5. Polymorphism, Selection and Phylogenetics

5.1 Population genetics

- Polymorphism in the genomes
  - Types of polymorphism
  - Measure of polymorphism
- Natural and artificial selection: the force shaping the genomes
  - Neutral tests
  - Domestication selection

5.2 Phylogenetics

- Distance method, MP and ML
Polymorphism in the genomes

- **Types of polymorphism**
  - Single nucleotide polymorphism (SNP)
  - Copy-number variation (CNV)
  - Insertion/deletion (indel)
  - Frame-shift
  - Presence and absence variation (PAV)

- **Measure of polymorphism**
  - $\pi$ (Tajima 1983): the average pairwise nucleotide diversity
  - $\theta$ (Watterson 1975): Watterson’s estimator
Selection

Selection skews the population frequency of genetic variants relative to NE expectations.

**Positive selection:**

- Positive selection is the process by which new advantageous genetic variants sweep a population.
- After a positive selection event, the frequency spectrum is skewed; there is an excess of rare polymorphisms as new variants accumulate after a selective sweep (Braverman et al. 1995).

**balancing selection**
Why selection is important?

- Loci or genes under positive selection or evolution present
  - the fixation of a favorable mutation
  - new function divergence

- Loci or genes under domestication selection or genetic improvement
  - important agronomy traits
Crop/animal breeding: domestication and subsequent genetic improvement

(Doebley et al. 2006)
Artificial selection

Artificial selection is a kind of recent positive selection (or adaptive /directional selection)

The most important evolutionary force shaping the crop genomes

Positive selection:
- Two major classes of methods are currently in use to detect positive selection: population and codon analysis methods
The effects of demography

The effects of domestication bottleneck on genetic diversity

Domestication bottleneck = domestication selection + demography effect

DNA diversity:
\( \pi \) (Tajima 1983)
\( \theta \) (Watterson 1975)

(Whitt and Gaut 2005)
Theoretical issues

- **Amount of diversity**
  - Reduction of nucleotide diversity

- **Frequency distribution of polymorphisms**
  - Selection skews the population frequency of genetic variants relative to neutral equilibrium model (NE) expectations
  - An excess of rare variants relative to NE expectations
  - Or, with recombination, an excess of high-frequency derived (non-ancestral) mutations

- **Degree of association between polymorphisms/linkage disequilibrium (LD)**
  - Selective sweep increase LD
In general, many positive selection events in the history of a species are not likely to be identified by population genetic approaches; instead, only a subset of relatively recent events are likely to be identified.

Domesticated plants are expected to be good models for detecting positive selection.

- Intense and recent selection and strong signature of selection
- Crop’s extant wild relative, pre-domestication population
Factors affecting the ability to detect positive selection

- Strength of selection
- The time since fixation of the beneficial mutation
- The amount of recombination between the selected and neutral sites
- The effects of demography
- Population structure
Key point:
– Artificial selection: a single gene and its linked regions
– Demography: all genes within the genome to some extent

The search for adaptive events becomes a search for genes that fulfill two criteria:
– Patterns of sequence diversity that deviate from the NE model
– Patterns of sequence diversity that are extreme for genes within that species
Natural and artificial selection

Natural test methods:
- Tajima’s $D$
- Fay and Wu’s $H$
- HKA (Hudson-Kreitman-Aguade)
- Fu and Li’s $D^*$ and $F^*$
- MK (McDonald-Kreitman)
- DAF
- Lewontin-Krackauer test
- Ka/Ks
- Empirical distribution (ranking)
- Coalescent simulation of domestication test
- EHH
- LRR
- iHS
- XP-EHH
A genome-wide association study (GWAS) is defined as any study of genetic variation across the entire genome that is designed to identify genetic associations with observable traits, or the presence or absence of a disease or condition.

Rapid advances in understanding the patterns of human genetic variation and maturing high-throughput, cost-effective methods for genotyping are providing powerful research tools for identifying genetic variants that contribute to health and disease.
**Basic Idea** A rather vague idea of a study design that involves genotyping cases and controls at a large number ($10^4 \rightarrow 10^6$) of SNP markers spread (in some unspecified way) throughout the genome. Look for associations between the genotypes at each locus and disease status.

```
0 1 1 1 1 1 0 1 0 2 1 2 2 0 1 0 0 0 1 1  Control
2 0 1 1 1 2 0 0 0 1 0 1 1 0 1 1 0 1 0 0  Control
2 0 1 2 2 0 1 2 1 0 0 1 1 1 0 1 0 0 1 1 1  Control
1 2 1 1 2 1 1 1 1 0 1 1 1 0 0 2 2 2 0 2  Control
1 1 2 1 0 1 2 1 1 1 1 2 1 2 1 2 1 2 1 1  Case
2 2 1 2 0 1 0 0 1 2 2 1 2 1 2 1 0 2 1  Case
0 1 1 0 0 2 1 0 0 2 1 1 1 2 1 1 2 0 1 0  Case
0 1 1 0 0 1 0 2 2 1 1 1 1 2 0 1 2 1 1 2  Case
```
5.2 Phylogenetic tree (系统发育树)

5.2.1 Introduction

5.2.2 Distance methods

5.2.3 Maximum parsimony (MP) and maximum likelihood (ML)
Why do a phylogenetic analysis?

- important for deciphering relationships in gene function and protein structure and function in different organisms
- helps to utilize genetic information of a model organism to analyze a second organism
- helps to sort out gene family relationships
- valuable tool for tracing the evolutionary history of genes
Evaluating sequence relationships

sequence A  ERKSIQDLFQSFTLFERRLLIEF
sequence B  ERLSISELIGSLRLYERRLIIIEY
sequence C  DRKSISDLIGSLRLALLIEF
sequence D  DRKDLISSSLRKALLIEW

1. Account for all column variations
   |A,B and C,D form similar groups based on col. 1
   |A,C,D based on col. 3
2. Count differences between sequences
   A,B 17/23 similar, 6/23 different
   C,D 21/23 similar, 2/23 different
What is a tree?

A graphical representation of the sequence similarities among a group of nucleic acid or protein sequences.

For example: number of differences between 3 sequences may be represented by ..
Phylogenetic tree (dendrogram)

Nodes: branching points
Branches: lines
Topology: branching pattern

Fig. 2.1 A simple tree and associated terms.
Branches can be rotated at a node, without changing the relationships among the OTU’s.
Rooted: unique path from root. Unrooted: degree of kinship, no evolutionary path.

Fig. 2.6 Rooted and unrooted trees for human (H), chimp (C), gorilla (G), orang-utan (O), and gibbon (B). The rooted tree (top) corresponds to the unrooted tree below.
Assumptions about rate of change in branches of tree

• may assume constant rate of change throughout tree - then branch length is proportional to no. of changes and we can easily root the tree using a simple algorithm OR
• may have variable rate and then root can be any branch

no ancestor sequence

ancestor sequence
Tree reconstruction as optimization

Number of possible trees is $2^S$. Considering branch lengths makes the problem harder. Almost as many unrooted trees.

Want the tree that maximizes some quality score. Score based on either

- Characters (e.g. sites in a sequence) directly, or
- Distances between character sets (e.g. alignment scores)

Another NP-hard optimization problem...
Number of possible phylogenetic trees

3 OTU’s: 1 unrooted tree
3 rooted trees

4 OTU’s: 3 unrooted trees
15 rooted trees.

实用分类单位 (operational taxonomic units, OTU)
<table>
<thead>
<tr>
<th>Number of OTUs</th>
<th>Number of rooted trees</th>
<th>Number of unrooted trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>105</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>954</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>10,395</td>
<td>954</td>
</tr>
<tr>
<td>8</td>
<td>135,135</td>
<td>10,395</td>
</tr>
<tr>
<td>9</td>
<td>2,027,025</td>
<td>135,135</td>
</tr>
<tr>
<td>10</td>
<td>34,459,425</td>
<td>2,027,025</td>
</tr>
<tr>
<td>11</td>
<td>654,729,075</td>
<td>34,459,425</td>
</tr>
<tr>
<td>12</td>
<td>13,749,310,575</td>
<td>654,729,075</td>
</tr>
<tr>
<td>13</td>
<td>316,234,143,225</td>
<td>13,749,310,575</td>
</tr>
<tr>
<td>14</td>
<td>7,905,853,580,625</td>
<td>316,234,143,225</td>
</tr>
<tr>
<td>15</td>
<td>213,458,046,676,875</td>
<td>7,905,853,580,625</td>
</tr>
<tr>
<td>16</td>
<td>6,190,283,353,629,375</td>
<td>213,458,046,676,875</td>
</tr>
<tr>
<td>17</td>
<td>191,898,783,962,510,625</td>
<td>6,190,283,353,629,375</td>
</tr>
<tr>
<td>18</td>
<td>6,332,659,870,762,850,625</td>
<td>191,898,783,962,510,625</td>
</tr>
<tr>
<td>19</td>
<td>221,643,095,476,699,771,875</td>
<td>6,332,659,870,762,850,625</td>
</tr>
<tr>
<td>20</td>
<td>8,200,794,532,637,891,559,375</td>
<td>221,643,095,476,699,771,875</td>
</tr>
</tbody>
</table>
Fig. 2.5 A phylogeny and the three basic kinds of tree used to depict that phylogeny. The cladogram represents relative recency of common ancestry; the additive tree depicts the amount of evolutionary change that has occurred along the different branches, and the ultrametric tree depicts times of divergence.
Newick (shorthand) format
- text based representation of relationships.

Fig. 2.4 A tree and its shorthand representation using nested parentheses.
Phylogenomics Tree

Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events

John E. Bowers, Brad A. Chapman, Junkang Rong & Andrew H. Paterson

Nature, 2003, VOL 422, 27
5.2.2 Distance methods

Basic idea:

- Employs the number of changes between each pair in a group of sequences to produce a phylogenetic tree of the group;
- For phylogenetic analysis, the distance score between two sequences is used. This score between two sequences is the number of mismatched positions in the alignment or the number of sequence positions that must be changed to generate the other sequence. Gaps may be ignored in these calculations or treated like substitutions.
- When a scoring or substitution matrix is used, the calculation is slightly more complicated, but the principle is the same.
**Theory of gene duplication**

**ORIGIN OF GENE FAMILIES**

- **One function**
  - Gene duplication
  - Speciation

- **Two functions**
  - Species A
  - Species B

**Orthologs**

**Paralogs**

(related by gene duplication)

- **Repeated duplications**

  - Species A
  - Species B

**Homoplasy:** Sequence similarity NOT due to shared ancestry, such as convergent or parallel evolution; horizontal transmission

All- by- all sequence analysis followed by single- linkage or maximal linkage clustering reveals orthologs and paralogs.

(rest are members of a paralogous family)
A
C
T
G
A
A
C
G
T
A
A
C
G
C
A
C
T
G
A→C→T
A
C→G
G
T→A
A
A→C→T
C
G
C
单一替换 (single substitution)
A
C
G
T
A
多重替换 (multiple substitutions)
A
C
→
A
C
T
A
同义替换 (coincidental substitutions)
G
T
→
A
A
A
→
T
趋同替换 (convergent substitutions)
C
G
C
→
T
→
C
反转替换 (back substitution)
A
C
T
C
祖先序列
遗传模型和序列距离

Judes-Cantor单参数模型 (1969)

**Figure 3.1** One-parameter model of nucleotide substitution. The rate of substitution in each direction is $\alpha$. 
Kimura两参数模型(Motoo Kimura, 木村资生, 1980): 转换 (transition, $\alpha$): 嘌呤间或嘧啶间的替换; 颠换 (transversion, $\beta$): 嘌呤和嘧啶间的替换。在大多数DNA片段中，转换出现的频率高于颠换。
在蛋白质编码基因中，仍为同义密码子的核苷酸替换称为同义或沉默替换（synonymous/silent substitution），而导致非同义密码子的替换，称为非同义或氨基酸更换替换（nonsynonymous/amino acid replacement substitution）。另外，导致形成终止密码子的突变，称为无义突变（nonsense mutation）。

Generally, genetic distance is thought of as related to the time and requires a genetic model specifying the processes such as mutation and drift causing the populations to diverge. Geometric distance: Euclidean distance

Judes and Cantor (1969) proposed the DNA sequence distance $K$ (earliest for amino acid sequence引入) calculation formula:

$$K = \frac{3}{4} \ln\left(\frac{3}{4q-1}\right) \approx 2\mu t$$

where $q$ is the probability of having the same base pair in the DNA sequence after $t$ generations, and $\mu$ is the base substitution frequency.
\[q_t = \frac{1}{4} + \frac{3}{4} \left(1 - \frac{8\mu}{3}\right)^t\]

Kimura在其两参数模型下证实，由于趋异变化，由转换造成差异( I 型变化)或由颠换造成差异( II 型变化)的碱基，随时间而变化。如果 \(k = a + 2\beta\) 是单位时间碱基替换的总频率，则

\[K = -\frac{1}{2} \ln[(1 - 2p_I - p_{II})\sqrt{1 - 2P_{II}}] \approx 2kt\]
DNA序列距离 $K$ 又可称为DNA序列间的分歧度，即序列间相异性的一个指标。蛋白质序列的分歧度分为两序列同义变化的分歧度($K_s$)和非同义变化的分歧度($K_A$)。根据Jukes-Cantor单参数模型和Kimura两参数模型等遗传模型，可以分别计算得到两序列的分歧度(或称为蛋白质序列间的距离)。
## A simple example

### A. Sequences

<table>
<thead>
<tr>
<th>Sequence</th>
<th>GAATTGCGGCAGATGGGCAAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>sequence A</td>
<td>ACGCGTTGGGGCGGATGGGCAAC</td>
</tr>
<tr>
<td>sequence B</td>
<td>ACGCGTTGGGGCGGACGGATAAT</td>
</tr>
<tr>
<td>sequence C</td>
<td>ACGCATTTGAAATGGATGATAAAT</td>
</tr>
<tr>
<td>sequence D</td>
<td>ACACATTGGAGTGATAATAAAT</td>
</tr>
</tbody>
</table>

### B. Distances between sequences, the number of steps required to change one sequence into the other.

<table>
<thead>
<tr>
<th>n_{AB}</th>
<th>n_{AC}</th>
<th>n_{AD}</th>
<th>n_{BC}</th>
<th>n_{BD}</th>
<th>n_{CD}</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>
C. Distance table

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>-</td>
<td>3</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>B</td>
<td>-</td>
<td>-</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>C</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

D. The assumed phylogenetic tree for the sequences A-D showing branch lengths. The sum of the branch lengths between any two sequences on the trees has the same value as the distance between the sequences.
Unweighted pair group method using arithmetic mean (平均连接聚类法).

A form of agglomerative clustering (successive pairing of closest nodes)

Method:
- Let initial organisms be leaves of the tree.
- Iteratively add a parent to the closest pair of existing nodes.
- Define the distance from an internal node to be the mean distance to its children.
实例：线粒体DNA序列

5个线粒体序列的差异核苷酸数(对角线下)和Jukes-Cantor距离(对角线上)

<table>
<thead>
<tr>
<th></th>
<th>人类(hu)</th>
<th>黑猩猩(ch)</th>
<th>大猩猩(go)</th>
<th>猩猩(or)</th>
<th>长臂猿(gi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>人类(hu)</td>
<td>–</td>
<td>0.015</td>
<td>0.045</td>
<td>0.143</td>
<td>0.198</td>
</tr>
<tr>
<td>黑猩猩(ch)</td>
<td>1</td>
<td>–</td>
<td>0.030</td>
<td>0.126</td>
<td>0.179</td>
</tr>
<tr>
<td>大猩猩(go)</td>
<td>3</td>
<td>2</td>
<td>–</td>
<td>0.092</td>
<td>0.179</td>
</tr>
<tr>
<td>猩猩(or)</td>
<td>9</td>
<td>8</td>
<td>6</td>
<td>–</td>
<td>0.179</td>
</tr>
<tr>
<td>长臂猿(gi)</td>
<td>12</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>–</td>
</tr>
</tbody>
</table>
最近的距离是人类和黑猩猩之间的，将它们合并为一个类。其它序列与这个新类之间的距离就是该序列到新类各成员间的平均距离：

\[
d_{(hu-ch),go} = \frac{1}{2}(d_{hu,go} + d_{ch,go}) = 0.037
\]

\[
d_{(hu-ch),or} = \frac{1}{2}(d_{hu,or} + d_{ch,or}) = 0.135
\]

\[
d_{(hu-ch),gi} = \frac{1}{2}(d_{hu,gi} + d_{ch,gi}) = 0.189
\]

<table>
<thead>
<tr>
<th></th>
<th>(hu-ch)</th>
<th>go</th>
<th>or</th>
<th>gi</th>
</tr>
</thead>
<tbody>
<tr>
<td>hu-ch</td>
<td>0.037</td>
<td>0.135</td>
<td>0.189</td>
<td></td>
</tr>
<tr>
<td>go</td>
<td></td>
<td>0.179</td>
<td></td>
<td></td>
</tr>
<tr>
<td>or</td>
<td></td>
<td></td>
<td>0.179</td>
<td></td>
</tr>
<tr>
<td>gi</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
其中人类—黑猩猩（**hu-ch**）与大猩猩（**go**）之间的距离最小。将它们合并为一类。新距离为:

\[
d_{(hu-ch-go), gi} = \frac{1}{3}(d_{hu,gi} + d_{ch,gi} + d_{go,gi}) = 0.185
\]

\[
d_{(hu-ch-go), or} = \frac{1}{3}(d_{hu,or} + d_{ch,or} + d_{gp,pr}) = 0.121
\]

<table>
<thead>
<tr>
<th></th>
<th>(hu-ch-go)</th>
<th>or</th>
<th>gi</th>
</tr>
</thead>
<tbody>
<tr>
<td>(hu-ch-go)</td>
<td>0.121</td>
<td>0.185</td>
<td></td>
</tr>
<tr>
<td>or</td>
<td>0.121</td>
<td>0.185</td>
<td>0.179</td>
</tr>
<tr>
<td>gi</td>
<td>0.121</td>
<td>0.185</td>
<td>0.179</td>
</tr>
</tbody>
</table>
平均连接聚类法系统树

0.092  0.060  0.019  0.007
Another agglomerative clustering method

Method:

- Let $A$ be active set, initially all leaves (organisms)
  Let $r(i)$ be the average distance from $i$ to all other leaves
- Define $D(i,j) = d(i,j) - (r(i) + r(j))$
- Pick $i,j$ with smallest $D(i,j)$. Create parent $k$ of $i$ and $j$.
- Set $d(i,k) = \frac{1}{2} [d(i,j) + r(i) - r(j)]$, and $d(j,k) = d(i,j) - d(i,k)$
- For all other $m \in A$, set $d(k,m) = \frac{1}{2} [d(i,m) + d(j,m) - d(i,j)]$
- Replace $i$ and $j$ in $A$ with $k$.
- Repeat until $|A| = 1$. 
邻接法的一般步骤

① 计算第$i$终端节点(即分类单位$i$)的净分歧度$r_i$

$$r_i = \sum_{k=1}^{N} d_{ik}$$

其中$N$为终端节点数，$d_{ik}$为节点$i$和节点$k$之间的距离，有$d_{ik} = d_{ki}$

② 计算并确定最小速率校正距离(rate-corrected distance) $M_{ij}$:

$$M_{ij} = d_{ij} - \frac{r_i + r_j}{N - 2}$$
③定义一个新节点\( u \)，\( u \)节点由节点\( i \)和\( j \)组合而成。节点\( u \)与节点\( i \)和\( j \)的距离为:

\[
S_{iu} = \frac{d_{ij}}{2} + \frac{r_i + r_j}{2(N-2)}
\]

\[
S_{ju} = d_{ij} - S_{iu}
\]

节点\( u \)与系统树其它节点\( k \)的距离为:

\[
d_{ku} = \frac{d_{ik} + d_{jk} - d_{ij}}{2}
\]

④从距离矩阵中删除列节点\( i \)和\( j \)的距离，\( N \)值(总节点数)减去1

⑤如果尚余2个以上终端节点，返回到步骤①继续计算，直至系统树完全建成。
同样以线粒体DNA为例

邻接法计算线粒体序列（图5.4）的距离d_{ij}（上对角线部分）和M_{ij}（下对角线部分）

<table>
<thead>
<tr>
<th></th>
<th>hu</th>
<th>ch</th>
<th>go</th>
<th>or</th>
<th>gi</th>
<th>净分歧度</th>
</tr>
</thead>
<tbody>
<tr>
<td>j=1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>r_i</td>
</tr>
<tr>
<td>j=2</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>j=3</td>
<td>0.015</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>j=4</td>
<td>0.045</td>
<td>0.143</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>j=5</td>
<td>0.198</td>
<td>0.179</td>
<td>0.350</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>hu</th>
<th>ch</th>
<th>go</th>
<th>or</th>
<th>gi</th>
<th>净分歧度</th>
</tr>
</thead>
<tbody>
<tr>
<td>i=1</td>
<td>0.235</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i=2</td>
<td></td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i=3</td>
<td>-0.204</td>
<td>-0.202</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i=4</td>
<td>-0.171</td>
<td>-0.171</td>
<td>-0.203</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i=5</td>
<td>-0.181</td>
<td>-0.183</td>
<td>-0.181</td>
<td>0.092</td>
<td>0.179</td>
<td>0.346</td>
</tr>
</tbody>
</table>

净分歧度计算如下：

$$r_i = \sum_{j=1}^{5} |d_{ij} - M_{ij}|$$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>净分歧度</th>
</tr>
</thead>
<tbody>
<tr>
<td>hu</td>
<td>i=1</td>
<td>0.000</td>
<td>0.015</td>
<td>0.045</td>
<td>0.143</td>
<td>0.198</td>
</tr>
<tr>
<td>ch</td>
<td>i=2</td>
<td>-0.235</td>
<td>0.000</td>
<td>0.030</td>
<td>0.126</td>
<td>0.179</td>
</tr>
<tr>
<td>go</td>
<td>i=3</td>
<td>-0.204</td>
<td>-0.202</td>
<td>0.000</td>
<td>0.092</td>
<td>0.179</td>
</tr>
<tr>
<td>or</td>
<td>i=4</td>
<td>-0.171</td>
<td>-0.171</td>
<td>-0.203</td>
<td>0.000</td>
<td>0.179</td>
</tr>
<tr>
<td>gi</td>
<td>i=5</td>
<td>-0.181</td>
<td>-0.183</td>
<td>-0.181</td>
<td>-0.246</td>
<td>0.000</td>
</tr>
</tbody>
</table>
第一步, or和gi之间的$M_{ij}$值最小, 则它们用节点1取代, 进入第2步, 则新节点(节点1)到这二个节点的距离为:

$$d_{gi, 节点1} = d_{or, gi} - d_{or, 节点1} = 0.122$$

$$d_{or, 节点1} = \frac{1}{2} d_{or, gi} + \frac{r_{or} - r_{gi}}{6} = 0.057$$
<table>
<thead>
<tr>
<th></th>
<th>hu</th>
<th>ch</th>
<th>go</th>
<th>节点1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>j=1</td>
<td>j=2</td>
<td>j=3</td>
<td>j=4</td>
</tr>
<tr>
<td>hu</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ch</td>
<td>i=2</td>
<td>-0.110</td>
<td>0.000</td>
<td>0.030</td>
</tr>
<tr>
<td>go</td>
<td>i=3</td>
<td>-0.086</td>
<td>-0.084</td>
<td>0.000</td>
</tr>
<tr>
<td>节点1</td>
<td>i=4</td>
<td>-0.085</td>
<td>-0.086</td>
<td>-0.110</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>go</th>
<th>节点1</th>
<th>节点2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>j=1</td>
<td>j=2</td>
<td>j=3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>go</td>
<td>i=1</td>
<td>0.000</td>
<td>0.046</td>
<td>0.030</td>
</tr>
<tr>
<td>节点1</td>
<td>i=2</td>
<td>-0.141</td>
<td>0.000</td>
<td>0.065</td>
</tr>
<tr>
<td>节点2</td>
<td>i=3</td>
<td>-0.141</td>
<td>-0.141</td>
<td>0.000</td>
</tr>
<tr>
<td>节点3</td>
<td>i=2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>go</th>
<th>节点3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>j=1</td>
<td>j=2</td>
</tr>
<tr>
<td>go</td>
<td>i=1</td>
<td>0.000</td>
</tr>
<tr>
<td>节点3</td>
<td>i=2</td>
<td></td>
</tr>
</tbody>
</table>
图 5.10  根据线粒体序列（图 5.4）构建邻接法无根系统树
Parsimony:
General scientific criterion for choosing among competing hypotheses that states that we should accept the hypothesis that explains the data most simply and efficiently.

Maximum parsimony method of phylogeny reconstruction:
The optimum reconstruction of ancestral character states is the one which requires the fewest mutations in the phylogenetic tree to account for contemporary character states.
First step: Identify all of the informative sites

**Invariant**: all OTU’s possess the same character state at the site. Any invariant site is **uninformative**.

For DNA sequences, an informative site is one where there are at least 2 different bases, and each different base occurs at least twice. A site with only one base that occurs in only one sequence is not informative, as it is due to a single base substitution along a branch leading to that sequence. This base substitution is compatible with any topology.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
</tbody>
</table>

**Invariant**: all OTU’s possess the same character state at the site. Any invariant site is **uninformative**.

For DNA sequences, an informative site is one where there are at least 2 different bases, and each different base occurs at least twice. A site with only one base that occurs in only one sequence is not informative, as it is due to a single base substitution along a branch leading to that sequence. This base substitution is compatible with any topology.
Two types of variable sites:

**Informative**: favors a subset of trees over other possible trees.

**Uninformative**: a character that contains no grouping information relevant to a cladistic problem (i.e. autapomorphies).
<table>
<thead>
<tr>
<th>Sequence</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>G</td>
<td>C</td>
<td>C</td>
<td>G</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
</tbody>
</table>

Uninformative: each tree 3 steps
2nd step: Calculate the minimum number of substitutions at each informative site

Informative: favors tree 1 over other 2 trees.
Final step: Sum the number of changes over all informative sites for each possible tree and choose the tree associated with the smallest number of changes.

Site 3

Tree I ((1,2),(3,4))

1 G ———> A

2 C ———> A A

Tree II ((1,3),(2,4))

1 G ———> A C

3 A ———> A

Tree III ((1,4),(2,3))

1 G ———> A C

4 A ———> A

Site 4

Tree I ((1,2),(3,4))

1 A ———> C

2 C ———> T

Tree II ((1,3),(2,4))

1 A ———> T

3 T ———> C

Tree III ((1,4),(2,3))

1 A ———> A C

4 G ———> C

Site 5

Tree I ((1,2),(3,4))

1 G ———> A

2 G ———> A

Tree II ((1,3),(2,4))

1 G ———> A

3 A ———> A

Tree III ((1,4),(2,3))

1 G ———> A

4 A ———> A

Site 9

Tree I ((1,2),(3,4))

1 T ———> A

2 T ———> T

Tree II ((1,3),(2,4))

1 A ———> A T

3 A ———> T

Tree III ((1,4),(2,3))

1 T ———> T

4 T ———> A

3 steps

3 steps

4 steps
Parsimony Search Methods

**Exhaustive search method:** searches all possible fully resolved topologies and guarantees that all of the minimum length cladograms will be found. (not a practical option, time consuming)

**Branch and bound methods:** begins with a cladogram. The length of starting cladogram is retained as an upper bound for use during subsequent cladogram construction. As soon as a length of part of the tree exceeds the upper bound, the cladogram is abandoned. If equal length, cladogram is saved as an optimal topology. If length is less, it is substituted for the original as the optimal upperbound. (good option for fewer than 20 taxa, time consuming)

**Heuristic methods:** approximate or “hill climbing technique”
Begin with a cladogram, add taxa and swap branches until a shorter length cladogram is found. Procedure can be replicated many times to increase chance of finding minimum length cladogram.
Different types of parsimony analyses:

**Unweighted parsimony**: all character state changes are given equal weight in the step matrix.

**Weighted parsimony**: different weights assigned to different character state changes.

**Transversion parsimony**: transitions are completely ignored in the analysis, only transversions are considered.
Consider Bayes rule:
\[ P(\text{model}|\text{data}) = \frac{P(\text{data}|\text{model}) \cdot P(\text{model})}{P(\text{data})} \]

\( P(\text{data}) \) is constant for any problem. If priors over model are uniform, the model that gives the highest likelihood for the data is the best.

If we have a parameterized statistical model of evolution, then we could find most likely set of parameters given a set of observations.

Now we have a numeric optimization problem to score even a single tree!
This method uses probability calculations to find a tree that best accounts for the variation in a set of sequences.

The method is similar to the maximum parsimony method in that the analysis is performed on each column of a multiple sequence alignment. All possible trees are considered. Hence, the method is only feasible for a small number of sequences.

The method can be used to explore relationships among more diverse sequences, conditions that are not well handled by maximum parsimony methods.
Given a tree topology, optimize score (likelihood) by setting lengths and parameters (e.g. evolution rates) for each edge.

Heuristic search through possible tree topologies.
- Greedy approaches (or beam search)
- Branch swapping
- Start with MP as approximate solution?
Statistical models of evolution

Various models of observable sequences, based on phylogenetic tree + randomness

- **IID models**: If characters (e.g. sites in an alignment) are independently and identically distributed, we can consider just a single site.

- **Markov models**: If probability of transitioning among character values depends only on current state, we have a Markov chain from ancestors to descendents
  - Discrete (transition matrix)
  - Continuous time (exponentiated instantaneous matrix * t)
More statistical models

Lots of 4 state (DNA) models
- Jukes-Cantor: all off-diagonal transition probabilities equal
- More freedom = more parameters (e.g. Kimura 2 param)

Cavender-Felsenstein (CF) model
- Simplest possible Markov model.
- Two states (0/1). Fixed root with specified state distribution
- Each edge specifies mutation probability <= 0.5
- Each site evolves i.i.d. according to a Markov process
- Equivalent to a model where each edge gets an expected number of changes that is Poisson distributed.
Mixture models
- Not a single Markov process, but a mixture (e.g. drawing at random from two distributions of characters)
- Non-iid: characters drawn from different classes of distributions (e.g. different positions in a codon)
- Classes with differently parameterized Poisson processes mixed based on prior distributions of parameters drawn from gamma distributions are widely used.

More complex models do not necessarily give better trees given limited data size!
Identifiability: a set of trees are identifiable under a model if the distributions of characters are disjoint for different tree topologies (with trivial exceptions, like 0 branch lengths)

Consistency: A method is consistent if it will recover the true tree with arbitrarily high probability given enough data (i.e. a long enough sequence)
Under iid Markovian models, trees are identifiable.

If a particular mixture is known and identical over all possible trees, then the trees are identifiable.

Unclear for more complex mixture models (like the Poisson / gamma distribution).

Identifiability
Maximum parsimony is not always statistically consistent, even in very simple models like CF. However, in simulation studies parsimony often performs well, so its consistency under biologically realistic conditions may be good. For many problems, maximum likelihood estimators are provably consistent.
MP is much faster, since scoring a tree is linear in tree size. Scoring an ML tree requires optimizing the edge weights (which may have multiple parameters per edge).

In some circumstances, MP and ML give identical answers

- When there are few mutations
- In a model where sites are independent, but not necessarily identically distributed, and there is no assumption of scaling across the sites
16S ribosomal trees

Pioneered by Woese
– Archaeal kingdom hypothesis came from 16S tree!

Exist in all organisms, highly conserved
Suitable for very broad phylogenies
Sequenced in tens of thousands of organisms
Not appropriate for close relatives...
Whole genome trees

- Characteristics over entire genomes
  - Presence/absence of homologous genes
  - Chromosomal rearrangements (inversions, translocations)

Good for distant taxa

 Mostly practical for microbes, since relatively few completely sequenced genomes of others yet.
Multiple hit correction

A site can change and then change back to where it was during the course of evolution.

Poisson correction: assume that changes after time t are independent of ones before t. \( D(i,j) = - \frac{1}{2} \log [1 - h(i,j)] \)

Other models are possible as well...
Ultrametric trees

In addition to the additive tree criterion, require that all distances to the root be equal. Idea is that differences in sequence should be proportional to the evolutionary time since divergence. Assumes that DNA evolves at a constant rate across lineages.
The Molecular Clock

Are mutation rates constant? If so, sequence distance = evolutionary time distance

Why not?

- Selective pressure. E.g. in coding regions, mutations that disrupt function will occur less often than neutral mutations
- Punctuated equilibrium. E.g. causative factors in mutation rates (e.g. UV light exposure) vary over evolutionary time
- Random changes in mutation rate (can model these!)

Apparently not true in general, but perhaps...

- Over relatively short intervals? In certain regions?
Comparing distance to state methods

Distance methods are suitable for continuous characters. For sequences, we need to define a distance function.

MP doesn't provide branch lengths.

Speed: distance < MP << ML
Tree merging

Consensus trees
- Multiple trees constructed from same data to evaluate robustness

Supertrees
- Trees constructed on overlapping sets of taxa and then combined to form larger trees without considering across-subtree interactions
Consensus trees

When many different trees have similar scores, look to see where they are identical.

Various consensus criteria

- Unanimous, majority, 2/3, score-weighted voting

Bootstrapping

- Resample characters (sequences) from inferred tree
- Compute new tree from resampled data
- Report on consensus between original and resampled data.
Supertrees

Compute trees from overlapping taxa

Only way to create very large trees

Each tree needs significant overlap (and consensus) with at least one other tree.

Warnow's DCM algorithm
When to use which...

- Continuous characters, lots of data, computational constraints: Neighbor Joining
- Discrete characters, moderate data, no homoplasy: maximum parsimony
- Discrete characters, limited sequence lengths, some homoplasy: maximum likelihood
- Many taxa: supertrees
- Complete genomes: whole genome methods