염기서열의 진화적 거리 추정(Ver.25.01.04)

극지연구소

서태건(seo.taekun@gmail.com)

(한국진화학회 겨울학교용 자료)

표 1은 길이가 20 염기인 두개의 염기 서열을 나타낸다. 이 두개의 염기서열간의 진화적 거리를 추정해보자. 염기서열의 진화적 거리는 통상 '사이트당 염기치환 수'로 정의된다. 20개의 사이트중에 4개의 사이트에서 염기치환이 관찰되었으므로 단순 비율을 계산하면, 4/20=0.2 치환/사이트가 되며 이를 p-distance라고 한다. 하지만, 실제 일어났으나 관찰이 안 된 치환이 있을 수 있다. 예를 들어 첫번째 사이트는 A와 G사이에 여러번 치환이 발생했을 수 있다. 또한 5–20번째 사이트는 두 염기서열이 우연히 동일한 염기로 치환되어 변이가 없는 것으로 보일 수도 있다. 이렇게 눈에 보이지 않는 잠재적인 치환까지 고려한다면 실제 치환수는 0.2보다 큰 값이어야 한다. 이처럼 과소평가된 p-distance를 보정하기 위한 적절한 방법이 필요하고 이를 위해 염기치환의 통계모형을 가정하게 된다. 염기치환의 통계모형은 이전 논문 (서태건 2022)을 참조하기 바란다. 여기에서는 가장 간단한 JC 모형(Jukes and Cantor 1969)을 이용한 진화적 거리 추정에 대해 알아보자.

진화적 거리가 t일때 JC 모형하에서 염기 치환 확률은 다음과 같이 유도할 수 있다(Yang 2014).

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-4t/3} & \text{if } i = j \\ \frac{1}{4} - \frac{1}{4}e^{-4t/3} & \text{if } i \neq j \end{cases}, \tag{1}$$

또한, 길이가 n인 두 염기서열에서 염기가 다른 사이트의 수가 k일때 가능도(likelihood) 함수는 다음과 같이 정의할 수 있다

$$L(t) := \left\{P_{ij}(t)\right\}^k \left\{P_{ii}(t)\right\}^{n-k} \tag{2}$$

이는 이항분포를 이용한 것이고 $n!/(k!(n-k)!)$은 t와 무관하므로 생략되었다. 실제로 가능도 함수를 다룰 때는 식 (2)과 같은 가능도에 로그를 취한 로그-가능도를 이용해 각종 계산을 하게 된다. 가능도를 최대로 하는 t를 구하는 방법이 최대가능도 추정법(maximum likelihood estimation)이고 JC 모형의 경우 다음 식을 통해서 해석적으로 얻을 수 있다.

$$\frac{d}{dt}\log L(t) = 0$$

이를 통해 얻어진 진화적 거리의 최대 가능도 추정량은 다음과 같다.

$$\widehat{t} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\cdot\frac{k}{n}\right)$$

이 식을 통해 얻은 표 1의 진화적 거리의 최대가능도 추정값(maximum likelihood estimate; MLE)은 $\widehat{t} \approx 0.2326$이다. 즉, 실제 관찰된 비율(p-distance)인 0.2보다 큰 값이 된다. 그림 1은 p-distance와 JC-distance의 차이를 보여준다.

그림 2은 20개의 사이트 중 4개의 사이트에서 치환이 관찰된 경우 (왼쪽)와 200개의 사이트 중 40개의 사이트에서 치환이 관찰된 경우 (오른쪽) 식 (2)로부터 얻어지는 로그가능도 함수를 나타낸 것이다. 얼핏 보기에는 같은 것처럼 보이나 Y-축의 스케일이 다르다. 이를 주목해서 보면 오른쪽 그래프에서 MLE 주위에서 더 뾰족함을 알 수 있다. 즉, MLE는 똑같은 $\widehat{t} \approx 0.2326$로 얻어지지만, 염기서열의 길이가 긴만큼 데이타는 더 많은 정보를 함유하고 있고 이는 뾰족한 형태의 로그가능도 함수로, 그리고 폭이 좁아진 신뢰구간 형태로 나타난다.

Taxon name	Site index 1 2 12345678901234567890
T1	**** ACTGACTGACTGACTGACTG
T2	GACTACTGACTGACTGACTG

Table 1. 가상적인 염기서열 데이타. 20개의 사이트중 네개의 사이트 (*로 표시)에서 치환이 관찰되었다.

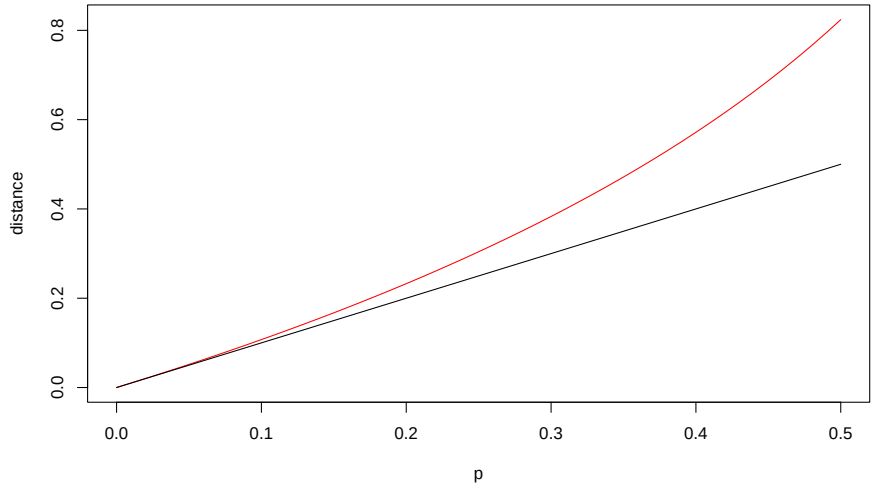

그림 1. p-distance(검은색)와 jc-distance(붉은색)의 관계

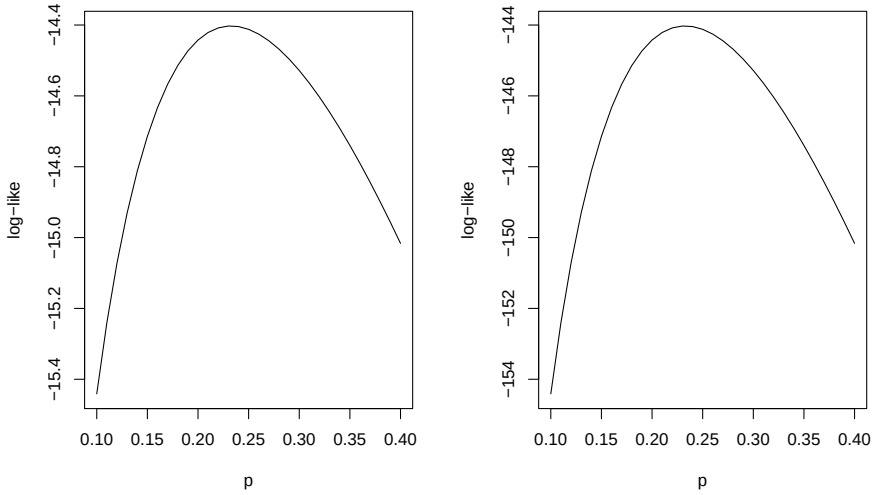

그림 2. log-likelihood curve 예시

참고문헌

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.

Yang Z. 2014. Molecular Evolution: A Statistical Approach. Oxford University Press.

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