

베이지안 분자계통수 추정(Ver.24.01.04)

서태건^{1,*}

요약: (한글 요약 작성중)(주의: 이 원고는 한국진화학회 투고 준비중인 원고입니다. 최종본이 아니므로 한국진화학회 겨울학교 교재로만 사용해주세요.)

키워드: 키워드1, 키워드2

¹ 인천광역시 연수구 송도동 송도미래로26 극지연구소

* Corresponding author: seo.taekun@gmail.com

1 서론

2 (서론 계속 업데이트중)¹

3 최근 베이지안 방법을 활용한 데이터 분석이 분자진화학 연구에 많이 사용되고 있다(예를 들어 문헌
4 XXXX). 베이지안 방법을 이용하여 데이터를 분석한 논문을 보면 진화생물학자로서는 생소한 용어와
5 개념이 많이 나온다. 사전분포(prior distribution), 사후분포(posterior distribution), MCMC(Markov chain
6 Monte Carlo) 알고리즘, burn-in step 등등... 본 논문은 베이지안 통계의 기본 지식을 설명하여 진화생물학
7 자들이 관련 논문을 읽거나 데이터 분석을 함에 있어서 어려움을 최소화 하게 하려는 목적으로 기획되었
8 다. 또한 본 논문은 MrBayes(Ronquist et al. 2012) 프로그램을 이용한 분자계통수의 베이지안 추정 과정을
9 설명하고 결과의 해석방법을 논의한다.

10 먼저 기본 용어에 대해 알아보자. 동전을 던져서 앞면이 나오는 확률을 구하는 예시를 들어 기본 용어
11 를 설명하겠다.

12 사전분포(prior distribution)

13 아직은 동전을 던지는 실험을 하는 전 단계이다. 즉, ‘동전을 던지는 실험’을 어떤 ‘사건(event)’라고 간
14 주하고 사건이 일어나기 전(‘사전事前’- 이라는 접두어가 붙는다) 확률 대해 생각하는 단계인 것이다.
15 아무런 추가 언급없이 “동전을 던져서 앞면이 나올 확률은 얼마인가?”라고 물으면 베이지안 통계학을
16 접한적이 없는 사람들 대다수는 1/2이라고 대답하려는 경향이 있다. 과연 합리적인 대답일까?

17 동전이 완벽하게(!) 좌우 대칭으로 만들어졌다면 1/2이라는 대답은 매우 합리적일 것이다. 그리고 우
18 리가 일상적으로 보는 동전 대부분은 좌우 대칭에 매우 근접한 형상을 띠고 있다.² 1/2이라고 대답하는
19 사람은 무의식중에 자신의 경험(좌우 대칭에 매우 근접한 동전만 보아온 경험)에 경도되어 그런 답변을
20 한 것이라. 하지만 질문자는 동전이 좌우 대칭으로 공평하게 만들어졌다고 말한적이 없다.

21 이제 답변자가 질문을 함부로 해석하지 않도록 질문자가 보다 명확하게 다시 묻는다 “내 손 안에는
22 비대칭적으로 왜곡되었을지도 모르는 동전이 있다. 이 동전을 던져서 옆면으로 설 경우 앞면 혹은 뒷면이

¹ 본 논문을 읽는 독자는 이전 논문 (서태건 2022)의 내용을 숙지했다고 가정한다. 따라서 AIC, BIC 같은 정보량기준이나 감마분포를 이용한 사이트간 진화속도 이질성 모형등, 기본적인 사항은 자세한 설명없이 기술하도록 하겠다.

² 물리적으로 아무리 정교하게 좌우대칭으로 동전을 제조한다 해도 오차가 필연적으로 존재하기 때문에 엄밀하게 말하면 좌우대칭이라 말할 수 없다.

23 나올때까지 다시 던진다. 이 동전을 던져 앞면이 나올 확률은 얼마인가?” 당신은 뭐라 답하겠는가?

24 동전이 완벽하게 좌우대칭이면 $1/2$ 이겠지만, 심하게 왜곡되어 있다면³ $1/2$ 과는 사뭇 다른 값을 갖게
25 될 것이다. 하지만 동전의 형태를 보기 전에는 $1/2$ 보다 클지 작을지 그리고 $1/2$ 보다 얼마나 크게 차이가
26 날지 알 수 없다.⁴ 이런 경우 앞면이 나올 확률을 어떻게 표현하는 것이 좋을까? 동전의 형태를 모르는
27 상황에서 $1/2$, $1/3$, $2/3$ 같은 어떤 특정한 “하나의 값”으로 대답하는 것은 그다지 합리적으로 보이지 않는
28 다. ‘하나의 값’이 아닌, 여러개의 값을 가질 수 있는 ‘불확실성(uncertainty)’을 고려하여 어떤 확률분포를
29 생각하는 것이 좋지 않을까?⁵

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

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

33 여기에서 $\Gamma(\cdot)$ 는 감마함수를 의미한다. $\text{beta}(\alpha_1, \alpha_2)$ 분포의 평균(μ)과 분산(σ^2)은 다음과 같다.

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

34 그림 1는 모수 α_1, α_2 에 따라 beta 분포의 확률밀도함수가 어떻게 달라지는가 나타낸다. 사전에 어떤
35 정보에 의하여 동전이 비교적 공평하게 만들어졌다는 것을 알게 될 경우에는 그림 1의 $\text{beta}(10, 10)$ 처럼
36 $1/2$ 주위에 분포된 형태를 생각할 수 있다. α_1, α_2 를 같은 값을 갖게 하며 더 증가시키면 $1/2$ 주위에 더
37 뾰족한 분포가 된다. 예컨대 어떤 사전 정보에 의하여 동전이 공평하다고 강하게 확신 할 수 있으면 더
38 뾰족한 분포를 생각할 수도 있다. 반대로, 앞면이 나오기 쉽게(혹은 어렵게) 만들어진 동전이라는 어떤
39 사전정보가 있다면 $\text{beta}(10, 1)$ (혹은 $\text{beta}(1, 10)$) 같은 분포를 θ 의 사전분포로 생각할 수 있다.

40 이처럼 동전을 던지는 실험을 하기 전에(데이터가 가진 정보를 반영하기 전에) 모수에 대한 정보를
41 반영하여 생각하는 분포를 사전분포(prior distribution), 이로부터 얻어진 확률밀도함수를 사전확률밀도
42 함수(prior probability density function), 이로부터 얻어진 확률을 사전확률(prior probability)라고 하는 등,
43 ‘사전(事前, prior)-’이라는 접두어를 사용하여 이러한 정보를 나타낸다. 사전분포를 지정함에 있어 다분히
44 주관성이 개입될 여지가 있어서 베이지안 방법론이 비판을 받기도 한다. 사전에 가진 정보가 없을 때는

³ 예를 들어 원뿔에서 뾰족한 부분을 밑면과 평행하게 잘라낸 형태를 가진, 앞면과 뒷면의 면적이 다른 동전을 생각해보자. 면적의 차이가 클수록 앞면과 뒷면의 확률 차이는 커질 것이다.

⁴ 다시 한번 강조하지만, 질문을 하는 시점은 본격적으로 동전을 던져 앞면의 횟수를 세는 실험을 하는 전단계이다. 그래서 사전(사전)분포, 사전(사전)확률처럼 ‘사전(사전)’이라는 접두어가 붙는다. 즉, 아무 정보없이, 뜬금없이 확률을 묻고 있는 것이다.

⁵ 이 질문에 “아니오”라고 대답하는 독자들이 있다면 베이지안 사고방식에는 어울리지 않을지도 모른다. 비(非)-베이지안(non-Bayesian) 방법을 이용하여 분석을 하는 것을 추천한다. 이는 선택의 문제일 뿐 잘못된 사고방식이라는 의미는 절대 아니다. 지난 논문들(서태건 2022, 2023)에서 사용된 최대가능도 추정법이 대표적인 비-베이지안 방법이다. ‘베이지안 방법’이라는 명시적인 언급이 없으면 비-베이지안 방법이라 보면 된다.

45 주관성의 논란을 최소화하기 위해 모든 가능성에 공평하게 비중을 둔 균등분포를 사전분포로 지정하는
 46 것이 일반적이다.⁶ 그림 1의 $\text{beta}(1,1)$ 분포는 θ 가 가질 수 있는 범위가 0과 1사이에서 균등하게 분포되어
 47 있다. 이처럼 모든 가능성을 균등하게 생각하는 사전 분포는 베이지안 방법론에서 흔히 사용된다.⁷

48 베이지안 사고방식과 비(非)-베이지안(non-Bayesian) 사고방식의 결정적인 차이는 미지의 모수 θ 에
 49 대한 불확실성을 다루는 방식이다.⁸ 베이지안 사고방식에서는 미지의 모수 θ 는 어떤 ‘고정된 값’이 아닌
 50 ‘유동적인 값’이다.⁹ 따라서 식 1과 같이 어떤 확률 분포를 이용하여 모형화 하는 것이 논리적으로 자연스
 51 럽다. 하지만, 비-베이지안 사고방식은 ‘ θ 는 (그 값이 어떤 값인지는 비록 모르지만) 어떤 하나의 값으로
 52 고정되어 있다.’라는 입장을 견지한다. ‘하나의 값’으로 고정되어 있으므로 식 1과 같은 분포를 생각하는
 53 것은 논리적으로 옳지 않다.

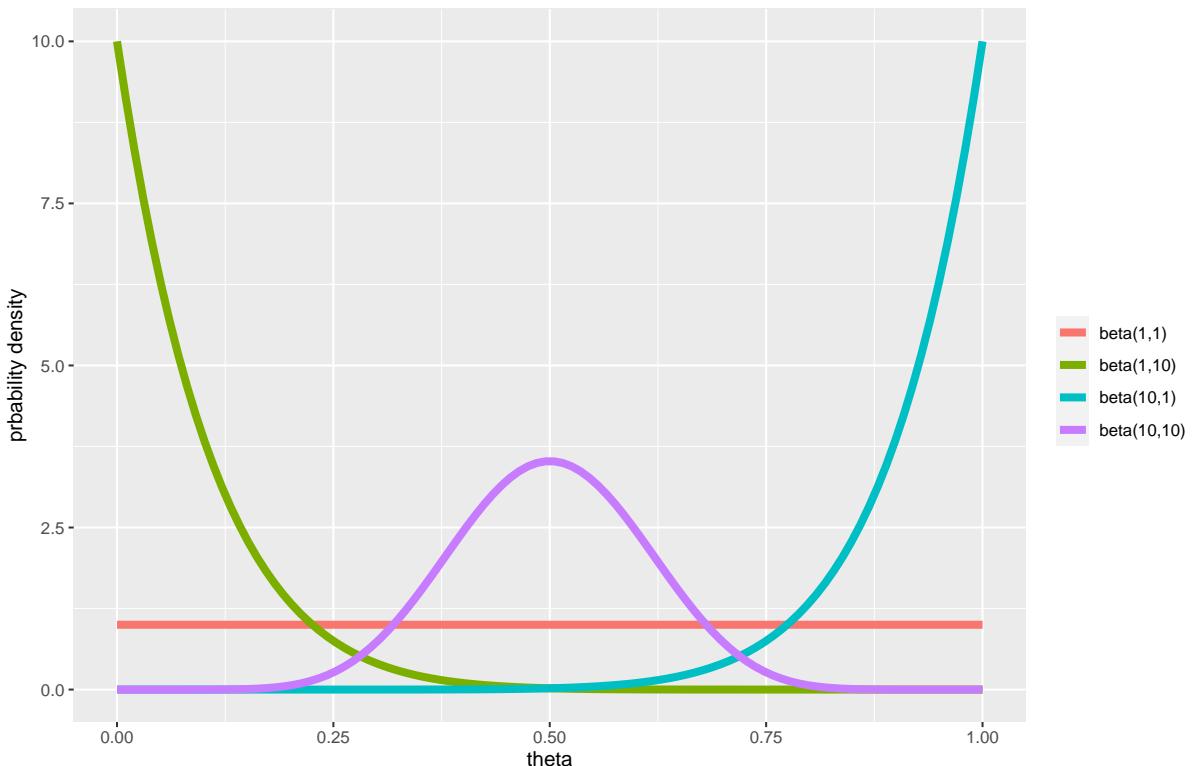


그림 1. Beta 분포 확률밀도함수의 여러 형태

54 가능도(likelihood)

55 이제 앞면이 나올 확률을 추정하기 위해 본격적으로 실험을 해보자. 실험은 간단하다. 동전을 n 번 던져
 56 앞면이 나온 횟수 k 를 세면 된다. 앞면이 나올 확률이 θ 일 때 n 번 동전을 던져 k 번 앞면이 나오는 확률은

⁶ 이를 무차별의 원칙(principle of indifference)라고 한다(문현).

⁷ 뒤에서 설명하지만 데이터의 양이 충분하면 어떤 사전 분포를 선택하더라도 결과에 큰 차이가 없다. 실제 데이터 분석에서 사전분포의 주관성은 “데이터의 양이 많다면” 큰 문제가 되지 않는다.

⁸ non-Bayesian framework을 frequentist framework이라 부르기도 한다.

⁹ 설명의 편의상 ‘값’이라는 단어를 사용하였으나, θ 는 실수의 모임인 벡터나 행렬이 될 수도 있고 본론에서 설명하는 계통수의 계통관계가 될 수도 있다. 일반적으로 우리가 추정하고자 하는 모수를 통칭한다고 이해하면 된다.

57 ○ 항분포(binomial distribution)를 이용하면 다음과 같이 표현할 수 있다.

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

58 이 단계(즉, 실험으로 데이터를 얻는 단계)에서는 앞에서 논의한 사전확률은 고려하지 말고 일단은 데이터
59 가 제공하는 정보에만 집중하여 살펴보도록 하자. 이 경우 식 (3)의 확률밀도함수는 가능도함수(likelihood
60 function)가 되고 (n, k) 가 주어졌을 때 가능도를 최대로 하는 θ 의 ML 추정량 (maximum likelihood estima-
61 tor)과 n 이 충분히 클 때 95% 신뢰구간(Confidence Interval; CI)은 다음과 같음을 기초통계학 교재 등을
62 통해 쉽게 확인할 수 있다(예를 들어, 문현).

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

63 가령 문제의 동전을 100번 던져 앞면이 60번 나왔다고 하자. 식 (4)를 적용하면 최대가능도 추정량은
64 $60/100, 95\%$ 신뢰구간은 $(0.504, 0.696)$ 과 같이 구해진다. 이 결과는 앞에서 논의한 사전확률을 전혀 고-
65 려하지 않은 비-베이지안 추정 결과이다.

66 사후분포(posterior distribution)

67 식 (1)의 사전분포와 식 (3)의 가능도를 베이지안 방식으로 결합한 사후분포를 생각해보자. ‘데이터를 관-
68 칠하는 사건’이 발생한 이후라는 의미에서 ‘사후(事後)-’라는 접두어를 사용하여 관련 용어를 정의한다.
69 자세한 내용은 본론에서 설명하도록 하고 요점만 간단히 말하면 “사후분포는 사전분포와 가능도의 곱에
70 비례하는 형태로 주어진다.”

$$\text{사후분포} \propto \text{가능도} \times \text{사전분포} \quad (5)$$

71 동전을 던져 앞면이 나올 확률 θ 에 대한 사후 분포를 위의 식에 준하여 고찰하고, 사후분포는 확률분포
72 이므로 θ 가 가질 수 있는 범위에서 확률분포의 적분값이 1이 되도록 정규화하면 사후확률분포는 다음과
73 같이 beta 분포로 주어짐이 알려져 있다(김우철 2021 p. 452).

$$p(\theta|X) \propto p(X|\theta)p(\theta) \quad (6)$$

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

74 즉, 요약하면 beta(α_1, α_2)를 사전분포로 가정하고 동전 던지기로 (n, k) 결과를 얻었을 때 사후분포는 beta($\alpha_1 +$
75 $k, n - k + \alpha_2$)로 주어진다. ¹⁰ 식 (2)를 이용하여 구한 사후분포의 평균은 $(\alpha_1 + k)/(\alpha_1 + \alpha_2 + n)$ 이 된다.

¹⁰ 이처럼 사전분포와 사후분포가 같은 분포형태를 갖을 때 사전분포를 컨디전트 분포(conjugate prior distribution)라고 하며, 이때 사후분포는 해석적으로 간단히 계산된다. 하지만 실제 데이터 분석(후술하는 분자계통수 분석과 같은)에서 컨디전트 분포

76 평균이 사전분포의 $\alpha_1/(\alpha_1 + \alpha_2)$ 로부터 데이터가 제공하는 정보에 영향을 받아 업데이트된 것이다.

77 그림 1에서 보인 사전분포 $\text{beta}(1, 1)$, $\text{beta}(1, 10)$, $\text{beta}(10, 1)$, $\text{beta}(10, 10)$ 각각에 대해 실험 결과 $(n, k) = (100, 60)$ 의 가능도를 결합해 얻는 사후분포는 각각 $\text{beta}(61, 41)$, $\text{beta}(61, 50)$, $\text{beta}(70, 41)$, $\text{beta}(70, 50)$ 이다(그림 2). 여기에서 주목할 점은 사전분포들 간에 보이는 큰 차이가 많이 완화되어 사후분포들끼리는 서로 유사하는 점이다. 이는 사전분포간의 평균의 차이와 사후분포간의 평균의 차이를 식 (2)를 이용하여 직접 계산/비교해보면 알 수 있다. 일반적으로 데이터가 충분히 클 경우 식 (5)에서 가능도가 사전분포를 암도하게 되어 사후분포는 거의 가능도에 의해 좌우된다¹¹

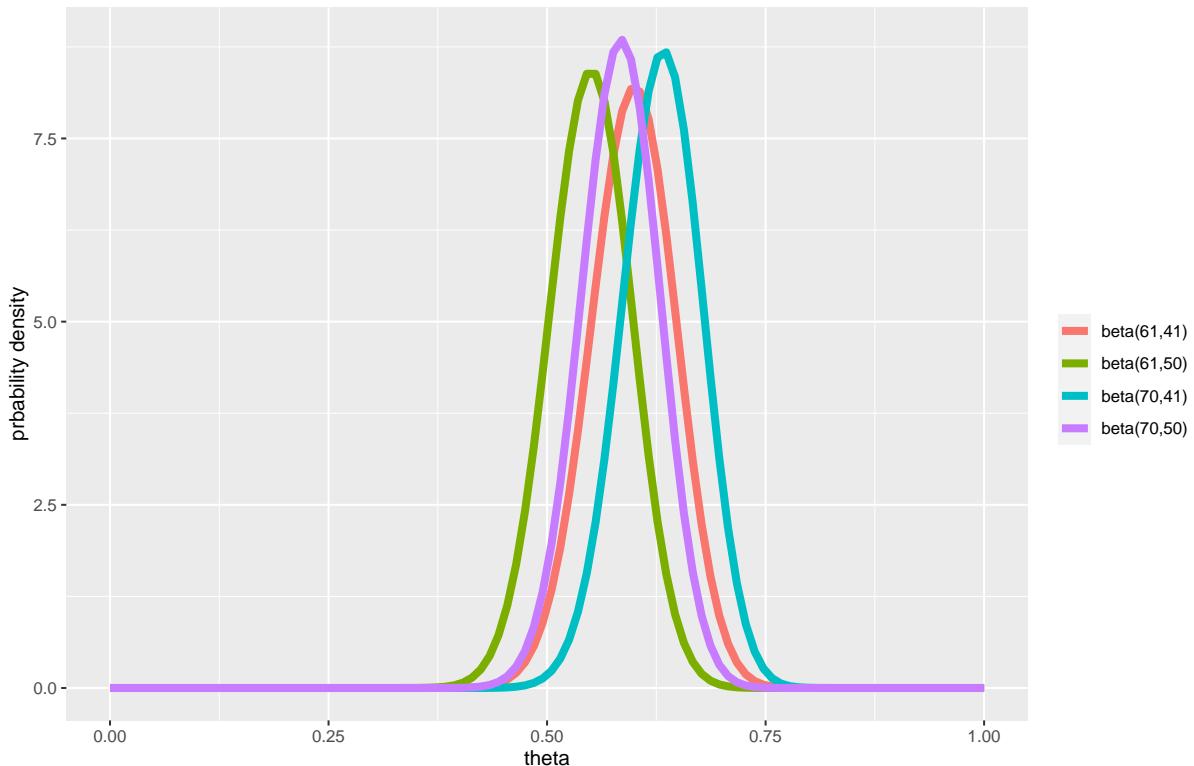


그림 2. 사후 Beta 분포 확률밀도함수의 여러 형태

83 베이지안 방식 vs. 빈도주의 방식

84 식 (5)처럼 사전분포와 가능도의 곱을 이용하여 사후분포를 생각하는 방식이 베이지안 방식(Bayesian framework)이다. 비-베이지안(non-Bayesian) 방식은 통상 빈도주의 방식(frequentist framework)이라고도 불리는데 빈도주의 방식에서는 사전확률을 고려하지 않고 가능도만을 고려해서 문제를 해결한다. 즉, 빈도주의 방식에서는 모수가 어떤 하나의 값으로 고정되어 있기 때문에 분포를 생각하는 것은 논리적으로 타당하지 않다. 따라서 $p(\theta)$ 없이, 단지 식 (3)으로 주어진 가능도 함수만을 이용하여 미지의 모수를

를 찾는 것은 불가능한 경우가 많아 일반적으로는 수치적으로(numerically) 사후분포를 구한다. 이때 사용되는 것이 후술하는 MCMC 알고리즘이다.

¹¹ 이런 현상을 흔히 “Likelihood dominates prior”라고 표현하기도 한다. 따라서 데이터의 양이 충분히 클 때는 어떠한 사전분포를 가정하더라도 사후분포는 크게 영향을 받지 않아 사전분포의 주관성을 크게 우려하지 않아도 된다. 본 실험의 경우 데이터의 양을 더 늘려 $(n, k) = (1000, 600)$ 을 얻었다고 하자. 앞면이 나온 비율은 전과 같이 600/1000으로 동일하지만 데이터가 가진 정보의 양은 훨씬 많다. 사후분포들의 형태와 평균값을 비교해보면 $(n, k) = (100, 60)$ 의 경우보다 훨씬 더 서로 유사해졌다는 것을 확인할 수 있다.

89 추정하는 것이다.¹² 반면 베이지안 방식에서는 사전에 실험자가 인식하고 있는 확률분포를 정하게 하고,
 90 실험에서 얻은 결과를 이용하여 사전에 갖고 있는 인식을 수정하게 하는 것이다. 수정된 인식은 사후확률
 91 분포로 구체화된다.

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

97 빈도주의 방식에서 ‘95% 신뢰구간’의 올바른 해석은 “동전을 100번 던져 앞면이 나오는 횟수를 측정
 98 하고 식(4)에 의해 신뢰구간을 구하는 작업을 무한히 반복하면 대략 100번중 95번의 비율로 신뢰구간이
 99 참값을 포함한다. (0.504, 0.696)은 이러한 무한히 많은 신뢰구간 중 하나이고 이 구간이 참값을 포함하
 100 는지 아닌지는 알 수 없다.”라는 것이 올바른 해석이다. 이는 직관적으로 잘 와닿지 않고 이해하기 힘든
 101 해석이다. 그럼 3는 θ 의 참값이 1/2일 때 동전을 100번 던져 식 (4)에 의해 신뢰구간을 구하는 과정을 100
 102 번 반복한 모의실험(simulation) 결과이다. 참값 $\theta = 1/2$ 를 신뢰구간이 포함하면 검은색, 포함하지 않으면
 103 붉은색으로 표시하였다. 이 모의실험에서는 100회중 검은색 신뢰구간이 93회 관찰되었으나 모의실험의
 104 횟수를 늘이면 점근적으로 95%에 근사하게 된다.

105 빈도주의 방식과 달리 베이지안 방식에서 신뢰구간 (credible interval)¹³은 직관적인 해석이 가능하
 106 다. 즉, “미지의 모수가 주어진 신뢰구간안에 포함될 확률이 95%이다”라는 해석이 가능한 것이다. 이는
 107 미지의 모수 θ 가 어떤 고정된 값이 아니라 랜덤하다는 사전분포의 가정으로 가능하게 된 것이다.¹⁴

108 본론

109 사후분포를 보다 정확히 이해하기 위해서 조건부 확률에 대해 살펴보자.

110 조건부 확률의 계산

111 그림 4는 100가구로 이루어진 어느 마을의 신문 구독 상황을 모식적으로 나타낸 것이다. 전체 100가구 중
 112 A 신문을 구독하는 가구는 25가구, B 신문을 구독하는 가구는 20가구, 두 신문 모두 구독하는 가구는 5
 113 가구, 두 신문 모두 구독하지 않는 가구는 60가구이다. 랜덤하게 선택한 가구가 A 신문을 구독자일 확률,
 114 B 신문 구독자일 확률, 두 신문 모두 구독자일 확률은 그림 4의 벤 다이어그램을 통해 쉽게 계산할 수 있다:
 115 $P(A) = 25/100 = 1/4$, $P(B) = 20/100 = 1/5$, $P(A \cap B) = 5/100 = 1/20$.

¹² 가능도를 최대로 하는 모수를 찾는 추정법이 최대가능도 추정법(maximum likelihood estimation)이다.

¹³ 베이지안 방식에서는 ‘credible interval’, 빈도주의 방식에서는 ‘confidence interval’, 머릿글자는 똑같이 CI로 쓰지만 영문용 어가 다른데 주의한다. 우리말로는 똑같이 ‘신뢰구간’으로 번역하며, 전자를 ‘신용구간’으로 번역하기도 한다.

¹⁴ 식 (4)에 대응되는 베이지안 신뢰구간은 일반적으로 간단한 수식으로 표현하기는 어려워 수치적으로 구한다.

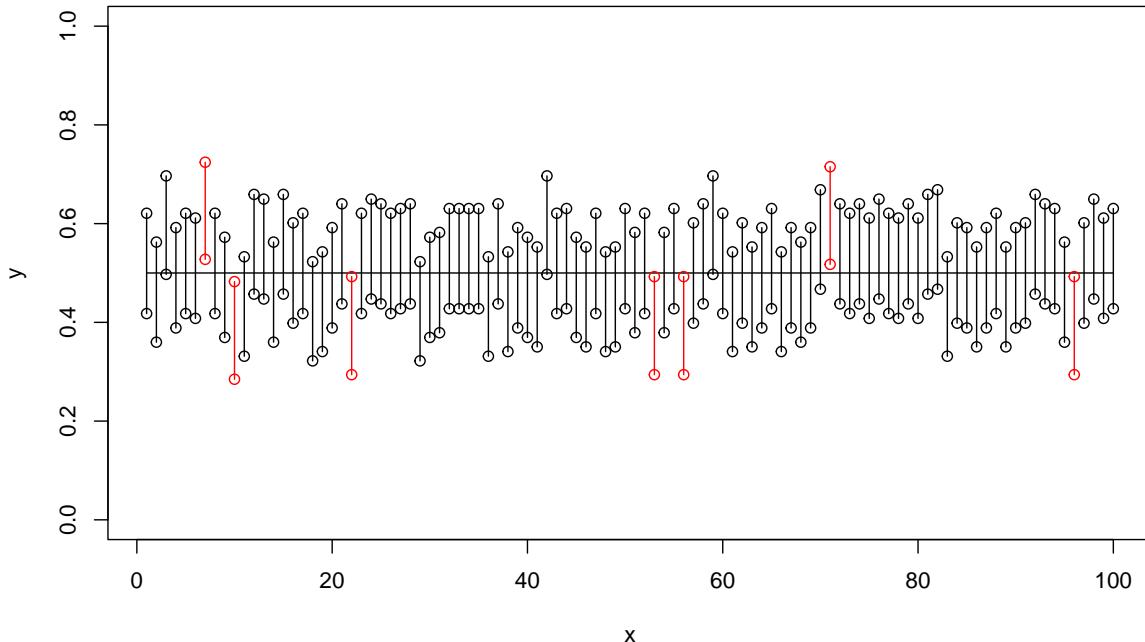


그림 3. 95% 신뢰구간(Confidence Interval)의 의미를 알아보기 위한 모의실험

116 두 신문 모두 구독자일 확률 $P(A \cap B)$ 은 다음과 같이 조건부 확률로 표현할 수 있다.

$$\begin{aligned} P(A \cap B) &= P(A)P(B|A) \\ &= P(B)P(A|B) \end{aligned} \tag{8}$$

117 즉, A가 선택될 확률에 A의 조건하에 B가 선택될 확률을 곱하거나 (첫째 줄), 혹은 B가 선택될 확률에 B
118 의 조건하에 A가 선택될 확률을 곱함으로써 표현할 수 있다(둘째 줄).

119 그림 4의 예에서는 $P(B|A) = P(B)$ 이 성립한다. 이처럼 A라는 조건하에서 B의 확률과 전체 집단에서 B
120 의 확률이 같을 때(마찬가지로 B라는 조건하에 A의 확률과 전체 집단에서 A의 확률) A와 B는 독립이라고
121 하고 다음이 성립한다.

$$P(A \cap B) = P(A)P(B) \tag{9}$$

122 베이즈 정리와 사후분포의 계산

123 데이터 D 가 주어졌을 때 미지의 모수 θ 에 대한 확률분포(probability distribution; 이하 간단히 ‘분포’라
124 하자)를 생각해보자. 식(8)을 응용하면 다음과 같이 표현할 수 있다.

$$P(\theta|D) = \frac{P(\theta \cap D)}{P(D)}$$

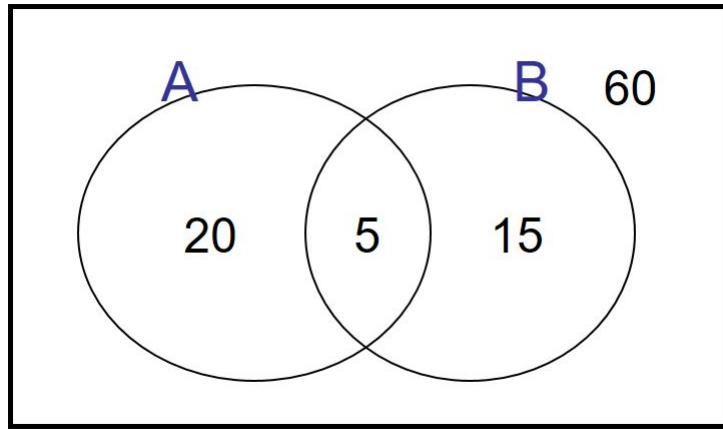


그림 4. 100가구의 신문 구독 상황 모식도

125 여기에 다시 분자를 다른 형태의 조건부 확률로, 분모를 조건부 확률의 합(혹은 적분)으로 나타내면 다음
126 과 같은 베이즈 정리를 얻는다.

$$\begin{aligned} P(\theta|D) &= \frac{P(\theta \cap D)}{P(D)} \\ &= \frac{P(D|\theta)P(\theta)}{\sum_{\theta} P(D|\theta)P(\theta)} \quad (\text{모수 } \theta \text{가 이산형인 경우}) \end{aligned} \quad (10)$$

$$= \frac{P(D|\theta)P(\theta)}{\int_{\theta} P(D|\theta)P(\theta)} \quad (\text{모수 } \theta \text{가 연속형인 경우}) \quad (11)$$

127 여기에서 분모의 $P(D)$ 는 데이터의 ‘확률’이다. 서론에서 설명한 동전 던지기 사례는 $P(D)$ 의 계산이 가능
128 한 특수한 경우이고, 일반적으로 $P(D)$ 는 매우 많은 항의 합이나 다중적분의 형태로 정의되어 직접적인
129 계산이 현실적으로 불가능하다. 이 문제를 해결하기 위해 개발된 것이 Markov chain Monte Carlo(MCMC)
130 알고리즘이다(Metropolis et al. 1953; Hastings 1970).¹⁵ 데이터 D 가 관찰된 이후에 D 는 변하지 않는 고정된
131 값이다. 우리의 주된 관심사는 θ 이고 $P(D)$ 는 θ 에 의존하지 않는다는 점에 착안하면 $P(D)$ 를 무시할 수
132 있고 식 (10,11)는 다음과 같이 간략하게 표현될 수 있다.

$$P(\theta|D) \propto P(D|\theta)P(\theta) \quad (12)$$

133 이는 서론에서 언급한 식 (5)과 같은 형태이다. MCMC 알고리즘은 $P(D)$ 가 θ 에 의존하지 않는다는 사실을
134 교묘하게 이용하여 $P(D)$ 를 계산하지 않고도 $P(\theta|D)$ 를 구할 수 있게 하는 훌륭한 알고리즘이다.

135 **MCMC** 알고리즘을 이해하기 위한 사고실험 1

136 당신이 앞을 볼 수 없는 시각장애인이라고 가정해 보자. 당신은 당신 집 근처에 있는 산의 형태를 알고
137 싶다. 어떻게 하면 될까? 먼저 산이 있는 영역을 대략 직사각형 형태로 특정하고 가로(x축)와 세로(y축)
138 를 적절하게 구획한다. 가령 가로, 세로 각각 99개의 구역으로 구획한다면 모두 10000($= 100 \times 100$)개의

¹⁵Metropolis–Hastings algorithm이라고도 불린다.

139 좌표가 생성된다. 각각의 좌표를 모두 방문하여 각 좌표의 지점에서 산의 상대적인 높이를 측정하고 이를
 140 종합하면 산의 형태를 알 수 있게 된다. 이처럼 2차원 평면 위를 탐색하는 경우 차원이 낮기 때문에 모든
 141 좌표지점을 방문하는 것이 그다지 어렵지 않다. 여기서 등장하는 x 축, y 축은 각종 데이터 분석에서 모수에
 142 비유할 수 있다. 모수가 2개일 경우에는 이처럼 모수의 구간을 잘개 나누어 모수의 조합 각각을 방문하여
 143 평가하는 것이 가능하다. 하지만 모수의 갯수가 많은 경우에는 이처럼 구획된 모든 좌표를 방문하는 것은
 144 불가능하다. 따라서 당신은 보다 더 일반적인 상황을 상정하여 좌표를 방문하는 전략을 계획하고 싶다.

145 **MCMC 알고리즘을 이해하기 위한 사고실험 2**

146 당신은 당신의 집앞에서 출발한다. 뒤에서 설명할 MCMC 용어를 사용하여 이 시점을 세대 0(generation
 147 0)라고 부르자. 다음 시점 (generation 1)위치의 후보를 정하기 위해서 4면으로 이루어진 주사위를 던진다.
 148 1이 나오면 동쪽으로 1미터, 2가 나오면 서쪽으로 1미터, 이런 식으로 다음 위치의 후보를 동서남북 중 하
 149 나, 거리는 1미터로 하여 정하자(후술하는 proposal step이다). 후보 위치가 정해졌으면 현재 위치와 후보
 150 위치의 상대적인 높낮이를 비교한다. 해발고도는 몰라도 상대적인 높이는 쉽게 측정할 수 있다.¹⁶ 만약
 151 후보위치가 현재보다 높거나 같으면 무조건 후보위치로 이동한다. 만약 후보위치가 현재보다 낮으면 확
 152 률적으로 간다. 즉, 현재보다 높이가 낮지만 차이가 미미하다면 이동할 확률이 높고, 현재위치보다 높이가
 153 현저하게 낮다면 그쪽으로 이동할 확률은 현저하게 낮아진다. 후보위치를 채택하면 그쪽으로 이동하고
 154 만약 확률적으로 채택되지 못했다면 현재위치에 계속 머무른다(후술하는 accept/rejection step이다). 이
 155 과정이 끝나면 ‘세대 1(generation 1)’이 되어 세대수가 하나 증가한다. 이런 전략으로 계속 이동한다면
 156 어떻게 될까? 처음에는 집 주위의 평지를 의미없이 맴도는 식으로 움직이겠지만 어느 순간에는 산의 밑
 157 자락에 도착할 것이고 그때부터는 산의 정상방향으로 이동하는 경향이 강할 것이다(그림 5). 세대수가
 158 증가하면 당신은 산의 정상 근처에서 동서남북으로 오르락 내리락 하는 움직임을 보일 것이다. 세대수가
 159 충분히 지난 후에 당신이 방문한 지점들을 모아 2차원 히스토그램을 그려보자. 대략 산의 형태와 유사
 160 할 것이다. 이 전략을 취하면 사고실험 1에서 지정된 좌표지점을 모두 방문하지는 못한다. 그리고 같은
 161 위치를 여러번 방문하게 되어 일견 비효율적으로 생각될 수도 있다. 하지만, 모수의 차원이 커져서 모든
 162 좌표를 방문할 수 없는 경우에는 아주 효율적인 좌표 방문 전략이 된다. 한번에 이동하는 거리를 1미터보
 163 다 더 작게 설정하고 움직이는 방향을 8방향 16방향 등으로 개선한다면 더 정밀한 산의 형태를 얻을 수
 164 있을 것이다.

165 **MCMC 알고리즘**

166 ‘사고실험 2’의 전략을 형식화 하여 다음과 같이 기술할 수 있다.

- 167 1. 0번째 세대에서 출발점이 되는 θ 는 적당히 임의로 정한다.흔히 사전확률분포 $P(\theta)$ 로부터
 168 랜덤하게 정한다.
- 169 2. 현재의 모수 θ 로부터 다음 세대의 모수 후보 θ' 를 $J(\theta'|\theta)$ 확률밀도 함수로 구한다.¹⁷

¹⁶해발고도(절대높이)를 모른다는 상황 설정은 식 (12)에서 $P(D)$ 를 모른다는 상황에 비유될 수 있다.

¹⁷이 단계를 proposal step 이라 부른다.

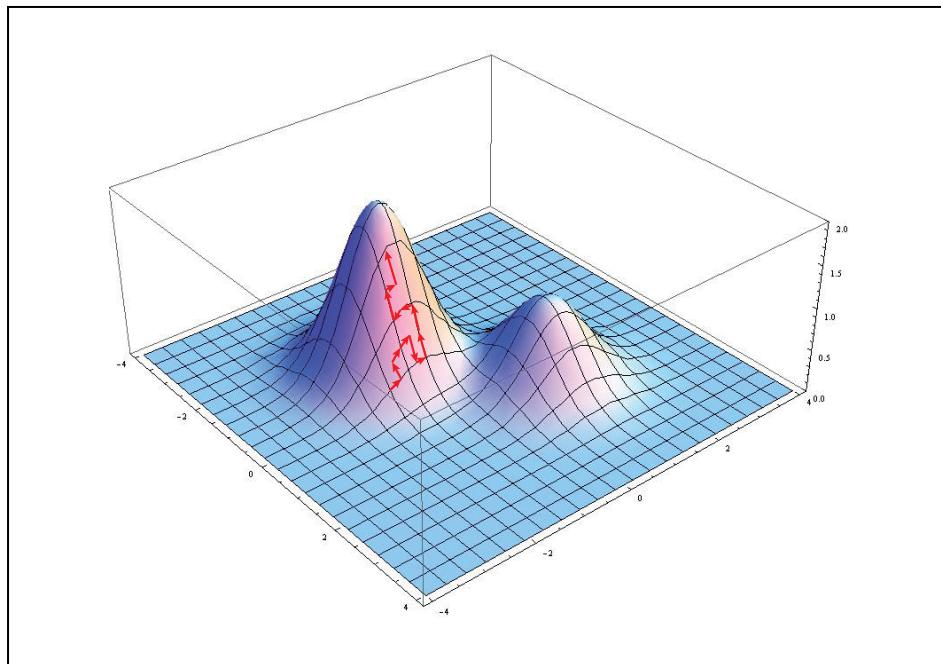


그림 5. 사고실험2의 등산 경로 모식도(붉은색 화살표). 고도가 낮은 곳으로 이동하기도 하지만 전반적으로 높은 곳으로 가려는 경향성이 강하다. 따라서 고도가 높은 곳은 그만큼 자주 방문하게 되어 방문 지점을 2차원 히스토그램으로 나타내면 산의 모양과 대략 비슷하게 된다.

170 $J(\theta'|\theta)$ 함수는 해결하려는 문제에 따라서 적절히 결정해야 성능이 좋아진다.

171 3. θ' 가 결정되면 아래의 값을 계산한다.

$$\alpha(\theta, \theta') := \min \left\{ \frac{P(D|\theta')P(\theta')J(\theta|\theta')}{P(D|\theta)P(\theta)J(\theta'|\theta)}, 1 \right\} \quad (13)$$

173 여기에서 $\min\{\cdot\}$ 함수는 둘 중 작은 수를 택하는 함수이다. 만약 $\alpha(\theta, \theta')$ 이 1이면 현재의
174 모수를 θ 로부터 θ' 로 업데이트한다(acceptance). 만약 $\alpha(\theta, \theta')$ 이 1보다 작으면 (0, 1) 구간의
175 균등분포로부터 난수를 생성한다. 생성된 난수가 $\alpha(\theta, \theta')$ 보다 작으면 θ 로부터 θ' 로 업데이
176 트한다(acceptance). 생성된 난수가 $\alpha(\theta, \theta')$ 보다 크면 θ 를 그대로 유지한다(rejection).¹⁸

177 4. 세대수를 하나 증가시킨다.

178 5. 사전에 설정한 세대수에 도달했으면 알고리즘을 종료한다. 그렇지 않으면 2번으로 간다.

179 MCMC 알고리즘을 실행하는데 중요한 세가지 용어에 대해 언급한다:(1)number of generation, (2)num-
180 ber of burn-in step, (3)sampling interval. 위의 2~4 단계를 반복하여 사전에 정한 세대수(이를 number of
181 generation이라 한다)만큼 방문한 θ 의 값들은 사후분포 $P(D|\theta)$ 로부터 얻은 랜덤 샘플로 간주될 수 있다.
182 하지만 연속된 세대의 θ 값들은 매우 높은 상관관계를 갖는다. 우리가 원하는 것은 서로 독립적인 랜덤
183 샘플이므로 일정한 세대 간격을 두고 (이것이 sampling interval이다) θ 값들을 모아 저장할 필요가 있다.
184 또한, MCMC 실행의 초기 세대에는 θ 값들이 사후분포와는 거리가 먼 영역을 배회한다.¹⁹ 따라서 초기의

¹⁸oi 단계를 acceptance/rejection step이라 부른다.

¹⁹사고실험 2에서 산 밑자락에 도달하기 전에 평지를 배회하는 것에 비유할 수 있다.

185 일정한 세대수(이)를 number of burn-in step이라 한다)를 샘플링에서 제외할 필요가 있다.

186 **MCMC** 알고리즘에 대한 고찰

187 그림 6 – 7은 현재의 모수 θ 에서 다음 세대의 모수 후보 θ' 를 제안하는 함수인 $J(\theta'|\theta)$ 가 중요하다는 것을
188 보여준다. 사후분포가 봉우리 두개로 이루어진 형태라고 가정하자. θ 와 θ' 간의 거리가 너무 가까우면(그
189 림 6) 산이 높고 골짜리가 깊은 경우 다른 산으로 이동하지 못해 두개의 봉우리중 하나의 봉우리 주위만
190 맴돌게 된다. 또한 거리가 너무 멀면(그림 7), θ 가 봉우리 정상 부근에 있을 경우 제안되는 θ' 의 위치가
191 너무 낮아 이로부터 계산되는 식 (13)의 $\alpha(\theta, \theta')$ 값이 매우 낮게되고 따라서 θ' 로의 업데이트가 일어나지
192 않아 계속 동일한 위치에 머무르게 된다. $J(\theta'|\theta)$ 를 지정함에 있어 θ 와 θ' 간의 거리가 너무 가깝지도 않고
193 멀지도 않게 설정하는 것이 중요하지만 이는 사전 분포와 가능도에 따라 달라지므로 $J(\theta'|\theta)$ 를 효율적으
194 로 설계하는 것은 쉽지 않다. 이를 해결하는 방법으로 cold chain과 hot chain을 동시에 실행시키는 방법이
195 있다. 우리가 추정하고자 하는 사후분포에 따라 θ 가 움직이는 것을 ‘cold chain’이라 명명하고 사후분포에
196 열을 가해 봉우리가 녹아내린 그림 8과 같은 가상의 사후분포에 따라 θ 가 움직이는 것을 ‘hot chain’이라
197 명명한다. θ 와 θ' 간의 거리에 민감하게 영향을 받는 cold chain에 비하여 hot chain에서는 θ 값들 비교적
198 자유롭게 이동할 수 있다. Cold chain과 hot chain을 동시에 실행시키며 정기적으로 θ 값들을 맞바꾸는
199 작업을 병행하면 cold chain이 가진 $J(\theta'|\theta)$ 함수 설계의 문제점을 개선할 수 있다.

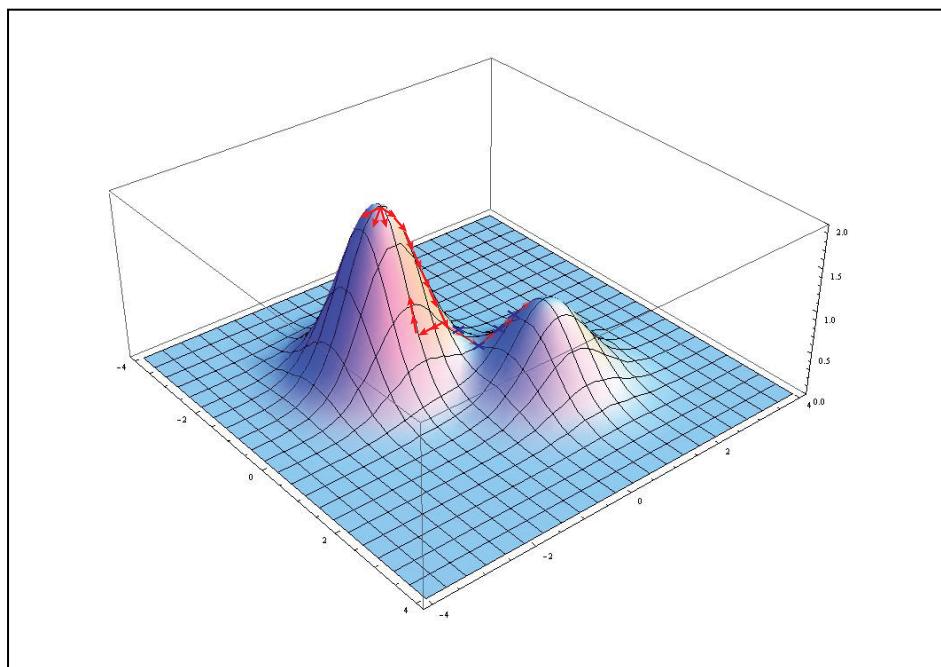


그림 6. $J(\theta'|\theta)$ 함수를 정의할 때 θ 와 θ' 간의 거리가 너무 가까우면 다른 산봉우리로 이동하지 못하고 하나의 산봉우리 주위에만 맴돌게 된다. 이 경우 서로 다른 seed number를 이용해 MCMC 알고리즘을 실행하면 실행할 때마다 다른 결과가 얻어지곤 한다.

201 MrBayes 프로그램(Ronquist et al. 2012)을 이용하여 데이터 분석의 예시를 보이겠다. 분석데이터는 Mr-
202 Bayes가 예시로 제공하는 데이터(Nylander et al. 2004)이다.

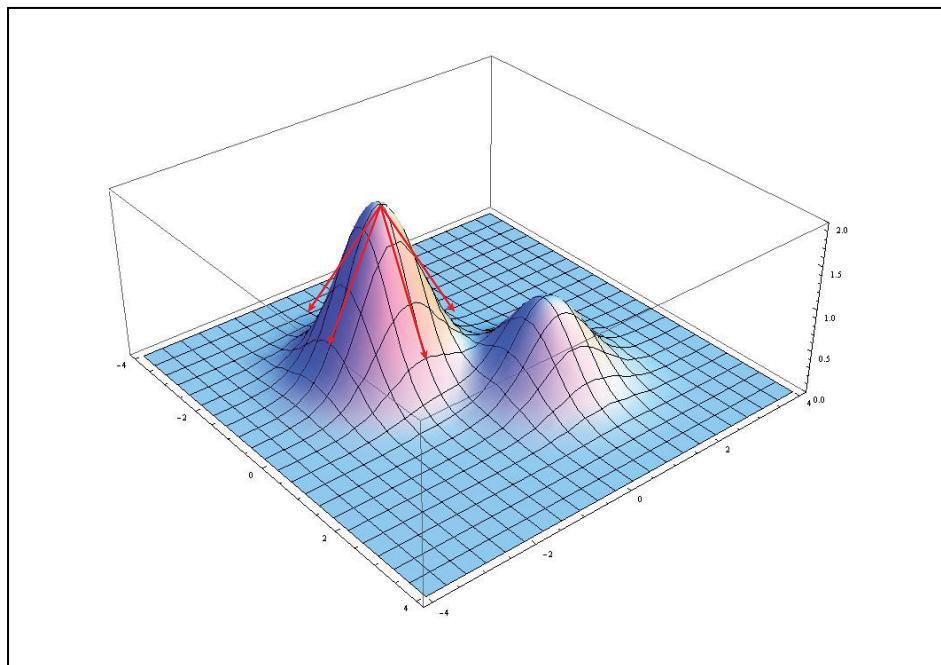


그림 7. $J(\theta'|\theta)$ 함수를 정의할 때 θ 와 θ' 간의 거리가 너무 멀면 θ 의 이동이 잘 이루어지지 않아 (“MCMC chain is not mixing well.”) θ 값들의 모임이 사후분포의 형태를 잘 대변해 주지 못하게 된다.

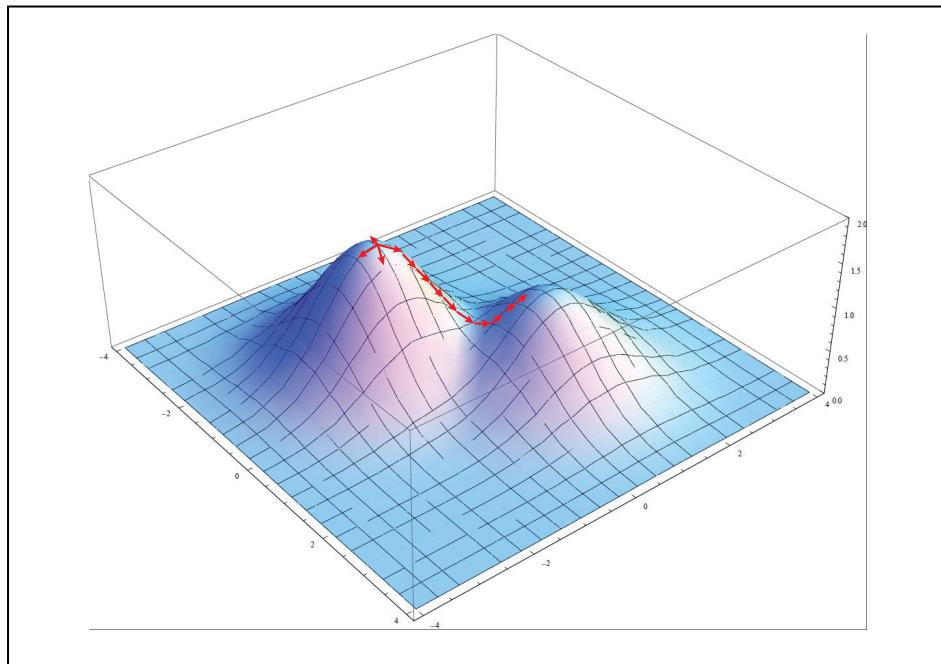


그림 8. 사후분포에 열을 가해 산봉우리가 녹아내려 산정상과 골짜기의 높이차가 줄어든 모식도. 이 경우 θ 의 움직임을 hot chain이라고 한다. θ 와 θ' 간의 거리에 크게 영향을 받지 않고 θ 가 양 봉우리를 쉽게 오갈 수 있다.

203 MrBayes 프로그램과 데이터 파일 (mb.3.2.7-win64.exe 와 cynmix.nex)을 적당한 폴더(예를들어 C:\temp)
 204에 복사한다. 명령프롬프트를 실행시켜(cmd.exe 명령어 실행) C:\temp로 이동한 후 MrBayes 프로그램을
 205 실행하면 다음과 같이 MrBayes 프로그램 내부의 프롬프트가 표시된다.

206 MrBayes>

207 이후 모든 명령어는 MrBayes 프롬프트에서 실행한다. 필수적인 것은 아니지만 seed와 swapseed를
 208 MrBayes 실행 직후에 다음과 같이 설정하면 분석결과를 완벽하게 재현할 수 있는 장점이 있다("1"대
 209 신 다른 양의 정수 설정 가능). 명령어 exec를 이용하여 데이터 파일 cynmix.nex를 읽어들이고 파티션을
 설정한다.

```
MrBayes> set seed=1 swapseed =1; ↵
MrBayes> exec cynmix.nex ↵
MrBayes> set partition = favored; ↵
```

그림 9. 씨드넘버 설정과 데이터 읽어들이기

210
 211 데이터파일 cynmix.nex는 넥서스 포맷의 파일로서 Windows 메모장위에 drag/drop하면 그림 10과 같
 212 이 파일 앞부분을 확인할 수 있다. 'datatype' 키워드를 이용해 데이터가 형태적 형질('Standard' 키워드)
 213 과 DNA로 구성되어 있음을 명시하고 각각의 시작위치와 끝위치를 지정한다. 데이터가 연속적으로 나열
 214 되어 있지 않으므로 'interleave=yes'로 지정하고, gap이나 missing에 해당하는 문자를 지정한다. 데이터가
 215 모두 나열되면 'Begin data;'로 시작한 데이터 블럭이 'End;'로 완성된다. 이후 그림 11과 같이 MrBayes
 216 명령어 블럭이 나열된다('begin mrbayes;'로 시작하여 'end;'로 끝남). 'charset' 키워드를 이용하여 DNA
 217 서열데이터의 파티션을 나누어 구획하고 적당히 이름을 붙여준다. 'partition' 키워드를 이용하여 파티션
 218 방식의 이름을 적당히 지정해 주고 파티션의 수를 명시한다. 여기에서 정의한 파티션을 그림 9의 'set'
 219 명령어를 이용하여 실행했다.

```
#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          0000000000000002-000000000000000?000000000000000100{01}0100001-00100000-0001
Synergus        1-1-1000000002021102010110101101000000000101210011201010101010000000011
Periclistus    1-1-1000000002021102010111011010000000001010010100010101001100000000011
Ceropptres     1-1-100010000202100201011101001000000000111000??101010100100000000011
Synophromorpha 1-1-00001000021-100101001111101000000000101001010001000100100000000011
Xestonhanes   1-1-00001000011-10-110001011010100000000101101010001000100100000000011
```

그림 10. cynmix.nex 파일의 앞부분.

220 그림 9와 같이 실행한 이후에 그림 12의 명령어들을 한 줄씩 차례대로 실행해보자. 이 명령어들은
 221 파일 cynmix.nex 내부에 이미 기록되어 있으나 각진 괄호('[', ']')로 감싸져 있어 데이터를 읽어들일때
 222 실행이 되지 않고 주석처럼 무시되었다. 실제 데이터 분석에서는 그림 12처럼 하나씩 실행하는 것 보다
 223 데이터 파일에서 각진 괄호를 삭제함으로써 자동실행으로 더 간편하게 분석을 진행할 수 있다. 그림 12

```

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 끝부분.

224 의 showmodel 명령어는 설정된 모형들을 보여주는 명령어로서 2-4번째 줄 명령라인을 사이 사이에 실행하여 설정이 어떻게 바뀌는지 하나씩 살펴보는 것도 학습 측면에서 좋을 것이다. 2-4번째 명령라인의 225 실행으로 첫번째 파티션은 DNA 치환의 JC 모형(Jukes and Cantor 1969; 서태건 2022)과 유사한 모형이, 226 2-5번째 파티션은 DNA 치환의 GTR 모형(Tavaré 1986; 서태건 2022)이 설정되었다. 사이트간 진화속 227 도의 이질성(rate heterogeneity among site(RHAS); Yang 1994)은 각각 gamma모형과 invgamma 모형이 228 선택되었다(자세한 설명은 서태건 2022 참조). 또한 unlink 명령어를 이용하여 각 파티션의 치환률 행렬 229 과 invgamma 모형의 inv 비율, gamma 모형의 모수등이 독립적으로 다루어지도록 설정하였다. 그림 12 230 의 첫번째 showmodel의 실행과 마지막 showmodel의 실행 결과의 일부를 그림 13에 비교하였다. 마지막 231 showmodel 실행이후 각 파티션의 모수에 1부터 21까지 번호가 부여되었는데 서로 다른 번호는 서로 다른 232 모수(혹은 모수의 집합)을 의미한다.

```

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.

234 그림 12의 명령어들을 실행한 후 ‘help lset’ 명령(그림 14)으로 설정 결과를 확인한다. 그림 14는 화 235 면에 표시된 결과의 일부(파티션 5)에 대한 내용이다. 본 논문에서 언급하지 않은 항목은 디폴트 설정에 236 맡겨 실행한다. 이후 ‘prset’ 명령어를 이용하여 사전분포를 설정한다. 여기에서는 그림 14의 두번째 줄처 237 럼 ratepr 항목만 설정을 수정한다. 이는 그림 13(b)의 19번 모수와 관련된 내용으로 각 파티션들이 서로 238 다른 치환속도를 갖게 하는 설정이다.

239 이제 MCMC 알고리즘 실행을 위한 설정을 진행한다. 그림 16처럼 ‘mcmc’ 명령어를 이용한다. 여기 240 서는 계산시간을 단축하기 위해 세대수(ngen)를 100만, 샘플간 간격(samplefreq)을 500, 화면 표시 간격 241 (prinfreq)을 100, MCMC가 잘 수렴했는지 진단하는 간격(diagfreq)을 각각 1000세대로 설정한다.²⁰ 설

²⁰ ngen과 samplefreq 설정값은 크면 클수록 좋다. 충분이 큰 값인지는 그림 22의 왼쪽 패널에 있는 ESS(Effective Sample Size;

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

(a) 그림 12의 첫번째 showmodel 명령어 실행 결과 일부

(b) 그림 12의 두번째 showmodel 명령어 실행 결과 일부

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

```
MrBayes> help lset ↵
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. ‘help prset’으로 사전분포 설정을 확인할 수 있다.

Model settings for partition 5:							
Parameter	Options	Current Setting					
Nucmodel	4by4/Doublet/Codon/Protein 1/2/6/Mixed	4by4 6					
Code	Universal/Vertmt/Invermt/Yeast/Mycoplasma/ Ciliate/Echinoderm/Euplotid/Metmt	Universal					
Ploidy	Haploid/Diploid/Zlinked	Diploid					
Rates	Equal/Gamma/LNorm/Propinv/ Invgamma/Adgammma/Kmixture	Invgamma					
Ngammacat	<number>	4					
Nlnormcat	<number>	4					
Nmixcat	<number>	4					
Nbetacat	<number>	5					
Omegavar	Equal/Ny98/M3	Equal	Clockvarpr	Strict/Upp/IKUz/Igr/Mixed	Strict		
Covarion	No/Yes	No	Oppratepr	Fixed/Exponential	Exponential(0.10)		
Coding	All/Variable/Informative/Nosingleton Noabsencesites/Nopresencesites/ Nosingletonabsence/Nosingletonpresence	All	Oppmultdevpr	Fixed	Fixed(0.40)		
Parsmodel	No/Yes	No	TK02varpr	Fixed/Exponential/Uniform	Exponential(1.00)		
			Igrvarpr	Fixed/Exponential/Uniform	Exponential(10.00)		
			Ratepr	Fixed/VariableDirichlet	Dirichlet(...,1.0,...)		
			Generatepr	Fixed/VariableDirichlet	Fixed		

(a) 그림 14의 help lset 명령어 실행 결과 일부

(b) 그림 14의 help prset 명령어 실행 결과 일부

그림 15. ‘help xxx(명령어)’로 각종 명령어로 설정한 옵션과 관련내용 설명을 확인할 수 있다.

242 정이후에 ‘help mcmcp’ 명령으로 설정을 확인한다(그림 17). 디풀트 설정은 MCMC를 2회 동시에 실행
 243 (Nruns 옵션)하고 각 실행에는 4개의 chain(Nchains 옵션; cold chain 1개 나머지는 hot chain)으로 되어
 244 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.

245 설정을 모두 마친 후 그림 18처럼 ‘mcmc’를 입력하면 그림 19과 같이 프로그램이 실행된다. 각 라
 246 인의 각진 괄호는 cold chain을 둘근 괄호는 hot chain을 의미하며 중앙의 ‘*’을 기준으로 좌우 두개의
 247 독립적인 MCMC run이 실행되고 있음을 알 수 있다. 그림 17의 설정에 따라 매 1000세대당 ‘Average
 248 standard deviation of split frequencies’ 메시지가 나오며 이 값은 0.01보다 작은 것이 바람직하다고 알려져
 249 있다(MrBayes 설명서 참조). 설정한 최종 세대수에 도달하면 계속 진행할지 물어보는데(그림 20) ‘no’를
 250 입력하여 MCMC 실행을 종료한다.

```
MrBayes> mcmc   ↵
```

그림 18. MrBayes 실행 명령어

251 이후 그림 21과 같이 burnin의 비율을 25%로 설정하여 MCMC 샘플로부터 각종 모수들의 사후분포를
 252 추정하고 (sump 명령어), 계통수의 사후분포를 주정한다 (sumt 명령어).
 253 명령어 ‘sump’ 실행으로 생성된 *.run*.p(본 분석에서는 cynmix.nex.run1.p, cynmix.nex.run2.p) 파일을

```

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) (-29086.294) (-28899.418) * [-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) * [-27827.888] (-27999.038) [-27469.711] -- 2:33:38
1400 -- [-27418.700] (-27388.986) (-27302.430) (-27307.089) * [-27520.325] (-27710.002) (-27462.591) [-27420.087] -- 2:34:32

```

그림 19. MrBayes 실행 초기 화면.

```

Average standard deviation of split frequencies: 0.005460

998100 -- (-26550.291) [-26526.585] (-26533.179) (-26548.298) * [-26535.568] (-26538.489) (-26532.042) (-26544.724) -- 0:00:19
998200 -- (-26544.423) [-26532.126] (-26535.483) (-26543.154) * [-26529.331] (-26536.281) (-26526.384) (-26545.223) -- 0:00:18
998300 -- (-26546.771) [-26532.941] (-26536.762) (-26554.364) * [-26532.636] (-26534.869) (-26528.616) (-26545.230) -- 0:00:17
998400 -- (-26553.216) [-26533.143] (-26546.007) (-26553.919) * [-26528.170] (-26550.687) (-26530.388) (-26540.282) -- 0:00:16
998500 -- (-26546.576) [-26533.058] (-26544.599) (-26559.953) * [-26538.783] (-26546.102) (-26528.781) [-26540.176] -- 0:00:15
998600 -- (-26539.610) [-26531.967] (-26541.835) (-26555.832) * [-26543.522] (-26551.674) [-26526.860] (-26547.567) -- 0:00:14
998700 -- (-26533.468) [-26523.269] (-26544.875) (-26552.799) * [-26537.804] (-26554.226) (-26531.164) (-26547.811) -- 0:00:13
998800 -- (-26532.194) [-26521.952] (-26550.870) (-26551.396) * [-26534.674] [-26553.409] (-26541.061) (-26544.463) -- 0:00:12
998900 -- (-26532.885) [-26522.219] (-26548.156) (-26551.624) * [-26534.607] (-26550.197) (-26540.621) (-26547.552) -- 0:00:11
999000 -- (-26539.707) [-26522.163] (-26547.147) (-26548.942) * [-26534.987] (-26548.281) (-26546.160) [-26550.281] -- 0:00:10

Average standard deviation of split frequencies: 0.005402

999100 -- (-26543.177) [-26517.502] (-26547.697) (-26546.545) * [-26522.247] (-26546.231) (-26548.351) (-26550.283) -- 0:00:09
999200 -- (-26534.871) [-26515.092] (-26555.376) (-26537.982) * [-26523.955] (-26546.769) (-26563.648) (-26543.853) -- 0:00:08
999300 -- (-26541.478) [-26512.130] (-26552.697) (-26532.226) * [-26523.747] [-26542.773] (-26562.142) (-26549.338) -- 0:00:07
999400 -- (-26538.384) (-26518.928) (-26555.241) [-26535.862] * [-26522.047] (-26546.667) (-26567.006) (-26543.911) -- 0:00:06
999500 -- (-26535.859) [-26519.080] (-26552.608) (-26540.379) * [-26522.902] (-26549.997) (-26562.522) (-26546.062) -- 0:00:05
999600 -- (-26532.493) [-26524.546] (-26545.202) [-26537.017] * [-26526.572] (-26552.777) (-26562.637) (-26544.703) -- 0:00:04
999700 -- [-26527.338] (-26528.291) (-26539.300) (-26530.230) * [-26526.252] (-26554.610) (-26563.025) (-26546.465) -- 0:00:03
999800 -- [-26532.021] (-26531.298) (-26542.708) (-26534.073) * [-26523.469] (-26553.909) (-26561.622) (-26544.778) -- 0:00:02
999900 -- (-26539.992) (-26532.001) (-26532.497) [-26530.797] * [-26529.943] (-26550.553) (-26559.568) (-26537.877) -- 0:00:01
1000000 -- (-26534.265) (-26533.214) (-26543.286) [-26531.288] * [-26523.106] (-26545.323) (-26554.454) (-26540.410) -- 0:00:00

Average standard deviation of split frequencies: 0.005449

Continue with analysis? (yes/no): no

```

그림 20. MrBayes 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.

254 Tracer 프로그램²¹ 으로 읽어 들이자(그림 22). 왼쪽 패널에서 선택한 모수에 대해서 오른쪽 패널에 세대별
255 변화양상(Trace 탭)을 볼 수 있다. 변화양상이 랜덤하게 보이는 것이 바람직하다.

256 왼쪽 패널의 ESS(Effective Sample Size)는 샘플링된 모수의 갯수가 몇개의 독립적인 샘플수에 필적
257 하는지 나타내는 수치이다. Sampling interval을 큰값으로 설정할수록 연속적으로 샘플링된 모수끼리의
258 상관관계는 줄어들지만 아무리 큰값을 지정하더라도 상관관계는 현실적으로 0이 되지는 않는다. 따라서
259 ESS는 실제 샘플수보다 작은 값을 갖게 된다. ESS값이 검은색이 아닌 항목은 샘플수가 충분하지 않다는
260 것을 의미하므로 총세대수 (mcmc 설정에서 ngen항목)를 증가시킬 필요가 있다.

261 (분석에 대한 고찰 계속 작성중)

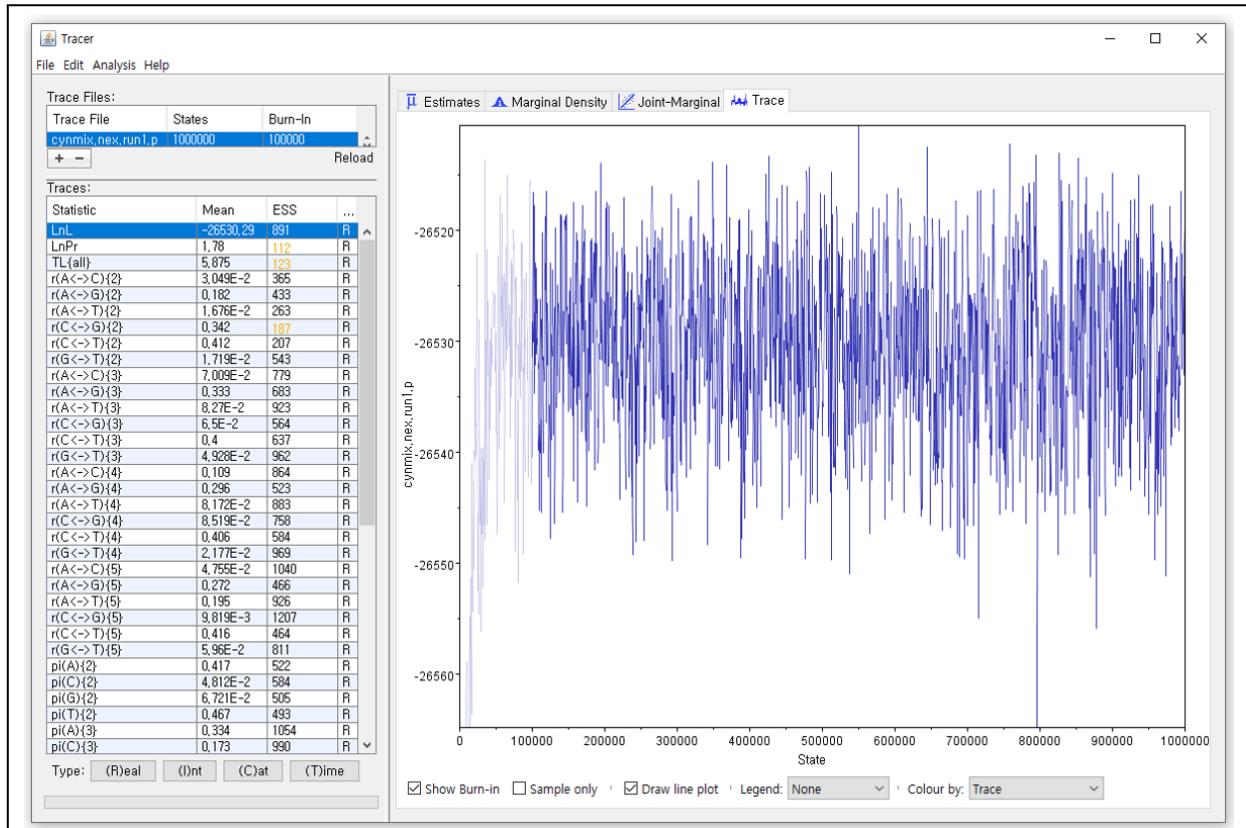


그림 22. Tracer프로그램으로 읽어들인 cynmix.nex.run1.p 파일

262 명령어 ‘sumt’ 실행으로 생성된 *.run*.t(본 분석에서는 cynmix.nex.run1.t, cynmix.nex.run2.t) 에는
263 사후분포로부터 샘플링된 계통수가 저장되어 있다. 이를 요약하여 각 노드별 사후확률을 추정하기 위해
264 TreeAnnotator 프로그램²²을 실행한다. 그림 23과 같이 burin-in 비율, 입력파일, 출력파일을 적절히 선택
265 하면 요약된 결과가 지정된 출력파일로 저장되고 출력파일을 FigTree 프로그램으로 열면 그림 24처럼 각
266 노드별 사후분포를 볼 수 있다.

267 결론

268 (결론 작성중)

²¹<http://tree.bio.ed.ac.uk/software/tracer/>

²²Beast 프로그램의 부속 프로그램이다. <https://www.beast2.org/download-windows/>

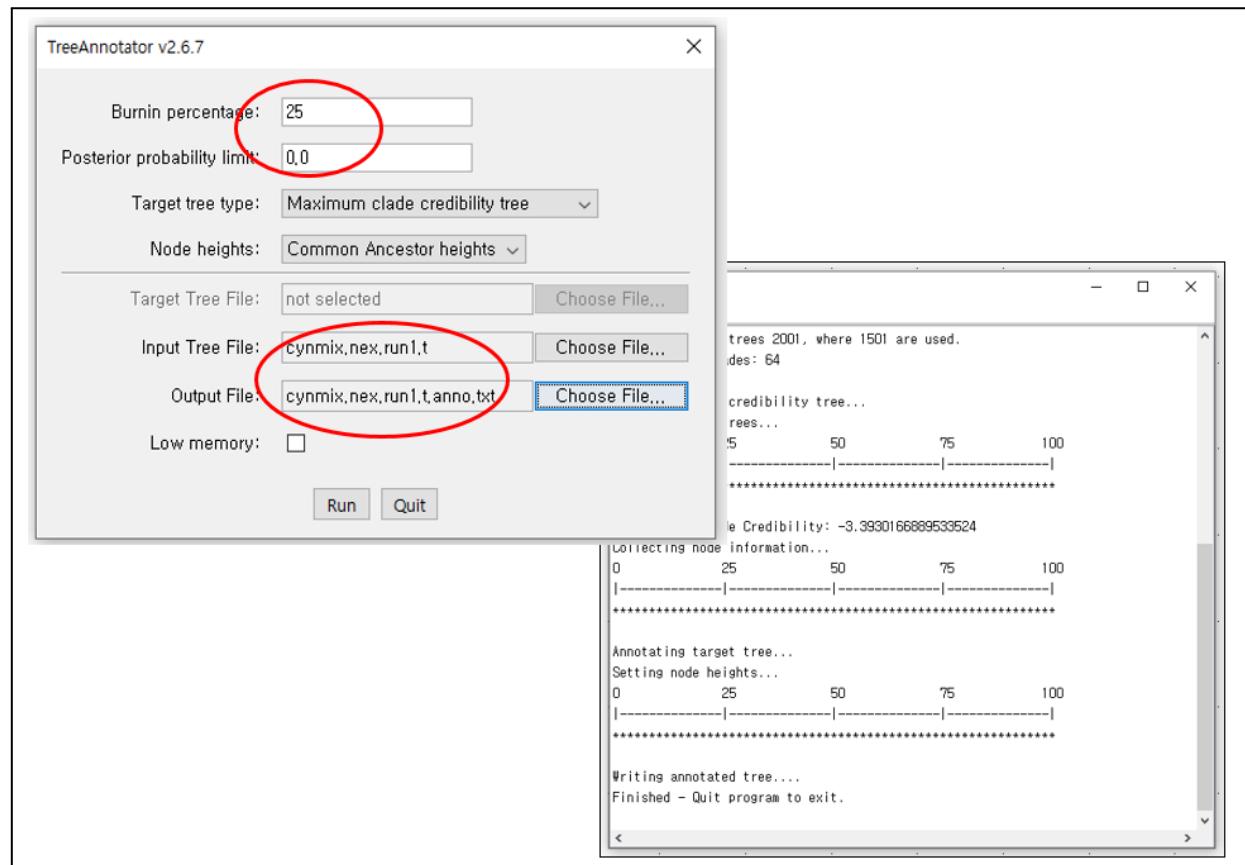


그림 23. TreeAnnotator 프로그램 실행

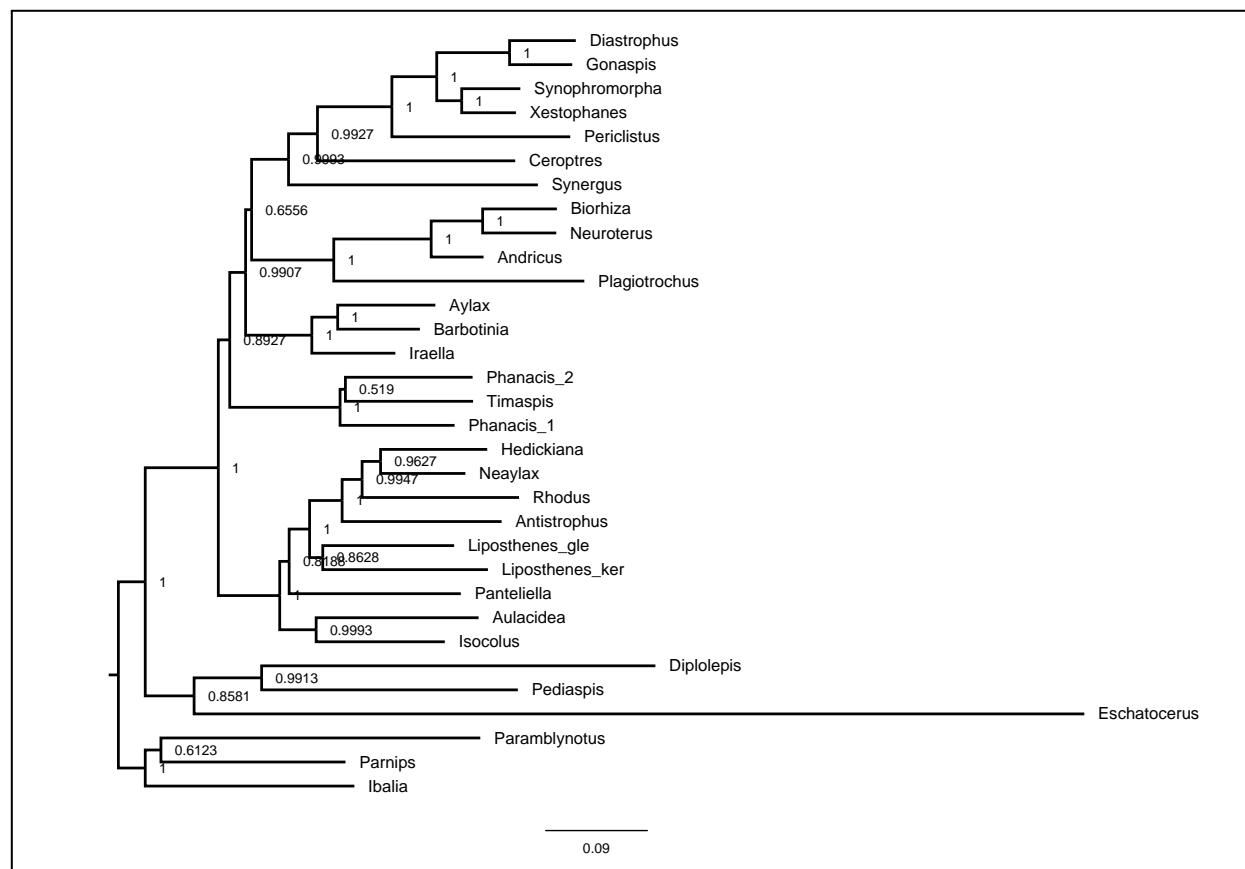


그림 24. 각 노드별로 사후확률이 명시된 계통수

269

감사의 글

270 본연구는 해양수산부의 재원으로 극지연구소의 지원을 받아 수행되었다 (과제번호: PE24xxx).

271

참고문헌

272 (참고문헌 작성중)

273 Gelman A., Carlin J.B., Stern H.S., Rubin D.B. 2004. Bayesian Data Analysis (2/e). Chapman & Hall/CRC.

274 Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97-109.

275 Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In Mammalian protein metabolism (ed. H. N. Munro), pp. 21–123. Academic Press, New York.

276 Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machine. J. Chem. Phys. 21:1087-1091.

277 Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL. 2004. Bayesian phylogenetic analysis of combined data. Systematic Biology 53:47-67.

278 Ronquist, F., M. Teslenko, P. van der Mark, D. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539-542. DOI: 10.1093/sysbio/sys029

279 Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.

280 김우철 2021. 수리통계학(개정판). 민영사.

281 서태건 2022. DNA 엔기치환 모형의 비교. 한국진화학회지 1:88-104.

282 서태건 2023. 아미노산 서열과 코돈 서열의 진화모형. 한국진화학회지 2(2):41–60.

290

영문초록

291 (작성중)