第二章
连锁分析和遗传图谱构建

王建康
中国农业科学院作物科学研究所
wangjiankang@caas.cn
http://www.isbreeding.net

连锁图谱在遗传研究中的重要性

- 重组率是指两个标记或基因座位之间发生奇数次交换的概率, 直观上反映了两个基因座位间的遗传距离. 重组率的估计是遗传研究中的经典问题. 连锁图谱是指基因或标记在染色体上的相对位置与遗传距离, 遗传距离一般以厘摩 (centi-Morgan, cM) 表示, 1%的重组率对应的遗传距离定义为1cM
- 世界上第一张遗传连锁图谱是利用5个形态特性标记构建的果蝇X染色体, 现在的连锁图谱一般都包含成百上千个标记. 建立在重组率估计之上的连锁图谱, 是开展遗传研究, 基因定位, 精细定位和克隆的前提

第二章 连锁分析和遗传图谱构建

§ 2.1 世代转移矩阵

§ 2.2 两个座位上各种基因型的理论频率

§ 2.3 两个标记/基因座位间重组率的估算

§ 2.4 不同遗传群体估计重组率的比较研究

§ 2.5 作图函数和遗传图谱构建

§ 2.6 随机交配群体的连锁分析

§ 2.1 世代转移矩阵

§ 2.1.1 世代转移矩阵的定义

§ 2.1.2 回交世代转移矩阵

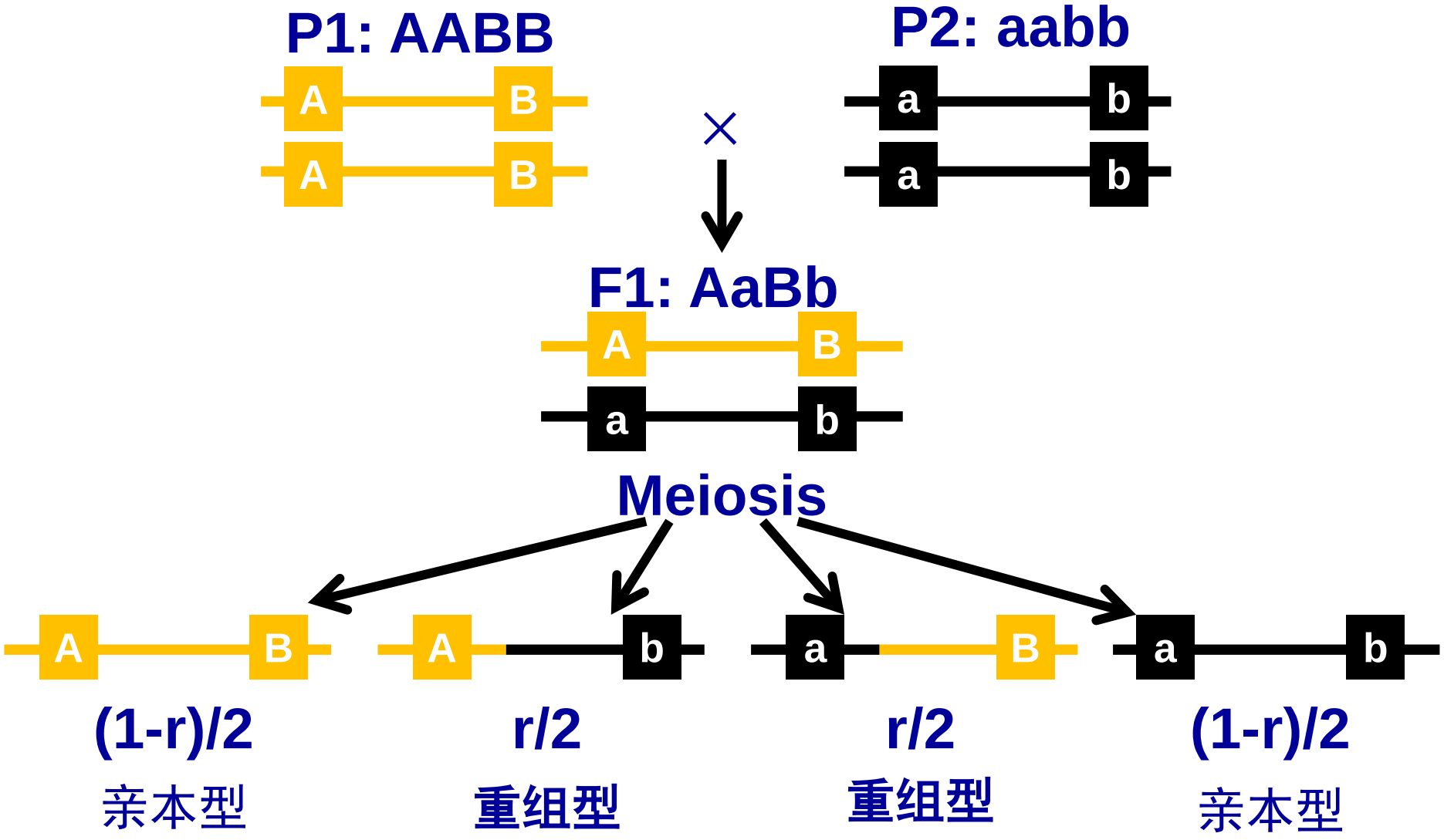
§ 2.1.2 自交世代转移矩阵

§ 2.1.3 加倍单倍体世代转移矩阵

§ 2.1.4 连续自交的世代转移矩阵

§ 2.1.5 基因型理论频率的矩阵表示

连锁和交换经典遗传定律



世代转移矩阵的定义

- 考虑两个座位上的等位基因A-a和B-b, 亲本的基因型为AABB和aabb, 后代中有九种可能的基因型. 给定重组率的大小, 每种基因型在特定群体中有特定的理论频率 (也称作期望频率). 基因型存在于群体中的理论频率是重组率估计的基础.
- 有些群体, 如回交一代, F1DH和F2, 是F_1群体通过适当的交配繁殖方式产生的. 由于F_1只有一种基因型, 因此, 容易计算经过一次回交, 加倍单倍体或一次自交之后, 产生出来的遗传群体中各种基因型的频率. 如果一个群体是由其他群体经过多次回交和自交而产生, 如BC1F2, BC2F2, F3和RIL等, 基因型理论频率的推算需借助转移矩阵.

两个座位上的10种基因型

- 双杂合基因型AB/ab和Ab/aB在重组率估计中是不能区分的. 虽然它们产生同样类似的配子, 但同一种配子的频率却是不同的. 在计算理论基因型频率时, 要区分对待. 在估计重组率时, 一般仅知道两种双杂型的观察值之和. 因此, 需要再把这两种基因型的频率进行合并
- 为推导不同群体中各种基因型的频率, 需考虑10种不同的基因型, 称为类型1, 类型2, …, 类型10

10种基因型的来历：从两个座位看

座位1	座位2		
	BB	Bb	bb
AA	AABB	AABb	AAbb
Aa	AaBB	AaBb	Aabb
aa	aaBB	aaBb	aabb

- 双纯型：AABB、AAbb、aaBB、aabb
- 单纯型（或单杂型）：AaBB、Aabb、AABb、aaBb
- 双杂型：AaBb (既可能是AB/ab、又可能是Ab/aB)

10种基因型的来历：从雌雄配子看

雌配子	雄配子			
	AB	**Ab**	**aB**	**ab**
AB	**AABB**	AABb	AaBB	**AB/ab**
Ab	AABb	**AAbb**	**Ab/aB**	Aabb
aB	AaBB	**Ab/aB**	**aaBB**	aaBb
ab	**AB/ab**	Aabb	aaBb	**aabb**

10种基因型的频率向量

- 两个基因座位上10种基因型的频率用行向量$\mathbf{f}^{(t)}$表示, 即,

$$\mathbf{f}^{(t)} = \left[f_{AABB}^{(t)} \quad f_{AABb}^{(t)} \quad f_{AAbb}^{(t)} \quad f_{AaBB}^{(t)} \quad f_{AB/ab}^{(t)} \quad f_{Ab/aB}^{(t)} \quad f_{Aabb}^{(t)} \quad f_{aaBB}^{(t)} \quad f_{aaBB}^{(t)} \quad f_{aaBB}^{(t)} \right]$$

- 如果把组成群体的不同个体视为从遗传群体中抽取的一组随机样本, 这组样本将服从频率为$\mathbf{f}^{(t)}$的多项分布. 10种基因型包含了一个随机样本所有可能的取值. 因此, 行向量$\mathbf{f}^{(t)}$的元素之和为1, 概率统计中称之为概率向量.

后续世代中10种基因型的频率向量

- 为表达方便, 把自交, 回交, 单倍体加倍统称为交配. 交配之后, 群体进入t+1世代, 交配后的基因型也有10种可能, 但它们的频率却发生了变化, 交配后群体的频率用行向量$\mathbf{f}^{(t+1)}$表示, 即,

$$\mathbf{f}^{(t+1)} = \left[f_{AABB}^{(t+1)} \quad f_{AABb}^{(t+1)} \quad f_{AAbb}^{(t+1)} \quad f_{AaBB}^{(t+1)} \quad f_{AB/ab}^{(t+1)} \quad f_{Ab/aB}^{(t+1)} \quad f_{Aabb}^{(t+1)} \quad f_{aaBB}^{(t+1)} \quad f_{aaBb}^{(t+1)} \quad f_{aabb}^{(t+1)} \right]$$

转移矩阵

- 世代$t+1$的基因型频率仅依赖于世代t的基因型频率, 而与世代t之前的基因型频率无关. 如果把不同世代群体中, 个体的基因型看作随机变量, 这些随机变量则形成一个马尔可夫链. **T**表示特定交配方式下, 一次交配的转移矩阵. 转移矩阵的每一行代表每种基因型产生的各种后代基因型的频率. 这个矩阵的每一行的元素之和为1, 概率统计中称为概率转移矩阵

转移矩阵的作用

- 一次交配发生后, 基因型的频率向量$\mathbf{f}^{(t+1)}$就能表示为交配前的频率向量$\mathbf{f}^{(t)}$与转移矩阵$\mathbf{T}$的乘积, 即,

$$\mathbf{f}^{(t+1)} = \mathbf{f}^{(t)}\mathbf{T}$$

- 因此, 如果知道了各种交配方式的转移矩阵, 就能得到一个群体交配后, 各种基因型的理论频率. 下面首先给出与P_1回交一代, 与P_2回交一代, 自交一代和加倍单倍体一代后的转移矩阵. 用$\mathbf{T}_{P1B}$表示与P_1回交一代的转移矩阵, $\mathbf{T}_{P2B}$表示与P_2回交一代的转移矩阵, $\mathbf{T}_S$表示自交一代的转移矩阵, $\mathbf{T}_D$表示加倍单倍体的转移矩阵.

AABB和AABb与亲本P1回交

- **基因型AABB与亲本P1 (AABB) 回交, 后代的基因型全部为类型1 (AABB),** 因此转移矩阵$\mathbf{T}_{P1B}$的第1行只有第1个元素为1, 其他为均0.
- **基因型AABb与亲本P1 (AABB) 回交, 后代的基因型只能为类型1 (AABB) 或类型2 (AABb), 两种基因型的频率均为1/2.** 因此, 转移矩阵$\mathbf{T}_{P1B}$第2行的前两个元素均为1/2, 其他为均0.

AAbb和AaBB与亲本P1回交

- **基因型AAbb与亲本P1 (AABB) 回交, 后代的基因型全部为类型2 (AABb).** 因此, 转移矩阵$\mathbf{T}_{P1B}$第3行的第2个元素为1, 其他为均0.
- **基因型AaBB与亲本P1 (AABB) 回交, 后代的基因型只能为类型1 (AABB) 或类型4 (AaBB), 两种基因型的频率均为1/2.** 因此, 转移矩阵$\mathbf{T}_{P1B}$第4行的第1和4两个元素均为1/2, 其他为均0.

AB/ab与亲本P1回交

- **基因型AB/ab与亲本P1 (AABB) 回交, 后代的基因型只能为类型1 (AABB), 类型2 (AABb), 类型4 (AaBB) 和类型5 (AB/ab) 四种可能. 类型1和5是非交换型配子AB和ab与P1的配子AB杂交产生的基因型, 频率均为 (1-*r*)/2. 类型2和4是交换型配子Ab和aB与P1的配子AB杂交产生的基因型, 频率均为 *r*/2.** 因此, 转移矩阵$\mathbf{T}_{P1B}$第5行的第1, 2, 4和5四个元素分别为 (1-*r*)/2, *r*/2, *r*/2 和 (1-*r*)/2 , 其他为均0.

Ab/aB与亲本P1回交

- **与基因型AB/ab类似, 基因型Ab/aB与亲本P1 (AABB) 回交, 后代的基因型只能为类型1 (AABB), 类型2 (AABb), 类型4 (AaBB) 和类型5 (AB/ab) 四种可能. 但是, 相对于Ab/aB来说, 配子AB和ab是交换型, Ab和aB是非交换型. 因此, 类型1和5是交换型配子AB和ab与P1的配子AB杂交产生的基因型, 频率均为 *r*/2. 类型2和4是非交换型配子Ab和aB与P1的配子AB杂交产生的基因型, 频率均为 (1-*r*)/2.** 因此, 转移矩阵$\mathbf{T}_{P1B}$第6行的第1, 2, 4和5四个元素分别为 *r*/2, (1-*r*)/2, (1-*r*)/2 和 *r*/2, 其他为均0.

Aabb和aaBB与亲本P1回交

- **基因型Aabb与亲本P1 (AABB) 回交, 后代的基因型只能为类型2 (AABb) 或类型5 (AB/ab), 两种基因型的频率均为1/2.** 因此, 转移矩阵$\mathbf{T}_{P1B}$第7行的第2和5两个元素均为1/2, 其他为均0.
- **基因型aaBB与亲本P1 (AABB) 回交, 后代的基因型全部为类型4 (AaBB).** 因此, 转移矩阵$\mathbf{T}_{P1B}$第8行的第4个元素为1, 其他为均0.

aaBb和aabb与亲本P1回交

- 基因型aaBb与亲本P1 (AABB) 回交, 后代的基因型只能为类型4 (AaBB) 或类型5 (AB/ab), 两种基因型的频率均为1/2. 因此, 转移矩阵$\mathbf{T}_{P1B}$第9行的第4和5两个元素均为1/2, 其他为均0.
- 基因型aabb与亲本P1 (AABB) 回交, 后代的基因型全部为类型5 (AB/ab). 因此, 转移矩阵$\mathbf{T}_{P1B}$的第10行只有第5个元素为1, 其他为均0.

与亲本P1回交的转移矩阵

$$
\mathbf{T}_{\mathrm{P1B}} = \begin{array}{c} \\ \text{AABB} \\ \text{AABb} \\ \text{Aabb} \\ \text{AaBB} \\ \text{AB/ab} \\ \text{Ab/aB} \\ \text{Aabb} \\ \text{aaBB} \\ \text{aaBb} \\ \text{aabb} \end{array}
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2}(1-r) & \frac{1}{2}r & 0 & \frac{1}{2}r & \frac{1}{2}(1-r) & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2}r & \frac{1}{2}(1-r) & 0 & \frac{1}{2}(1-r) & \frac{1}{2}r & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

与亲本P2回交的转移矩阵

$$
\mathbf{T}_{\mathrm{P2B}} = \begin{array}{c} \\ \mathrm{AABB} \\ \mathrm{AABb} \\ \mathrm{Aabb} \\ \mathrm{AaBB} \\ \mathrm{AB/ab} \\ \mathrm{Ab/aB} \\ \mathrm{Aabb} \\ \mathrm{aaBB} \\ \mathrm{aaBb} \\ \mathrm{aabb} \end{array} \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2}(1-r) & 0 & \frac{1}{2}r & 0 & \frac{1}{2}r & \frac{1}{2}(1-r) \\ 0 & 0 & 0 & 0 & \frac{1}{2}r & 0 & \frac{1}{2}(1-r) & 0 & \frac{1}{2}(1-r) & \frac{1}{2}r \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}
$$

自交世代转移矩阵：双纯型

- 无杂合座位, 即两个座位上的基因型都纯合. 纯合基因型的自交后代的基因型与亲代相同, 四种纯合基因型分别对应于类型1 (AABB), 类型3 (AAbb), 类型8 (aaBB) 和类型10 (aabb). 因此, 转移矩阵$\mathbf{T}_S$第1行的第1个因素为1, 其余因素为0; 第3行的第3个因素为1, 其余因素为0; 第8行的第8个因素为1, 其余因素为0; 第10行的第10个因素为1, 其余因素为0.

自交世代转移矩阵：单纯或单杂型

- 一个座位纯合, 一个座位杂合. 在杂合座位上, 自交后代的基因型按照1:2:1的比例分离, 即频率分别为1/4, 1/2和1/4. 以类型2 (AABb)为例, 自交后代的基因型为类型1 (AABB), 类型2 (AABb) 和类型3 (AAbb), 频率分别为 , 和 . 因此, 转移矩阵$\mathbf{T}_S$第2行的第1, 2和3个元素1/4, 1/2和1/4, 其余为0. 类型4 (AaBB), 类型7 (Aabb) 和类型9 (aaBb) 与类型2类似.

自交世代转移矩阵：双杂型AB/ab

雌配子	雄配子			
	AB, (1-*r*)/2	Ab, *r*/2	aB, *r*/2	ab, (1-*r*)/2
AB, (1-*r*)/2	AABB, $(1-r)^2/4$	AABb, $r(1-r)/4$	AaBB, $r(1-r)/4$	AB/ab, $(1-r)^2/4$
Ab, *r*/2	AABb, $r(1-r)/4$	AAbb, $r^2/4$	Ab/aB, $r^2/4$	Aabb, $r(1-r)/4$
aB, *r*/2	AaBB, $r(1-r)/4$	Ab/aB, $r^2/4$	aaBB, $r^2/4$	aaBb, $r(1-r)/4$
ab, (1-*r*)/2	AB/ab, $(1-r)^2/4$	Aabb, $r(1-r)/4$	aaBb, $r(1-r)/4$	aabb, $(1-r)^2/4$

自交世代转移矩阵：双杂型Ab/aB

雌配子	雄配子			
	AB, $r/2$	Ab, $(1-r)/2$	aB, $(1-r)/2$	ab, $r/2$
AB, $r/2$	AABB, $r^2/4$	AABb, $r(1-r)/4$	AaBB, $r(1-r)/4$	AB/ab, $r^2/4$
Ab, $(1-r)/2$	AABb, $r(1-r)/4$	AAbb, $(1-r)^2/4$	Ab/aB, $(1-r)^2/4$	Aabb, $r(1-r)/4$
aB, $(1-r)/2$	AaBB, $r(1-r)/4$	Ab/aB, $(1-r)^2/4$	aaBB, $(1-r)^2/4$	aaBb, $r(1-r)/4$
ab, $r/2$	AB/ab, $r^2/4$	Aabb, $r(1-r)/4$	aaBb, $r(1-r)/4$	aabb, $r^2/4$

自交世代转移矩阵

$$
\mathbf{T}_S = \begin{array}{c} \\ \text{AABB} \\ \text{AABb} \\ \text{Aabb} \\ \text{AaBB} \\ \text{AB/ab} \\ \text{Ab/aB} \\ \text{Aabb} \\ \text{aaBB} \\ \text{aaBb} \\ \text{aabb} \end{array}
\begin{array}{c}
\begin{array}{cccccccccc}
\text{AABB} & \text{AABb} & \text{Aabb} & \text{AaBB} & \text{AB/ab} & \text{Ab/aB} & \text{Aabb} & \text{aaBB} & \text{aaBb} & \text{aabb}
\end{array} \\
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & \frac{1}{2} & \frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 \\
\frac{1}{4}(1-r)^2 & \frac{1}{2}r(1-r) & \frac{1}{4}r^2 & \frac{1}{2}r(1-r) & \frac{1}{2}(1-r)^2 & \frac{1}{2}r^2 & \frac{1}{2}r(1-r) & \frac{1}{4}r^2 & \frac{1}{2}r(1-r) & \frac{1}{4}(1-r)^2 \\
\frac{1}{4}r^2 & \frac{1}{2}r(1-r) & \frac{1}{4}(1-r)^2 & \frac{1}{2}r(1-r) & \frac{1}{2}r^2 & \frac{1}{2}(1-r)^2 & \frac{1}{2}r(1-r) & \frac{1}{4}(1-r)^2 & \frac{1}{2}r(1-r) & \frac{1}{2}r^2 \\
0 & 0 & \frac{1}{4} & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\end{array}
$$

加倍单倍体世代转移矩阵

$$
\mathbf{T}_{\mathrm{D}} = \begin{array}{c}
\text{AABB} \\ \text{AABb} \\ \text{Aabb} \\ \text{AaBB} \\ \text{AB/ab} \\ \text{Ab/aB} \\ \text{Aabb} \\ \text{aaBB} \\ \text{aaBb} \\ \text{aabb}
\end{array}
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 \\
\frac{1}{2}(1-r) & 0 & \frac{1}{2}r & 0 & 0 & 0 & 0 & \frac{1}{2}r & 0 & \frac{1}{2}(1-r) \\
\frac{1}{2}r & 0 & \frac{1}{2}(1-r) & 0 & 0 & 0 & 0 & \frac{1}{2}(1-r) & 0 & \frac{1}{2}r \\
0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
$$

连续自交世代转移矩阵

$$
\mathbf{T}_{\mathrm{R}} = \begin{array}{c}
\text{AABB} \\ \text{AABb} \\ \text{Aabb} \\ \text{AaBB} \\ \text{AB/ab} \\ \text{Ab/aB} \\ \text{Aabb} \\ \text{aaBB} \\ \text{aaBb} \\ \text{aabb}
\end{array}
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 \\
\frac{1}{2}(1-R) & 0 & \frac{1}{2}R & 0 & 0 & 0 & 0 & \frac{1}{2}R & 0 & \frac{1}{2}(1-R) \\
\frac{1}{2}R & 0 & \frac{1}{2}(1-R) & 0 & 0 & 0 & 0 & \frac{1}{2}(1-R) & 0 & \frac{1}{2}R \\
0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
$$

基因型理论频率的矩阵表示

- 利用这五种转移矩阵, 20种双亲群体中, 各种基因型的理论频率都可以用杂种F_1的频率与转移矩阵的乘积来表示. F_1的基因型为AB/ab, 频率为1, 其余类型的频率为0, 即,

$$\mathbf{f}^{(0)} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- 利用这些关系就能推导出这些群体中基因型理论频率, 进而用于重组率的极大似然估计.

基因型理论频率与杂种F_1的频率$f^{(0)}$和转移矩阵的关系（1-10）

群体编号	群体名称	基因型理论频率
1	P1BC1F1	$f^{(0)}\times T_{P1B}$
2	P2BC1F1	$f^{(0)}\times T_{P2B}$
3	F1DH	$f^{(0)}\times T_{D}$
4	F1RIL	$f^{(0)}\times T_{R}$
5	P1BC1RIL	$f^{(0)}\times T_{P1B}\times T_{R}$
6	P2BC1RIL	$f^{(0)}\times T_{P2B}\times T_{R}$
7	F2	$f^{(0)}\times T_{S}$
8	F3	$f^{(0)}\times T_{S}\times T_{S}$
9	P1BC2F1	$f^{(0)}\times T_{P1B}\times T_{P1B}$
10	P2BC2F1	$f^{(0)}\times T_{P2B}\times T_{P2B}$

基因型理论频率与杂种F_1的频率$f^{(0)}$和转移矩阵的关系（11-20）

群体编号	群体名称	基因型理论频率
11	P1BC2RIL	$f^{(0)}\times T_{P1B}\times T_{P1B}\times T_R$
12	P2BC2RIL	$f^{(0)}\times T_{P2B}\times T_{P2B}\times T_R$
13	P1BC1F2	$f^{(0)}\times T_{P1B}\times T_S$
14	P2BC1F2	$f^{(0)}\times T_{P2B}\times T_S$
15	P1BC2F2	$f^{(0)}\times T_{P1B}\times T_{P1B}\times T_S$
16	P2BC2F2	$f^{(0)}\times T_{P2B}\times T_{P2B}\times T_S$
17	P1BC1DH	$f^{(0)}\times T_{P1B}\times T_D$
18	P2BC1DH	$f^{(0)}\times T_{P2B}\times T_D$
19	P1BC2DH	$f^{(0)}\times T_{P1B}\times T_{P1B}\times T_D$
20	P2BC2DH	$f^{(0)}\times T_{P2B}\times T_{P2B}\times T_D$

§2.3 两个标记/基因座位间重组率的估算

§2.3.1 DH群体中重组率的极大似然估计

§2.3.2 重组率极大似然估计的一般形式

§2.3.3 F2群体中一个共显性座位和一个显性座位间的重组率估计

§2.3.4 Newton迭代算法中初始值的选取

§2.3.5 F2群体中重组率估计的EM算法

§2.3.6 奇异分离对重组率估计的影响

DH群体中重组率的极大似然估计

- 利用杂种F_1植株上的配子培养的DH群体, 具有最简单的遗传结构, 所谓遗传结构就是一个遗传群体中的基因和基因型频率
- 首先以DH群体为例, 介绍重组率的极大似然估计的基本原理. 假定亲本P_1和P_2的标记基因型分别为AABB和aabb, 两个标记间的重组率为r
- 杂种F_1的基因型为AB/ab, 在F_1将产生基因型为AB, Ab, aB和ab的四种配子类型. AB和ab称为亲本配子型, Ab和aB称为交换配子型

DH群体中重组率的极大似然估计

- 根据遗传学的交换原理, 亲本型的频率等于1-r, 交换型的频率为r. F_1群体中, 每个等位基因的频率均为0.5, AB和ab出现的频率相同, Ab和aB出现的频率相同.
- 因此, 四种配子类型AB, Ab, aB和ab的频率分别为(1-r)/2, r/2, r/2和 (1-r)/2. 同时, 这些频率也就是DH群体中四种基因型AABB, AAbb, aaBB和aabb的频率.
- 用n_1和n_4表示亲本基因型的DH家系数, n_2和n_3表示重组基因型的DH家系数, 总的观测个体数为$n=n_1+n_2+n_3+n_4$.

DH群体中的期望基因型频率和观测值

基因型	AABB	AAbb	aaBB	aabb
基因型编码	(2, 2)	(2, 0)	(0, 2)	(0, 0)
期望或理论频率	$f_1=(1-r)/2$	$f_1=r/2$	$f_1=r/2$	$f_1=(1-r)/2$
观测样本量	n_1	n_2	n_3	n_4
Act8A 和 OP06 的样本量	64	8	7	61

重组率的极大似然估计和显著性检验

1、建立重组率*r*的似然函数

- 上中的观测次数n_1, n_2, n_3和n_4服从频率为f_1, f_2, f_3和f_4的多项分布. 因此似然函数为,

$$L(r) = \frac{n!}{n_1!n_2!n_3!n_4!}[\tfrac{1}{2}(1-r)]^{n_1}(\tfrac{1}{2}r)^{n_2}(\tfrac{1}{2}r)^{n_3}[\tfrac{1}{2}(1-r)]^{n_4}$$

$$= C(1-r)^{n_1+n_4} r^{n_2+n_3}$$

$$C = \frac{n!}{n_1!n_2!n_3!n_4!}(\tfrac{1}{2})^{n_1+n_2+n_3+n_4}$$ 与待估的重组率*r*无关

重组率的极大似然估计和显著性检验

2、建立对数似然函数

- 对似然函数直接求解有时很困难, 这时, 往往对似然函数求自然对数, 即,

$$\ln L(r) = \ln C + (n_1 + n_4)\ln(1-r) + (n_2 + n_3)\ln(r)$$

重组率的极大似然估计和显著性检验

3、对对数似然函数求导

- 求对数似然函数对重组率*r*的一阶和二阶导数

$$[\ln L(r)]' \hat{=} \frac{d \ln L}{dr} = -\frac{n_1 + n_4}{1-r} + \frac{n_2 + n_3}{r}$$

$$[\ln L(r)]'' = \frac{d^2 \ln L}{d^2 r} = -\frac{n_1 + n_4}{(1-r)^2} - \frac{n_2 + n_3}{r^2}$$

重组率的极大似然估计和显著性检验

4、求解重组率*r*的极大似然估计

- 令一阶导数等于0（得到的等式称为似然方程），求解得到重组率估计的极大似然估计为，

$$\hat{r} = \frac{n_2 + n_3}{n_1 + n_2 + n_3 + n_4} = \frac{n_2 + n_3}{n}$$

重组率的极大似然估计和显著性检验

5、求重组率估计值的方差

- 极大似然估计的方差一般从Fisher信息量获得, Fisher信息量*I*等于对数似然函数二阶导数的相反数, 一般可作为估计量方差的估计. 因此,

$$I = -[\ln L(r)]''|_{r=\hat{r}} = [-\frac{n_1+n_4}{(1-r)^2} - \frac{n_2+n_3}{r^2}]|_{r=\hat{r}} = \frac{n}{\hat{r}(1-\hat{r})}$$

$$V_{\hat{r}} = \frac{1}{I} = \frac{\hat{r}(1-\hat{r})}{n}$$

重组率的极大似然估计和显著性检验

6、重组率显著性的似然比检验

- 显著性检验的零假设是H_0: r=0.5, 即两个基因座位间不存在连锁关系. 备择假设是H_A: r<0.5, 即两个基因位点间存在连锁关系.
- 似然比统计量 (likelihood ratio test, LRT) 定义为备择假设和零假设两种情形下, 极大似然函数比值的自然对数的2倍.
- LRT统计量在大样本的情况下, 近似服从于卡方分布, 卡方分布的自由度等于两种假设下独立参数个数间的差异, 此时为1.

重组率的极大似然估计和显著性检验

6、重组率显著性的似然比检验

$$\max L(H_0) = L(r = 0.5) = C(\tfrac{1}{2})^n$$

$$\max L(H_a) = L(r = \hat{r}) = C(1-\hat{r})^{n_1+n_4}(\hat{r})^{n_2+n_3}$$

$$LRT = -2\ln\frac{\max L(H_0)}{\max L(H_A)} = -2\ln\frac{(\frac{1}{2})^n}{(1-\hat{r})^{n_1+n_4}(\hat{r})^{n_2+n_3}}$$

$$= 2(n_1+n_4)\ln[2(1-\hat{r})] + 2(n_2+n_3)\ln(2\hat{r}) \sim \chi^2(1)$$

大麦DH中标记Act8A和OP06之间的重组率

基因型	AABB	AAbb	aaBB	aabb
基因型编码	(2, 2)	(2, 0)	(0, 2)	(0, 0)
Act8A和OP06的样本量	n_1=64	n_2=8	n_3=7	n_4=61

$$\hat{r} = 0.1071 \qquad SE(\hat{r}) = 0.0261$$

$$LRT = 98.44 \quad (P=2.88\times10^{-23})$$

F_2和F_3群体中一个共显性座位A和一个显性座位B间6种基因型的理论频率

基因型	F2群体理论频率	F3群体理论频率
AAB$_*$	$(1-r^2)/4$	$(3/2-r-r^2+2r^3-r^4)/4$
AAbb	$r^2/4$	$r[1+r(1-r)^2]/4$
AaB$_*$	$(1-r+r^2)/2$	$(1-2r+4r^2-4r^3+2r^4)/4$
Aabb	$r(1-r)/2$	$r(1-r)(1-r+r^2)/2$
aaB$_*$	$r(2-r)/4$	$r[1+(1-r)(2-r+r^2)]/4$
aabb	$(1-r)^2/4$	$[2(1-r)+(1-r)^4+r^4]/8$

重组率MLE的Newton迭代算法

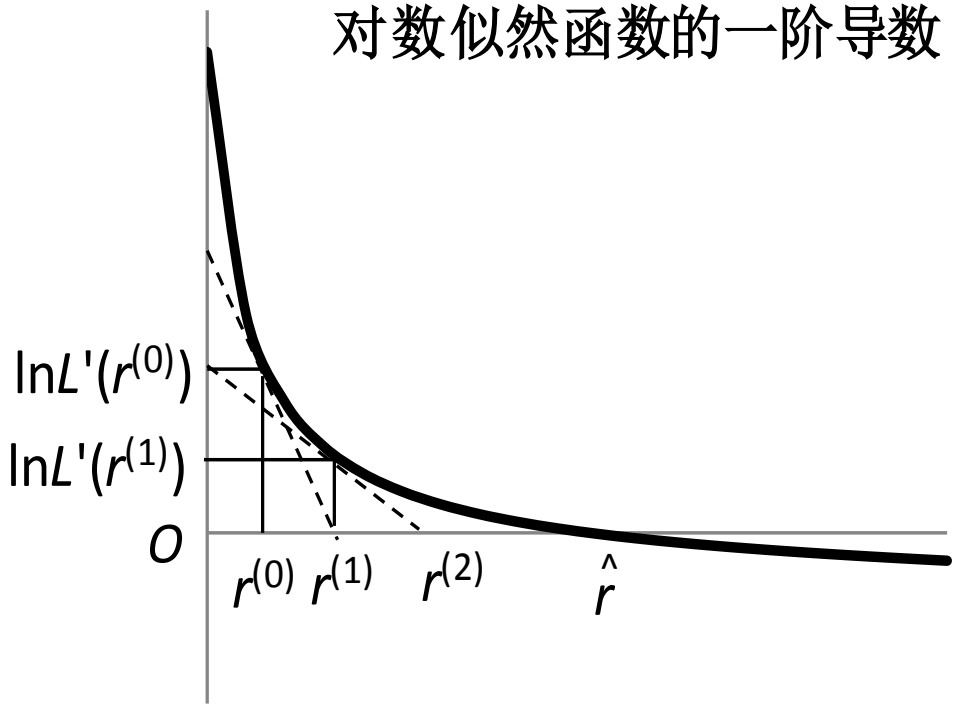
- 有些群体中, 令一阶导数 (公式2.3.10) 等于0 (称为似然方程), 可以直接计算出重组率, 如DH, RIL, BC1F1等. 还有些群体, 难以对似然方程直接求解, 这时需采用迭代算法.
- 当一个函数的一阶和二阶导数有明显的表达式时, Newton迭代算法 (也称Newton-Raphson算法) 是通用的求解方法. 首先选定一个重组率的起始值, 利用下面的公式计算一个新的重组率,

$$r^{(1)} = r^{(0)} - \frac{[\ln L(r)]'|_{r=r^{(0)}}}{[\ln L(r)]''|_{r=r^{(0)}}}$$

- 重复这一过程, 当两次迭代间重组率之差的绝对值小于事先设定的允许误差ε时, 则停止迭代. 并把最后一次的迭代值, 作为的重组率的极大似然估计值. 允许误差ε可取10^{-4}或更小的数字.

Newton迭代算法的几何原理

- Newton迭代算法的收敛性和收敛速度与初始值的选取有关, 当初始值接近真实值时, Newton迭代算法的收敛速度很快; 当初始值超过真实值太远时, Newton迭代算法可能不收敛. 下图给出Newton迭代算法的几何解释.



Newton迭代算法的初始值选择

- 对于小于极大似然估计$\hat{r}$的初始值$r^{(0)}$, Newton迭代能很快收敛到$\hat{r}$. 当要计算的$\hat{r}$很小时, 选取较大的正数如0.2作为初始值时, Newton迭代算法可能收敛不到 . 这时应该逐渐减小初始值, 如选取0.2的一半, 即0.1, 作为新的初始值进行迭代. 研究表明 (Sun et al., 2012), 选取较小的一个正数作为初始值, 如$r^{(0)}$=0.01或0.001, Newton迭代算法在绝大多情况下都能收敛到极大似然估计$\hat{r}$, 对于较大的$\hat{r}$, 只不过迭代次数多些而已.

F_2群体中重组率估计的EM算法

- 以两个共显性标记为例, 说明F_2群体中重组率估计的EM算法. 下表给出一个F_2群体中两个共显性标记九种基因型的观测值.

基因型	观测值	理论频率	重组单倍i型的频率
AABB	n_1=10	$f_1=(1-r)^2/4$	p_1=0
AABb	n_2=2	$f_2=r(1-r)/2$	p_2=0.5
AAbb	n_3=1	$f_3=r^2/4$	p_3=1
AaBB	n_4=1	$f_4=r(1-r)/2$	p_4=0.5
AaBb	n_5=21	$f_5=(1-2r+2r^2)/2$	$p_5=r^2/(1-2r+2r^2)$
Aabb	n_6=3	$f_6=r(1-r)/2$	p_6=0.5
aaBB	n_7=0	$f_7=r^2/4$	p_7=1
aaBb	n_8=1	$f_8=r(1-r)/2$	p_8=0.5
aabb	n_9=17	$f_9=(1-r)^2/4$	p_9=0

F_2群体中重组率估计的EM算法

- E-步骤: 根据重组率的初始值计算各种标记基因型属于重组型的期望概率. 给定初始重组率r, 一般可以让初始重组率r＝0.25. 根据前表最后一列, 计算各种标记基因型的重组频率p_i, i表示不同的标记基因型.
- M-步骤: 在E-步骤得到的各种基因型重组概率的基础上, 重新计算重组率的极大似然估计. 根据标记基因型属于重组基因型的概率重新计算重组率r,

$$r' = \frac{1}{n} \sum_{i=1,2,\cdots,9} n_i p_i$$

- 利用M-步骤计算出的重组率作为新的起始值, 重复上述过程, 直到指定的精度为止. 例如, 当两次迭代间重组率差值的绝对值小于10^{-4}, 则停止迭代.

三种重组率初始值下, EM算法六次迭代的结果

重组率初始值	迭代次数					
	1	2	3	4	5	6
0.01	0.0804	0.0832	0.0834	0.0834	0.0834	0.0834
0.25	0.1179	0.0869	0.0837	0.0835	0.0834	0.0834
0.5	0.2679	0.1246	0.0878	0.0838	0.0835	0.0834

EM迭代算法的初始值选择

- 表2.3.5给出0.01, 0.25和0.5三种初始值的迭代结果. 可以看出, EM算法经过六次迭代, 得到重组率的估计值为0.0834. 说明该算法有很快的收敛性, 收敛性和收敛到的极值点不依赖于初始值, 同时不用计算似然函数的一阶和二阶导数. F_2群体中, 当标记不是共显性时, 也能利用EM算法.
- 但对有些群体, 如F3, BC1F2, BC2F1, BC2F2等, 由于存在多次减数分裂过程, E-步骤重组型的期望频率难以计算. 因此, EM算法在有些群体中难以实现. 另外, 如果想要通过Fisher信息量获得重组率估计值的方差, 仍然要计算二阶导数.

奇异分离现象

- 奇异分离一般是由于不同基因型有不同的适合度 (用*w*表示) 造成的. 假定两种基因型AA和aa各100个个体, AA个体的繁殖成活率为1, aa个体为0.9. 那么, 我们就说aa相对于AA的适合度0.9. 所以, 适合度是指某基因型间能繁殖成活后代的相对能力, 其值在0和1之间.
- 当基因型的个数多于两个时, 繁殖成活率最高的基因型的适合度设为1, 其他基因型的适合度为各自的繁殖成活率与最高繁殖成活率的比值. 1-*w*在群体遗传学中称为选择系数, 用*s*表示.
- 奇异分离现象几乎存在于所有的遗传群体, 一个座位上的奇异分离会引起连锁标记或基因出现奇异, 从而导致基因型偏离孟德尔分离比, 基因型偏离孟德尔分离比会影响群体的遗传方差, 从而影响基因定位的功效.

奇异分离对重组率估计的影响

- 但是, 奇异分离对重组率估计的影响却很小, 在此我们以最简单的DH群体为例说明这一现象.

基因型	无奇异分离的理论频率	选择系数	选择后的频率
AABB	(1-*r*)/2	*s*	(1-*r*)(1-*s*)/2
AAbb	*r*/2	1	*r*/2
aaBB	*r*/2	*s*	*r*(1-*s*)/2
Aabb	(1-*r*)/2	1	(1-*r*)/2
总和	1		(2-*s*)/2

- 利用选择后频率得到的重组率：

$$r' = \frac{\frac{1}{2}r + \frac{1}{2}r(1-s)}{\frac{1}{2}(2-s)} = r$$

基因型aa相对于AA的选择系数*s*取不同值时重组率的估计值

标记Act8A	标记OP06	s=0	s=0.5	s=0.75	s=1
AA	BB	64	64	64	64
AA	bb	8	8	8	8
aa	BB	7	4	2	0
aa	bb	61	31	15	0
重组基因型个数		15	12	10	8
观测值之和		140	107	89	72
重组率估计值		0.1071	0.1122	0.1124	0.1111

§2.5 作图函数和遗传图谱构建

§2.5.1 遗传干涉和干涉系数

§2.5.2 作图函数

§2.5.3 标记分群算法

§2.5.4 标记排序算法

遗传连锁图谱

- 连锁图谱是指基因或标记在染色体上的相对位置与遗传距离. 通过连锁图谱可以大致了解基因和标记之间的相对位置, 了解哪些基因更靠近着丝粒, 哪些更靠近端粒等.
- 连锁图谱的构建是很多遗传研究的基础, 使用的标记越多, 遗传连锁图谱的分辨率就越高. 但是标记数目增加之后, 也给标记的分群和排序带来难度.
- 因此, 高密度连锁图谱的构建方法也一直是遗传学研究的一个热点问题.

三点分析

- 对于三个连锁的基因座M_1, M_2和M_3, 根据§2.3的内容, 可以估计三个成对座位间的重组率.
- 用r_{12}, r_{23}和r_{13}表示标记区间M_1-M_2, M_2-M_3和M_1-M_3上的重组率. 根据这三个重组率的估计值, 就能够判断这三个基因座在染色体上的相对位置. 例如, 如果r_{13}的估计值大于r_{12}和r_{23}, 三个基因座排列顺序可能为M_1-M_2-M_3.

无干涉时三个重组率的关系

- 假定连锁图上三个座位的排列顺序为M_1-M_2-M_3, 标记区间M_1-M_2和M_2-M_3上不存在干涉时, 即交换独立发生, 三个重组率的关系为,

$$(1-r_{13})=(1-r_{12})(1-r_{23})+r_{12}r_{23}$$

$$r_{13}=r_{12}(1-r_{23})+(1-r_{12})r_{23}=r_{12}+r_{23}-2r_{12}r_{23}$$

遗传干涉和干涉系数

- 对于完全干涉, 即区间M_1-M_2 (或M_2-M_3) 上的交换将完全阻止区间M_2-M_3 (或M_1-M_2) 上交换的发生, 这时,

$$r_{13} = r_{12} + r_{23}$$

- 一般一般情况下, 用δ表示干涉系数, 则有,

$$r_{13} = r_{12} + r_{23} - 2(1-\delta) r_{12} r_{23}$$

干涉系数的计算

- 当δ=0时, 等式 (2.5.3) 与 (2.5.1) 相同. 因此, δ=0表示两个区间上的交换是独立的.
- 当δ=1时, 等式 (2.5.3) 与 (2.5.2) 相同. 因此, δ=1表示两个区间上的交换是完全干涉. 这时, 一个区间上的交换完全阻止另外一个区间上的交换的发生.
- 如果三个连锁的基因座的顺序为M_1-M_2-M_3, 干涉系数可利用三个重组率进行估计, 即,

$$\delta = 1 - \frac{r_{12} + r_{23} - r_{13}}{2 r_{12} r_{23}}$$

大麦DH群体中1H染色体上14个标记的成对重组率估计值

标记	Act8A	OP06	aHor2	MWG943	ABG464	Dor3	iPgd2	cMWG733A
OP06	0.107							
aHor2	0.111	0.076						
MWG943	0.419	0.429	0.419					
ABG464	0.475	0.485	0.458	0.128				
Dor3	0.457	0.460	0.459	0.308	0.184			
iPgd2	0.438	0.468	0.419	0.321	0.214	0.036		
cMWG733A	0.451	0.482	0.448	0.370	0.283	0.101	0.070	
AtpbA	0.437	0.482	0.455	0.390	0.304	0.122	0.105	0.036
drun8	0.500	0.532	0.529	0.467	0.436	0.262	0.241	0.175
ABC261	0.483	0.507	0.511	0.441	0.410	0.236	0.222	0.155
ABG710B	0.493	0.525	0.530	0.496	0.475	0.317	0.294	0.227
Aga7	0.479	0.504	0.515	0.504	0.500	0.355	0.331	0.266
MWG912	0.464	0.489	0.481	0.504	0.529	0.400	0.376	0.317

干涉系数可正可负

- 以表2.5.1前三个标记为例, Act8A与OP06之间重组率的估计值为0.107, OP06与aHor2之间为0.076, Act8A与aHor2之间为0.111. 从这三个估计值可以看出标记OP06应该排序在Act8A与aHor2之间. 根据公式 (2.5.4) 得到干涉系数δ=-3.422, 说明区间Act8A - OP06与区间Act8A - aHor2可能存在负干涉, 即双交换的频率大于无干涉时的频率$r_{12}r_{23}$.
- 再以第5-7个标记为例, ABG464与Dor3之间重组率的估计值为0.184, Dor3与iPgd2之间为0.036, ABG464与iPgd2之间为0.214. 从这三个估计值可以看出标记Dor3应该排序在ABG464与iPgd2之间. 根据公式 (2.5.4) 得到干涉系数δ=0.617, 说明区间ABG464 - Dor3与区间 Dor3 - iPgd2可能存在正干涉, 即双交换的频率小于无干涉时的频率$r_{12}r_{23}$.

作图函数

- 由于遗传干涉的存在, 重组率一般不满足可加性. 而距离一般是可加的, 对于遗传图谱来说, 希望图谱上的距离也满足可加性.
- 设连锁图上有排列顺序为M_1-M_2-M_3的3个座位, M_1与M_3之间的图距用m_{13}表示, M_1与M_2之间的图距用m_{12}表示, M_2与M_3之间的图距用m_{23}表示. 根据距离的可加性,

$$m_{13} = m_{12} + m_{23}$$

作图函数

- 前面的公式中, m是两个位点间的遗传距离, 称为图距. 图矩的单位为摩尔根 (用M表示)或厘摩 (用cM表示), 1M=100cM.
- 图距m是交换率r的函数, 即 , 称f为作图函数. 交换率r=0.01的两个位点间的图距大约为1cM.
- 在连锁作图研究中, 有不同的作图函数, 可以把重组率转换为图距, 这里介绍常用的三种作图函数.

Morgan作图函数

- 由Morgan在1928年和Sturtevant (1931) 提出, 它将重组率的百分数作为图距, 即m=100×r, 单位为cM. 对于紧邻的两个区间, 可以采用求和的办法计算图距. 例如顺序排列的3个位点M_1-M_2-M_3, M_1-M_2间的重组率为0.02, 即图距为2cM; M_2-M_3间的重组率为0.01, 即图距为1cM.
- 根据Morgan作图函数, M_1-M_3间的图距为3cM. Morgan作图函数没有考虑大标记区间中存在多重交换的可能, 且假定干涉系数δ=1. 事实上, 一个较长的染色体区间上可能存在双交换甚至多次交换, 使得重组率不具有线性可加性的. 因此, Morgan作图函数不能应用于比较长的染色体区段.

Haldane作图函数

- 对于顺序排列的3个位点M_1-M_2-M_3, 在没有干涉的情况下, 即假定M_1-M_2间的交换和M_2-M_3间的交换独立发生, 并考虑到一个区间可以发生多次交换, Haldane (1919) 给出下面的作图函数,

$$m = f(r) = -\tfrac{1}{2}\ln(1-2r) \qquad r = \tfrac{1}{2}(1-e^{-2m})$$

- 其中, m的单位为M. 实际中, m常用cM为单位, 这时,

$$m = f(r) = -50\ln(1-2r) \qquad r = \tfrac{1}{2}(1-e^{-m/50})$$

Kosambi作图函数

- 考虑到遗传干涉的存在，提出干涉系数应是重组率的函数. 即, 染色体区间越短, 干涉的程度越大; 染色体区间越长, 干涉系数越小. 由此建立的作图函数为,

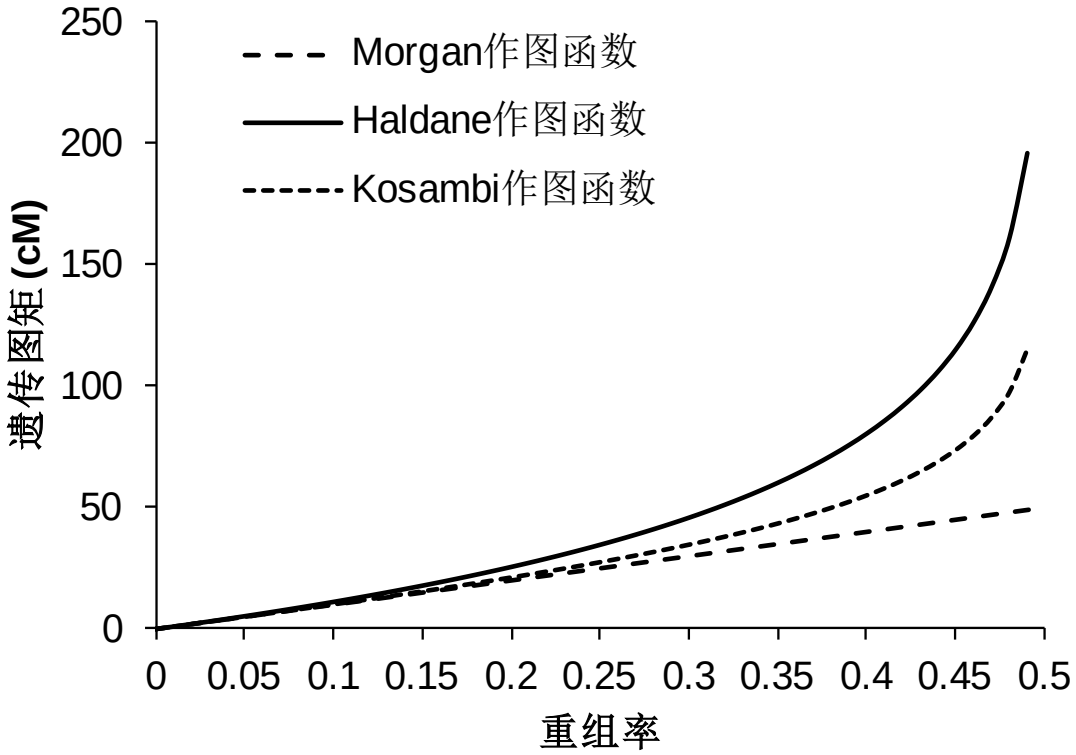
$$m = \frac{1}{4}\ln\frac{1+2r}{1-2r} \qquad r = \frac{1}{2}\frac{e^{4m}-1}{e^{4m}+1}$$

- 其中, m的单位为M. 实际中, m常用cM为单位, 这时,

$$m = 25\ln\frac{1+2r}{1-2r} \qquad r = \frac{1}{2}\frac{e^{m/25}-1}{e^{m/25}+1}$$

不同作图函数的区别和使用

- 上述三种作图函数, Haldane和Kosambi作图函数用得较多. 对于给定的重组率, Haldane作图函数给出的图距最大, Morgan函数给出的图距最小. 当重组率r<0.1时, 三种作图函数得到非常相近的图距.



遗传连锁图谱构建的步骤

- 分群
- 排序
- 调整
- 输出

QTL IciMapping中的分群算法

(i) a threshold of LOD score

(ii) a threshold of recombination frequency

(iii) a threshold of marker distance (cM)

(iv) anchor information

(v) a given number of group (added in version 4.1)

连锁图谱构建步骤一：分群

- 建立连锁图谱的第一步是将来自不同染色体的标记进行分群. 理想的情况是, 有多少条染色体, 就把标记分成多少个群, 一个标记群代表一条染色体.
- 分群时采用的标准可以是检测连锁的LOD统计量, 也可以是重组率的估计值, 还可以是根据重组率转换成的图距.
- 现以LOD分群标准为例, 说明分群的过程. 设定一个LOD临界值, n个待分群标记用集合的形式表示为$G_0=\{M_1, M_2, \ldots, M_n\}$. 分群后标记用$k$个非空集合$G_1, \ldots, G_k$表示. 分$k=0$和$k>0$两种情形讨论.

情形1: k=0, 即当前没有任何标记群

- 在G_0中, 确定一对优先分群标记M_{j1}和M_{j2} (即连锁最紧密的两个标记), 满足:

$$D_{j_1 j_2} = \mathrm{Max}\{LOD(\mathrm{M}_{i_1}, \mathrm{M}_{i_2}); i_1, i_2 = 1,2,\cdots,n, i_1 \neq i_2\}$$

- 如果D_{j1j2}大于指定的LOD临界值, 则生成第一个群G_1, 将M_{j1}和M_{j2}分入G_1中; 否则, 生成2个群G_1和G_2, 将M_{j1}和M_{j2}分别分入G_1和G_2中.
- 将M_{j1}和M_{j2}从G_0中删除.

情形2: k>0, 即已经产生一些标记群, 适用于有锚定标记的分群

- 在G_0中, 确定一个优先分群标记M_j, 方法如下: 对G_0中的任意$M_{j'}$, 计算

$$C_{j'} = \mathrm{Max}\{LOD(\mathrm{M}_{j'}, \mathrm{G}_{xy}); x = 1,2,\cdots k, y = 1,2,\cdots n_x\}$$

- 确定M_j优先分进的群G_i, 方法如下: 对任意$G_{i'}$, 计算

$$D_{i'} = \mathrm{Max}\{LOD(\mathrm{M}_{j}, \mathrm{G}_{i'y}); y = 1,2,\cdots n_{i'}\}$$

- 确定M_j是否应该分入G_i中, 方法如下: 如果D_i>指定的LOD临界值, 则把M_j分进G_i中; 否则, 生成1个新群G_{k+1}, 并把M_j分进新群G_{k+1}中.
- 将M_j从G_0中删除. 如果G_0=ø, 则分群完成; 否则重复上述过程.

分群结果的不确定性

- 最后得到的G_1, G_2, … 就是对这*n*个标记的分群.
- 如选择重组率或图距作为分群标准, 公式 (2.5.10) (2.5.11) 和 (2.5.12) 中的最大化改为最小化, 判断标准改为小于即可.

- 为分群结果的不确定性，软件中提供了根据给定分群个数的分群方法，这个方法类似聚类分析

QTL IciMapping中的排序算法

(i) **SER:** SERiation (Buetow and Chakravarti, 1987. Am J Hum Genet 41:180–188)

(ii) **RECORD:** REcombination Counting and ORDering (Van Os et al., 2005. Theor Appl Genet 112: 30–40)

(iii) **nnTwoOpt:** nearest neighbor was used for tour construction, and two-opt was used for tour improvement, similar to Travelling Salesman Problem (TSP) (Lin and Kernighan, 1973. Oper. Res. 21: 498–516.

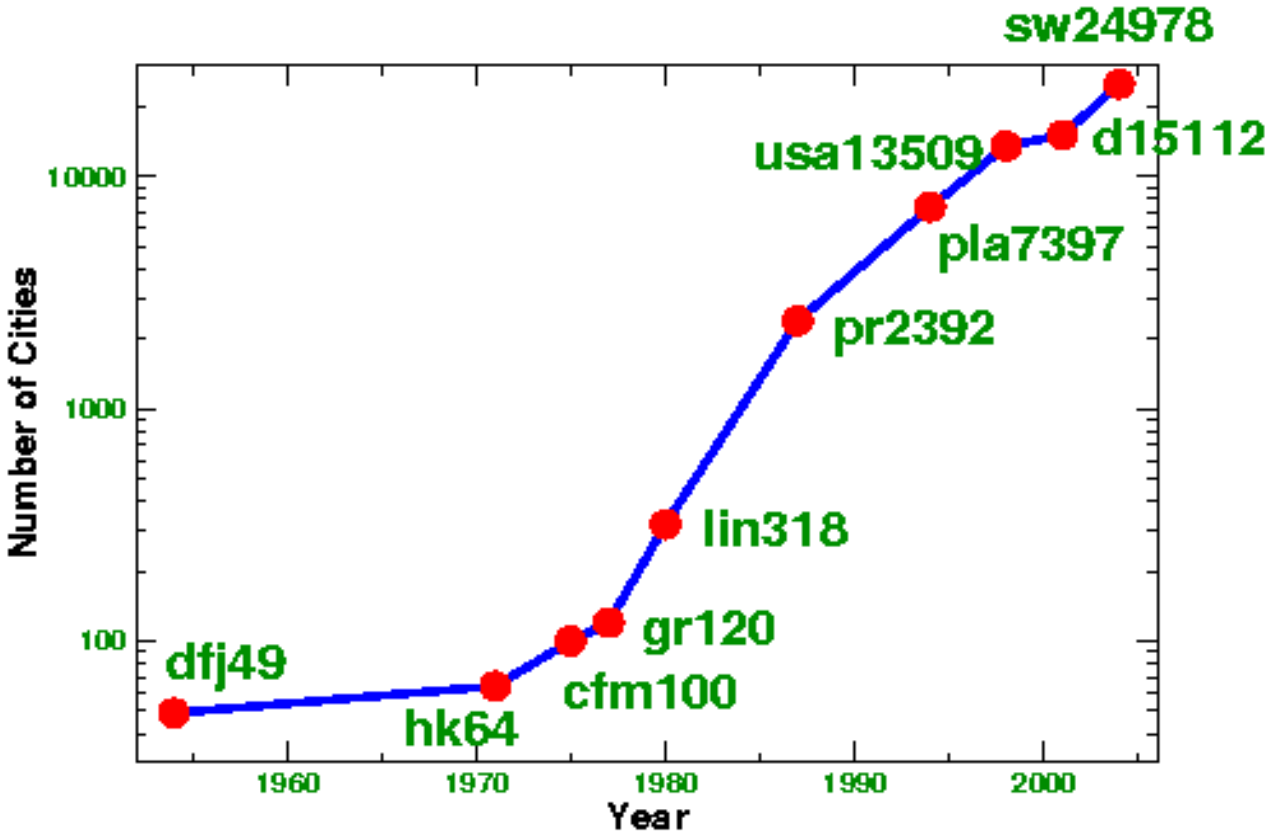
(iv) **By Input**: the order in input file will be used. Say we know the order from physical map for GBS markers

(v) **By Anchor Order:** Only order those markers with no anchor information. No effect on the anchor marker order.

组合数学中的旅行商问题 (traveling salesman problem, TSP)

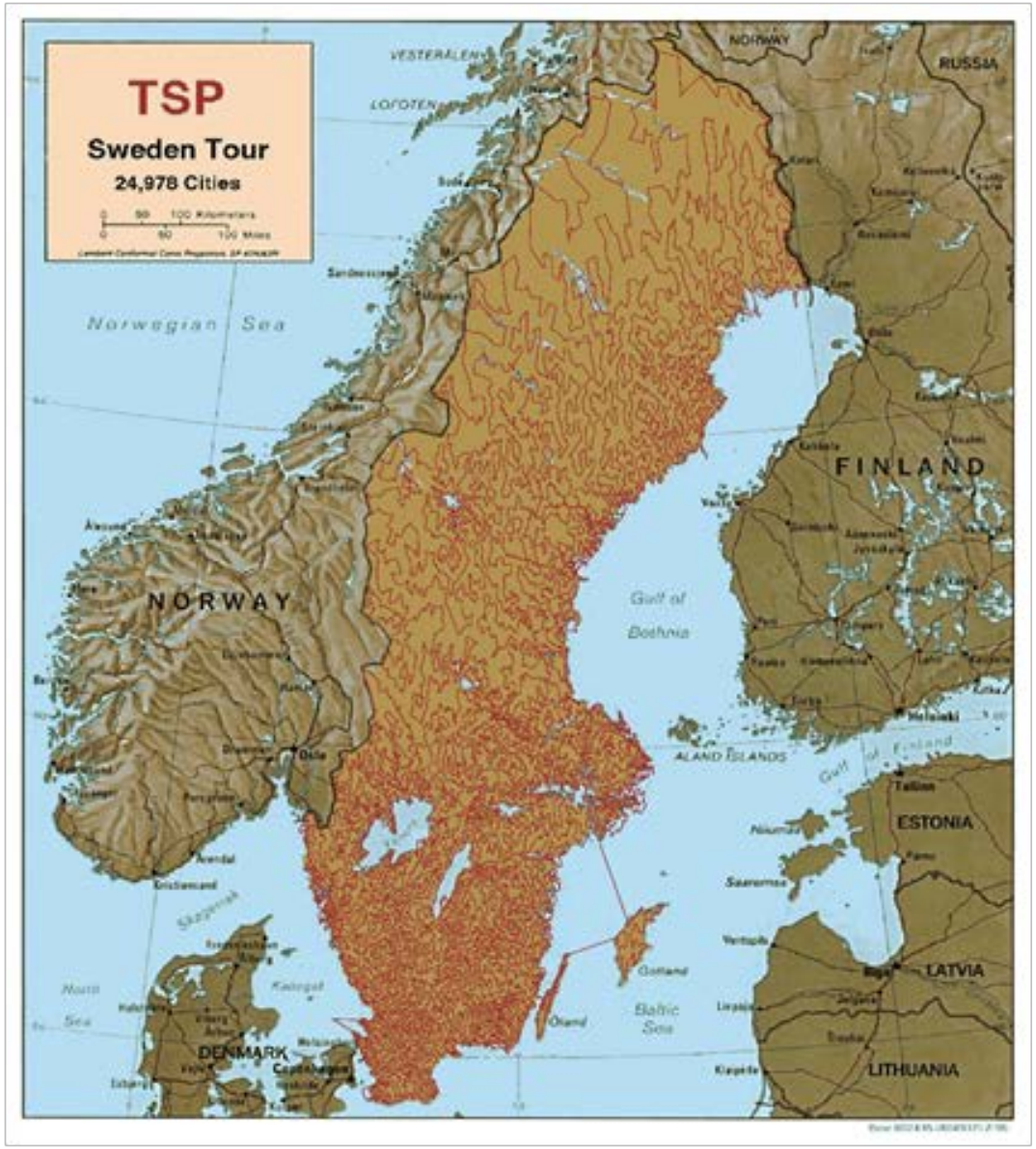
- 已知*n*个城市之间的直线距离, 有一个旅行商需要遍访这*n*个城市, 并且每个城市只能访问一次, 最后返回出发城市, 这就是组合数学中的旅行商问题 (traveling salesman problem, TSP).
- 求解TSP问题, 就是要选择一条路程最短的旅行路线. 数学上已经证明, TSP是运筹学, 图论和组合优化中的一个NP难题 (non-deterministic poly-nominal time hard, NP-hard). 当城市数较大时, 不存在全局最优解的精确算法, 所有的解都是近似最优解. 但可喜的是, 目前已研究出多种有效求解TSP问题的近似算法 (Lin and Kernighan, 1973; Laporte, 1992).

Finding the solutions



- TSP is represented by some letters plus the number of cities. For example, there are 24978 Sweden cities in TSP “sw24978”.

The solution for TSP “sw24978”



标记排序与TSP问题求解之间的相似性

- 连锁图谱构建过程中, 排序的目的是寻求图距最短的一个标记顺序. 当一个群中有*n*个标记时, 所有可能的排序有 种. 当*n*=50时, n!/2=1.52×10^{64}. 因此, 要比较所有可能的顺序几乎是不可能的.
- 高通量分子标记可以构建超高密度遗传连锁图谱, 但同时对连锁图谱构建算法提出巨大的挑战. 一些传统的方法如顺序排列法 (Buetow and Chakravarti, 1987),重组计数排序法 (van Os et al., 2005), 单向生长算法 (Tan and Fu, 2006) 等, 存在时间复杂度过高, 排序准确度差等问题.
- 连锁图谱构建与TSP问题求解之间存在极大的相似性. 成对标记间的重组率或图距可看作TSP问题中两两城市间的路程. 但两者之间又有一定区别, 遗传距离的估计受群体类型, 群体大小, 标记缺失等诸多因素的影响, 估计值有一定误差. 而TSP中的物理距离一般没有误差, 或者误差很小.

TSP问题求解步骤1：构造一个起始序列

- 构造算法也有很多, 这里介绍最近邻居算法, 也称为贪婪算法. 算法从距离最短的两个标记开始, 然后在待排标记中, 依次加入与已排顺序具有最短距离的标记. 实际计算中, 可以从任意一个标记开始, 构造出不同的顺序, 然后选择具有最短距离的一个, 作为起始序列.

最近邻居算法构造起始序列

- 假定群G中有*n*个标记, 用集合的形式表示为G={M_1, M_2, …, M_n}表示. 构造算法如下:

  (1.1) 对于G中的任一标记M_i, 将M_i作为起始序列的起始标记, 同时M_i也是序列的终止标记, 并将M_i从中G中删除, 删除后的标记群用G_0表示.

  (1.2) 在G_0中寻找与已有序列的终止标记具有最短距离的标记M_j, 作为新的终止标记, 同时并将M_j从中G_0中删除.

  (1.3) 如果G_0=ø, 则从 (1.1) 循环, 直到G中最后一个标记; 否则从 (1.2) 循环.

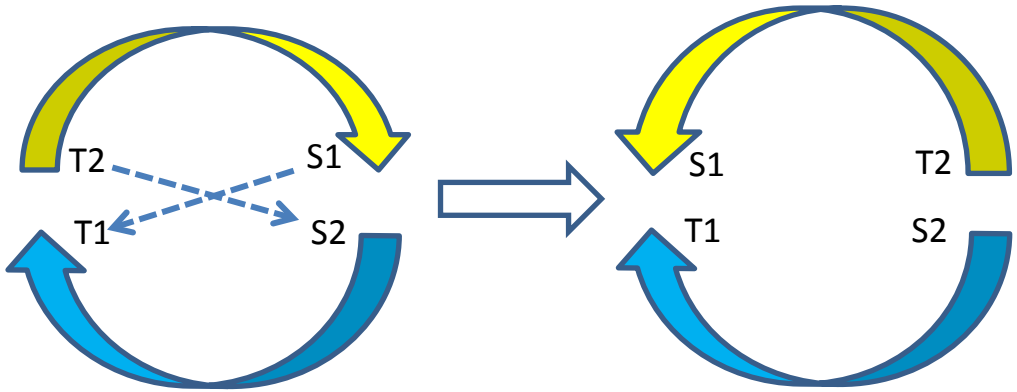
  (1.4) 从上述的*n*个序列中, 选择最短的一个.

TSP问题求解步骤2：序列改进的Two-opt算法

- 把步骤1构造出的序列首尾相连, 形成一个类似TSP问题的回路. 将回路从任意两个位置上断开, 颠换后对接, 如颠换对接后有更短的图距, 则把对接后的回路作为新的回路继续改进, 直到回路不再变短为止.
- 如果新的回路与之前的回路相比有较短的路程, 则在新回路的基础上重复Two-opt改进算法. 如果新的回路与之前的回路相比没有较短的路程, 则在旧回路的其他位置上重复Two-opt改进算法. 对最终得到的回路, 把首尾相连的两个标记重新分开, 或从最长的区间上将回路断开, 就得到我们想要的连锁图.

Two-opt改进算法示意图

左图为交换前的回路, 右图为交换后的回路



TSP问题的解与图谱之间的差异

- TSP is a closed route. There is no starting and ending points.
- Linkage map is an open route. It has a starting point and an ending point.
- Criteria for the shortest map
  - Convert a TSP route to a map by breaking the TSP route from the longest interval
  - Map length = TSP route length – the longest interval
- The above criteria are used to determine whether a shorter map is found

TSP问题求解步骤3：图谱调整算法

- 对于只包含数十个标记的连锁群, 通过步骤1和2一般就能得到最优的标记顺序.
- 当标记更多时, 步骤1和2得到的标记顺序还有进一步改进的必要. 具体过程如下:
  - 选定一个窗口大小w, w一般在5~10个标记之间. 窗口太小, 改进效果不明显; 窗口太大, 则耗时较长. 假定一个连锁群体上有n个标记, 一般认为$n>>w$.
  - 对i=1, 2, …, n-w做循环. 对w个标记M_i, M_{i+1}, …, M_{i+w}的所有w!个可能排列中, 寻找最短的一种排列顺序. 例如w=5时, w!=120; w=8时, w!=40320.

QTL IciMapping中调整算法的标准

- Four rippling criteria are
  - (i) SARF (Sum of Adjacent Recombination Frequencies)
  - (ii) SAD (Sum of Adjacent Distances)
  - (iii) SALOD (Sum of Adjacent LOD scores)
  - (iv) COUNT (number of recombination events)
- **The only criteria: shortest map!**

大麦DH群体中1H染色体上14个标记的排序

标记编号	标记名称	与下一个标记的间距 (cM)	染色体上位置 (cM)	两个相邻区间上的干涉系数
1	Act8A	10.88	0	0
2	OP06	7.64	10.88	0
3	aHor2	60.59	18.52	0.1738
4	MWG943	13.07	79.11	0.9282
5	ABG464	19.28	92.18	0.6166
6	Dor3	3.55	111.46	0.0581
7	iPgd2	7.04	115.01	0.9
8	cMWG733A	3.56	122.05	1
9	AtpbA	13.61	125.61	0
10	drun8	4.88	139.22	0
11	ABC261	7.09	144.10	1
12	ABG710B	3.50	151.19	1
13	Aga7	5.70	154.69	
14	MWG912		160.39	

连锁图谱与物理图谱

Species	Size of haploid genome (kb)	Size of linkage map (cM)	kb/cM
Yeast	2.2×10^4	3700	6
Neurospora	4.2×10^4	500	80
Arabidopsis	7.0×10^4	500	140
Drosophila	2.0×10^5	290	700
Tomato	7.2×10^5	1400	510
Human	3.0×10^6	2710	1110
Wheat	1.6×10^7	2575	6214
Rice	4.4×10^5	1575	279
Corn	3.0×10^6	1400	2140