

lec14.1 Hierarchical Clustering

层次聚类

- 不需要指定簇的数量 k :

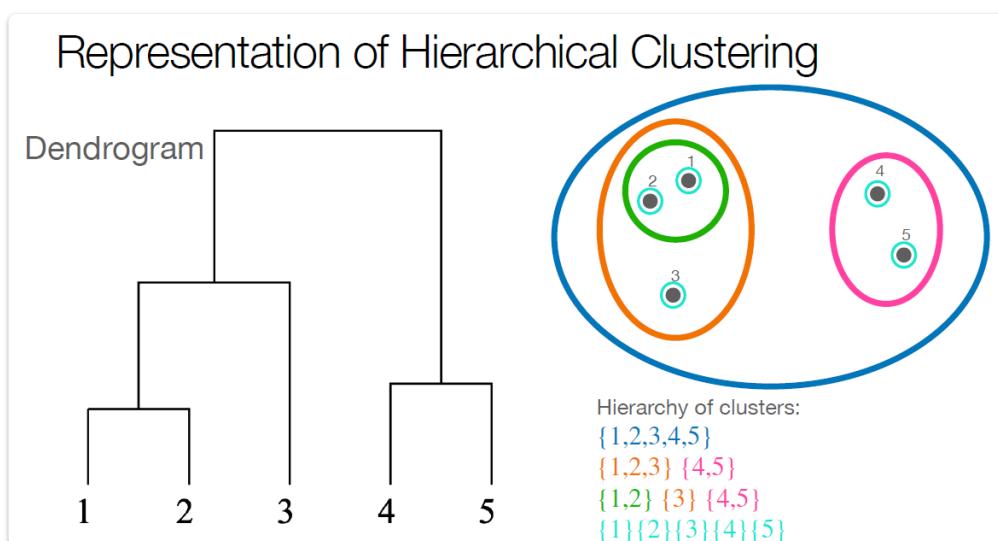
层次聚类不需要在开始时指定簇的数量 k 。这是一个显著的优点，因为在许多情况下，预先确定适当的簇数是非常困难的。因此层次聚类对于不确定最优簇数的数据集来说非常有用

- 同时指定所有粒度的聚类:

层次聚类能够同时生成不同粒度的聚类结果。这意味着它能够为数据提供不同层次的聚类，从细粒度的聚类（更多的小簇）到粗粒度的聚类（更少的大簇），并且所有这些结果都是在一次聚类过程中生成的

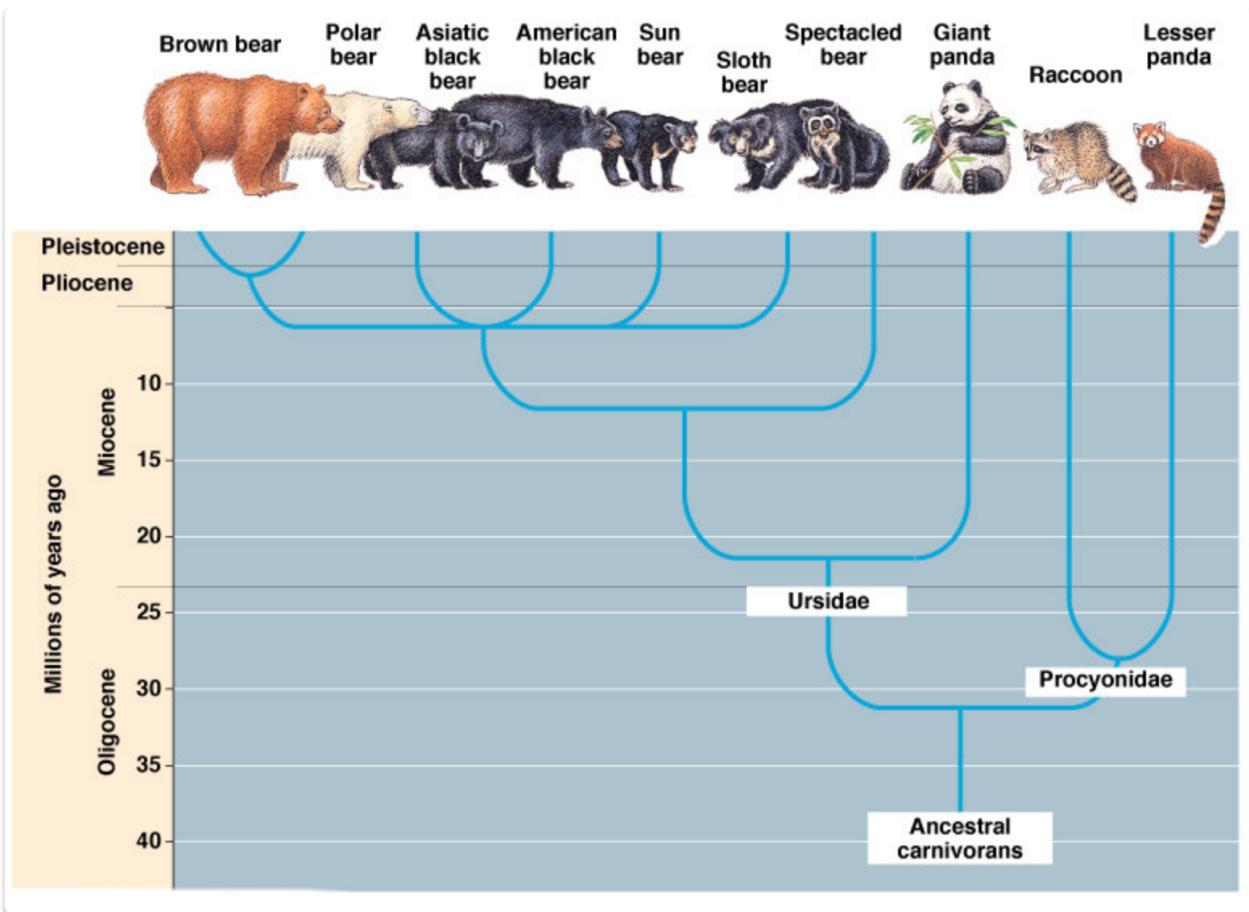
- 可以用于组织数据，构建用于可视化（抽象）目的的层次结构:

层次聚类能够将数据组织成一个包含多个层次的层次结构。这种层次结构非常有助于数据的可视化和理解，可以用于抽象和总结数据，帮助我们看到数据的总体结构和模式



示例 1：通过系统发育树（Phylogenetic Trees）绘制进化过程

- 生成 DNA 序列
- 计算所有序列之间的编辑距离
- 基于编辑距离计算 DNA 相似性
- 构建系统发育树（Phylogenetic Tree）
通过这种方法，可以使用系统发育树来描述不同物种或基因的进化关系



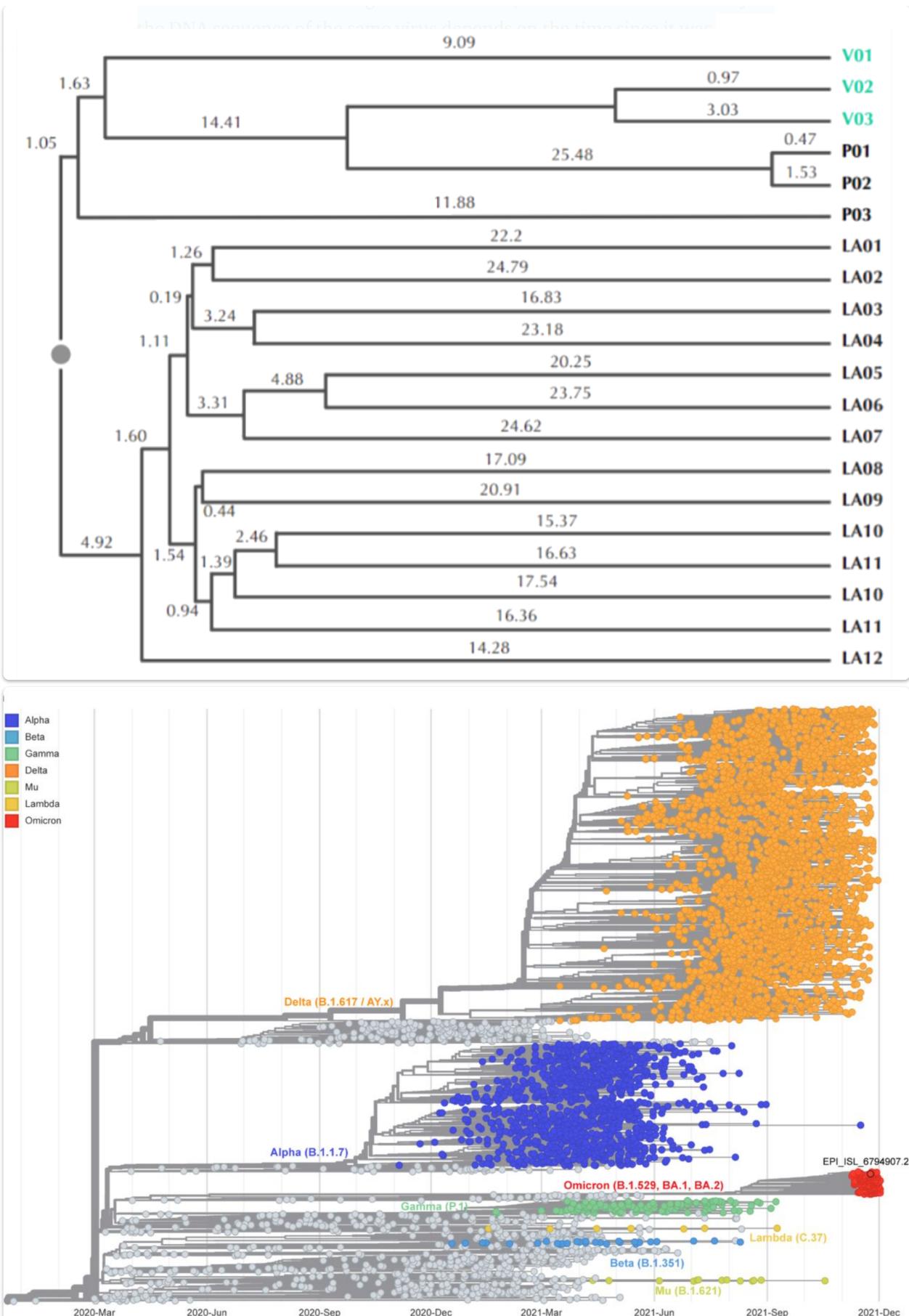
示例 2：通过系统发育树跟踪病毒传播

1. 高突变率的病毒（如 HIV）：

- 由于 HIV 等病毒具有高突变率，其 DNA 序列的相似性取决于传输时间。这可以用于追踪病毒传播路径

2. 法庭案例中的应用：

- 这种方法曾被用作法庭证据，其中受害者的 HIV 病毒株与被告的病毒株相比，发现更为相似



层次聚类的方法

将在后两个文档中介绍

凝聚层次聚类 (Agglomerative Clustering)

- **自底向上的方法:**
 - 从单个数据点开始（每个点各自一个簇），迭代地合并最相似的两个簇，直到所有数据点合并为一个簇

分裂层次聚类 (Divisive Clustering)

- **自顶向下的方法:**
 - 从一个大簇开始（包含所有数据点），逐步分裂簇，直到每个数据点单独成一个簇