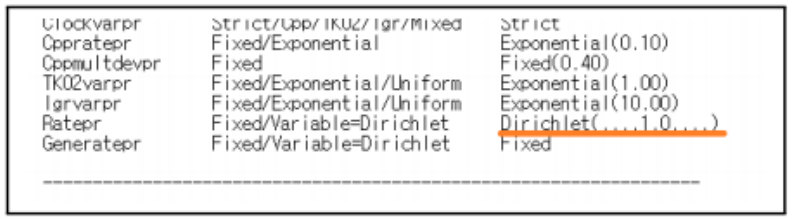
그림 14. 'help lset'으로 가능도 모형 설정을 확인할 수 있다. prset 명령어로 각 파티션별 상대적인 진화속도의 사전분포를 지정한다.



Model settings for partition 5:

Parameter	Options	Current Setting
Nucmodel	4by4/Doublet/Codon/Protein	4by4
Nst	1/2/6/Mixed	6
Code	Universal/Vertmt/Invermt/Yeast/Mycoplasma/ Ciliate/Echinoderm/Euplotid/Metmt	Universal
Ploidy	Haploid/Diploid/Zlinked	Diploid
Rates	Equal/Gamma/LNorm/Propinv/ Invgamma/Adgamma/Kmixture	Invgamma
Ngammacat	<number>	4
Nlnormcat	<number>	4
Nmixtcat	<number>	4
Nbetacat	<number>	5
Omegavar	Equal/Ny98/M3	Equal
Covarion	No/Yes	No
Coding	All/Variable/Informative/Nosingletons Noabsencesites/Nopresencesites/ Nosingletonabsence/Nosingletonpresence	All
Parsmodel	No/Yes	No

(a) 그림 14의 help lset 명령어 실행 결과 일부



Clockvarpr	Strict/Upp/TK02/Igr/Mixed	Strict
Cppratepr	Fixed/Exponential	Exponential(0.10)
Cppmultdevpr	Fixed	Fixed(0.40)
TK02varpr	Fixed/Exponential/Uniform	Exponential(1.00)
Igrvarpr	Fixed/Exponential/Uniform	Exponential(10.00)
Ratepr	Fixed/Variable=Dirichlet	Dirichlet(...,1.0,...)
Generatepr	Fixed/Variable=Dirichlet	Fixed

(b) 그림 14의 help prset 명령어 실행 결과 일부; prset ratepr=variable 실행 전후 밑줄 부분이 바뀌었음에 주목한다.

그림 15. 'help + 명령어'로 각종 설정의 현재 상황과 관련내용을 확인할 수 있다.

```
MrBayes>  mcmcp ngen=1000000 samplefreq=500 printfreq=100
diagnfreq=1000;
MrBayes>  help mcmcp;
```

그림 16. MCMC 알고리즘 실행을 위한 옵션 설정.

Parameter	Options	Current Setting
Ngen	<number>	1000000
Nruns	<number>	2
Nchains	<number>	4
Temp	<number>	0.100000
Reweight	<number>,<number>	0.00 v 0.00 ^
Swapfreq	<number>	1
Nswaps	<number>	1
Samplefreq	<number>	500
Printfreq	<number>	100
Printall	Yes/No	Yes
Printmax	<number>	8
Mcmcdiagn	Yes/No	Yes
Diagnfreq	<number>	1000
Diagnstat	Avgstddev/Maxstddev	Avgstddev
Minpartfreq	<number>	0.10
Allchains	Yes/No	No

그림 17. help mcmcp 실행 결과 일부.

```
MrBayes>   mcmc   ↩
```

그림 18. MrBayes 프로그램 MCMC 실행 명령어

```
Chain results (1000000 generations requested):

    0 -- [-36420.775] (-36612.898) (-36521.562) (-36315.410) * [-36398.170] (-36535.731) (-36347.760) (-36231.370)
  100 -- [-33056.218] (-33453.289) (-34107.536) (-34076.557) * (-33764.603) (-33392.604) [-32423.715] (-32517.073) -- 2:46:39
  200 -- [-31774.765] (-32229.865) (-32461.079) (-32043.488) * (-31670.658) (-31390.711) (-31251.369) [-31072.383] -- 2:46:38
  300 -- [-30321.925] (-31283.716) (-31156.318) (-31389.299) * (-30856.835) (-30666.852) (-30596.379) [-29999.069] -- 2:46:37
  400 -- [-29977.023] (-30688.639) (-30607.615) (-30533.192) * (-29903.261) (-30231.651) (-30199.993) [-29381.049] -- 2:46:36
  500 -- [-29035.889] (-29767.380) (-29943.057) (-29817.365) * (-29423.371) (-29873.605) (-30000.738) [-28721.935] -- 2:46:35
  600 -- [-28667.397] (-29342.569) (-29331.651) (-29442.260) * (-28927.953) (-29207.586) (-29742.622) [-28528.715] -- 2:46:34
  700 -- [-28331.464] (-28718.483) (-29086.234) (-28989.418) * (-28727.895) (-28890.244) (-29191.674) [-28460.043] -- 2:46:33
  800 -- [-28196.837] (-28321.949) (-28999.631) (-28779.440) * (-28395.527) (-28482.847) (-28850.931) [-28227.453] -- 2:25:43
  900 -- [-27934.312] (-28255.836) (-28340.623) (-28570.761) * [-27935.677] (-28297.463) (-28552.078) (-27958.409) -- 2:28:00
 1000 -- [-27774.970] (-28129.473) (-28226.602) (-28451.800) * [-27815.150] (-28148.527) (-28323.983) (-27842.900) -- 2:29:51

Average standard deviation of split frequencies: 0.202031

 1100 -- [-27643.928] (-27947.658) (-28033.841) (-28168.699) * [-27711.066] (-28103.560) (-28150.524) (-27769.080) -- 2:31:20
 1200 -- [-27562.198] (-27795.880) (-27965.956) (-28081.829) * [-27634.301] (-27934.538) (-28070.805) (-27604.640) -- 2:32:35
 1300 -- [-27472.963] (-27475.010) (-27888.699) (-27898.726) * (-27585.441) (-27827.888) (-27999.038) [-27469.711] -- 2:33:38
 1400 -- (-27418.700) [-27388.986] (-27702.430) (-27707.089) * (-27520.325) (-27740.902) (-27962.591) [-27420.087] -- 2:34:32
```

그림 19. MrBayes 프로그램 MCMC 실행 화면.

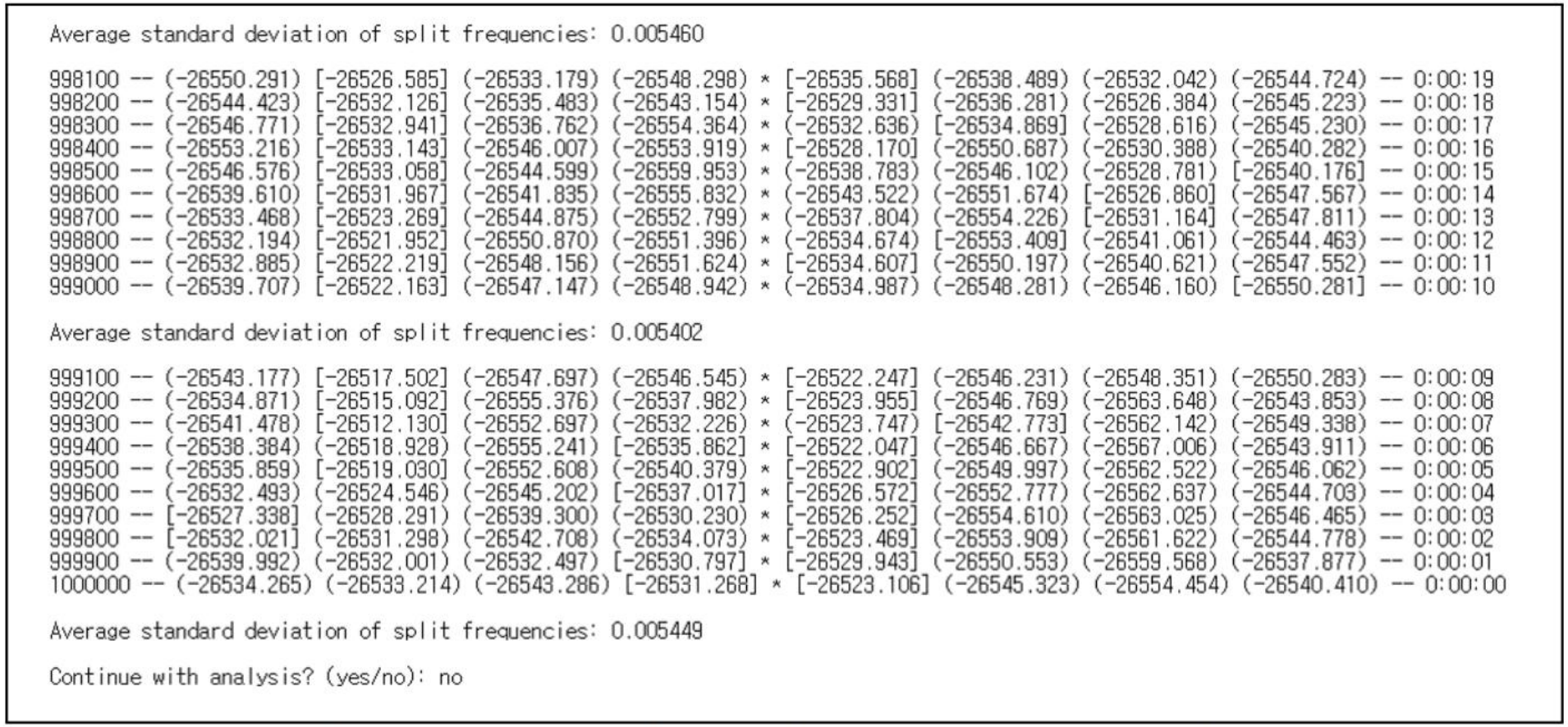


그림 20. MrBayes 프로그램 MCMC 실행으로 마지막 세대에 도달한 장면.

```
MrBayes>  sump relburnin=yes burninfrac=0.25 ;
MrBayes>  sumt relburnin=yes burninfrac=0.25 conformat=simple;
```

그림 21. sump와 sumt 명령어를 이용한 결과의 요약.

* ESS(Effective Sample size)가 작을 때는 붉은(혹은 주황)색으로 표시됨 →.Posterior sample 수를 늘릴 필요가 있음.

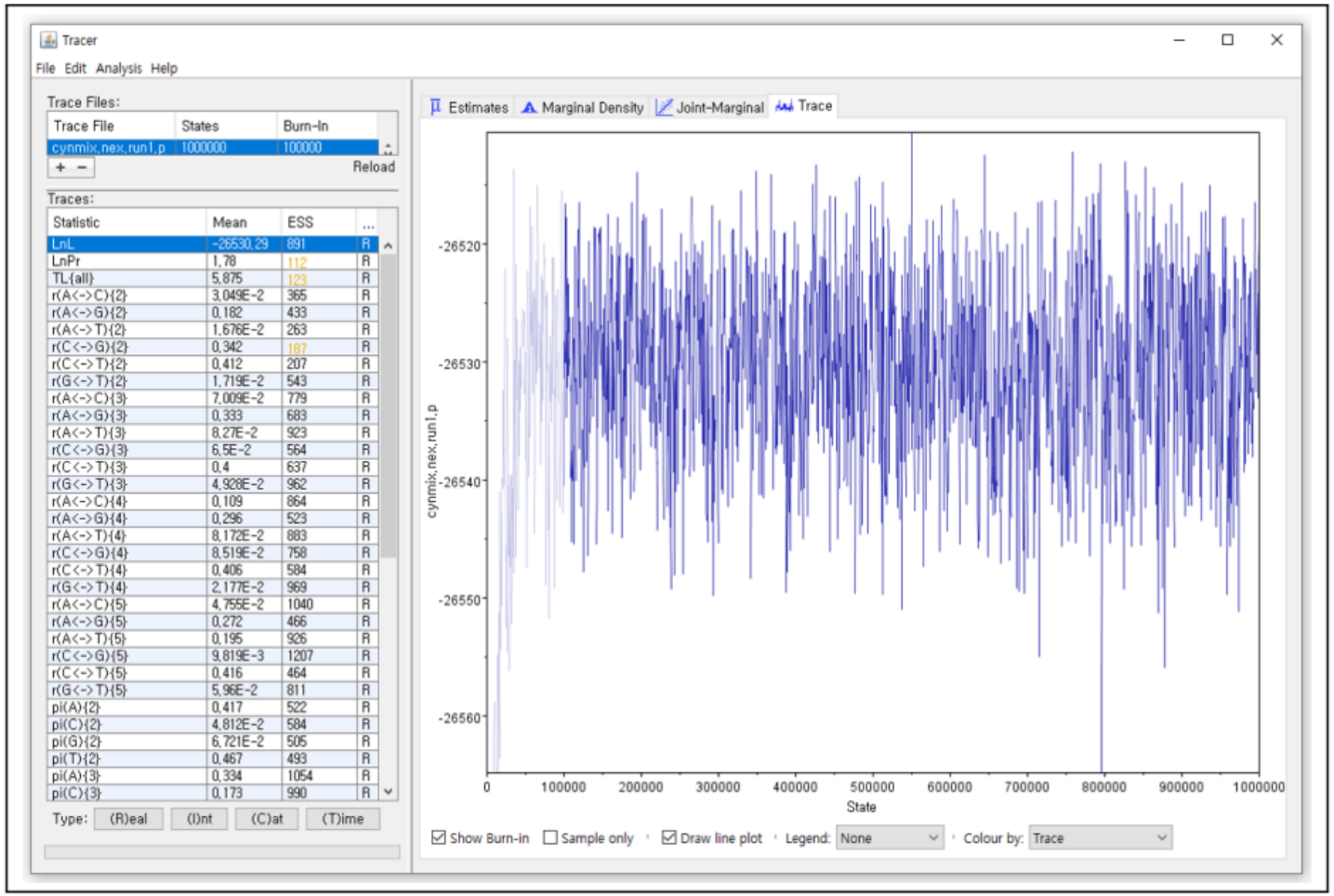


그림 22. Tracer프로그램으로 읽어들인 cynmix.nex.run1.p 파일

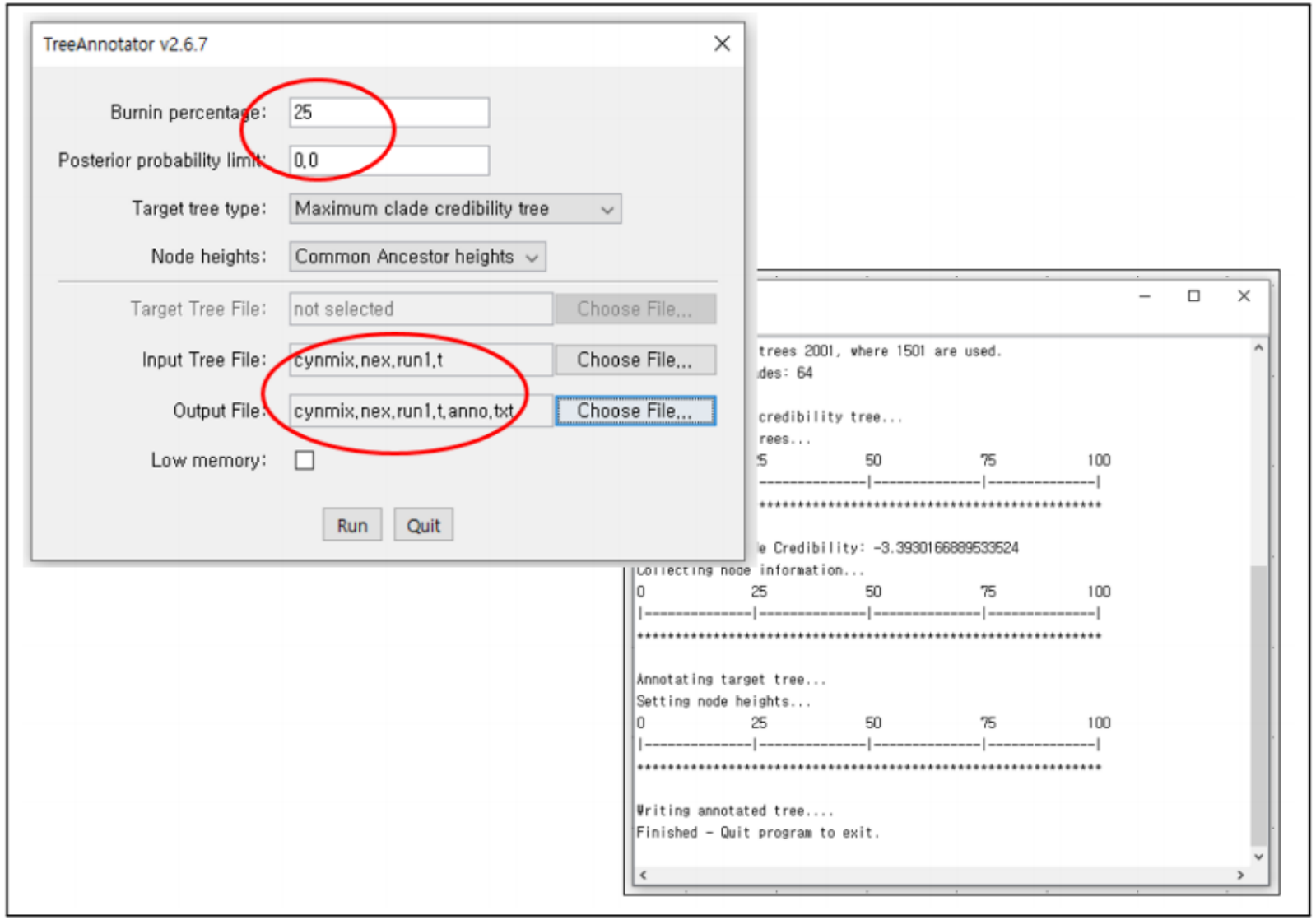
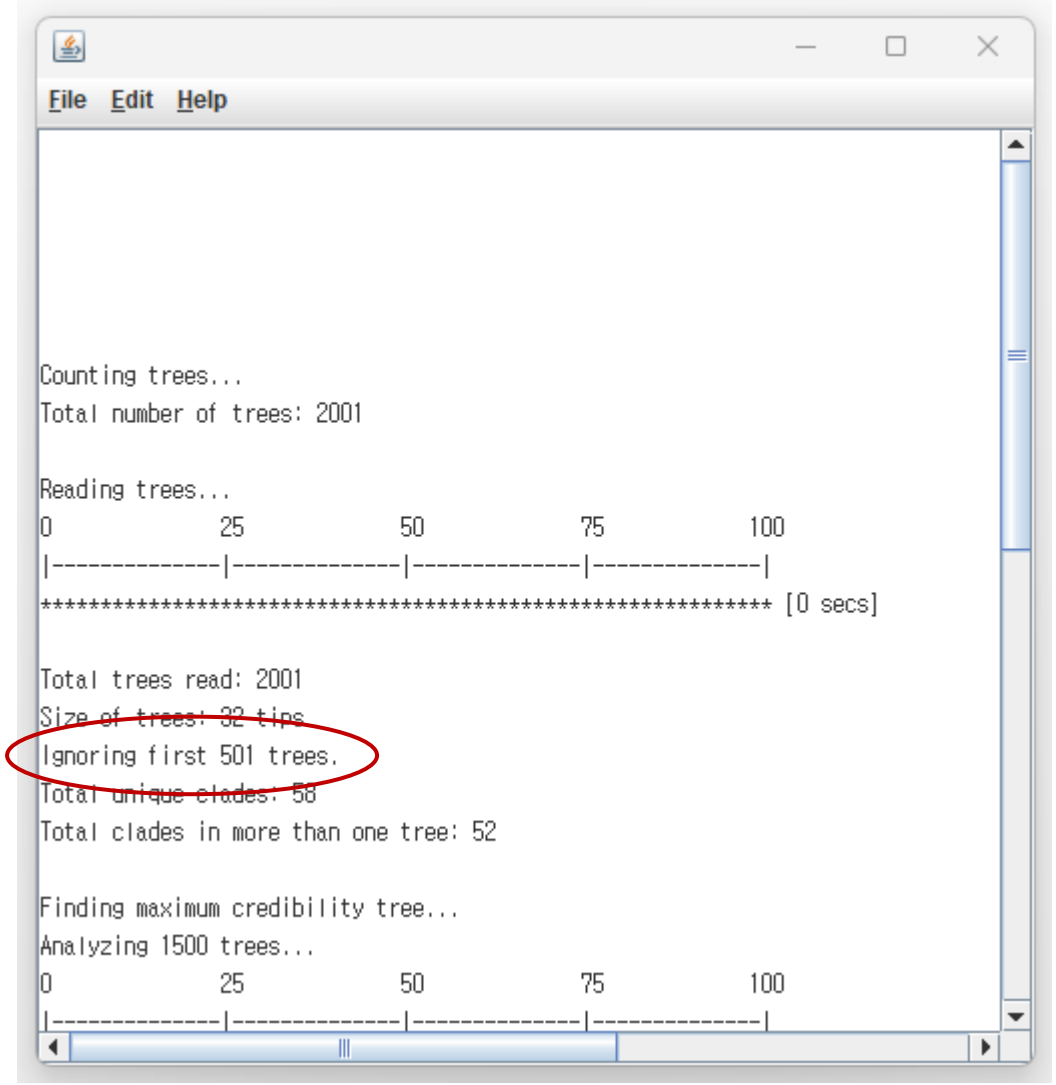
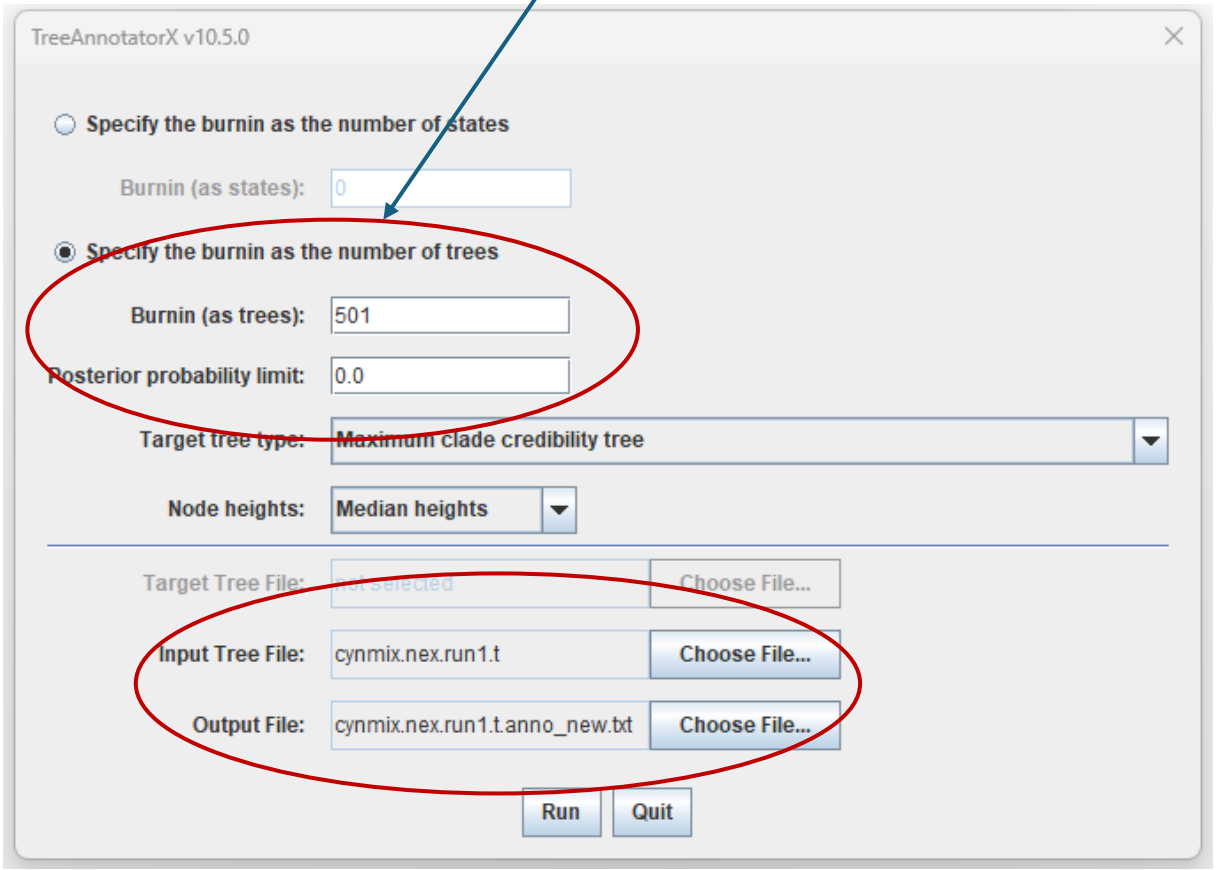


그림 23. TreeAnnotator 프로그램 실행 장면. MrBayes 프로그램 실행으로 얻은 각각의 계통수 샘플로부터 하나의 대표적인 계통수를 얻기 위해 "Target tree type: Maximum clade credibility tree"를 지정하였다.

(TreeAnnotator 버전: 2.7.6)

전체 계통수 샘플은 2001개, 501개를 무시하면 대략 25% 무시하게 됨





(TreeAnnotator 버전: X.v10.5.0; 버전에 따라 인터페이스가 약간 다름)

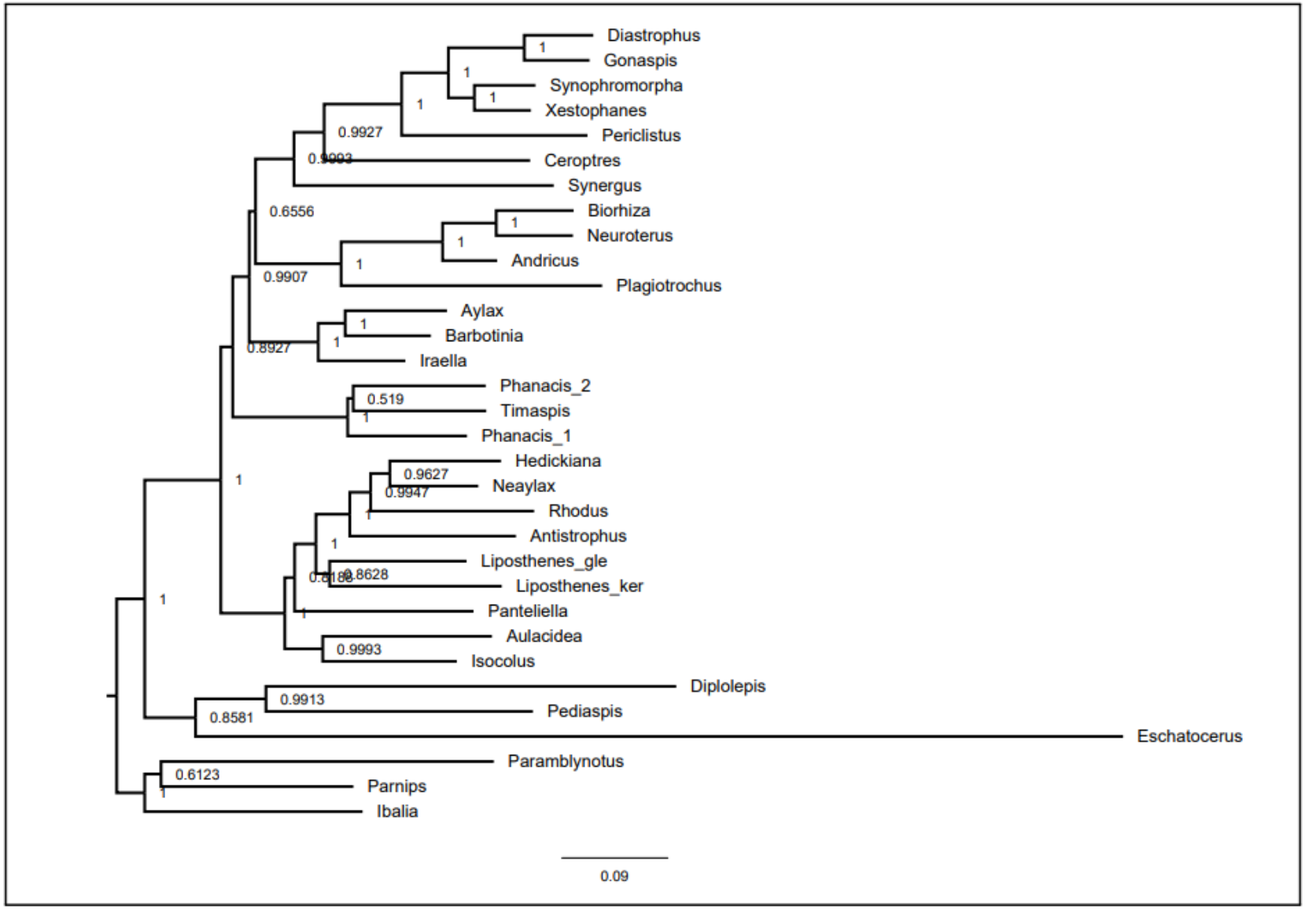


그림 24. 각 노드별로 사후확률이 명시된 계통수. TreeAnnotator 프로그램의 결과를 FigTree 프로그램으로 읽어들인 것이다.

FigTree 프로그램의 왼쪽 패널, Node labes / Display에서 "posterior" 선택