한국진화학회지**(2022) VOL. 1(1):88-104**
Review article

DNA 염기치환 모형의 비교

서태건[1]*

요약**:** 분자진화 데이터 분석에 있어서 염기치환 모형은 최대가능도 추정법이나 베이지안 추정법 적용에 매우 중요한 역할을 한다. DNA 치환 모형의 경우, 네 종류 염기 간의 치환에 대해 상대적인 비율을 규정한 것으로 가장 단순한 JC모형부터 가장 복잡한 GTR모형에 이르기까지 4×4 치환율 행렬만 203가지가 존재하고, 여기에 사이트 간 진화 속도의 이질성 모형, 불변사이트 모형과의 조합을 고려하면 주어진 데이터에 적용할 수 있는 염기치환 모형의 수는 비약적으로 증가하게 된다. 본 논문에서는 다양한 염기치환 모형의 후보군으로부터 최적의 모형을 선택하는 정보량기준에 관해 기술하고 대표적인 정보량기준인 AIC, BIC가 실제 분자진화 데이터 분석에 어떻게 적용이 되는지 IQ-TREE 프로그램의 예제를 이용하여 설명한다.

키워드**:** 염기치환, DNA 모형, AIC, BIC, LRT

[1] *인천광역시 연수구 송도동 송도미래로26 극지연구소*

***Corresponding author**: seo.taekun@gmail.com

서론

DNA 염기서열 혹은 아미노산 서열데이터와 같은 '분자데이터 (molecular data)'를 이용한 '분자진화 (molecular evolution)'연구는 현대 진화학 발전에 매우 중요한 역할을 하고 있다. 분자데이터는 형태적인 데이터 (morphological data)에 비해 다량으로 확보 가능하고 정량적이고 객관적인 데이터 처리가 가능하다는 장점이 있다. 20세기 후반부터 비약적으로 발전한 DNA 염기서열 결정 기술에 힘입어 분자진화 연구도 비약적으로 발전을 거듭하고 있다.

생물의 진화 연구에 있어 가장 기본이 되는 작업은 생물들 간의 '비교'이다. 비교를 통해 생물들이 무엇이 얼마큼 다른가를 정량적으로 측정하고 '다름'의 정도를 수치화하여 '거리'를 측정하는 것은 진화 연구의 가장 기본이 되는 일이라고 할 수 있다. 분자진화의 관점에서 생물 간의 거리 , 즉 '진화거리 (evolutionary distance)'는 흔히 "염기서열 데이터의 사이트당 발생한 DNA치환 [1] 의 횟수"로 나타낸다. 분자데이터의 진화 과정 중 발생하는 치환 현상은 일반적으로 직접 관찰할 수 없어 진화 과정의 결과물인 염기서열 데이터를 이용해 과거의 사건을 추정하곤 한다. 이 과정에서 DNA 염기치환 모형을 이용한 통계분석 방법이 유용한 도구로서 사용된다.

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

[1] 코돈 치환이나 아미노산 치환을 기반으로 하는 진화거리도 있으나 이에 대한 논의는 별도의 논문에서 하기로 하고 본 논문에서는 DNA 염기치환에만 한정하여 서술한다(결과 및 고찰 참조).

[2] A↔C, A↔G, A↔T, C↔G, C↔T, G↔T 여섯가지이다.

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.

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

[3]분기연대 추정은 치환 모형에 크게 영향을 받지 않는다는 보고도 있다 (Tao et al. 2020)

[4]시뮬레이션같은 통제된 상황에서는 '옳은 모형'이 존재할 수 있으나, 자연 상태에서 얻은 실제 데이터 (real data) 분석에 있어서 '옳은 모형'은 일반적으로 있을 수 없다.

[5]AIC의 경우 $\rho = 1$ 이고 BIC의 경우 $\rho = \log(\text{데이터 갯수})/2$ 이다.

[6]본 논문에 등장하는 영문 통계학 용어는 한국통계학회의 번역과 김우철(2021)을 참고하여 번역하였다. http://www.kss.or.kr/bbs/board.php?bo_table=psd_sec

항이다. 이 두번째 항은 모수의 갯수 p_m에 비례한다. 따라서 모형이 복잡할수록, 첫 번째항은 커지지만 두 번째 항도 역시 커지게 되고, 모형이 어느 정도 이상 복잡해지면 첫번째 항의 증가분은 둔화하는 반면 두 번째 항은 지속적으로 증가하여 어느 정도 수준에서 모형의 적합성 (첫 번째 항)과 모수 추정의 불확실성 (두 번째 항)의 trade-off 관계가 형성된다. 결과적으로 식 (1)이 최대가 되도록 하는 모형이 가장 좋은 모형으로 결정되는 것이다. [7]

본 논문에서는 DNA 염기치환 모형 중 대표적인 몇 개를 선별하여 염기치환 모형이 정의되고 해석되는 방법을 설명한다. 많은 분자진화 분석프로그램은 자동화가 되어 있어 사용자가 직접 AIC, BIC 스코어 등을 직접 계산하지 않고도, 심지어는 정보량 기준에 대한 이해가 전혀 없어도, 기계적으로 최적의 모형을 선택할 수 있도록 개발되어 있다. 하지만, 정보량기준의 정의를 명확히 이해하고 스코어를 직접 계산하여 확인하는 작업은 모형에 대한 이해를 높이고 염기치환모형이 설명하는 분자진화 메카니즘에 대한 이해를 증진시키므로 연구에 직간접적으로 큰 도움이 되리라 생각한다. 본 논문은 이러한 목적을 위해 기획되었다. 또한, 분자진화 연구에 널리 사용되는 IQ-TREE 프로그램의 사용 예를 소개하여 실제 데이터 분석에도 도움이 되도록 구성하였다.

DNA 치환모형의 정량적인 비교

DNA 염기치환모형

너무 단순하지도 복잡하지도 않은 중간 정도의 복잡도를 가진 모형인 HKY모형(Hasegawa et al. 1985)으로부터 논의를 시작하도록 하자. 이 모형은 매우 짧은 시간에 일어나는 치환의 상대적인 발생율 (전이율)을 아래와 같이 규정한다.

$$\mathbf{R}^{(HKY)} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & - & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & - & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & - \end{bmatrix} \end{array} \tag{2}$$

행렬 **R** 의 (i,j) 원소 R_{ij} 는 염기 i 가 염기 j로 치환되는 순간치환율을 나타낸다. $\pi_j (j \in \{A,C,T,G\})$ 는 염기 j의 빈도를 나타낸다. '−' 기호로 표시된 대각원소는 $R_{ii} = -\sum_{j \neq i} R_{ij}$로 정의된다. 즉, 대각원소는 각 행의 비대각원소의 합에 음의 부호를 취한 것이다.[8] 염기치환은 고장난 전구의 교체에 비유하면 이해하기 쉽다. DNA 염기서열은 네가지 색상 (A,C,G,T)의 전구가 불이 켜진 채 일렬로 늘어선 장치에 비유할 수 있고, 염기치환은 전구가 고장났을 때 새로운 전구로 교체하는 작업으로 비유할 수 있다. 고장난 전구를 빼내고 새롭게 투입되는 전구의 색상은 그 전 색상과 반드시 일치할 필요는 없고 랜덤하게 선택된다고 가정하자. 그러면 새롭게 투입되는 전구는 재고 비율에 비례하여 선택이 될 것이다. 마찬가지로 어떤 염기가 다른 염기로 치환될때 새롭게 투입되는 염기가 네 종류 염기 중에 랜덤하게 결정된다고 가정한다면

[7] AIC와 BIC는 식 (1)에 {-2}를 곱한 것으로 정의된다. 따라서 AIC와 BIC는 스코어값이 작을수록 좋은 모형이 된다.

[8] 대각원소의 정의는 미분방정식을 이용한 전이확률 계산을 위해 필요한 것이다. 자세한 설명은 본 논문의 수준을 넘는다.

치환율은 새롭게 투입되는 염기의 빈도에 비례한다고 가정하는 것이 자연스럽다. 따라서 치환율이 π_j에 비례하는 것은 직관적으로 이해하기 어렵지 않다. 또한 퓨린(purine; A와 G)끼리, 피리미딘(pyrimidine; C와 T)끼리는 분자구조가 비슷하여 A가 빠진 자리는 G로 대체될 가능성이 크고 C가 빠진 자리에는 T로 대체될 가능성이 크다. 따라서 퓨린끼리의 치환 혹은 피리미딘끼리의 치환 [9]이 퓨린과 피리미딘간의 치환 [10]보다 더 빈번하게 일어남을 가정하는 것이 합리적으로 보인다. 이를 표현하는 모수가 κ이며 κ는 두 치환의 비(ratio)로 정의된다. 식 (2)와 같은 치환율행렬이 주어지면 t 시간동안 발행하는 염기치환의 확률은 $\exp\{t\mathbf{R}\}$ 로 계산된다. 계통수 (phylogenetic tree) 상의 여러 가지 (branch) 들에 대해 염기치환 확률을 구하고 이를 이용하여 가능도를 계산한다. 가능도를 계산하는 효율적인 방법으로 prunning algorithm (Felsenstein 1981)이 알려져 있다.

식 (2)에서 $\kappa = 1$로 치환할 경우를 생각해 보자. 이 경우 HKY모형은 transition과 transversion의 치환율에는 차이가 없다고 가정하는 F81(Felsenstein 1981)모형이 된다. 치환율행렬은 다음과 같이 주어지고 각 염기치환율은 치환되는 염기의 빈도에만 비례한다.

$$\mathbf{R}^{(F81)} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & \pi_C & \pi_G & \pi_T \\ \pi_A & - & \pi_G & \pi_T \\ \pi_A & \pi_C & - & \pi_T \\ \pi_A & \pi_C & \pi_G & - \end{bmatrix} \end{array} \tag{3}$$

한편, 식 (2)의 HKY모형에서 염기의 빈도가 동일하다고 가정해보자. 이는 각 π_j를 동일하게 1/4로 치환하는 것이다. 그런데 전이율행렬은 가능도 계산과정중에 정규화[11]를 실행하기 때문에 동일한 상수 (예컨대 1)로 치환해도 무방하다. 그러면 transition과 transversion의 치환율 차이는 가정하지만 (κ로 표현) 네 종류의 염기의 비율은 동일하다고 가정하는 K80 모형(Kimura 1980)[12]이 된다. K80 모형에 추가로 $\kappa = 1$ 조건을 가정하면 모든 염기치환이 동일한 비율로 발생한다고 가정하는, 염기치환 모형중 가장 단순한 모형인 JC모형(Jukes and Cantor 1969)이 된다.

$$\mathbf{R}^{(K80)} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & 1 & \kappa & 1 \\ 1 & - & 1 & \kappa \\ \kappa & 1 & - & 1 \\ 1 & \kappa & 1 & - \end{bmatrix} \end{array}, \quad \mathbf{R}^{(JC)} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & 1 & 1 & 1 \\ 1 & - & 1 & 1 \\ 1 & 1 & - & 1 \\ 1 & 1 & 1 & - \end{bmatrix} \end{array} \tag{4}$$

[9] 이를 transition 이라 한다.

[10] 이를 transversion 이라 한다.

[11] 단위시간당 일어나는 치환수가 진화거리가 되도록 $\{-\Sigma_i \pi_i R_{ii}\}^{-1}$를 전이율행렬의 모든 원소에 곱해주는 작업이다.

[12] K80 모형의 전이율 행렬에서 '1'의 자리에 α를, κ 자리에 β를 표시하고 Kimura's two parameter model 이라고 부르는 경우도 있다. 전이율행렬에 두개의 모수가 있는 것처럼 보이지만 상대적인 비율이 중요하므로 실제로 전이율행렬의 모수는 한개 ($\beta/\alpha = \kappa$)이다.

이제 식 (2)의 HKY모형보다 복잡한 모형을 살펴보자. 식 (2)는 두 종류의 transition, A↔G, C↔T를 구분하지 않고 각각의 순간치환율에 동일한 모수 κ를 할당한 모형이다. 하지만 이 두 가지 transition이 다른 치환율을 가질 것이라고 생각하는 것은 지극히 자연스럽다. 따라서 두가지 transition을 구분하여 별도의 모수 κ_1, κ_2 을 할당한 모형이 제안되었는데 이것이 TN 모형(Tamura-Nei 1993)이다. HKY모형에서 TN모형으로 확장하는 것과 같은 방식으로 여섯가지 염기치환 쌍을 단계적으로 구분지어 별도의 모수를 할당해 가면 최종적으로 여섯가지 염기치환에 모두 다른 모수를 할당하는 GTR 모형(General Time Reversible model; Tavaré 1986)에 도달하게 된다. 앞에서 언급한 바와 같이 가능도를 계산할때는 전이율 행렬의 정규화가 행하여지고 상대적인 치환율이 중요하므로 여섯가지 모수중 하나를 1로 고정시킬 필요가 있고 여기에서는 편의상 G↔T 쌍에 1을 할당하겠다. 그러면 GTR모형은 다섯개의 모수를 갖게 되고 각 염기치환은 치환되는 염기의 빈도에 비례하며, G↔T 치환의 치환율에 비해 다른 염기치환은 각각 a,b,c,d,e배 빠르거나 느리다는 것을 가정하게 된다.

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

위에서 설명한 여섯가지 모형의 관계를 그림 1에 나타냈다(TIM2 모형은 후술하는 데이터 분석에서 등장하는 모형이다). 위에서 아래로 갈수록 모수가 증가하여 복잡한 모형이 되며, 수직선 혹은 대각선으로 연결되어 있는 경우 복잡한 모형의 일부 모수를 특정 값으로 치환하면 단순한 모형이 됨을 의미한다.[13] 화살표 방향을 따라가며 연결될 수 없는 경우, 이러한 모형 간의 포함관계는 성립하지 않는다. 예컨대, K80모형과 F81모형은 화살표 방향을 따라가며 연결될 수 없어 포함관계가 성립하지 않는다. 즉, 모수를 특정 값으로 고정함으로써 어느 한쪽에서 다른 쪽으로 변환되지 않는다.

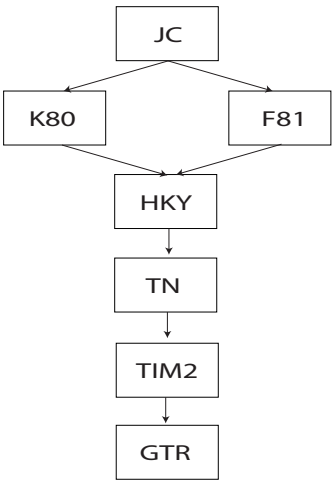

그림 1. DNA 치환모형들 간의 관계. 가장 단순한 JC 모형부터 가장 복잡한 GTR모형에 이르기까지 203개의 모형이 존재하나 이 그림에서는 몇 개만 선별하여 표시하였다.

위에서 설명한 모형중 JC, K80모형을 제외한 모형은 염기의 빈도가 각각 다르다는 것을 가정하고 있다. 이를 명시적으로 표시하기 위해 흔

[13] 이때 단순한 모형을 'nested model' '내포모형'이라 한다.

히 '+F' 태그를 모형 이름에 붙이는 경우도 있다.
(HKY+F, GTR+F 등). 또한, 논문의 출판 연도를 명시하여 TN모형을 TN93, HKY모형을 HKY85 모형으로 부르는 경우도 있다.

그 외 주목해야 할 가정들

식 (2) – (5)의 모형들은 염기치환 양상에 대해 여러가지 가정을 하고 있다. 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.

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 $\sim$ 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).

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

얼라인된 염기서열 데이터를 관찰하면 사이트별로 치환의 정도가 많이 다른 것을 흔히 볼 수 있다. 어떤 사이트는 치환이 전혀 일어나지 않아 동일한 염기가 관찰되는 반면, 어떤 사이트는 치환이 왕성하게 일어나 네 종류 염기가 모두 관찰되기도 한다. 염기치환은 확률적으로 일어나는 사건이므로 모든 사이트가 동일한 치환의 정도를 보이지 않는 것은 당연하다. 하지만 사이트 간 진화 속도가 동일하다는 전제 하에 확률적으로 설명할 수 있는 범위를 넘어서는 차이[14]를 보이는 경우가 있다. 이러한 사이트 간 진화 속도의 이질성(Rate Heterogeneity among Sites; RHAS)을 정량적으로 기술하기 위해 흔히 감마모형이 이용된다 (Yang 1994b). 감마모형은 통상 두개의 모수로 형태가 결정된다. 분자진화 분석을 위해서는 편의상 이 두

[14]예컨데, 변이가 없는 사이트가 너무 많다든지, 혹은 변이가 빈번한 사이트가 너무 많다든지 하는 사이트 간 차이

모수를 역수의 관계로 설정하여 미지의 모수 수를 하나로 줄이고 (이를 흔히 α로 표현한다) 감마분포의 평균이 1이 되도록 설정한다. 그리고 각 염기서열 사이트는 기준이 되는 진화속도에 감마분포에서 얻은 난수를 곱한 값을 진화속도로 갖는다는 설정이다.

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.

실제로 진화속도의 이질성 추정에 있어서 연속형 감마분포를 그대로 이용하는 것은 계산시간이 많이 소요되어 비현실적이다. 따라서 감마분포를 이산형 감마분포 (discrete gamma distribution)로 근사시켜 로그가능도를 계산한다. 가령 K개의 카테고리를 가진 이산형 감마분포를 가정한다고 하면, 감마분포의 변수가 취할 수 있는 범위 $(0,\infty)$ 를 1/K 확률로 균등하게 나눈 후, 각 구간의 확률변수를 그 구간의 평균 혹은 중앙값으로 고정하는 것이다. 이렇게 고안된 이산형 감마분포를 이용한 근사는 $K = 4$인 경우 연속형 감마분포에 필적하는 좋은 퍼포먼스를 보여준다(Yang 1994b). 이산형 감마분포는 위에서 언급한 여러 치환 모형 (식 (2) – (5))과 조합하여 적용할 수 있으며 '+GK' 태그를 모형명에 덧붙여 표현한다. 가령, $K = 4$ 인 이산형 감마분포를 HKY모형과 조합하면 'HKY+G4', GTR모형과 조합하면 'GTR+G4', 이런식으로 모형명을 나타낸다. 카테고리수 K가 달라지더라도 추정하는 모수는 해당하는 연속적인 감마분포의 모수 한개 (α)이다. 카테고리수가 주어질때 각 카테고리가 가질수 있는 변수는 α 이외의 다른 모수의 영향없이 자동적으로 결정되기 때문이다.

이산형 감마분포 가정에 의하면 모든 사이트는 빠르든 느리든 어느 정도의 진화속도를 가지며 진화를 한다. 하지만 염기치환이 전혀 일어나지 않은 것으로 추정되는 사이트도 실제로 다수 관찰된다. 이를 고려한다면 사이트의 일정 비율은 '불변사이트'라고 가정 할 수 있다 (invariable site model). 치환이 전혀 일어나지 않는 사이트의 비율을 나타내는 모수 (p)로 이 모형을 설명하며 '+I' 태그를 모형명에 추가한다. 예를 들어 'HKY+G4' 모형에 추가로 불변사이트모형을 가정한다면 해당모형은 'HKY+G4+I'가 된다.

감마분포가 아닌 다른 형태의 이산형 확률분포를 생각할 수도 있다. α가 주어지면 K개의 카테고리의 속도가 자동으로 결정되는 이산형 감마분포와는 달리 각 카테고리가 서로 다른 진화속도와 서로 다른 확률을 갖는다는 모형이 제안되었고 (Yang 1995; Soubrier et al. 2012) 이를 흔히 자유속도모형 (free rate model)이라 부른다. 속도평균이 1이 되고 각 카테고리의 상대빈도의 합이 1이 되는 필요조건을 가하면

[15] 모수가 boundary에 있는 상황으로 가능도비검정(likelihood ratio test; LRT)의 자유도 결정에 주의가 요구된다.

$(K-1)$ 개의 상대빈도 모수와 $(K-1)$ 개의 상대속도 모수, 도합 $2(K-1)$개의 모수를 추가로 더 추정하게 된다(이하 IQ-Tree를 이용한 분석 예 참조). 자유속도모형에서는 모형명으로 '+RK' 태그를 추가한다. 가령 카테고리 수가 네 개인 자유속도모형을 가정할 경우 모형명에 '+R4'를 추가한다 (예: HKY+R4, GTR+R4 등).

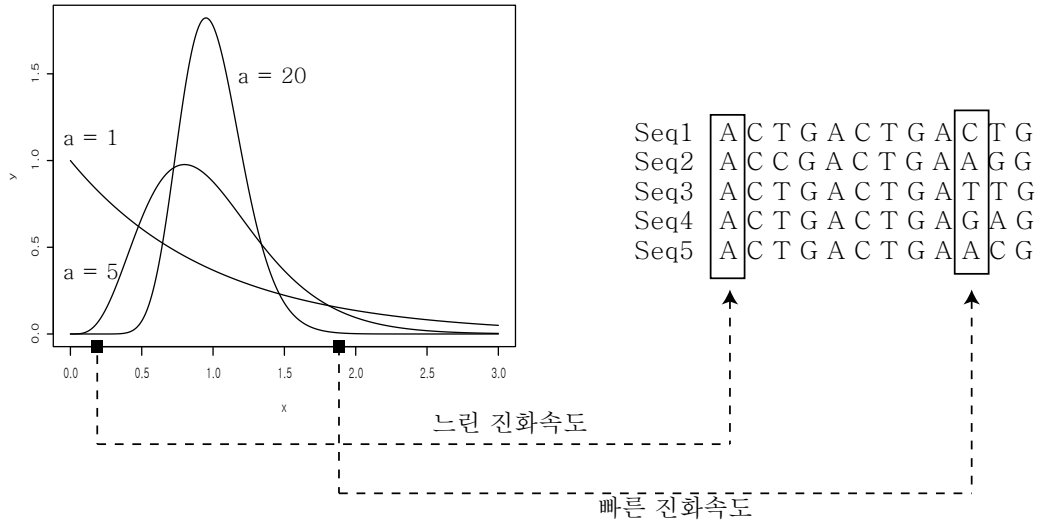

그림 2. 감마분포를 이용한 진화속도 이질성 모형화.

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, 복잡한모형} 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.

환모형의 비교에서 활용이 다소 제한적이다. 이런 문제점을 극복할 수 있는 것은 Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) 같은 정보량기준이다. 즉, 단순한 모형이 복잡한 모형의 내포모형이 아닌 경우에도 이 두 정보량기준은 적용이 가능하다.

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.

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

이 식에서 두번째 항이 AIC스코어를 보정하는 값인데, 주어진 p에 대해 n이 커지면 커질수록 0에 가까와져 염기서열의 길이가 충분히 클때는 AIC와 AICc의 정의는 동일한 값으로 수렴하게 되어 모형 선택의 결과에 큰 차이가 없게된다. 여기서 주의해야 할 점은 AICc는 선형회귀분석 모형을 위해 개발된 측도라는 것이다. 분자진화 모형은 선형회귀분석 모형과는 엄연히 다르고 식 (8)의 AIC 보정항이 마찬가지로 유용하다는 이론적 근거는 알려진 바 없다. 실제로 Susko and Roger (2019) 는 여러가지 모형설정에 따라 AICc가 달리 정의될 수 있음을 보이기도 했다.

데이터 분석의 예

IQ-TREE를 이용한 모형 비교의 예

DNA 염기치환모형을 비교하기 위한 프로그램이 다수 개발되어 있으나[18] 여기에서는 비교적 최근에 개발되었고 많이 사용되는 IQ-TREE 프로그램(Nguyen et al. 2014; version 1.6.12) 을 이용하여 설명하겠다. IQ-TREE는 입력한 염기서열데이터를 분석해 최적의 모형을 자동적으로 결정해준다. 따라서, 사용자는 AIC, BIC등의 정보량 기준을 이해하지 못하더라도 프로그램이 선택해준 모형을 이용할 수는 있다. 하

[17] 여기서 '가깝다'라는 의미는 Kullback-Leibler divergence를 기준으로 가깝다는 의미이다(Konishi and Kitagawa 2008). 모형간의 '거리'를 측정하는 측도는 다른 방식으로 정의할 수도 있다.

[18] https://evolution.genetics.washington.eduphylipsoftware.html#Modelselection 이곳에 관련 프로그램이 잘 정리되어 있다.

지만, 분자진화나 집단유전등의 데이터 분석에서는 프로그램이 계산한 로그가능도를 이용하여 사용자 본인이 직접 AIC, BIC를 계산하여 모형을 비교해야 하는 상황이 종종 발생한다. 또한 각종 논문에서 기술된 통계모형의 정당성 주장을 쉽게 이해하기 위해서는 위에서 설명한 AIC, BIC등의 정의에 맞춰서 구체적인 스코어가 계산되는 과정을 한번 학습해 보는 것이 필요하다 생각된다.

데이터 분석에는 IQ-TREE프로그램이 제공하는 예제파일 example.phy를 이용한다. 이 파일에는 17개 포유류종의 미토콘드리아 r-DNA로부터 추출한, 길이 1998 염기의 염기서열 데이터가 포함되어 있다. 편의상 윈도우 버전의 프로그램을 중심으로 설명한다.

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

그림 4는 그림 3 의 명령어 실행 이후에 생성된 로그파일(example.phy.log)의 일부를 나타낸 것이다. 각종 염기치환 모형과 RHAS 옵션의 조합에 따라 적용가능한 280여가지의 모형에 대해 로그가능도 스코어(음의 부호를 붙여 양수가 된 값이 '-LnL'열에 표시된다), 각각의 모형에 대한 모수의 갯수 ('df'열에 표시된 값이다), 이를 이용하여 얻은 AIC, AICc, BIC 스코어가 출력되고 최종적으로 세가지 정보량기준에 의해 선택된 최적의 (best-fit) 모형이 하단에 출력된다.

```
 ...
ModelFinder will test 286 DNA models (sample size: 1998) ...
No. Model        -LnL       df AIC        AICc       BIC
1   JC           23650.090 31 47362.181 47363.190 47535.778
2   JC+I         22582.953 32 45229.905 45230.980 45409.102
3   JC+G4        22261.254 32 44586.508 44587.583 44765.705
4   JC+I+G4      22247.806 33 44561.612 44562.755 44746.409
5   JC+R2        22284.451 33 44634.902 44636.044 44819.699
...
199 TIM2+F+I+G4 21164.169 39 42406.339 42407.932 42624.735
...
278 GTR+F+R2     21247.034 41 42576.069 42577.829 42805.665
279 GTR+F+R3     21157.735 43 42401.470 42403.406 42642.266
280 GTR+F+R4     21157.665 45 42405.330 42407.451 42657.326
Akaike Information Criterion: GTR+F+R3
Corrected Akaike Information Criterion: GTR+F+R3
Bayesian Information Criterion: TIM2+F+I+G4
Best-fit model: TIM2+F+I+G4 chosen according to BIC
...
```

그림 4. example.phy.log 일부분

식 (6)–(8)의 정의에 따라 AIC, AICc, BIC 스코어를 계산해보자. IQ-TREE는 염기서열 데이터로부터 maximum parsimony 계통수를 구한 후, 이 계통수를 모든 모형에 적용하여 로그가능도를 계산한다. 주어진 데이터는 taxa 수가 T=17이고 이에 해당하는 bifurcating unrooted 계통수의 branch 갯수는 2T-3=31개이다. JC모형은 전이율 행렬에 미지의 모수를 갖지 않아 (식 (4)), branch length가 유일한 미지의 모수이다. 따라서, 그림 4의 1번 모형의 df는 31이 되고 로그가능도값(=-23650.090)과 정의 (6)–(8)을 이용하면 AIC 스코어는 $2\times 23650.090+2\times 31=47362.180$, AICc값은 $47362.180+\frac{2p^2+2p}{n-p-1}\approx 47362.180+1.01=47363.181$, BIC값은 $2\times 23650.090+31\log(1998)\approx 47535.78$가 되어, AIC, AICc, BIC스코어가 정의에 따라 계산되었음을 확인 할 수 있다. AIC와 AICc 기준으로 최적의 모형으로 선택된 GTR+F+R3 모형(279번)의 모수 갯수를 세어보자. 계통수의 branch length 31개가 있고, 식 (5)의 GTR모형의 전이율행렬의 a부터 e까지 5개, "+F" 옵션에 따른 네 종류 염기의 빈도에 해당하는 모수 3개 (빈도의 합이 1이 되어야 한다는 제한에 따라 4개가 아니라 3개이다), "+R3" 옵션에 따른 3개의 속도이질성 그룹으로부터 $2\times(3-1)=4$개의 모수가 있다. 따라서 279번 모형의 df는 31+5+3+4 = 43이다. 이를 이용하여 AIC, AICc, BIC를 계산하면 그림 4의 결과와 일치함을 확인 할 수 있다.

BIC 기준에 의해 최적의 모형으로 선택된 것은 199번의 TIM2+F+I+G4 모형이다. TIM2 모형은 GTR 모형(식 (5)) 에서 A↔C, A↔T에 동일한 치환율을, C↔G와 G↔T치환에 동일한 치환율을, 그리고 A↔G와 C↔T에는 각각 다른 치환율을 할당하는 모형에 해당한다 (IQ-TREE 설명문서 참조). 알기 쉽게 전이율행렬로 표현하면 다음과 같다.

$$\mathbf{R}^{(TIM2)}=\begin{array}{c} \\ A \\ C \\ G \\ T \end{array}\begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} - & \kappa_3\pi_C & \kappa_1\pi_G & \kappa_3\pi_T \\ \kappa_3\pi_A & - & \pi_G & \kappa_2\pi_T \\ \kappa_1\pi_A & \pi_C & - & \pi_T \\ \kappa_3\pi_A & \kappa_2\pi_C & \pi_G & - \end{bmatrix} \end{array} \tag{9}$$

이를 식 (5)의 TN 모형과 비교해보면, TIM2 모형은 transversion을 두가지 타입으로 구분한 후 한쪽 (C↔G, G↔T)은 치환율을 1로, 다른 한쪽(A↔C, A↔T)은 치환율을 κ_3로 지정한 것임을 알 수 있다. Transition을 두가지 타입으로 구분한 것은 TN 모형과 같다. TIM2 모형은 TN모형보다 모수가 하나 더 많고 GTR 모형보다는 두개가 적어 그림 1에 표현된 것처럼 TN과 GTR 사이의 복잡도를 가진 모형이다. TIM2+F+I+G4 모형의 모수 갯수는 31(계통수로부터) + 3 (염기 빈도로부터) + 3 (κ 모수들) + 1 (invariant site model의 p에 해당) + 1 (이산형 감마분포의 α에 해당) = 39이고 이는 그림 4의 199번 모형의 df값과 일치한다. 이를 이용하여 TIM2+F+I+G4 모형의 AIC, AICc, BIC도 쉽게 계산할 수 있다. '+G' 옵션은 '+R' 옵션과 달리 미지의 모수는 한개임에 유의한다. 추정해야 할 모수는 α 하나이고 카테고리 수는 사전에 지정된다.

그림 4에서 알 수 있듯이 AIC와 AICc 기준에 의해 얻어진 최적의 모형과 BIC 기준에 의해 얻어진 최적의 모형은 다르다. 하지만 서로 다른 기준에서 1위를 차지한 모형들을 동일한 기준에서 스코어를 비교하면 그 차이는 크지 않음을 알 수 있다. 예컨데 BIC 기준으로 1위 모형인 TIM2+F+I+G4 모형의 AIC

스코어는 42406.339로서 AIC 기준으로 1위 모형인 GTR+F+R3 모형의 스코어 42401.470과 매우 유사한 값을 갖는다. 또한, AIC 기준으로 1위를 차지한 GTR+F+R3 모형의 BIC 스코어는 42642.266은 BIC 스코어의 최소값(TIM2+F+I+G4 모형의 BIC 스코어) 42624.735와 큰 차이를 보이지 않는다. 이처럼 서로 다른 기준으로 1위를 차지한 모형들을 동일한 기준하에 비교하면 매우 유사한 값을 가짐을 알 수 있다.

그림 4의 example.phy.log 화일 마지막 부분에는 TIM2+F+I+G4 모형을 이용하여 추정한 최대가능도 계통수, 전이율행렬의 모수 값 (식 (9)), 염기빈도, RHAS 모수 α의 추정값과 로그가능도가 저장되어 있다.

```
  ...
---------------------------------------------------------------------
| FINALIZING TREE SEARCH |
---------------------------------------------------------------------
Performs final model parameters optimization
Estimate model parameters (epsilon = 0.010)
1. Initial log-likelihood: -21152.525
Optimal log-likelihood: -21152.523
Rate parameters: A-C: 5.58055 A-G: 7.64567 A-T: 5.58055 C-G: 1.00000
C-T: 22.62532 G-T: 1.00000
Base frequencies: A: 0.355 C: 0.228 G: 0.192 T: 0.225
Proportion of invariable sites: 0.157
Gamma shape alpha: 0.736
Parameters optimization took 1 rounds (0.084 sec)
BEST SCORE FOUND : -21152.523
Total tree length: 4.218
...
```

그림 5. example.phy.log 마지막부분

여기에 표시된 로그가능도 값은 -21152.523이고 이는 그림 4의 199번 모형에 나와 있는 로그가능도값 -21164.169과는 다르다.[19] 그 이유는 그림 4의 모형 비교에 사용된 계통수와 최종분석에서 "TIM2+F+I+G4" 모형으로 얻어진 계통수는 다르기 때문이다. 엄밀히 말하면 그림 4 의 1번부터 280번 까지의 모형 비교에서 사용되는 계통수는 각각의 모형하에서 얻어진 최적의 계통수이어야 한다. 하지만, 이는 매우 많은 계산 시간을 요하기 때문에 공통된 계통수로 비교를 하는 방법이 흔히 사용된다. 앞에서 언급한 바와 같이 IQ-TREE는 공통된 계통수로 maximum parsimony tree를 사용한다. BIC로 최적의 모형이 결정되면 IQ-TREE는 최적의 모형으로 계통수 탐색을 다시 실행한다.

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

[19] 그림 4에는 로그가능도에 음의 부호를 붙여 양수로 표기되었다.

"GTR+F+R3" 모형을 이용하여 추정한 계통수와 전이율행렬의 모수, RHAS 의 α, invariant site 비율등이 example.phy.iqtree 화일에 저장된다. 데이터의 결과에 대한 고찰은 생략하고 여기서는 최대 로그가능도값에 대해서만 간단히 언급한다. 새로 생성된 example.phy.log 화일에는 최대로그가능도 값이 -21147.004이다. 이는 그림 4의 279번 모형에 나와 있는 로그가능도 값 -21157.735와 약 10.7 유닛 정도 차이가 난다. 이는 모수 약 11개에 해당하는 기여분으로 최적의 계통수는 아니지만 최적에 근접한 계통수 (=maximum parsimony tree)를 적용시키는 것만으로도 이 정도의 모수 기여분 차이를 가져올 수도 있음을 보여주는 예시라고 할 수 있다.

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.

기준 스코어를 계산한 그림 4와는 달리 30여 개의 모형에 대해서만 조사를 한 것을 알 수 있다 (데이터 생략).

결론 및 고찰

이 논문에서는 대표적인 정보량기준인 AIC와 BIC에 대해 설명하고 이를 DNA 염기치환 모형 비교에 적용해보았다. IQ-TREE 프로그램의 예제 데이터에서 보았듯이 어떤 정보량기준을 적용시키는가에 따라 선택되는 최적의 모형도 달라지게 된다. 그렇다면 어떤 정보량기준을 적용시켜야 할까? IQ-TREE는 여러 모형을 비교한 끝에 BIC 기준으로 선택된 모형으로 계통관계와 branch length, 각종 모수등을 다시 한번 추정하여 결과를 출력해준다. 이는 BIC가 AIC보다 더 좋은 정보량기준이라는 것을 의미하는가? 반드시 그렇지는 않다. AIC와 BIC는 서로 다른 의미를 갖는 정보량기준일 뿐, 둘 사이에 우열 관계는 일률적으로 말할 수 없다. AIC는 미지의 데이터 생성 메카니즘과 주어진 모형 사이의 Kullback-Leibler Divergence를 최소화하는 모형을 찾기 위해 고안된 정보량 기준이고 (Konishi and Kitagawa 2008), BIC는 데이터의 주변확률을 최대로 하는 모형을 찾기 위한 정보량 기준이다(Schwarz 1978). 즉, 바라보는 관점이 다른 것이다. 식 (1)과 같이 공통된 형태를 갖고 있고 (구체적인 정의는 식 (6), (7)) BIC는 두번째 페널티 항이 데이터의 크기 (염기서열의 길이) n 이 증가함에 따라 증가하는 형태를 갖고 있어, BIC는 AIC보다 더 단순한 모형

을 선택하려는 경향이 있다. 이 논문에서 설명한 AIC와 BIC 이외에도 다양한 정보량 기준이 존재하는데 (Diak et al. 2020), 모수의 속성을 잘 인식하여 복잡한 모형을 선택하려는 경향인 'sensitivity' 와 불필요한 모수를 생략하고 단순한 모형을 선택하려는 경향인 'specificity' 어느 쪽에 큰 비중을 둘 것인가에 따라 정보량 기준의 선택이 달라지게 되는 것일뿐 일률적으로 어느 하나가 다른 것보다 우월하다고는 말할 수 없다.

IQ-TREE 프로그램의 예제 데이터 분석에서 보았듯이, BIC로 TIM2+F+I+G4 모형을 선택한 후에 얻는 ML (maximum likelihood) 계통수와, AIC로 GTR+F+F3 모형을 선택한 후에 얻는 ML계통수는 위상 (topology; 계통수의 branch length 차이는 무시하고 계통관계만을 고려)이 동일하다 (데이타 생략). 하지만, 위상이 다르다면 어떻게 대응해야 할까? 일반적으로 서로 다른 모형을 적용시켰을때 위상이 다른 계통수가 얻어질 수 있음이 잘 알려져 있다. 하지만 그 차이는 계통관계가 애매한 부위 (즉, bootstrap 확률 혹은 사후확률등이 낮은 internal node)에서 주로 보여지고 계통관계가 명확한 부위는 비교적 로버스트 (robust)함이 알려져 있다 (Ripplinger and Sullivan 2008). 따라서 서로 다른 모형을 적용했을 때 계통수 위상의 공통점이 보이는 부위는 그만큼 확실하게 데이터로부터 지지를 받는다는 의미가 되고 신뢰할 수 있는 계통관계라 할 수 있을 것이다[20]. 차이가 보여지는 부위가 연구의 주된 관심사라면 데이터가 그 부분에 대해서는 충분한 정보를 갖고 있지 못할 가능성을 포함해서 정보량 기준의 선택 방법등을 다시 한번 면밀히 검토할 필요가 있을 것이다.

데이터 샘플수가 적을때 AICc (corrected AIC)가 흔히 쓰인다. 최대 로그가능도가 로그가능도의 기대값이 되기 위해서는 식 (1) 두번째 항만큼의 보정이 필요한데 데이터수가 적을때는 보정항의 정밀도를 높이기 위해 식(8)의 두번째 항만큼의 여분의 보정이 필요하다는 것이다. 주의할 점은 식 (8)의 여분의 보정항은 회귀분석 모형으로부터 유도된 값이라는 것이다(Konishi and Kitagawa 2008). 이 보정항이 분자진화 염기치환 모형에도 적절한지는 알려진 바 없고 보정항의 형태는 모형마다 다르다. Susko and Rogers (2019)는 여러 모형에 대해 보정항이 식 (8)의 두 번째 항과 다른 형태로 유도됨을 보였다.

위에서 언급하였듯이 AIC는 분포간의 거리 (Kullback-Liebler divergence)를 기준으로 유도되었고 BIC는 데이터의 주변가능도로부터 유도 되었다. 하지만, 이렇게 얻어진 최적의 모형이 반드시 계통수 branch length 값 추정에 최고의 정확도를 가진다고는 말할 수 없다. 확률분포간의 차이가 최소가 된다는 것이 혹은 데이터의 주변가능도가 최대가 된다는 것이 반드시 branch length 추정의 정확도가 최대가 되는 것을 의미하는 것은 아니기 때문이다. 이에 대한 해결책으로 branch length 추정의 정확도를 기준으로 모형을 선택하는 결정이론 (decision theory; DT)에 기반을 둔 모형 선택방법도 제안되었다 (Minin et al. 2003). Cross-validation 기법을 이용해 반복적으로 모수를 추정하고 이로부터 정확도를 계산하여 비교하는 방법이다. AIC나 BIC에 비하여 계산시간이 많이 소요되어 현재 DT방법보다는 AIC, BIC를 이용하는 방법이 주류를 이룬다.

AIC, BIC에 의해 선택된 모형을 이용하여 추정한 결과 (분자진화분석의 경우, branch length 뿐만 아니라, 계통관계, RHAS 모수 α, 분기연대등)가 반드시 옳다는 보장은 없다. 오히려 그다지 좋지 않은

[20] 보다 정확히는 bootstrap 확률, 사후확률을 계산하든가 각종 통계적 검정방법을 적용시켜 정량적으로 평가해야 할 것이다.

모형을 사용해서 얻은 결과가 최적으로 선택된 모형보다 더 합리적이고 진실에 가까운 결과를 유도하는 경우도 간혹 발생한다. 계통관계 위상이나 공동조상 염기서열의 추정에 있어서는 모형선택을 거쳐 결정된 최적의 모형에 의한 정확도와, GTR+I+G 모형에 의한 정확도가 거의 차이가 없어 추정의 목적 여하에 따라 모형선택 과정이 반드시 필수적인 것은 아니라는 주장이 최근에 제기된 바도 있다(Abadi et al. 2019). 하지만, 최적의 모형을 이용했을때 진실에 가까운 결과를 얻을 가능성이 '평균적으로' 혹은 '확률적으로' 더 클것으로 예상하는것이 합리적이다. 여기에 객관적이고 정량적인 모형선택의 의미가 있는 것이다. 또한, 모형 선택은 다른 분석을 위한 선행작업, 기초작업이라는데 의미가 있다. 선택된 모형을 이용하여 계통관계를 보다 더 심도 있게 조사한다든지, 진화거리를 정확하게 추정하여 이를 이용하여 분기연대를 추정한다든지, 중간단계에서 행해지는 여러가지 의사결정의 합리성과 객관성을 설명하는 도구로서 의미가 있다 하겠다.

염기서열의 분자진화를 기술하는 수리모형으로서 본 논문에서 설명한 DNA치환 모형뿐만 아니라, 아미노산치환 모형, 코돈치환 모형이 있다(Felsenstein 2004; Yang 2006). DNA 치환 모형은 식 (2) – (5)처럼 4×4 전이율행렬을 정의하는 반면 아미노산 모형은 20×20 전이율행렬, 코돈모형은 61×61 전이율행렬[21]을 정의한다. 아미노산치환 모형과 코돈치환 모형은 DNA치환 모형으로는 알수 없는 다른 측면을 바라볼 수 있는 장점이 있다. 이에 대한 상세한 설명은 별도의 논문으로 정리하도록 하겠다.

감사의 글

본연구는 해양수산부의 재원으로 극지연구소의 지원을 받아 수행되었다 (과제번호: PE22140). 논문의 완성도를 높이기 위해 귀중한 조언을 해준 익명의 리뷰어에게 감사드린다.

참고문헌

Abadi S., Azouri D., Pupko T., Mayrose I. 2019. Model selection may not be a mandatory step for phylogeny reconstruction. Nature Communications 10:934.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contr. 19:716–723.

Bettisworth B. Stamatakis A. 2021. Root Digger: a root placement program for phylogenetic trees. BMC Bioinformatics 22:225

Box G.E.P. 1976. Science and statistics. J. Am. Stat. Assoc. 71:791-799.

Dziak J.J., Coffman D.L., Lanza S.T., Li R., Jermiin L.S. 2020. Sensitivity and specificity of information criteria. Briefings in Bioinformatics. 21(2):553–565.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 394: 783–791.

[21] 표준코돈이 아닌 경우, 전이율행렬의 차원은 달라질 수 있다. 가령 포유류의 미토콘드리아 코돈의 진화를 기술하는 전이율행렬은 60×60이다.

Felsenstein J. 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, MA.

Galtier N., Gouy. M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol.15:871-879.

Hasegawa H., Kishino H., Yano T. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

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.

Kimura M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

Konishi S., Kitagawa G. 2008. Information Criteria and Statistical Modeling. Springer.

Minin V., Abdo Z., Joyce P., Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52: 674-683.

Nguyen L-T, Schmidt H.A., von Haeseler A., Minh B.Q. 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol. Biol. Evol. 32(1): 268–274.

Posada, D. and Buckley, T. R. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793–808.

Posada, D. and Crandall, K. A. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817–818.

Ripplinger J., Sullivan J. (2008) Does choice in model selection affect maximum likelihood analysis? Syst. Biol. 57(1):76-85.

Schenk J.J., Larry Hufford L. (2010) Effects of Substitution Models on Divergence Time Estimates: Simulations and an Empirical Study of Model Uncertainty Using Cornales. Systematic Botany 35(3):578-592

Schierup M.H., Hein H. (2000) Consequences of recombination on traditional phylogenetic analysis. Genetics 156(2): 879–891.

Schwarz G. 1978. Estimating the dimension of a model. Ann. Stat. 6:461–464.

Sullivan J., Joyce P. (2005) Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36:445–466

Soubrier J., Steel M., Lee M.S.Y., Sarkissian C.D., Guindon S., Ho S.Y.W., Cooper A. 2012. The Influence of Rate Heterogeneity among Sites on the Time Dependence of Molecular Rates Mol. Biol. Evol. 29(11): 3345–3358.

Sullivan J., Markert J. A., Kilpatrick C. W. (1997) Phylogeography and molecular systematics of the Peromyscus aztecus species group (Rodentia: Muridae) inferred using parsimony and likelihood. Syst. Biol. 46:426–440.

Susko E., Roger A.J. 2019. On the use of information criteria for model selection in phylogenetics. Mol. Biol. Evol. 37(2):549-562

Stuart A., Ord J.K. 1991. Kendall's advanced theory of statistics (vol 2). Edward Arnold. pp. 861 – 864.

Tamura K., Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512–526.

Tao Q., Jose Barba-Montoya J., Huuki L.A., Durnan M.K., Kumar S. (2020) Relative Efficiencies of Simple and Complex Substitution Models in Estimating Divergence Times in Phylogenomics. Mol. Biol. Evol. 37(6):1819–1831

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

Yang, Z. 1994a. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.

Yang, Z. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39:306–314.

Yang Z. 1995. A space-time process model for the evolution of DNA sequences. Genetics. 139(2):993–1005.

Yang Z. 2006. Computational Molecular Evolution. Oxford University Press.

김우철 2021. 수리통계학(개정판). 민영사.

영문초록

Title: Comparing nucleotide substitution models.

Abstract: Nucleotide substitution models play an important role in the evolutionary analysis using maximum likelihood and Bayesian methods. There are 203 rate matrices from Jukes-Cantor model to General Time Reversible model. Furthermore, combination of these models with rate heterogeneity among sites and invariable site models drastically increase the number of applicable model candidates. In this article, we overview some nucleotide substitution models, describe a basic idea of information criteria for selecting optimal models, and explain how AIC and BIC are applied in the data analysis by using IQ-TREE program.

Authors: Tae-Kun Seo [1,$]

Affiliation: [1] Division of Life Sciences, Korea Polar Research Institute, 26 Songdomirae-ro, Yeonsu-gu, Incheon 21990,Republic of Korea

Corresponding author: [$] seo.taekun@gmail.com