계통생물학 기본 용어 해설(Ver.25.01.15)

극지연구소
서태건(seo.taekun@gmail.com)

(한국진화학회 겨울학교용 자료)

생물들간의 진화적인 유연관계를 그래프 형태로 나타낸 것을 계통수(**phylogeny**, **phylogenetic tree**)라고 한다. 특히 염기서열과 같은 분자데이터(molecular data)의 관계를 나타낼 때 분자계통수 (**molecular phylogeny**)라고 한다. 그래프의 vertex 부분을 계통생물학에서는 통상 node라고 부르며 edge를 branch 혹은 lineage라고 부른다. 조상의 위치를 나타낸 **rooted tree**, 조상의 위치를 규정하지 않고 계통관계만을 나타내는 **unrooted tree**가 있다. 말단 노드를 **tip**, **taxon**(복수형 **taxa**), **terminal node**, **OTU (operational taxonomic unit)**라고 하고 내부 노드를 **internal node**라고 한다. 조상에서 후손 방향으로 진화할때 가지가 둘로 나뉘는 노드를 **bifurcating** node, 셋 이상으로 나뉘는 노드를 **multifurcating** node라고 한다.

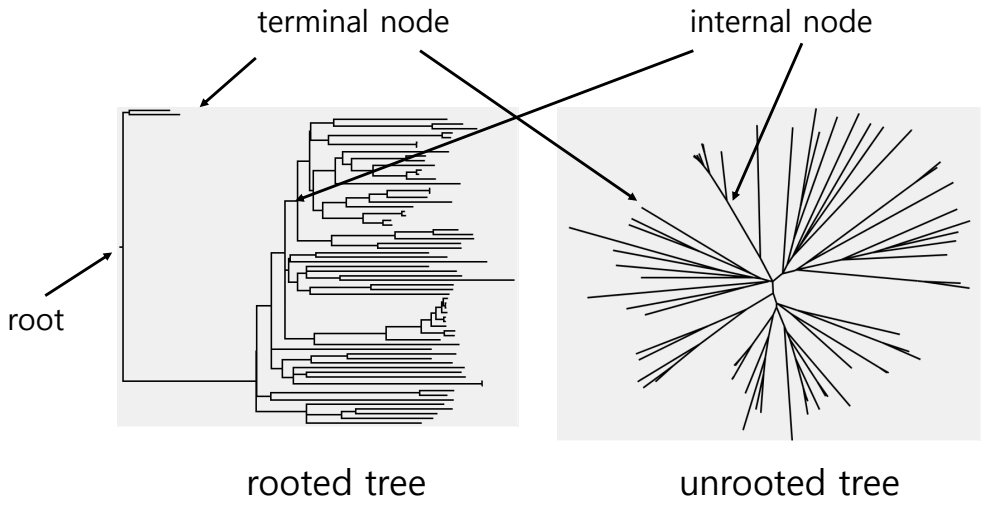

그림 1. 계통수 기본용어

주어진 taxa들의 공동조상중 가장 최근의 공동조상을 **most recent common ancestor (MRCA)**라고 한다(그림 1). 계통분석에서 주 관심사가 되는 taxa들의 모임을 **ingroup**이라 하고 ingroup의 root 위치를 파악하는데 사용되는 ingroup 이외의 taxa를 **outgroup**이라 한다. 주어진 taxa들의 MRCA의 모든 후손들이 주어진 taxa들과 동일할 때 주어진 taxa들을 **monophyletic group**이라 하고 동일하지 않을 때 **paraphyletic group**이라 한다. Paraphyletic group의 대표적인 예로 파충류(reptile)를 들 수 있다. 파충류의 MRCA의 후손에는 조류가 포함된다.

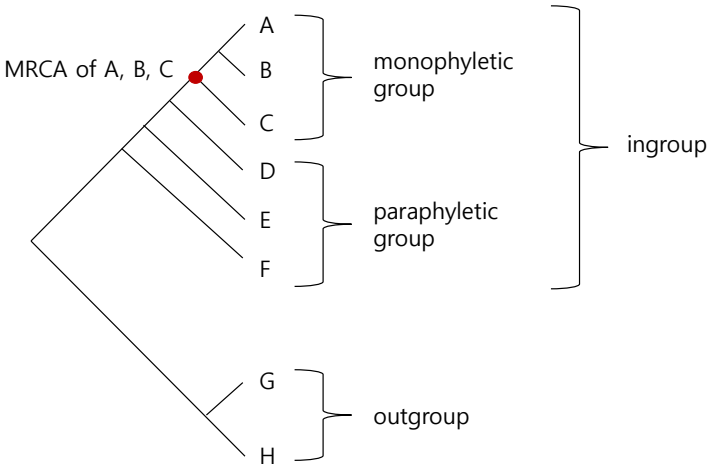

그림 2. 계통수 기본용어

종분화(speciation)로 생성된 상동 염기서열을 **ortholog**, 유전자중복(gene duplication)으로 생성된 상동 염기서열을 **paralog**라고 한다. 옳바른 계통관계 규명을 위해서는 반드시 orthologous sequences를 사용해야 한다(그림 3).

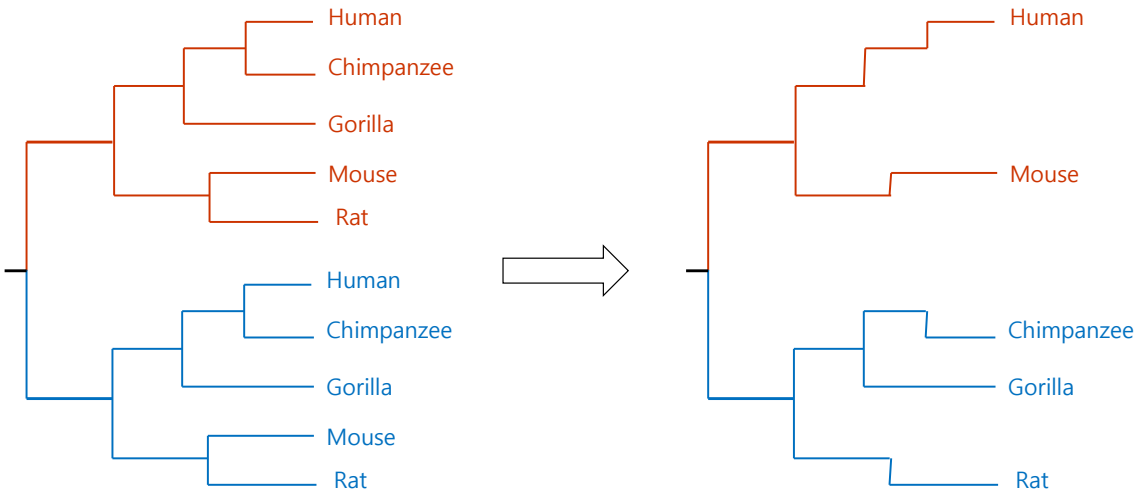

그림 3. Human-mouse 간의 종분화가 일어나기 전에 gene duplication으로 빨간색, 파란색 유전자가 생성된 가상적인 상황이다. 빨간색 염기서열끼리, 파란색 염기서열끼리는 orthologous sequences이고, 빨간색과 파란색 염기서열끼리는 paralogous sequences이다. Paralog를 이용하여 계통관계를 추정할 경우 잘못된 결과가 얻어질 수 있다.

Gene tree vs. Species tree

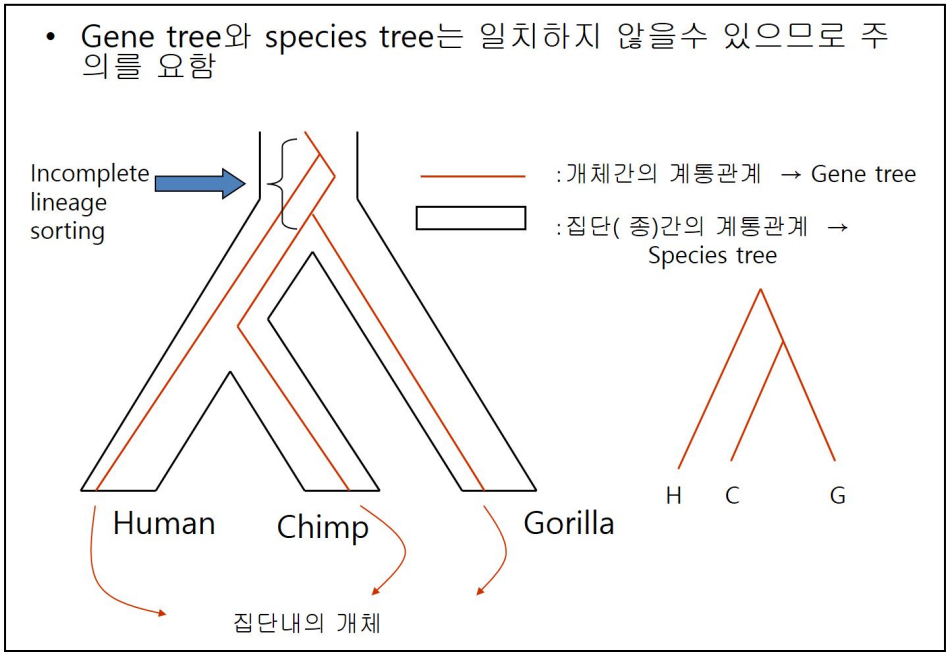

그림 4. Gene tree와 species tree의 비교

염기서열 데이터의 포맷

계통분석을 할 때 자주 사용되는 염기서열 데이터 포맷으로 **phylip포맷**과 **nexus포맷**이 있다.

```
4 1000
TaxonA   GCTGCTATATTTTATTTTTGGGATTTGATCCGGAATAATTGGATCTAGACTAAGAA...
TaxonB   GCTGCTATATTTTATTTTTGGGATTTGATCCGGAATAATTGGATCTAGACTAAGAA...
TaxonC   GCTGCTATATTTTATTTTTGGGATTTGATCCGGAATAATTGGATCTAGACTAAGAA...
TaxonD   GCTGCTATATTTTATTTTTGGGATTTGATCCGGAATAATTGGATCTAGACTAAGAA...
```

그림 5. phylip포맷

```
#NEXUS
Begin data;
       Dimensions ntax=4 nchar=1000;
       Format datatype=dna interleave=no gap=-;
       Matrix
TaxonA   GCTGCTATATTTTATTTTTGGGATTTGATCCGGAATAATTGGATCTAGACTAAGAA...
TaxonB   GCTGCTATATTTTATTTTTGGGATTTGATCCGGAATAATTGGATCTAGACTAAGAA...
TaxonC   GCTGCTATATTTTATTTTTGGGATTTGATCCGGAATAATTGGATCTAGACTAAGAA...
TaxonD   GCTGCTATATTTTATTTTTGGGATTTGATCCGGAATAATTGGATCTAGACTAAGAA...
;
End;
```

그림 6. nexus포맷

서열데이터의 포맷 변환에는 아래 사이트가 유용하다.

https://www.hiv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html

계통수 파일의 포맷

계통수 포맷으로 **newick포맷**과 **nexus포맷**이 있다(그림 7, 8).

```
((TaxonA, TaxonB),TaxonC, TaxonD);
```

그림 7. 계통수 newick포맷

```
#NEXUS
Begin taxa;
       Dimensions ntax=4;
         Taxlabels
           TaxonA
           TaxonB
           TaxonC
           TaxonD
;
End;
(이하 생략)
```

그림 8. nexus포맷

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.

$$\mathrm{E}[R] = \frac{\alpha}{\beta} \tag{1}$$

$$\mathrm{Var}[R] = \frac{\alpha}{\beta^2} \tag{2}$$

로그정규분포는 변수에 로그를 취한 값이 평균 μ, 분산 σ^2인 정규분포를 따른다고 가정한다(그림 10).

$$\log r \sim N(\mu, \sigma^2) \tag{3}$$

이 분포의 확률밀도함수는

$$f(r|\mu,\sigma^2) = \frac{1}{r\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\log r - \mu}{\sigma}\right)^2\right\}$$

이며 평균과 분산은

$$\mathrm{E}[R] = \exp\left\{\mu + \sigma^2/2\right\} \tag{4}$$

$$\mathrm{Var}[R] = \left\{\exp(\sigma^2) - 1\right\}\mathrm{E}[R]^2 \tag{5}$$

이다(Johnson et al. 1994).

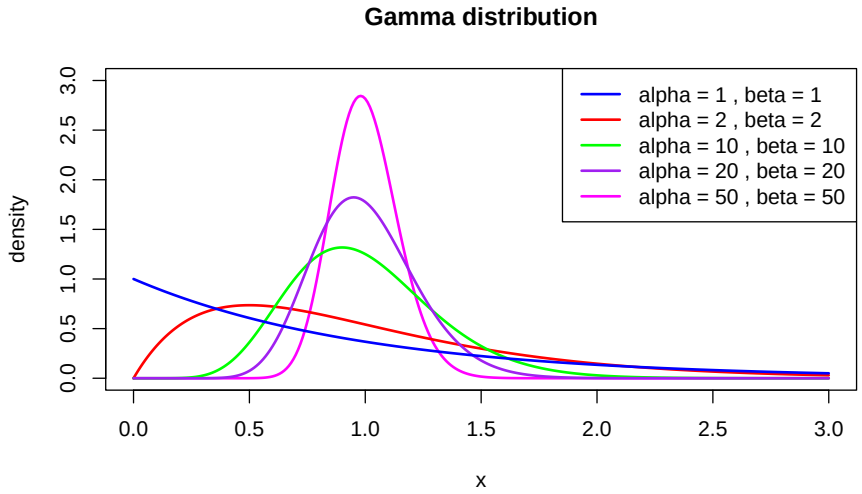

그림 9. 감마분포 확률밀도 함수

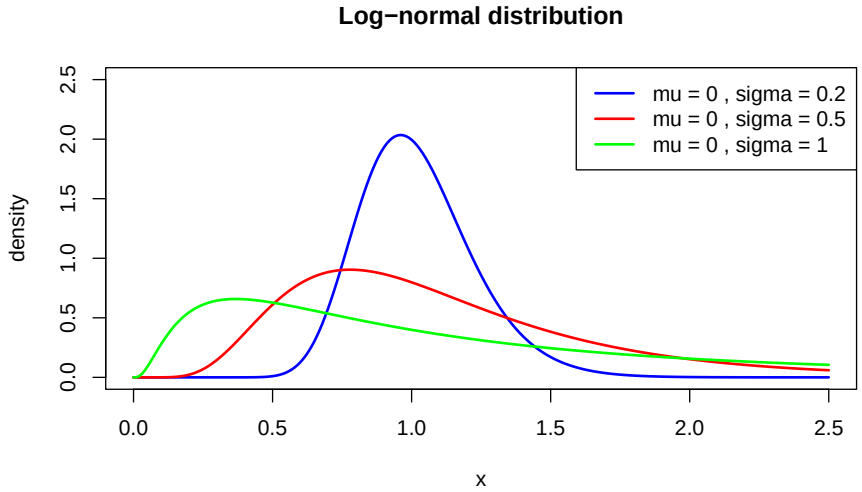

그림 10. 로그정규분포 확률밀도 함수

참고문헌

Johnson N.L., Kotz S. Balakrishnan N. 1994. Continuous univariate distribution, Vol 1. Wiley. Chapter 14.

김우철 2021. 수리통계학(개정판). 민영사.