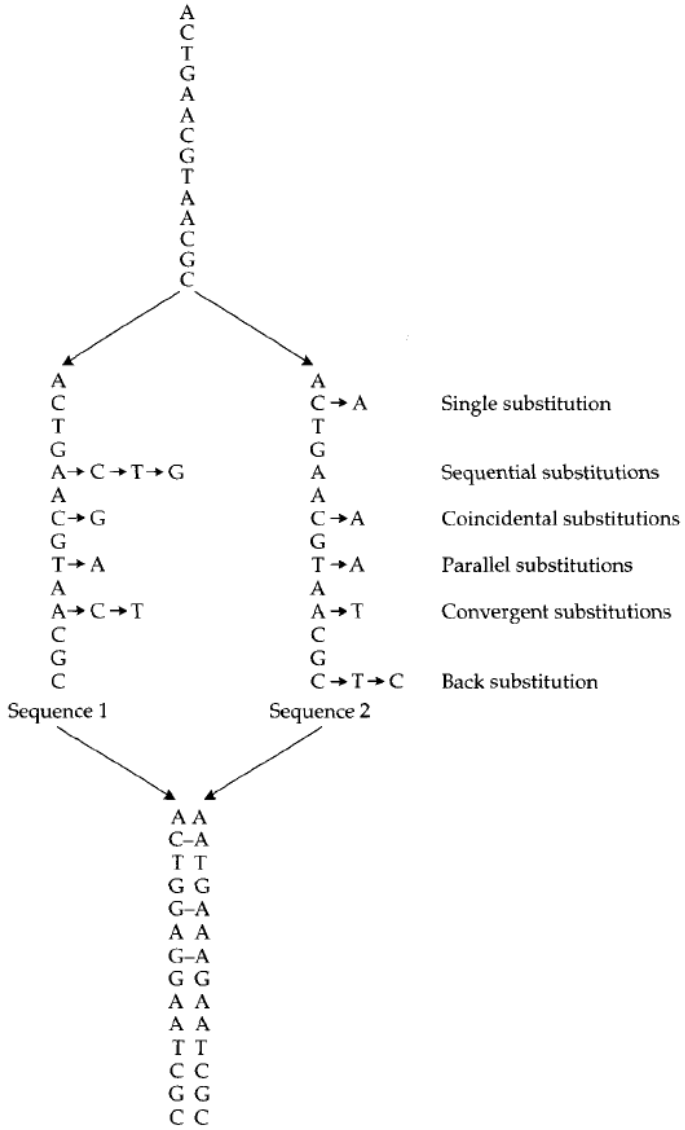
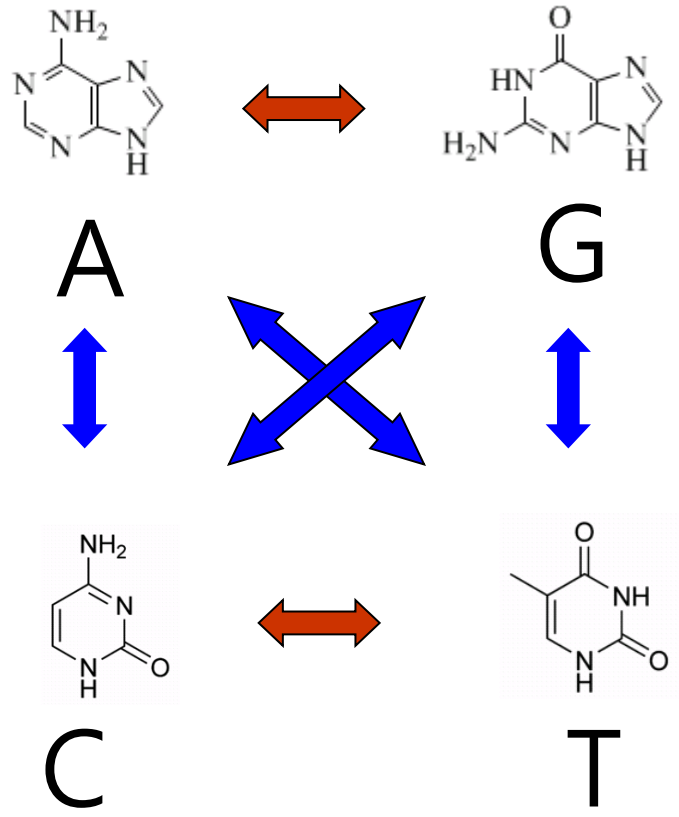
DNA 염기치환 모형 비교

(DNA 염기치환 모형(서태건 2022) 논문의 요약 및 추가 설명입니다) ; ver.2026.01.27

DNA서열 진화 모식도





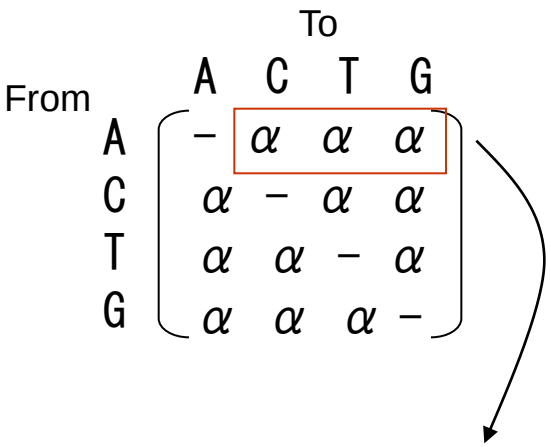
Transition: 염기치환이 빠름 (염기의 입체구조가 비슷)

Transversion: 염기치환이 느림

DNA염기치환 모형

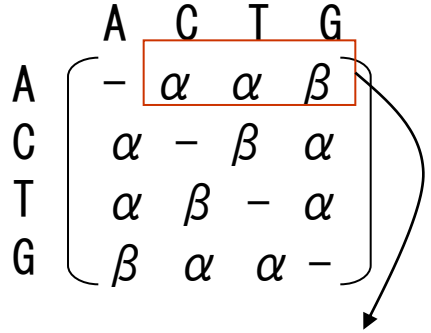
- 매우 짧은 시간동안의 염기치환 상대적인 속도를 파라미터로 표현
- 복잡한 분자진화 현상을 단순화 하여 몇 개의 파라미터로 설명 (예) Jukes-Cantor model (JC), Kimura's two parameter model (K2,K80), Hasegawa-Kishino-Yano model (HKY85), etc.

JC



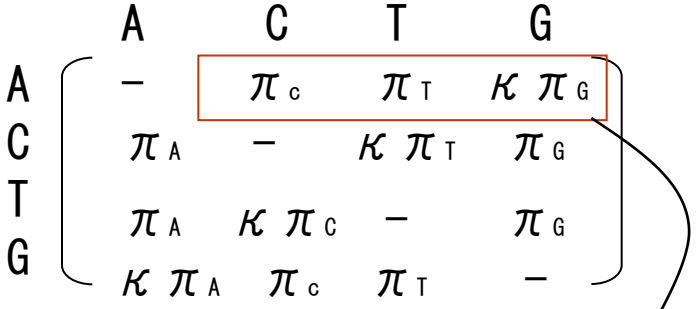
치환의 상대적 속도가 동일(α)

K2



Transition의 속도(β)와 transversion의 속도(α)가 다르다. 일반적으로 β>α

HKY85



염기치환속도가 염기의 빈도(π)에 비례. Transition이 transversion보다κ배 빠름.

$$\mathbf{R}^{(TN)} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & \pi_C & \kappa_1\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa_2\pi_T \\ \kappa_1\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa_2\pi_C & \pi_G & - \end{bmatrix} \end{array}, \quad \mathbf{R}^{(GTR)} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & - & \pi_T \\ c\pi_A & e\pi_C & \pi_G & - \end{bmatrix} \end{array} \tag{5}$$

Tamura Nei (1993) 모형

GTR (1986) 모형; General Time-Reversible

DNA 염기치환은 전후 순서를 무시하고 순서쌍만 고려하면 여섯 가지가 존재한다. [2] DNA 염기치환 모형은 여섯가지 염기치환의 순간적인 발생 비율을 규정한 것으로, 가장 단순한 JC 모형 (Jukes and Cantor 1969)부터 가장 복잡한 GTR 모형 (Tavaré 1986)에 이르기까지 다수의 모형이 개발되어 왔다 (자세한 내용은 Felsenstein 2004, Yang 2006 에 잘 정리되어 있다).

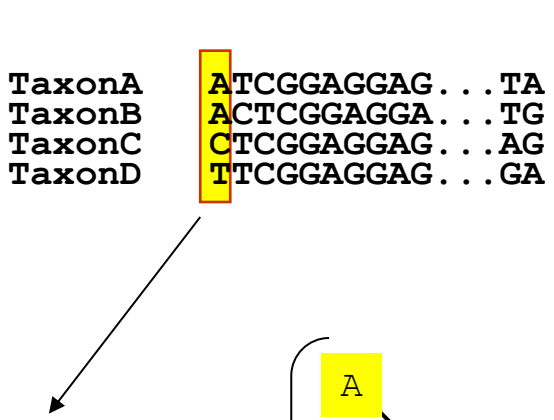
분자데이터 분석에 적용할 수 있는 DNA 염기치환 모형의 수가 증가함에 따라, 어떻게 모형들을 비교하고 가장 좋은 모형을 선택해서 데이터 분석에 이용해야 하는가하는 문제가 자연스럽게 등장한다. 염기치환 모형은 생물의 계통관계 추정뿐만 아니라, 분기연대의 추정, 자연선택의 검출 등의 데이터 분석 결과에 영향을 미친다. 부적절한 모형의 선택으로 잘못된 계통수가 도출된다든지(Sullivan et al. 1997; Ripplinger and Sullivan 2008), 진화거리의 부정확한 추정으로 분기연대 추정값이 부정확해진다든지(Schenk and Hufford 2010)[3]하는 사례들이 다수 보고된 바 있다. 따라서 모형 선택은 객관적이고 정량적이어야 하고, 이를 위한 합리적인 방법론의 연구도 많이 수행되었다(Sullivan and Joyce 2005).

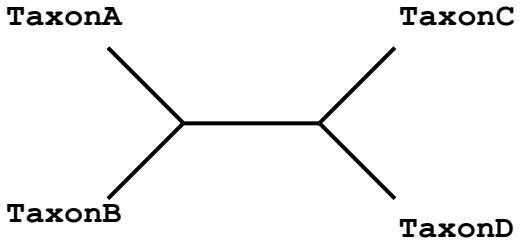
주의해야 할 점은 DNA 염기치환 모형은 복잡한 자연현상(즉, 복잡한 분자진화양상)을 어디까지나 단순화하고 근사시킨 '틀'에 불과하다는 것이다. 복잡한 자연현상을 모수 몇 개로 정리하고 요약하여 설명하려는 것이다. 이러한 시도는 자연현상을 쉽게 이해하고 설명하게 하는 장점도 있지만, 한편으로는 현실 세계와의 괴리가 필연적으로 존재한다는 단점도 있다. 모든 통계적 모형은 근본적으로 '틀린 모형 (wrong model)'이다 (Box 1976). '옳은 모형 (correct model)'이란 것은 존재할 수 없다 .[4] '(다른 모형보다) 더 좋은 모형 (better model)'이 있을 뿐이다.

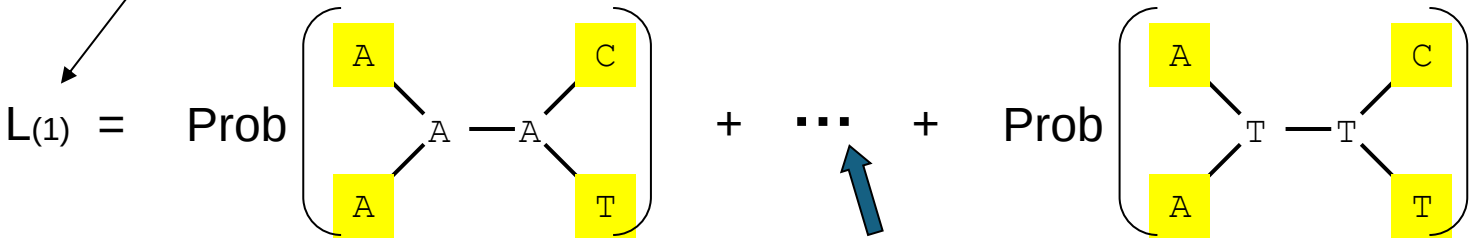
Maximum likelihood (ML) method

- 각 계통수별로 likelihood를 계산, likelihood가 최대가 되는 계통수 및 모델파라미터를 구하는 방법

(likelihood: 데이타가 모델에 어느정도 잘 맞는가 나타내는 수치. 「데이타가 생성될 확률」과 밀접한 관계가 있음.)







Internal node의 모든 조합(이 경우 16가지) 에 대하여 합을 계산

$$L = L_{(1)} \times L_{(2)} \times L_{(3)} \times \cdots \times L_{(n)}$$

$$\log L = \log L_{(1)} + \log L_{(2)} + \log L_{(3)} + \cdots + \log L_{(n)}$$

Log-likelihood가 최대가 되는 계통수, branch length, 모델 파라미터등을 추정한다.

Transition probabilities

$$\mathbf{R} := \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{T} \\ \mathtt{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{T} & \mathtt{G} \end{array} \\ \begin{pmatrix} - & \alpha & \alpha & \beta \\ \alpha & - & \beta & \alpha \\ \alpha & \beta & - & \alpha \\ \beta & \alpha & \alpha & - \end{pmatrix} \end{array}$$

Transition probabilities after time t are

$$\mathbf{P}(t) := e^{t\mathbf{R}} = \begin{pmatrix} P_{AA}(t) & P_{AC}(t) & P_{AT}(t) & P_{AG}(t) \\ P_{CA}(t) & P_{CC}(t) & P_{CT}(t) & P_{CG}(t) \\ P_{TA}(t) & P_{TC}(t) & P_{TT}(t) & P_{TG}(t) \\ P_{GA}(t) & P_{GC}(t) & P_{GT}(t) & P_{GG}(t) \end{pmatrix}$$

그렇다면 여러 가지 후보 모형 중에서 더 좋은 모형을 어떻게 판단할 수 있을까? 염기치환 모형의 정량적인 비교를 위해 AIC (Akaike Information Criterion; Akaike 1974), BIC (Bayesian Information Criterion; Schwarz 1978)와 같은 정보량기준 (Information Criterion; IC) 혹은 이들로부터 파생된 정보량기준이 흔히 쓰인다. 정보량기준은 공통적으로 아래와 같은 행태를 띄고 있다 (Dziak et al. 2019).

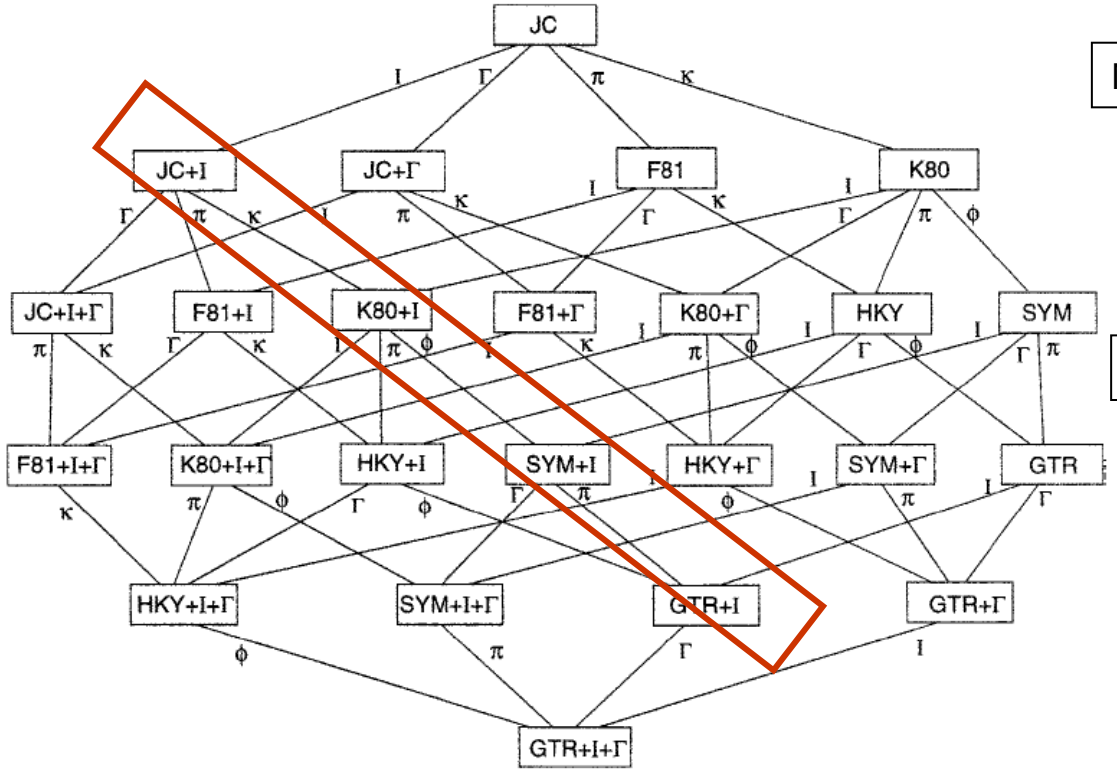
$$\mathrm{IC} := l_m - \rho \cdot p_m, \tag{1}$$

여기서 l_m은 모형 m 을 적용시켜 얻은 로그가능도 (log-likelihood), p_m은 모형 m이 포함하고 있는 미지의 모수의 갯수, ρ는 정보량기준에 의해 사전에 정해진 값이다.[5] 기호 ':='는 좌변의 의미를 우변으로 정의한다는 것을 말한다. 로그가능도 [6] 는 확률들의 곱의 형태로 표현되는 가능도 (likelihood)에 로그를 취한 것으로 모형이 데이터에 얼마나 잘 적합하게 들어 맞는지 나타내는 수치이다. 로그가능도가 크면 클수록 데이터는 해당 모형에 잘 부합한다는 것을 의미한다. 일반적으로 모형의 모수 수가 증가할수록 적합도가 증가하여 l_m은 커지게 된다. 하지만, 적합도가 증가한 만큼 단점도 존재하는데, 그 단점은 모수 추정의 불확실성이 증가한다는 것이다. 즉, 모형은 기존 데이터를 잘 설명하지만, 미지의 데이터 예측에는 신뢰성이 떨어지는 문제가 발생한다. 따라서 마냥 복잡한 모형 (l_m이 큰 모형)을 선택할 수는 없고 어느 정도 이상 복잡해지지 않도록 억제하는 '페널티'가 존재해야 하는데 이 역할을 하는 것이 식 (1) 우변의 두 번째

AIC: $\rho = 2$
BIC: $\rho = \log(n)/2$

Likelihood 스코어를 이용한 모델 비교 (likelihood ratio test; LRT)

Hierarchy of DNA models



Nested Model

(special case of nesting model)

Nesting Model

두 모델의 비교에 있어서 복잡한 모델의 파라미터를 특정한 값으로 고정시키면 단순한 모델이 될 경우 Likelihood ratio test (LRT) 를 적용할수 있다

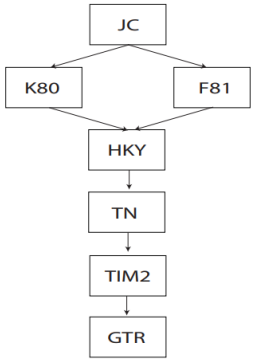


그림 1. DNA 치환모형들 간의 관계. 가장 단순한 JC 모형부터 가장 복잡한 GTR모형에 이르기까지 203개의 모형이 존재하나 이 그림에서는 몇 개만 선별하여 표시하였다.

Posada and Crandall (2001)

모형의 정량적인 비교

어떤 모형과 그 모형의 내포모형 (nested model)을 비교할때 전통적으로 가능도비검정 (Likelihood Ratio Test; LRT) 이 사용된다. 내포모형을 m_1, 복잡한 모형을 m_2라고 하고 각각의 모수 수를 p_1, p_2라 하자. 모형 m_1와 m_2 하에서 얻은 최대로그가능도를 각각 l_1, l_2라고 하면 귀무가설 (내포모형이 참이라고 가정하는 가설)하에서 로그가능도 차이의 두배는 자유도가 $(p_2 - p_1)$ 인 카이제곱분포를 따른다는 것이 알려져 있다 (Stuart and Ord 1991).

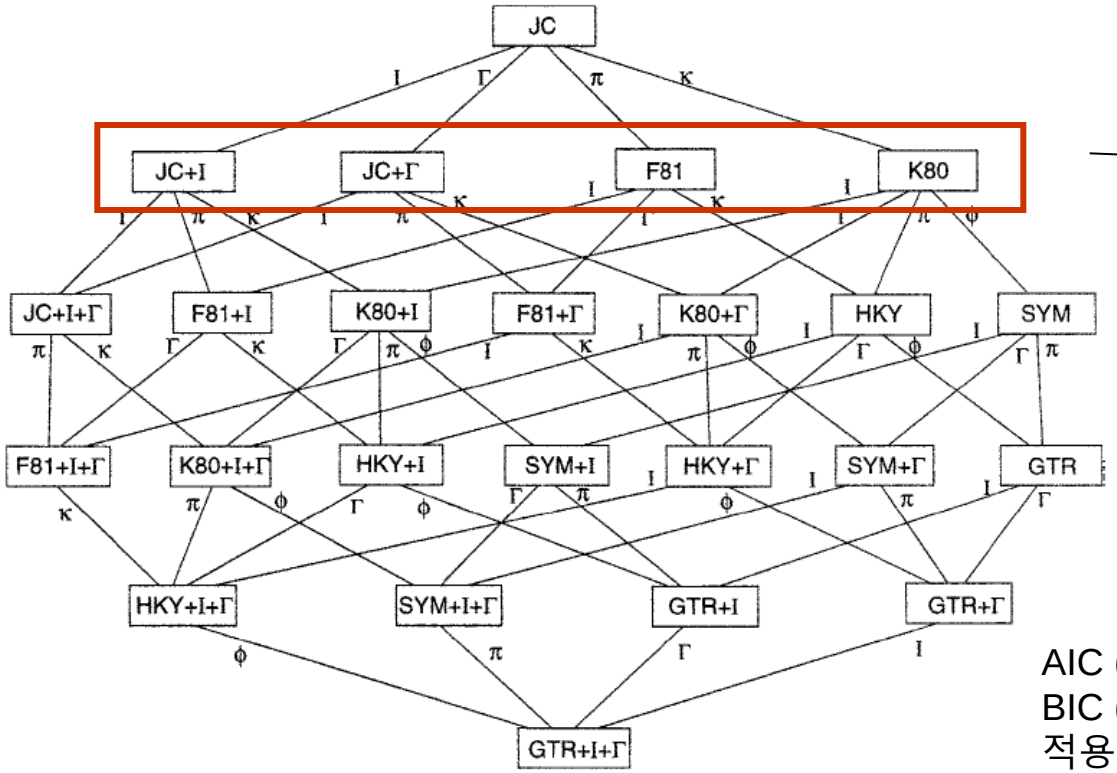
$$2\{l_2 - l_1\} \sim \chi^2_{(p_2-p_1)}$$

이 통계학적 사실을 DNA 치환 모형에 적용시켜, 가장 단순한 JC 모형부터 출발하여 {단순한모형, 복잡한 모형} 의 순서쌍을 하나하나 검정해가는 계층적 가능도비 검정 (hierarchical Likelihood Ratio Test; hLRT) 법이 제안되었다 (Posada and Crandall 1998). 하지만, hLRT에 의해 선택되는 최적의 모형은 순서쌍을 정하는 경로에 영향을 받기도 하고, 반복적으로 시행되는 통계적 가설검정에 따른 다중검정(multiple testing)의 문제, 무엇보다도 내포모형의 관계가 아닌 경우[16]에는 적용 불가능하다는 등의 문제가 있어 DNA 치

[16]예를 들어 그림 1에서 F81과 K80 모형은 내포관계가 성립하지 않는다.

Likelihood 스코어를 이용한 모델 비교 (AIC, BIC, 등)

Hierarchy of DNA models



Likelihood ratio test (LRT) 를 적용 할수 없는 일반적인 경우

AIC (Akaike Information Criterion)
BIC (Bayesian Information Criterion)
적용 가능함

Posada and Crandall (2001)

AIC 와 BIC는 식 (1)에 '-2'를 곱한 형태를 가지며 아래와 같이 정의된다.

$$\text{AIC} = -2 \times \text{log-likelihood} + 2p \tag{6}$$

$$\text{BIC} = -2 \times \text{log-likelihood} + p \log n \tag{7}$$

여기에서 log-likelihood는 최대로그가능도, p는 모수의 수, n은 염기서열의 길이를 의미한다. 비교대상이 되는 모형중에서 가장 적은 AIC 혹은 BIC 스코어를 가지는 모형이 가장 좋은 모형으로 선택이 된다. AIC 기준으로 선택된 모형은 미지의 '데이터 생성 메카니즘'과 '가장가까운[17] 모형'이다. BIC 기준으로 선택된 모형은 데이터의 주변확률밀도(marginalized probability density)가 가장 큰 모형이다(Konishi and Kitagawa 2008).

AIC 정보량기준은 데이터의 수가 매우 크고 비교되는 모형이 미지의 데이터 생성 메카니즘과 어느정도 비슷할 때 좋은 퍼포먼스를 보여준다. 하지만 실제 데이터 분석에서는 데이터의 수가 작은 경우가 흔히 있고 선형회귀분석 모형의 경우 작은 데이터수의 영향을 보정하는 아래와 같은 AICc (corrected AIC)가 제안되었으며(Konishi and Kitagawa 2008) 이 정보량 기준은 DNA 치환 모형에서도 흔히 사용된다.

$$\text{AICc} = \text{AIC} + \frac{2p^2 + 2p}{n - p - 1} \tag{8}$$

사이트 간 진화속도의 이질성 **(Rate Heterogeneity among Sites; RHAS)**

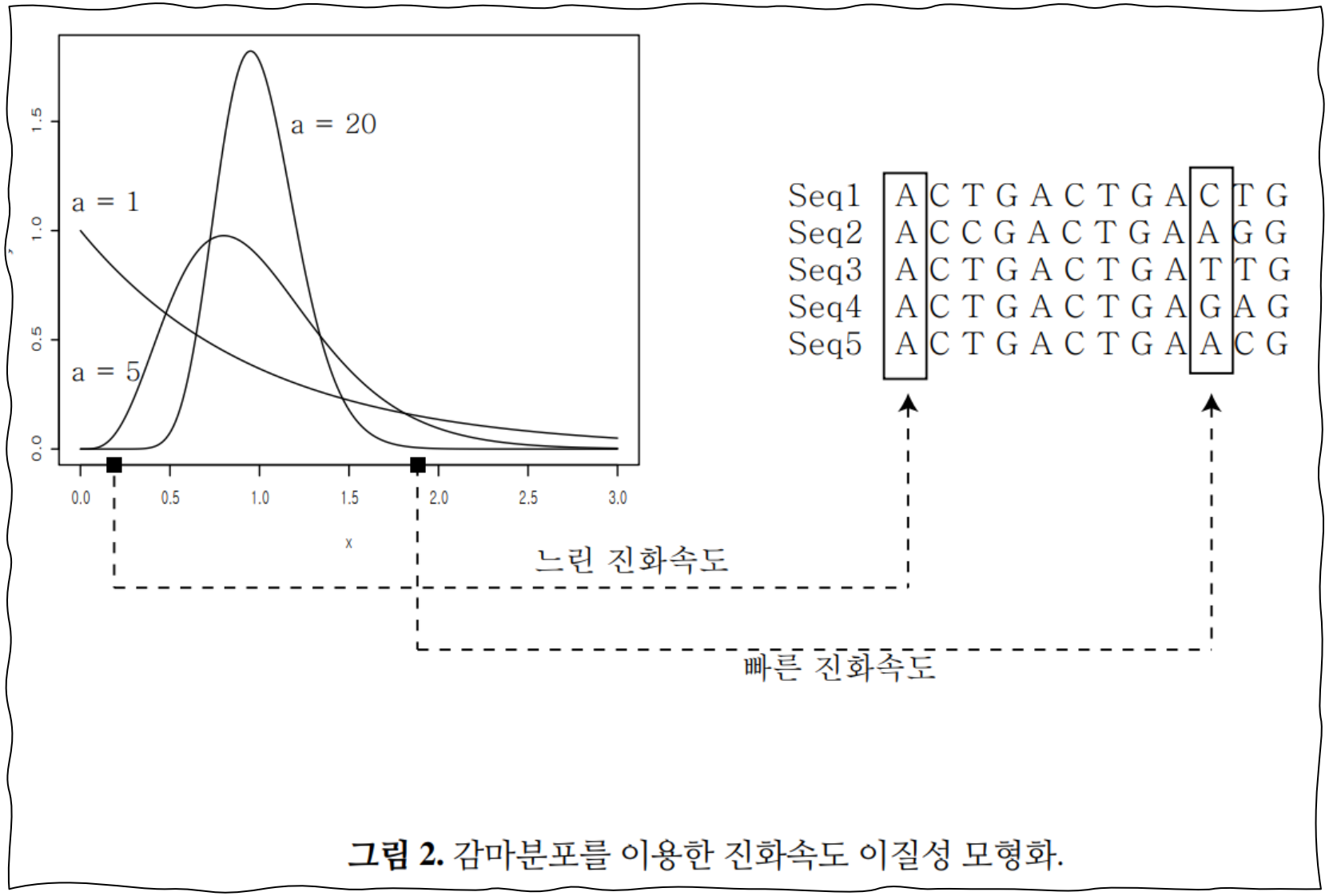


그림 2. 감마분포를 이용한 진화속도 이질성 모형화.

(β=α 로 제한)

그림 2는 감마분포를 이용한 진화속도 모형화를 모식적으로 나타낸 것이다. 각 사이트의 진화속도는 기준속도의 r배를 갖게되고 r은 감마분포를 따르는 확률변수이다. r이 우연히 작은 값을 가지면 그 사이트는 진화속도가 느려 치환이 덜 관찰될 것이고 r이 우연히 크면 진화속도가 빨라 치환이 많이 관찰될것이다. 감마분포는 α 모수로 형태가 결정되는데 이 모수가 크면 클수록 감마분포는 1주위에 밀도높게 분포하는 형태를 보이고 (그림 2 왼쪽) 추출되는 r도 1와 매우 유사한 값이 되어 사이트간 진화속도의 이질성이 감소한다. 극단적으로 $\alpha = \infty$ 인 경우 감마분포는 1에서 피크를 보이는 확률밀도함수를 갖게되고 이는 진화속도의 이질성 (RHAS) 을 가정하지 않는 모형과 동등한 모형이 된다. 따라서 RHAS를 가정하지 않는 모형은 RHAS를 가정한 모형의 내포모형이다. 위에서 식 (2) – (5)에서 특정 모수를 고정된 값으로 치환시키면 단순한 모형이 되는 것과 비슷한 상황이지만 α가 유한한 값이 아니라 무한대[15]이므로 비교에 있어 주의가 요구된다.

분자계통수 분석에서 자주 사용되는 통계 모형

감마분포(김우철 2021; 그림 9)의 확률밀도함수는 모수 α, β가 주어졌을 때 다음과 같다.

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

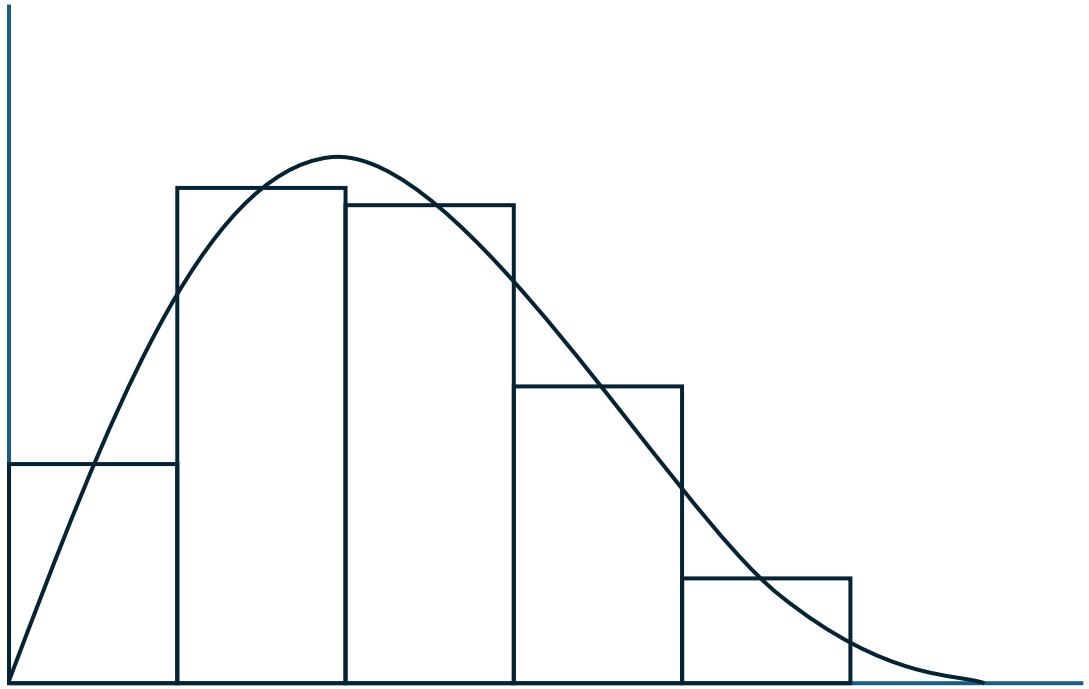
평균과 분산은 다음과 같다.

$$\mathrm{E}[R] = \frac{\alpha}{\beta} \quad (1)$$

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

계산시간을 줄이기 위해 연속형 감마분포를 이산형 감마분포로 근사한다.

(예) HKY+G5, GTR+G5 처럼 "+Gk" 형태로 나타냄

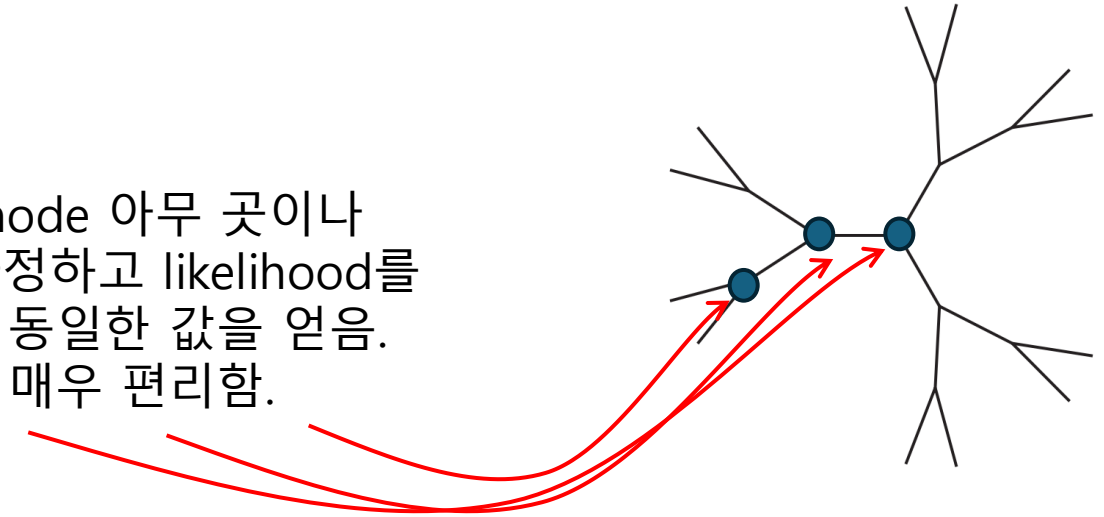


이산형 감마분포대신 "Free rate 모형"을 사용할 수도 있으나 ("+Rk" 형태의 태그; +R3, +R4, 등등) 일반적으로 이산형 감마분포가 널리 쓰인다.

Time-reversibility 가정

한 가지 주목할만한 가정은 시간가역성(time reversibility)이다. 시간가역성은 $\pi_i R_{ij} = \pi_j R_{ji}$라는 성질을 가짐을 의미한다. 시간가역성 가정은 계통수의 가능도를 계산할때 어느 노드를 공동조상으로 지정하여 계산을 해도 동일한 결과를 얻게 하여 모수의 추정을 빠르게 할 수 있는 장점이 있다. GTR모형의 GTR은 General Time Reviersible의 머릿글자로 GTR 모형의 모든 하위모형(내포모형)은 시간가역성을 가정하는 모형이 된다. 이 가정은 어디까지나 계산상의 편리함을 위한 가정일뿐 생물학적인 근거는 희박하다. 시간가역성을 가정하지 않는 모형도 소개되었으나 (Yang 1994a; Bettisworth and Stamatakis 2021) 계산시간 부담이 크기 때문에 많은 데이터 분석에서는 시간가역성을 가정하는 모형이 사용되고 있다.

Internal node 아무 곳이나 root로 가정하고 likelihood를 계산해도 동일한 값을 얻음. 최적화에 매우 편리함.



또 한 가지 염두해 두어야 할 가정은 '얼라인 된 각각의 염기서열 사이트는 독립적이고 동일한 분포를 따르는 (independently and identically distributed) 랜덤샘플'이라는 가정이다. 이러한 가정은 여러가지 통계학 이론 적용을 가능하게 하며 특히 bootstrap 방법(Felsenstein 1985)을 적용함에 있어 정당성을 부여해준다.

전이율행렬에 등장하는 모수(GTR모형의 a ~ e 그리고 HKY, TN 모형의 κ 모수들)가 계통수 전체에서 동일하다(homogenous)는 것도 널리 사용되는 가정이다. 이러한 가정을 변형시켜 특정생물군에 다른 전이율행렬 모수를 할당하는 연구도 있으나(Galtier and Gouy 1998), 생물간의 전이율 행렬의 차이가 주요관심사가 아닌 경우에는 일반적으로 정상성(stationarity)을 가정하여 가능도를 계산한다.

동일한 진화과정을 따르는 유전자 혹은 파티션 내에서는 재조합(recombination)이 일어나지 않는다는 가정도 진화거리를 계산하는데 중요한 가정이다. 분자계통수의 가능도 계산에서는 재조합을 고려하지 않기 때문에 재조합이 실제 일어났을 경우 분자진화 거리 추정에 편의(bias)가 발생하게 된다(Schierup and Hein 2000).

실습

예제파일과 IQ-TREE 실행파일(iqtree.exe와 libiomp5md.dll) 을 동일한 폴더(C:\temp)에 복사한 후 그림 3과 같이 실행한다. '-s' 옵션은 염기서열 데이터 파일을 지정하는 옵션이다. 염기서열 데이터를 지정하는 것 이외에 아무 옵션도 지정하지 않으면 IQ-TREE는 가능한 모든 염기치환 모형에 대해 AIC, BIC, AICc 스코어를 계산해주고 각각의 기준으로 선택된 최적의 모형을 출력해준다.

```
C:\temp>  iqtree.exe -s example.phy
```

그림 3. IG-TREE에 포함된 예제파일로 모형비교를 실행

명령 프롬프트

```
c:\temp>iqtree3.exe -s example.phy
IQ-TREE version 3.0.1 for Windows 64-bit built May  5 2025
Developed by Bui Quang Minh, Thomas Wong, Nhan Ly-Trong, Huaiyan Ren
Contributed by Lam-Tung Nguyen, Dominik Schrempf, Chris Bielow,
Olga Chernomor, Michael Woodhams, Diep Thi Hoang, Heiko Schmidt

Host:    T7820 (AVX512, FMA3, 255 GB RAM)
Command: iqtree3.exe -s example.phy
Seed:    317251 (Using SPRNG - Scalable Parallel Random Number Generator)
Time:    Fri Jan 09 10:15:56 2026
Kernel:  AVX+FMA - 1 threads (80 CPU cores detected)

HINT: Use -nt option to specify number of threads because your CPU has 80 cores!
HINT: -nt AUTO will automatically determine the best number of threads to use.

Reading alignment file example.phy ... Phylip format detected
Alignment most likely contains DNA/RNA sequences
Constructing alignment: done in 0.0033045 secs
Alignment has 17 sequences with 1998 columns, 1152 distinct patterns
```

참고문헌에서는 version 1의 결과가 소개되어 있다. 설정의 미세한 차이에 의해 Version 2, 3의 결과도 미세하게 다르나 주요 결과는 크게 차이 없다.

여기서는 version 3으로 실행한 경우를 보인다

명령 프롬프트

```
Gamma shape alpha: 0.739
Parameters optimization took 1 rounds (0.057 sec)
Time for fast ML tree search: 0.338 seconds

NOTE: ModelFinder requires 6 MB RAM!
ModelFinder will test up to 968 DNA models (sample size: 1998 epsilon: 0.100) ...
```

No.	Model	-LnL	df	AIC	AICc	BIC
1	JC	23662.322	31	47386.644	47387.653	47560.241
2	JC+I	22587.297	32	45238.594	45239.668	45417.791
3	JC+G4	22258.086	32	44580.171	44581.246	44759.368
4	JC+I+G4	22245.546	33	44557.092	44558.234	44741.889
5	JC+R2	22280.896	33	44627.792	44628.935	44812.589
6	JC+R3	22234.529	35	44539.059	44540.343	44735.056
7	JC+R4	22234.316	37	44542.631	44544.066	44749.828
14	JC+I+R2	22253.511	34	44575.021	44576.233	44765.418
15	JC+I+R3	22236.280	36	44544.560	44545.919	44746.157
16	JC+I+R4	22236.048	38	44548.095	44549.608	44760.891
26	JC+I+G4	22245.546	33	44557.092	44558.234	44741.889
28	JC+R3	22234.529	35	44539.059	44540.343	44735.056
48	F81+I+G4	22245.390	33	44556.780	44557.923	44741.577
50	F81+R3	22234.423	35	44538.847	44540.131	44734.843
70	F81+F+I+G4	22019.012	36	44110.023	44111.382	44311.620
72	F81+F+R3	22009.710	38	44095.420	44096.933	44308.216
92	K2P+I+G4	21833.754	34	43735.507	43736.720	43925.904
94	K2P+R3	21823.497	36	43718.994	43720.353	43920.591
114	K2P+I+G4	21833.754	34	43735.507	43736.720	43925.904
116	K2P+R3	21823.497	36	43718.994	43720.353	43920.591
136	HKY+I+G4	21833.723	34	43735.446	43736.658	43925.842
138	HKY+R3	21823.375	36	43718.750	43720.109	43920.347
158	HKY+F+I+G4	21475.608	37	43025.216	43026.651	43232.413

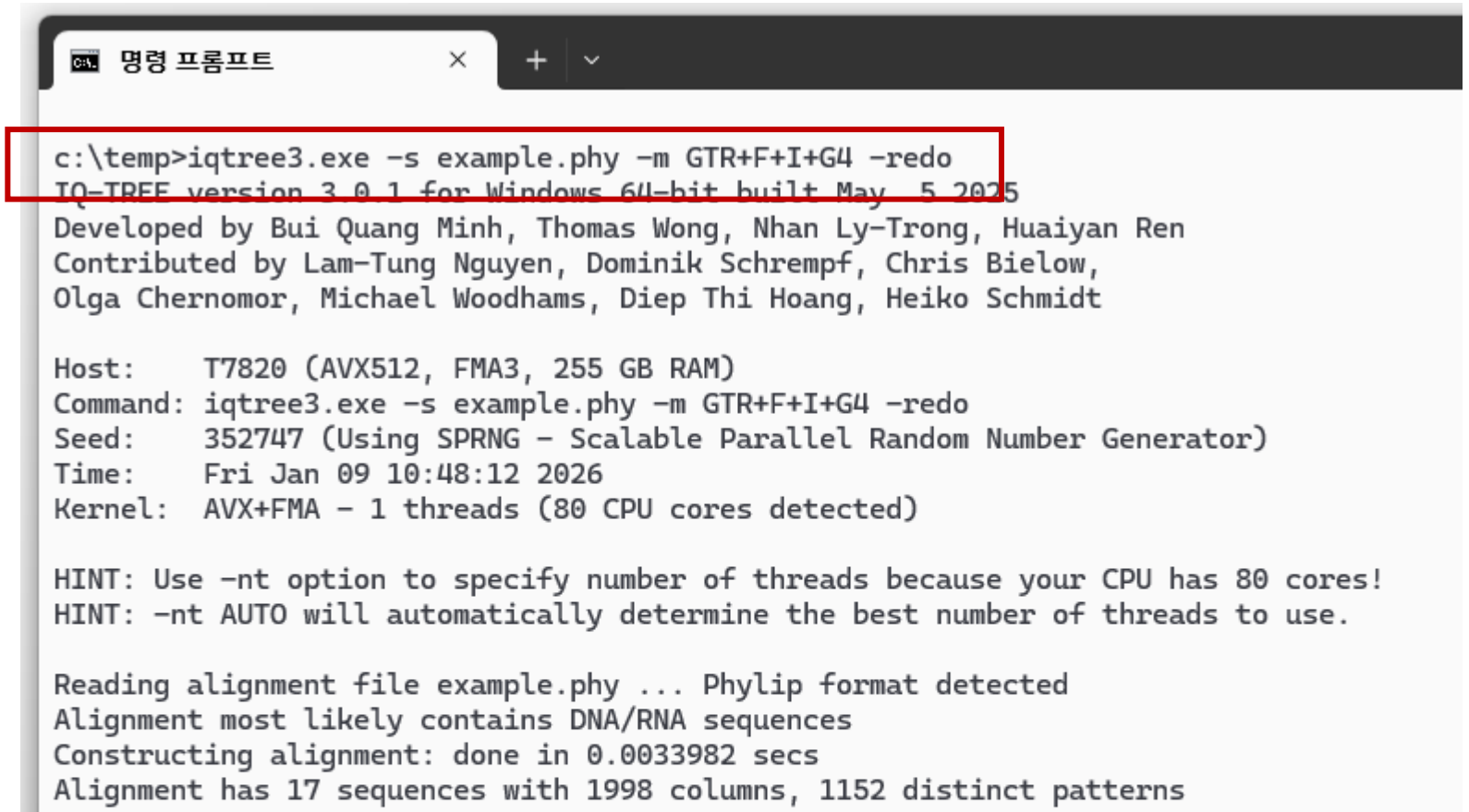
```
796  TVMe+I+G4      21407.848    37  42889.696    42891.130    43096.892
798  TVMe+R3        21401.157    39  42880.314    42881.907    43098.710
818  TVMe+I+G4      21407.848    37  42889.696    42891.130    43096.892
820  TVMe+R3        21401.157    39  42880.314    42881.907    43098.710
840  TVM+I+G4       21407.817    37  42889.635    42891.069    43096.831
842  TVM+R3         21401.104    39  42880.208    42881.801    43098.604
862  TVM+F+I+G4     21287.937    40  42655.873    42657.549    42879.869
864  TVM+F+R3       21279.783    42  42643.567    42645.414    42878.762
884  SYM+I+G4       21306.024    38  42688.048    42689.561    42900.844
886  SYM+R3         21300.830    40  42681.661    42683.337    42905.657
906  SYM+I+G4       21306.024    38  42688.048    42689.561    42900.844
908  SYM+R3         21300.830    40  42681.661    42683.337    42905.657
928  GTR+I+G4       21306.021    38  42688.043    42689.556    42900.839
930  GTR+R3         21300.830    40  42681.660    42683.336    42905.656
950  GTR+F+I+G4     21148.956    41  42379.912    42381.673    42609.508
952  GTR+F+R3       21147.067    43  42380.134    42382.071    42620.930
Akaike Information Criterion:           GTR+F+I+G4
Corrected Akaike Information Criterion: GTR+F+I+G4
Bayesian Information Criterion:         TIM2+F+I+G4
Best-fit model: TIM2+F+I+G4 chosen according to BIC

All model information printed to example.phy.model.gz
CPU time for ModelFinder: 7.891 seconds (0h:0m:7s)
Wall-clock time for ModelFinder: 8.829 seconds (0h:0m:8s)

NOTE: 2 MB RAM (0 GB) is required!
```

미묘한 설정 차이로 Version 3의 결과는 version 1의 결과와 약간 다르다 . → 큰 문제되지 않음

한편 AIC 혹은 AICc 기준으로 최적의 모형으로 판명된 “GTR+F+R3” 모형으로 계통수와 각종 모수 등을 추정해보자. 그림 6처럼 “-m GTR+F+R3” 옵션을 지정하여 프로그램을 실행한다. 위에서 이미 한번 IQ-TREE를 실행시켰기 때문에 여러 결과화일들이 생성되었는데 ‘-redo’ 옵션은 이들 파일에 새로운 결과를 덧씌우도록 하는 옵션이다.

```
C:\temp>  iqtree.exe -s example.phy -m GTR+F+R3 -redo
```

그림 6. AIC에 의해 선택된 모형으로 IGTREE를 실행

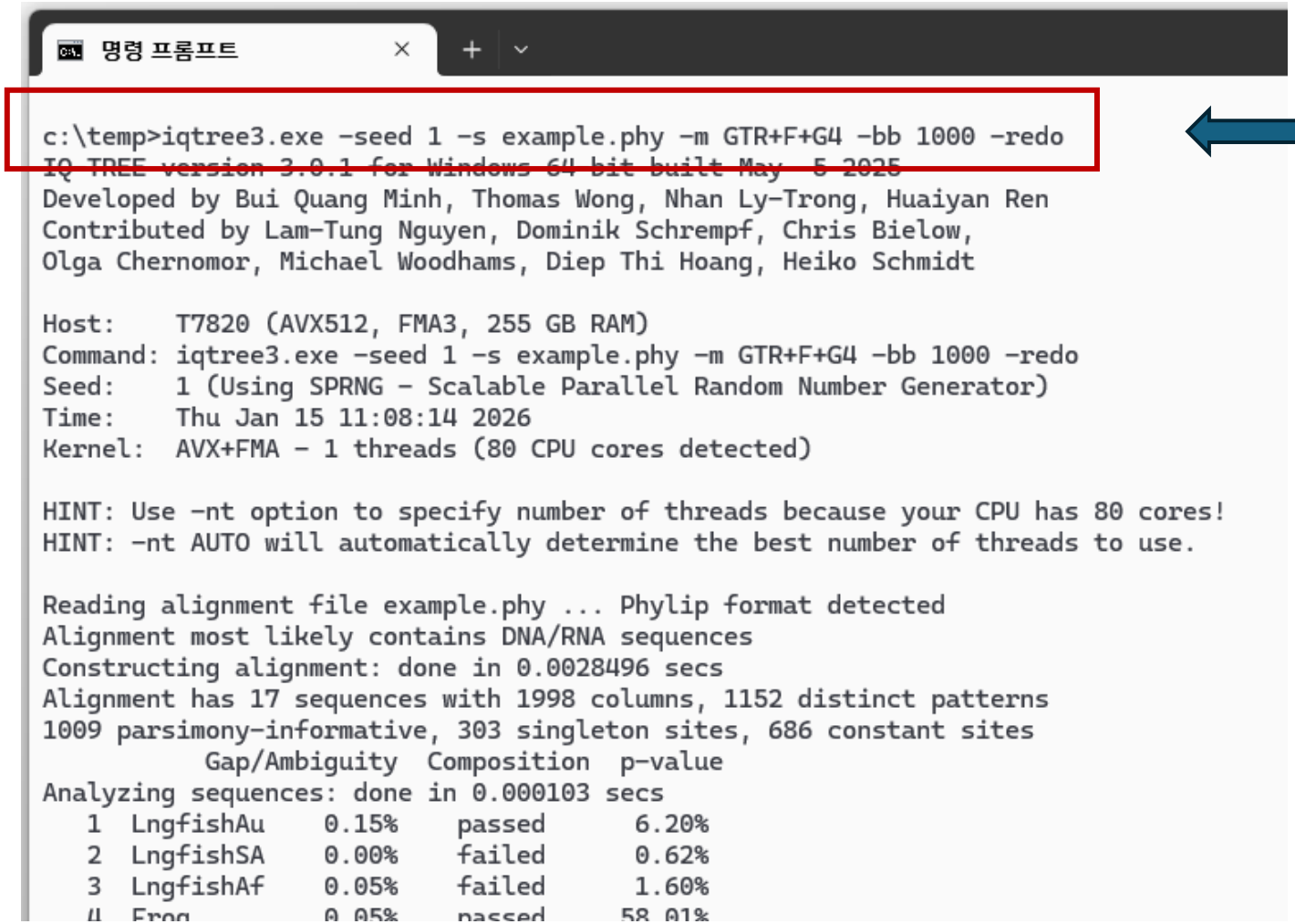


AIC 혹은 AICc 기준으로 최적의 모형으로 판명된 “GTR+F+I+G4” 모형으로 계통수와 각종 모수 등을 추정해보자

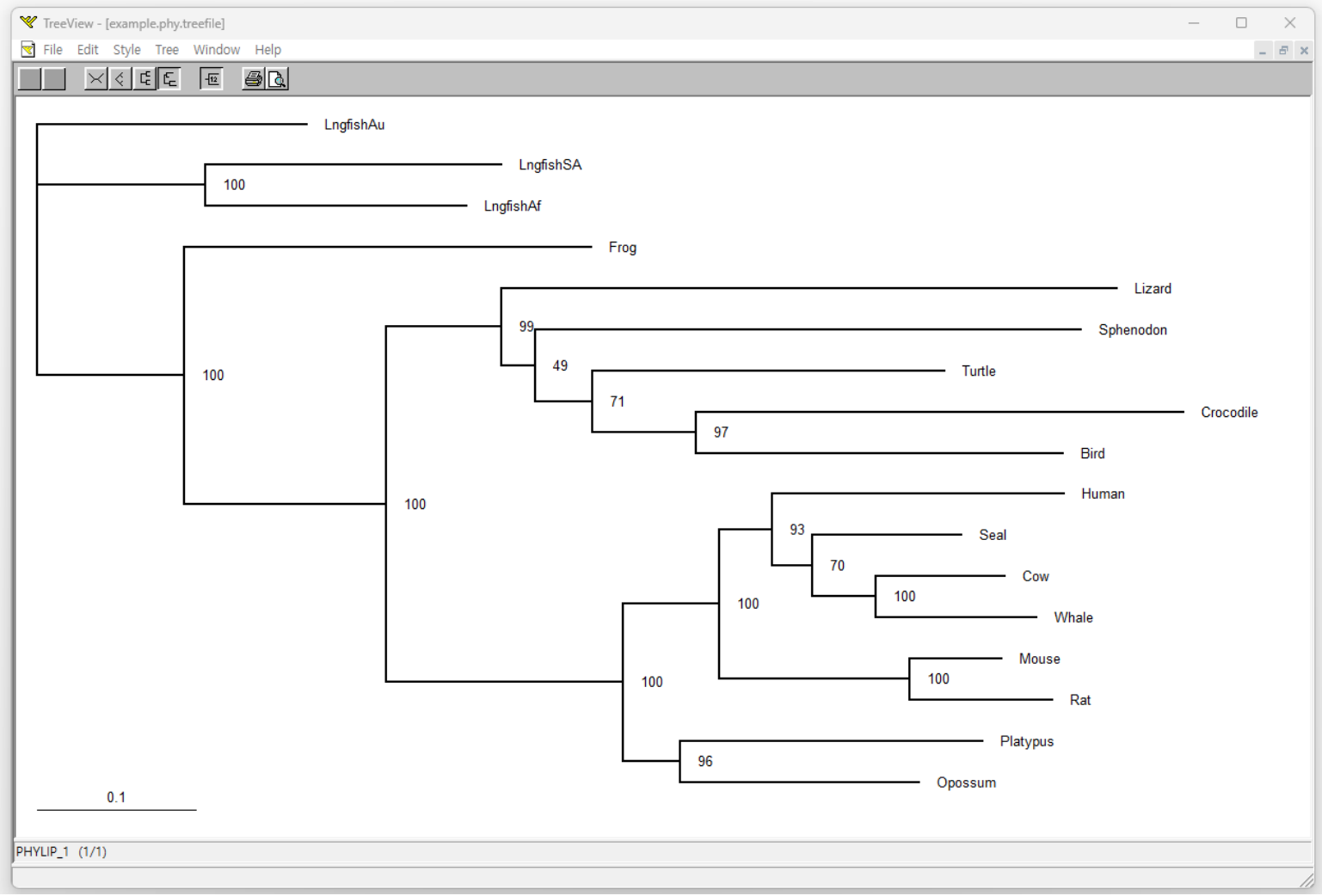
그림 3은 생각할 수 있는 모든 모형에 대해 AIC, BIC등의 스코어를 계산한후에 최적의 모형을 선택하는 과정을 보여준다. 사용된 예제화일은 17개 종에 불과하고 염기서열의 길이도 1998염기로 짧은 편이라서 짧은 시간안에 실행이 가능하지만, 데이터의 규모가 큰 경우 모든 모형에 대해서 스코어를 계산하는 것은 비효율적이다. 예컨데 JC, K80 같은 모형은 네 종류의 염기가 모두 같다고 가정하는데 이런 가정은 너무나도 비현실적이어서 거의 대부분의 데이터분석에서는 기각되는 모형이다. 따라서 이렇게 비현실적인 모형을 제외하고, 비교적 현실을 잘 반영한다고 알려진 모형만 한정해서 모형비교를 하는 것이 효율적일 수 있다. 이럴때 아래와 같이 실행할 수 있다. '-m MF'는 모형비교를 하는 옵션이고 '-mset GTR,HKY,TN '은 대상이 되는 모형세트를 나열하는 옵션이다. 즉 GTR, HKY, TN 모형과 그 변종들에 대해서만 모형 비교를 한다. 생성된 로그화일 (example.phy.log)을 보면 280여 개의 모형에 대해 정보량

```
C:\temp>  iqtree.exe -s example.phy -m MF -mset GTR,HKY,TN -redo
```

그림 7. GTR, HKY, TN 모형에 한정해서 모형선택을 수행함.



모형 비교를 생략하고 일반적으로 계통수 추정에 적용할 수 있는 명령어로 기억해 두자. (bootstrap 확률까지 같이 계산)



TreeView는 계통수의 형태를 빠르게 확인하거나 편집할 때 유용하다.

논문에 사용될 근사한(?) 그림을 위해서는 FigTree 프로그램을 사용하는 것이 좋다.