連鎖解析におけるS-PLUSの利用

慶應義塾大学大学院
理工学研究科基礎理工学専攻
修士2年　菅谷勇樹

近年のゲノム研究

- 大量のゲノム関連データが利用可能に
  - ヒトゲノムの全塩基配列の解読完了(2003年)

- これらのデータを使った研究へ
  - 塩基配列の違いが個人の違いにどのように影響しているか？
    - 疾患の有無
    - 薬剤の効果，副作用
    - etc…

    （これらを**形質**と呼ぶ）

- ゲノム創薬
- オーダーメイド医療

特に，塩基配列（遺伝子）と疾患との関係を探る

連鎖解析

報告内容

- Ⅰ　連鎖解析
  - i.　連鎖解析とは
  - ii.　現在の連鎖解析の問題点
- Ⅱ　新しい連鎖解析の計算アルゴリズムの提案
  - i.　確率継承アルゴリズム
  - ii.　確率継承アルゴリズムのS-PLUSによる実装
- ⊕　まとめと今後の課題

ヒトゲノムの全塩基配列の解読が終わり，大量のゲノム関連データが安価に利用できるようになった近年の時代背景を受け，ゲノム研究は塩基配列と形質との関係を明らかにすることに興味が移ってきている．その中でも，特に疾患と塩基配列（遺伝子）との対応を調べる手法のひとつに連鎖解析がある．

連鎖解析では，減数分裂の過程で相同染色体間で組み換えが起こる確率（θ）の家系図データに基づく尤度を計算する必要がある．しかし，既存の尤度計算アルゴリズムは親から子へという遺伝の流れを無視した形で尤度を求めるので，多数マーカーかつ大家系の解析は不可能である．そこで，遺伝の流れに素直な新たな尤度計算アルゴリズムとして**確率継承アルゴリズム**を提案し，S-PLUSで実装する．

本報告では，連鎖解析におけるS-PLUSの利用の一例を示す．

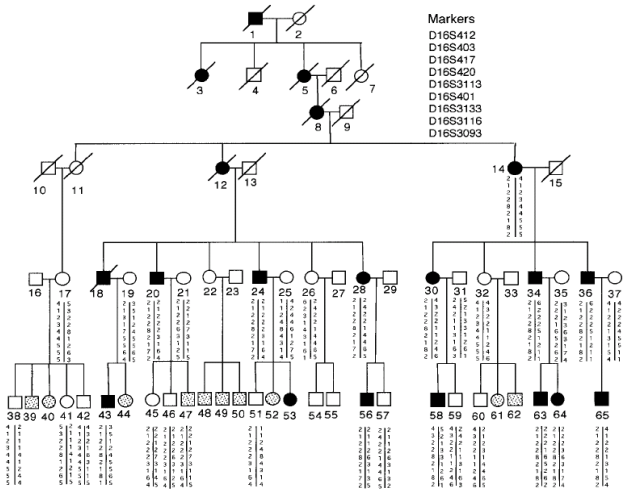
連鎖解析の概略

- 目的
  - 遺伝的疾患の原因遺伝子の場所を探る

    遺伝子の場所がわかる⇒新しい治療，新薬の開発
- 方法
  - 家系図を利用
    - 疾患の発症の有無
    - マーカー情報
  - 個人の塩基配列の違い（これをマーカーとする．観測可能）と，疾患原因遺伝子との連鎖を手がかりに探していく

    ↓

    メンデルの独立の法則の例外



連鎖

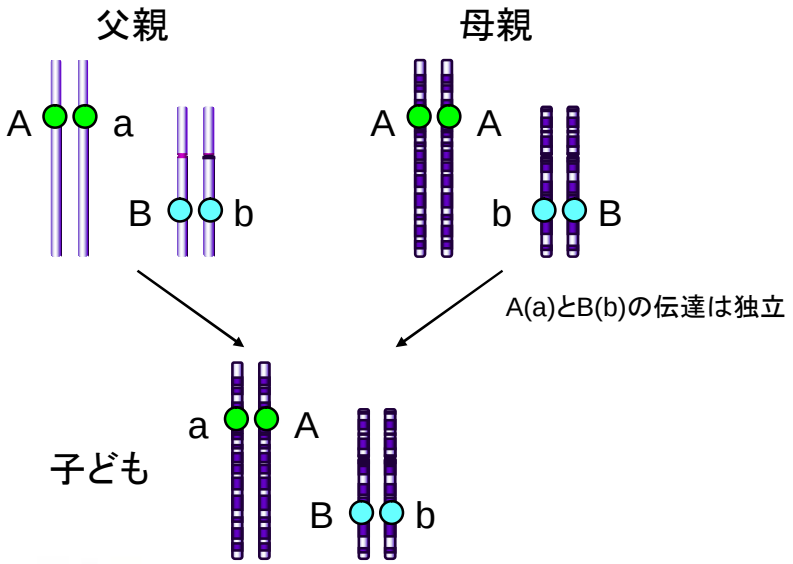
- メンデルの独立の法則

「性質の異なる形質は独立に遺伝する」

メンデル（Mendel）

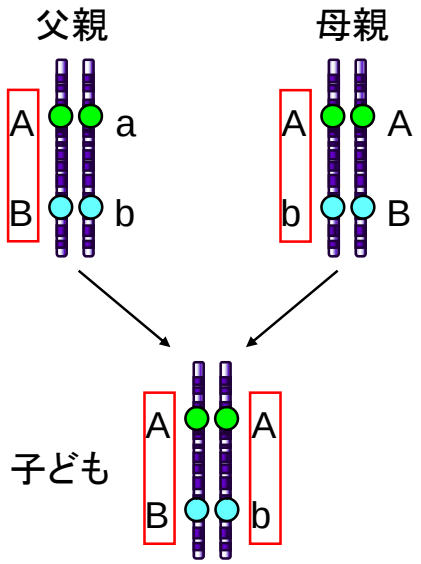
独立遺伝の場合

A,a ⇒ 形質1の原因遺伝子
B,b ⇒ 形質2の原因遺伝子



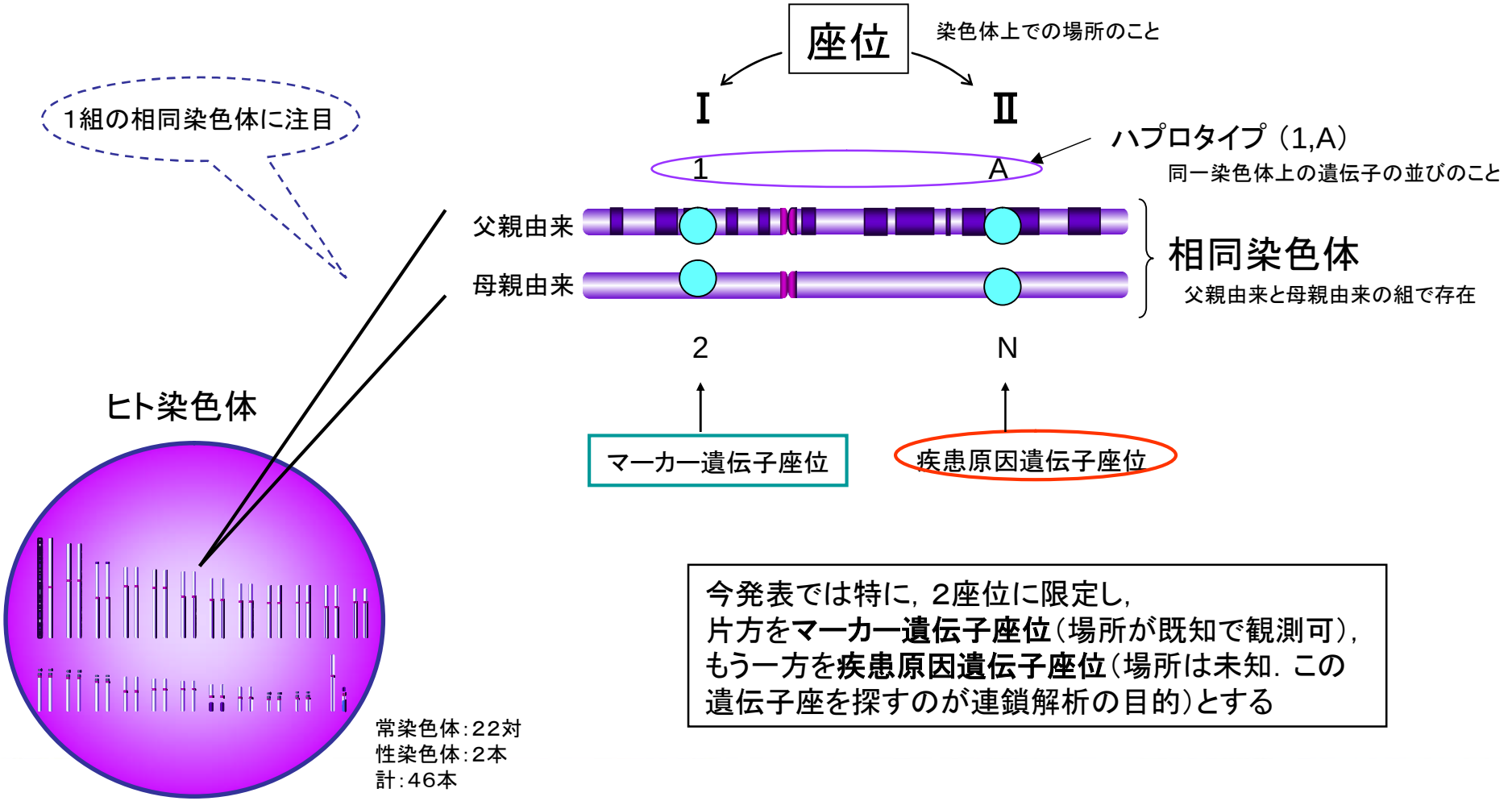
2つの形質は独立に遺伝する

独立法則の例外



原因遺伝子が染色体上の近くに存在するため，2つの形質の遺伝は独立ではない ⇒ **連鎖**

相同染色体

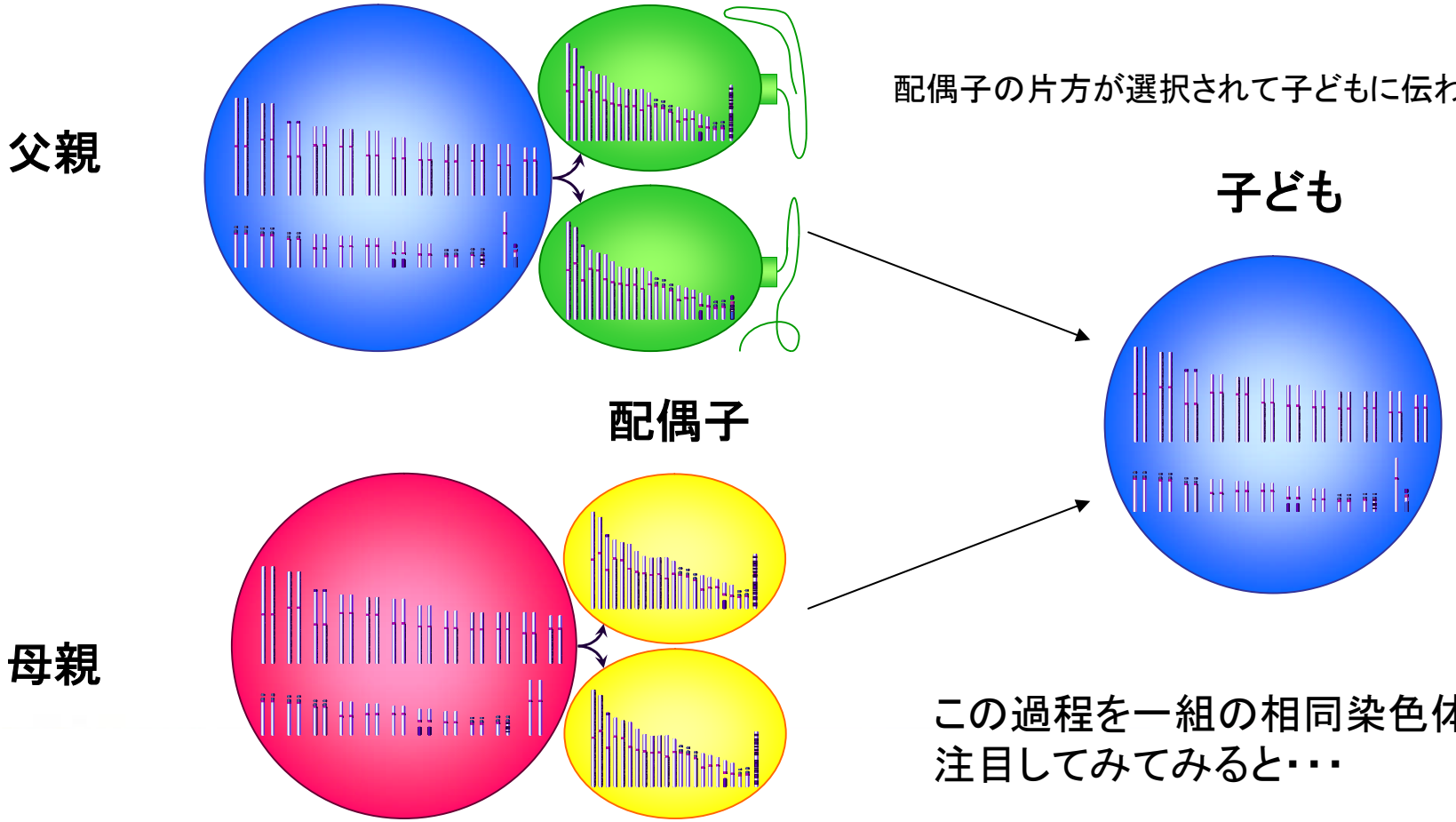


今発表では特に，2座位に限定し，
片方を**マーカー遺伝子座位**（場所が既知で観測可），
もう一方を**疾患原因遺伝子座位**（場所は未知．この遺伝子座を探すのが連鎖解析の目的）とする

http://www.tokyo-med.ac.jp/genet/chp.htm

親から子への配偶子伝達

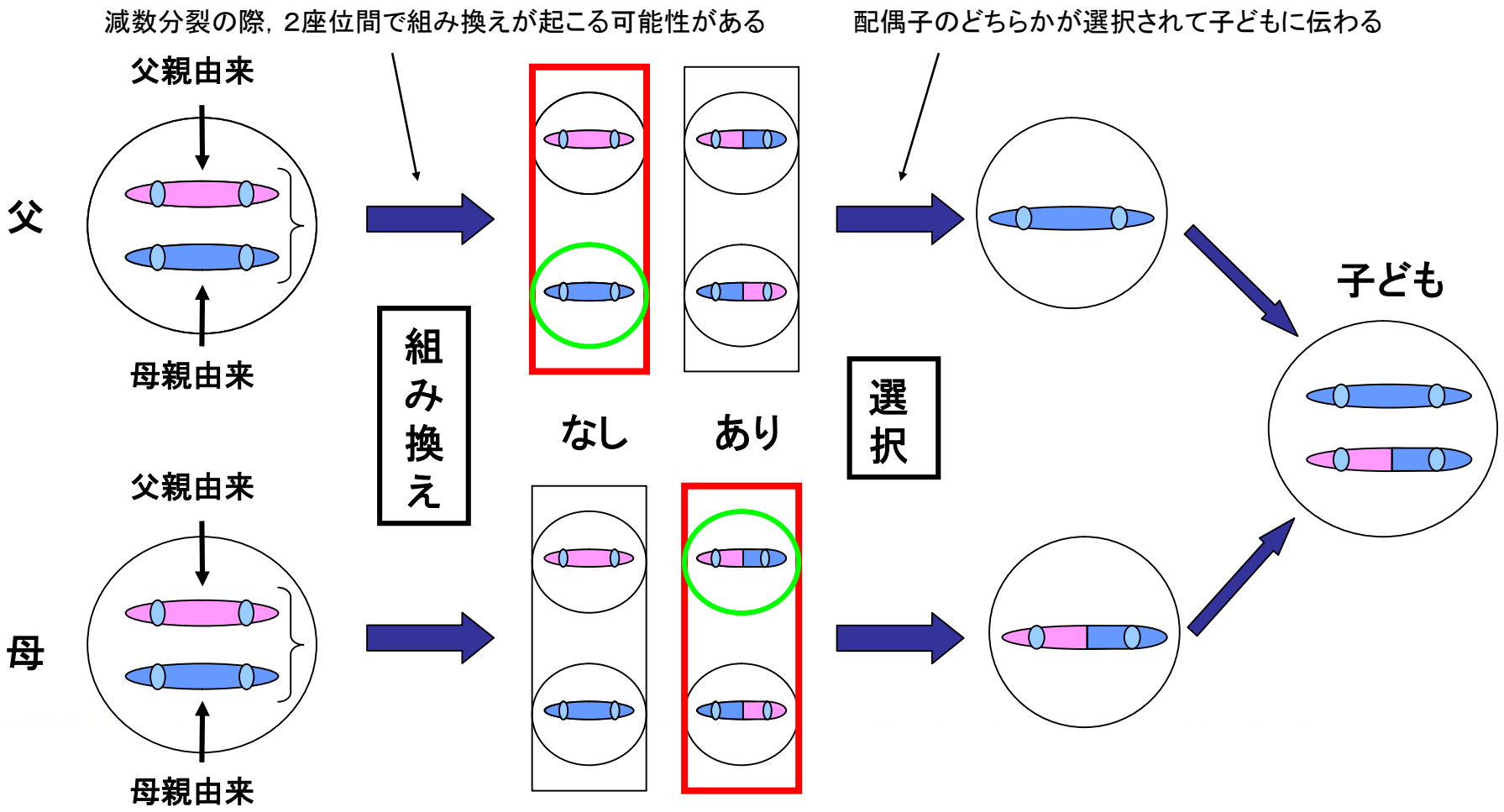
減数分裂の過程で，相同染色体は別れて配偶子に入る



この過程を一組の相同染色体に
注目してみてみると･･･

親から子へのハプロタイプ伝達

2座位に注目（マーカー座位と疾患原因座位）



交差と組み換え

組み換えとは奇数回の交差の結果である

交差の回数

0

組み換えなし

1
（奇数回）

組み換えあり

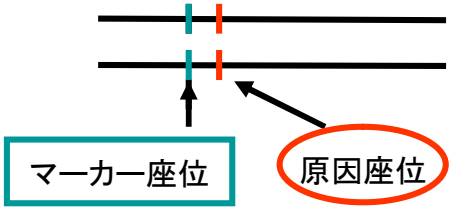
2
（偶数回）

組み換えなし

組み換え確率（θ）と座位の関係

2つの座位の間で組み換えが起こる確率

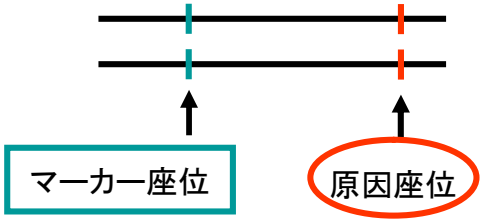
- 座位が近い



⇒ 交差は起こりにくい

⇒ $\theta \approx 0$

- 座位が遠い



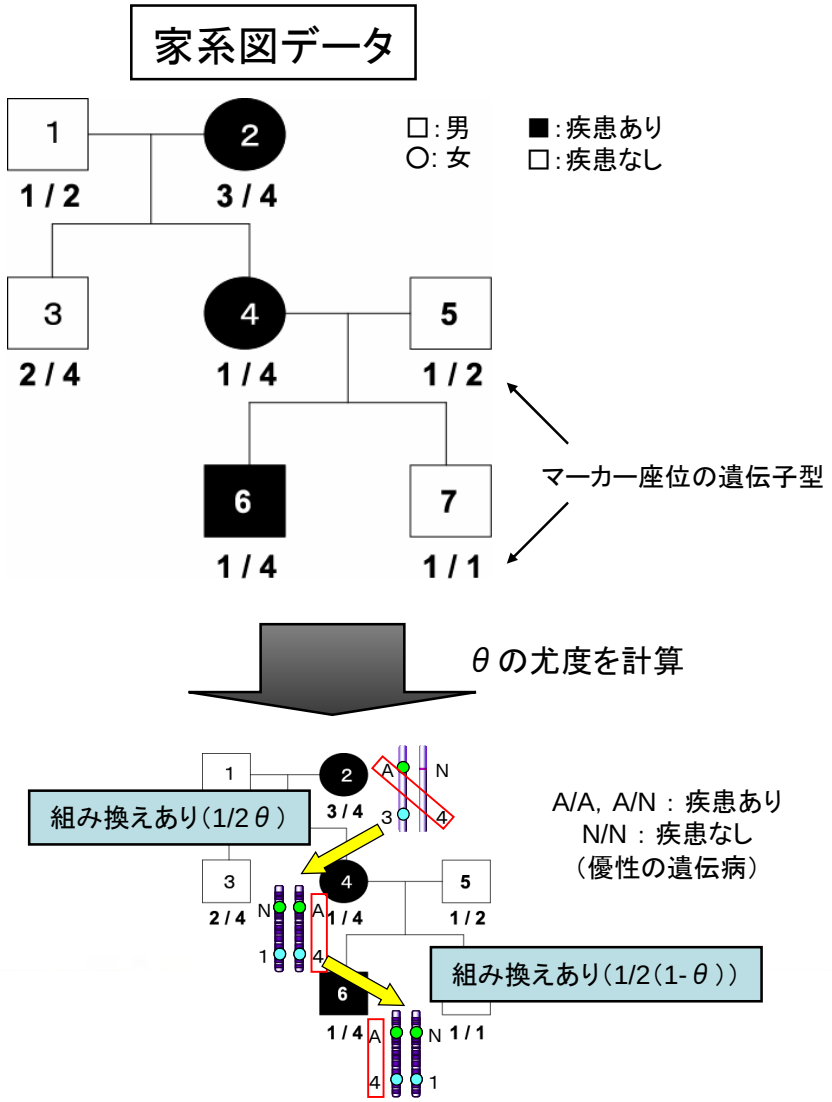
⇒ 複数回の交差により組み換えが起こるかどうかは等確率に近い

⇒ $\theta \approx 0.5$

組み換え確率（θ）が０に近い ⇒ 2つの座位は近くに存在

連鎖解析では，組み換え確率（θ）を推定し，０に近いかどうかで疾患原因座位がマーカー座位に近いかどうかを判断している

家系図データと連鎖解析



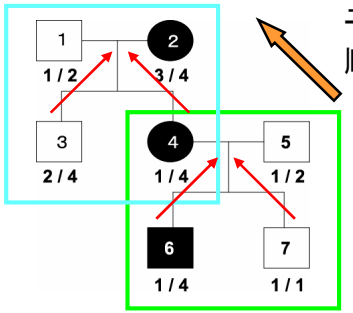
連鎖解析

- あるマーカーに注目
  - そのマーカーの近くに疾患原因遺伝子は存在するか？
- 家系図データに基づく組み換え確率（θ）の尤度を計算
- 尤度を最大にするθを推定
  - 推定値は0 に近いか？
  - 0 に近い⇒そのマーカーの近くに存在

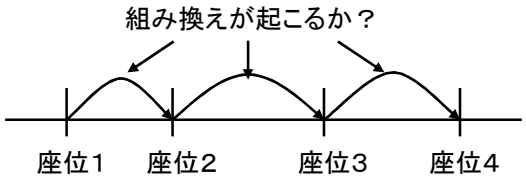
この過程を，マーカーを変えて繰り返すことにより，疾患原因遺伝子の位置を絞り込んでいく

家系図尤度の計算法とその問題点

- 既存の尤度計算アルゴリズム
  - Elston-Stewart アルゴリズム(1971)
    - backforward（子孫→祖先）
    - 大家系，少数マーカー向け
    - Linkage packageに実装
  - Lander-Green アルゴリズム(1987)
    - 継承ベクトル
    - 隠れマルコフモデル
    - 小家系，多数マーカー向け
    - Genehunterに実装



子どもの側から
順に尤度を計算



組み換えが座位ごと順々に起こると仮定

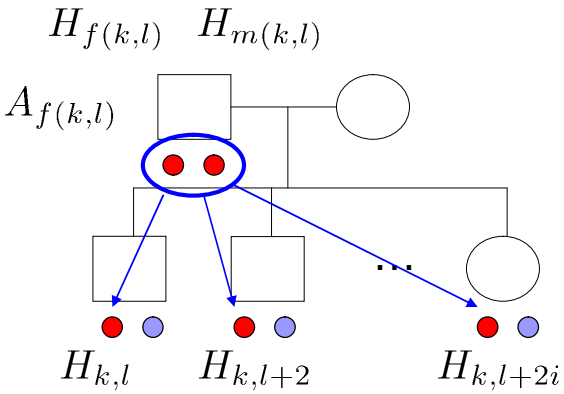
どちらのアルゴリズムも，親から子どもへという遺伝の流れに逆らう形で尤度を計算
⇒多数マーカーかつ大家系の解析は不可能

遺伝の流れを素直に記述

確率継承アルゴリズム

確率継承アルゴリズム

- 子どものハプロタイプは親のハプロタイプのみに依存して決まる
  - マルコフ性
- 親からハプロタイプが継承されるとともに親のもつ確率も継承
- 尤度は一世代の継承の積み重ねで計算可能
- 祖先→子どもの向きで尤度を計算（遺伝の流れと同じ）
- 創始者（家系図のtopの人）の確率は母集団のハプロタイプ頻度で置き換える



一世代の継承の確率

$$p_{k,(l,\cdots l+2i)}(h_0,\cdots,h_i,\mathbf{a}_{f(k,l)},\mathbf{m}_{f(k,l)})$$

$$= \sum_{(h_f,h_m)} \prod_{j=0}^{i} p_{k,l+2j}(h_j \mid \mathbf{a}_{f(k,l)},\mathbf{m}_{f(k,l)},h_f,h_m)\; p(\mathbf{a}_{f(k,l)},\mathbf{m}_{f(k,l)},h_f,h_m)$$

$$= \sum_{(h_f,h_m)} \prod_{j=0}^{i} p_{k,l+2j}(h_j \mid h_f,h_m)\; p(a_{f(k,l)},m_{f(k,l)} \mid h_f,h_m)$$

$$\times p_{f(k,l)}(h_f,\mathbf{a}_{f^2(k,l)},\mathbf{m}_{f^2(k,l)})\; p_{m(k,l)}(h_m,\mathbf{a}_{fm(k,l)},\mathbf{m}_{fm(k,l)})$$

- 再帰的に計算可能
- 親の確率が決まれば子どもの確率が決まる

- (k,l) ： ハプロタイプの番号
- $f(k,l)$ ： (k,l) の父親由来のハプロタイプの番号
- $f(k,l)$ ： (k,l) の母親由来のハプロタイプの番号
- $H_{k,l}$ ： ハプロタイプを表す確率変数
- $A_{k,l}$ ：疾患を確率変数
- $\mathscr{A}_{k,l}$ ： 祖先の発症を表す確率変数の集合
- $M_{k,l}$ ：マーカーを表す確率変数
- $\mathscr{M}_{k,l}$ ： 祖先のマーカーを表す確率変数の集合

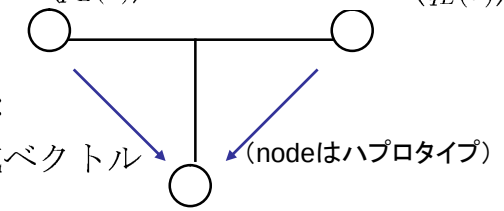
- $p_{k,(l,\cdots,l+2i)}(h_0,\cdots,h_i,\mathbf{a}_{f(k,l)},\mathbf{m}_{f(k,l)})$
  $= P(H_{k,l}=h_0,\cdots,H_{k,l+2i}=h_i,$
  $\mathscr{A}_{f(k,l)}=\mathbf{a}_{f(k,l)},\mathscr{M}_{f(k,l)}=\mathbf{m}_{f(k,l)})$

- $p(a_{f(k,l)},m_{f(k,l)} \mid h_f,h_m)$
  $= P(A_{f(k,l)}=a_{f(k,l)},M_{f(k,l)}=m_{f(k,l)} \mid$
  $H_{f(k,l)}=h_{f(k,l)},H_{m(k,l)}=h_{m(k,l)})$

一世代の継承のプログラム(1)

親のもつハプロタイプの確率は，どのように計算されて子どもに継承されるか？

$$\boldsymbol{p}(\theta)=\begin{pmatrix}p_1(\theta)\\p_2(\theta)\\\vdots\\p_L(\theta)\end{pmatrix}\qquad \boldsymbol{q}(\theta)=\begin{pmatrix}q_1(\theta)\\q_2(\theta)\\\vdots\\q_L(\theta)\end{pmatrix}$$



$\boldsymbol{p}(\theta),\boldsymbol{q}(\theta)$:
θ の多項式ベクトル

(nodeはハプロタイプ)

子どものハプロタイプの確率？

L2:マーカー遺伝子の種類
L:ハプロタイプの種類(2×L2)
(疾患原因遺伝子はA or N)

ハプロタイプの番号・・・A1～AL2，N1～NL2の順に1からつける

子どもにｉ 番目のハプロタイプが伝わる確率

$$\Rightarrow \quad \boldsymbol{p}(\theta)(A_i+\theta B_i)\boldsymbol{q}(\theta)$$

ただし，A_i, B_i は以下のような行列である

$$(A_i)_{jk}=\begin{cases}1 & j=k=i\\ \frac{1}{2} & (j=i,k\neq i)\mathrm{or}(j\neq i,k=i)\\ 0 & otherwise\end{cases}$$

(B_{ijk}) : B_i のブロック行列$(L2\times L2)$
$l1=(i-1)\div L2$の商$+1$
$l2=(i-1)\div L2$の剰余$+1$

- $(j=l1,k=l1)$
$$(B_{ijk})_{lm}=0$$

- $(j=l1,k\neq l1)$
$$(B_{ijk})_{lm}=\begin{cases}-\frac{1}{2} & l\neq l2,m=l2\\ \frac{1}{2} & l=l2,m\neq l2\\ 0 & otherwise\end{cases}$$

- $(j\neq l1,k=l1)$
$$(B_{ijk})_{lm}=\begin{cases}-\frac{1}{2} & l=l2,m\neq l2\\ \frac{1}{2} & l\neq l2,m=l2\\ 0 & otherwise\end{cases}$$

一世代の継承のプログラム(2)

Lemma θ の多項式行列 $\boldsymbol{p}(\theta), \boldsymbol{q}(\theta)$ の係数行列ををそれぞれ P, Q とする.
多項式 $(\boldsymbol{p}(\theta))^T A \boldsymbol{q}(\theta)$ において θ の v 次の係数は, $\Lambda = P^T A Q$ とすると,

$$\sum_{k+l-2=v} \Lambda_{kl} \qquad (v = 0, \ldots, n+m)$$

Corollary 多項式 $\theta(\boldsymbol{p}(\theta))^T A \boldsymbol{q}(\theta)$ の θ の v 次の係数は,

$$\sum_{k+l-2=v} \Lambda_{kl} \qquad (v = 1, \ldots, n+m+1)$$

親のもつハプロタイプの確率(P, Q)から, 子どものハプロタイプの確率を計算するプログラム

```
> Inherit
function(P, Q)
{
        R.row = function(i, P, Q)
        {
                Coef.v = function(Mat, P, Q)
                {
                        Lam = t(P) %*% Mat %*% Q
                        tapply(Lam, row(Lam) + col(Lam), sum)
                }
                c(Coef.v(Lambda1[[i]], P, Q), 0) +
c(0, Coef.v(Lambda2[[i]], P, Q))
        }
        t(sapply(1:L, R.row, P = P, Q = Q))
}
```

$\theta P^T B_i Q$ の計算

$P^T A_i Q$ の計算

家系図のデータ形式

既存のデータ形式

マーカー座位が増えるごとに列が追加される

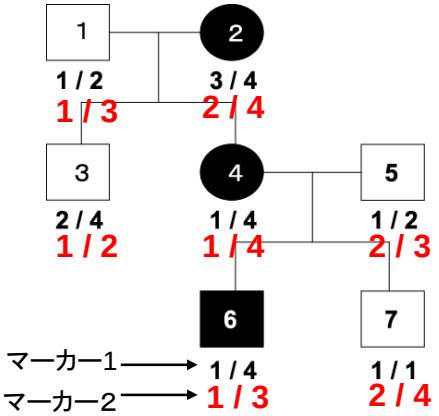
						マーカー1		マーカー2	
家系 ID	ID	父親 ID	母親 ID	疾患	性別	アレル 1	アレル 2	アレル 1	アレル 2
1	1	0	0	1	1	1	2	1	3
1	2	0	0	2	2	3	4	2	4
1	3	1	2	1	1	2	4	1	2
1	4	1	2	2	2	1	4	1	4
1	5	0	0	1	1	1	2	2	3
1	6	5	6	2	1	1	4	1	3
1	7	5	6	1	1	1	1	2	4

…

性別：男→1，女→2
疾患：疾患なし→1，疾患あり→2

分ける

非正規な形の表



プログラムで用いるデータ形式

familyデータ

家系 ID	ID	父親 ID	母親 ID	疾患	性別
1	1	0	0	1	1
1	2	0	0	2	2
1	3	1	2	1	1
1	4	1	2	2	2
1	5	0	0	1	1
1	6	5	6	2	1
1	7	5	6	1	1

genotypeデータ

家系 ID	ID	マーカー	アレル 1	アレル 2
1	1	マーカー 1	1	2
1	2	マーカー 1	3	4
1	3	マーカー 1	2	4
1	4	マーカー 1	1	4
1	5	マーカー 1	1	2
1	6	マーカー 1	1	4
1	7	マーカー 1	1	1

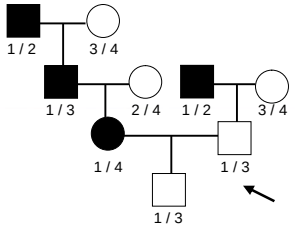
実装したプログラムでは，
2枚の表に分け入力データとする

尤度計算プログラムとその実行例

- 入力データ
  - familyデータ
  - genotypeデータ
  - マーカー遺伝子の種類（L2）
  - マーカー名
  - 創始者のハプロタイプ頻度
  - 疾患の形式
    - 優性遺伝の疾患か劣性遺伝の疾患か
- 出力
  - 尤度関数の係数
  - 尤度関数を微分し0とおいた方程式の根と，その根の関数値（polyrootを使用）
  - 尤度関数の最大値とそのときの θ の値（近似的な値．1/100単位）
  - 尤度関数のプロット

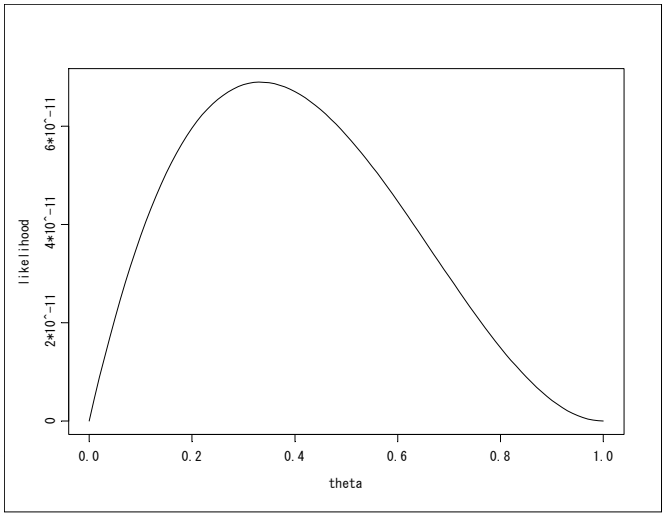
尤度計算プログラムの実行例 （既存のアルゴリズムとは異なる結果）

```
> Like.2(testfam,testgeno,4,"mark1",rep(1/8,8),dominant=T)
$likelihood:  #尤度関数の係数
[1]  0.000000000e+000  4.656612873e-010 -9.313225746e-010  4.656612873e-010

$theta:  #尤度関数を微分し0とおいた方程式の根
[1] 0.3333333333-2.019483917e-028i 1.0000000000+2.019483917e-028i

$value:  #上で求めた根における関数値
[1] 6.898685738e-011-5.220243574e-054i 0.000000000e+000+0.000000000e+000i

$max:
$max[[1]]:  #尤度関数を最大にする θ の値
[1] 0.33

$max[[2]]:  #尤度関数の最大値
[1] 6.898166612e-011
```



例に用いた家系図

尤度関数のプロット



まとめと今後の課題

- まとめ
  - 遺伝の流れに素直な確率継承アルゴリズムを提案した.
  - S-PLUSを用いて確率継承アルゴリズムを実装した.
  - 確率継承アルゴリズムは，創始者の確率の与え方が既存のアルゴリズムと違うため結果も異なる
- 今後の課題
  - 本報告では，連鎖解析のなかでも1つのマーカーに注目する単点解析のみを取り扱ったが，多数マーカーを用いた多点解析についても検討し，既存の尤度計算アルゴリズムとの比較を行なう.
  - ある性質をもつ家系や大きさの小さな家系においてはθの推定に偏りがあることがわかっている．今後は家系図のもつθに関する情報量についても検討する予定である.

参考文献

- Haldane JBS.(1919).The combination of linkage values and the calculation of distances between the loci of linkage factors.J Genet 8,299-309.
- Kosambi DD. (1944). The estimation of map distances from recombination values. Ann Eugen 12, 172-175.
- Karlin S. (1984). Theoretical aspects of genetic map functions in recombination processes. In Human Population Genetics, 209-228.
- Karlin S. (1994). Theoretical recombination processes incorporation interference effects. Theorical population biology 46, 198-231
- Kamatani, N. et al. (2000). Localization of a gene for familial juvenile hyperuricemic nephropathy causing underexcretion-type gout to 16p12 by genome-wide linkage analysis of a large family. Arth. Rheum 43, 925-929
- K.R.Gabriel (1959). The distribution of the number of successes in a sequence of dependent trials. Biometrika 46, 454-460
- Jurg Ott. Analysis of Human Genetic Linkage ．

- R.C.Elston, J.Stewart. (1971). A General Model for the Genetic Analysis of Pedigree Data. Human Heredity 21, 523-542.
- E.S.Lander, P.Green. (1987). Construction of multilocus genetic linkage maps in humans. Genetics 84, 2363-2367.
- Jurg Ott. (1974). Estimation of the Recombination in Human Pedigree:Efficient Computaion of the Likelihood for Human Linkage Dtudies. Am J Hum Genet 26, 588-597.
- R.M.Idury, R.C.Elston. (1997). A Faster and More General Hidden Markov Algorithm for Multipoint Likelihood Calculations. Human Heredity 47, 197-202.
- Leonid Kruglyak, Mark J.Daly, Eric S.Lander. (1995). Rapid Multipoint Linkage Analysis of Recessive Traits in Nuclear Families, Including Homozygosity Mapping. Am.J.Hum.Genet 56, 519-527.
- Audustine Kong. at el.(2002). A high-resolution recombination map of the human genome. Nature Genetics 31, 241-247
- 鎌谷直之(2001).ポストゲノム時代の遺伝統計学.羊土社.