Markov Chain Modelling of Recombination Fraction of Chromosomes

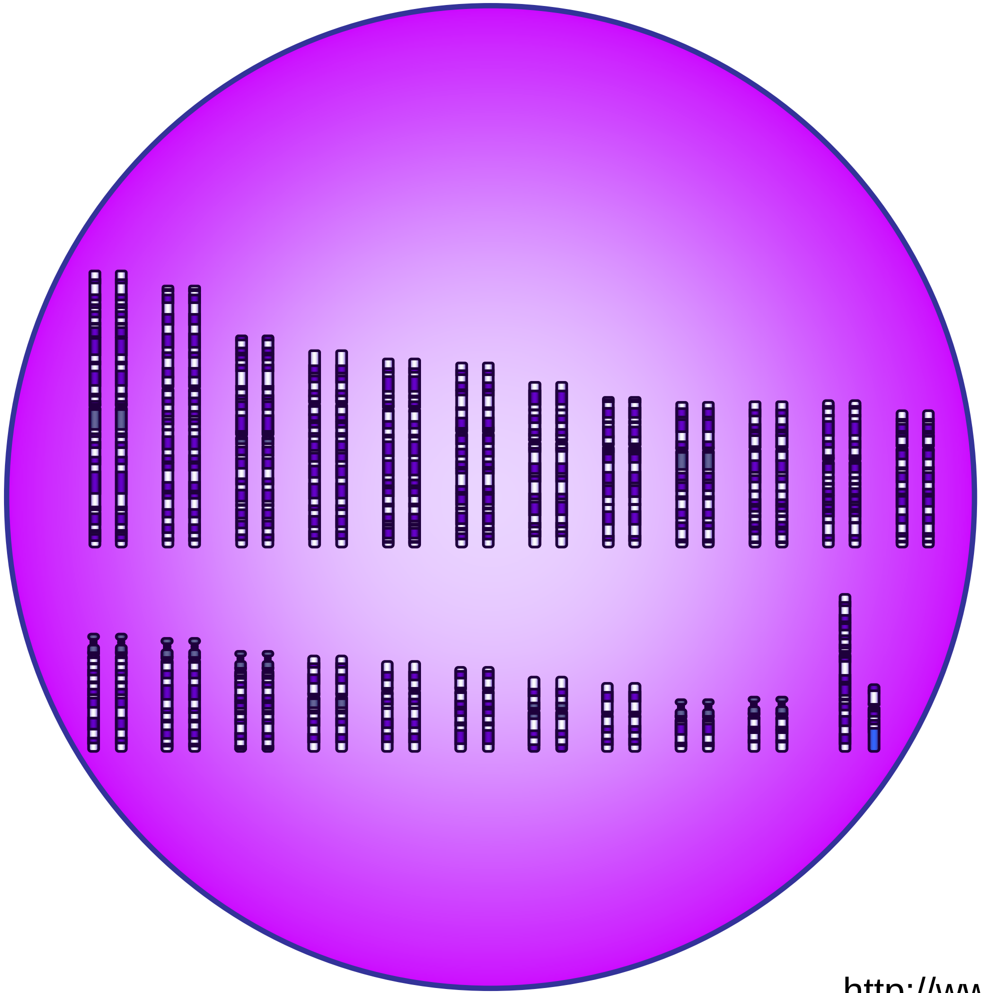
Yuki Sugaya
Ritei Shibata
Keio University, Japan

Shigeo Kamitsuji
StaGen Co., Ltd.

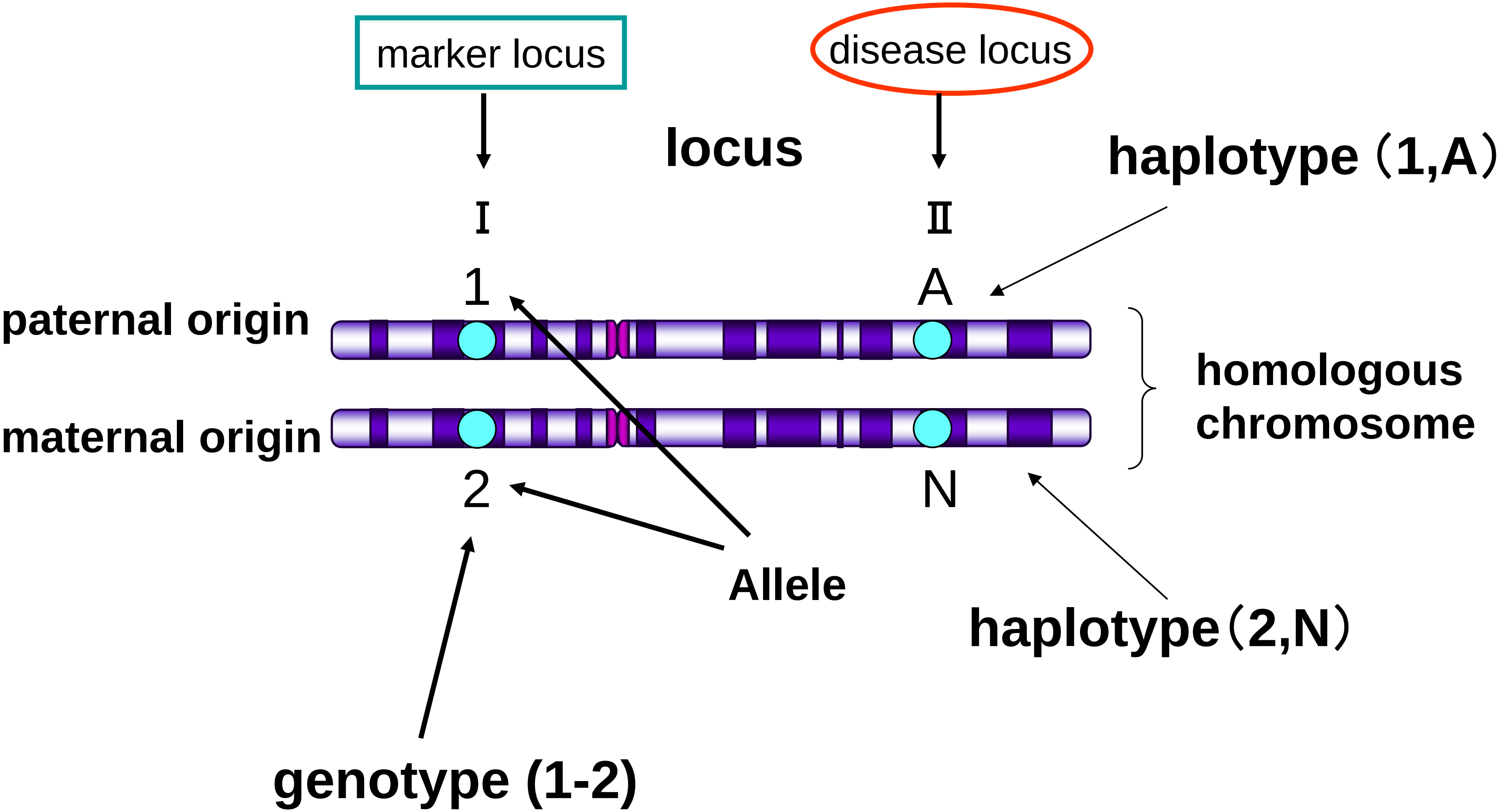
Objective

- Markov Chain modelling of recombination fraction
- A new algorithm for linkage analysis
  - “Probability inheritance” algorithm

Chromosomes in nuclear



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



marker locus

disease locus

locus

haplotype（1,A）

Ⅰ

Ⅱ

1

A

paternal origin

maternal origin

homologous chromosome

2

N

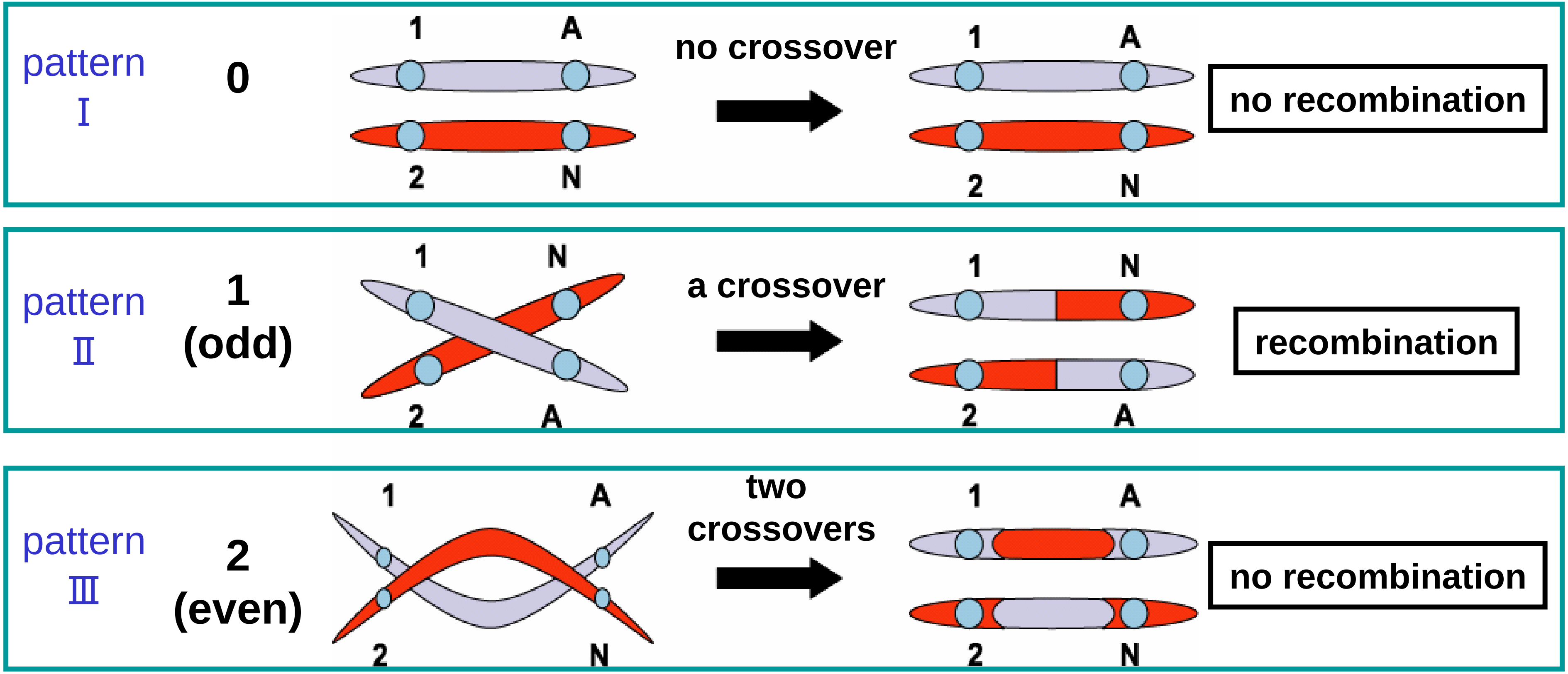
Allele

haplotype（2,N）

genotype (1-2)

Crossovers and Recombination

The number of crossovers



pattern Ⅰ — **0** — **no crossover** → **no recombination**

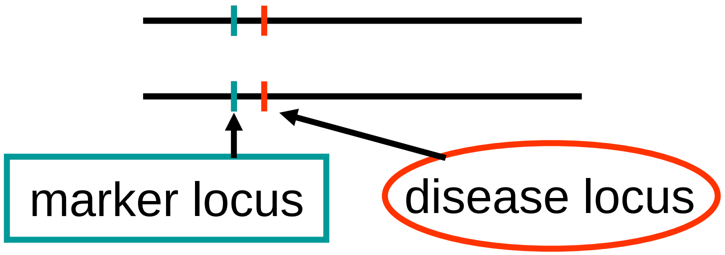
pattern Ⅱ — **1 (odd)** — **a crossover** → **recombination**

pattern Ⅲ — **2 (even)** — **two crossovers** → **no recombination**

Recombination fraction (θ)

Probability of recombination between two loci

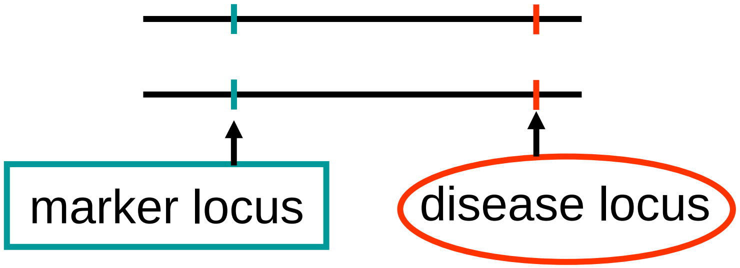
- Two loci are close



Crossover is not apt to occur

$\theta \approx 0$

- Two loci are far



It is believed that $\theta \approx 0.5$

Map function $\theta(d)$

- d : genetic distance
  - the expected number of crossovers occurring on a single chromosome between two loci
- Haldane(1919),Kosambi(1944),Carter-Falconer(1951), Sturt(1976), Felsenstein(1979), Karlin(1984)
  - Example

Haldane(1919)

$$\theta(d) = \frac{1 - \exp^{-2d}}{2}$$

Kosambi(1944)

$$\theta(d) = \frac{1}{2}\tanh 2d$$

Probabilistic model of recombination fraction –independent case-

Recombination fraction θ (n) between *n* loci

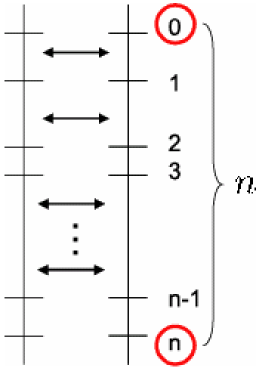
Assumption **At most a crossover between the (*i-1*)-th and the *i*-th loci**

$$X_i = \begin{cases} 1 & (\text{crossover between } i-1 \text{ and } i) \\ 0 & (\text{no crossover between } i-1 \text{ and } i) \end{cases}$$

$$P(X_i = 1) = p_0$$
$$P(X_i = 0) = 1 - p_0$$

$$X_i \perp\!\!\!\perp X_j$$

$$\theta(p_0, n) = P\left(\sum_{i=1}^{n} X_i = odd\right)$$
$$= \sum_{k:odd} \binom{n}{k} p_0^k (1-p_0)^{n-k}$$
$$= \frac{1}{2}\left\{1 - (1-2p_0)^n\right\}$$



Markov chain modelling of recombination fraction -dependent case -

- Assumption
  - The direction of crossover :

    the *0*-th locus ⇒ the *n*-th locus
  - Markov property of crossover

$$\boxed{\begin{array}{ll} P(X_i = 1 | X_{i-1} = 0) = p_1 & P(X_i = 1 | X_{i-1} = 1) = q_1 \\ P(X_i = 0 | X_{i-1} = 0) = 1 - p_1 & P(X_i = 0 | X_{i-1} = 1) = 1 - q_1 \end{array}}$$

$$\theta = P\left(\sum_{i=1}^{n} X_i = odd \right)$$

complicated

Second order Markov modelling

$$Y_i = \begin{cases} 0 & \text{(paternal origin)} \\ 1 & \text{(maternal origin)} \end{cases} \quad \left(\begin{array}{l} X_i = 0 \Leftrightarrow (Y_{i-1} = 0, Y_i = 0) \text{ or } (Y_{i-1} = 1, Y_i = 1) \\ X_i = 1 \Leftrightarrow (Y_{i-1} = 0, Y_i = 1) \text{ or } (Y_{i-1} = 1, Y_i = 0) \end{array} \right)$$

$$p_{tu|st} = P(\ Y_i = t, Y_{i+1} = u\ | Y_{i-1} = s, Y_i = t\ )$$

No distinction between paternal and maternal origins

$$\begin{cases} p = p_{01|00} = p_{10|11} & 1 - p = p_{00|00} = p_{11|11} \\ q = p_{10|01} = p_{01|10} & 1 - q = p_{11|01} = p_{00|10} \end{cases}$$

$$\begin{cases} P(Y_0 = 0) = P(Y_0 = 1) = \dfrac{1}{2} \\ P(Y_0 = 0,\ Y_1 = 0) = P(Y_0 = 1,\ Y_1 = 1) = \dfrac{1-r}{2} \\ P(Y_0 = 0,\ Y_1 = 1) = P(Y_0 = 1,\ Y_1 = 0) = \dfrac{r}{2} \end{cases}$$

Transition matrix

$$T = \begin{array}{c} \\ (00) \\ (10) \\ (01) \\ (11) \end{array} \begin{array}{c} \begin{array}{cccc} (00) & (10) & (01) & (11) \end{array} \\ \begin{pmatrix} 1-p & 0 & p & 0 \\ 1-q & 0 & q & 0 \\ 0 & q & 0 & 1-q \\ 0 & p & 0 & 1-p \end{pmatrix} \end{array}$$

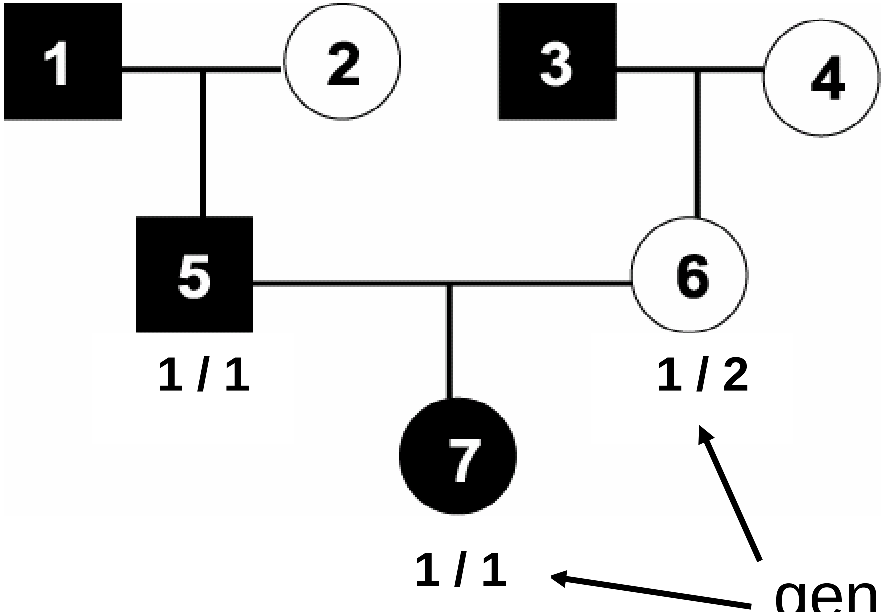
Recombination fraction derived from Markov model

$$
\begin{aligned}
\theta(p,q,r,n) &= P\big(Y_n = 1 \,|Y_0 = 0\big) \qquad \big(= P\big(Y_n = 0 \,|Y_0 = 1\big)\big) \\
&= \sum_{k,l} \frac{P\big(Y_n = 1, Y_{n-1} = k, Y_1 = l, Y_0 = 0\big)}{P\big(Y_0 = 0\big)} \\
&= \sum_{k,l} P\big(Y_n = 1, Y_{n-1} = k \,|Y_1 = l, Y_0 = 0\big) \,\frac{P\big(Y_1 = l\ , Y_0 = 0\big)}{P\big(Y_0 = 0\big)} \\
&= \frac{1}{2} + \frac{1}{2(\lambda_1 - \lambda_2)} \Big\{ (2r-1)(\lambda_1^n - \lambda_2^n) + \lambda_1 \lambda_2 (\lambda_1^{n-1} - \lambda_2^{n-1}) \Big\} \ \ (n \geq 1)
\end{aligned}
$$

$$
\left(\lambda_1 = \tfrac{1-p-q+\sqrt{(1-p-q)^2 - 4(p-q)}}{2}, \quad \lambda_2 = \tfrac{1-p-q-\sqrt{(1-p-q)^2 - 4(p-q)}}{2} \right)
$$

- $(p = q = r) \quad \theta(p,n) = \dfrac{1}{2}\Big\{ 1 - (1-2p)^n \Big\}$
- $\theta(p,q,r,n) \to \dfrac{1}{2} \ \ (n \to \infty)$

Pedigree data

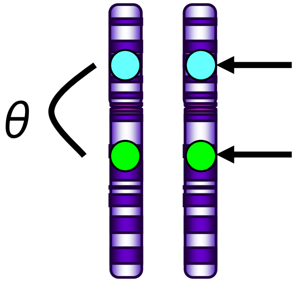


□ : male

○ : female

■ : affected, □ : not affected
(phenotype)

genotype at a DNA marker locus



marker locus (known)

disease locus (only phenotype is known)

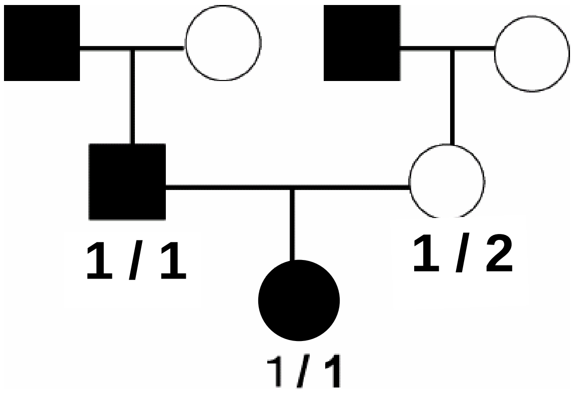
Evaluation of the likelihood

- Existing Algorithm
  - Elston-Stewart algorithm(1971)
    - backward algorithm
    - Good for large pedigree and small number of markers
    - Implemented in the LINKAGE program
  - Lander-Green algorithm(1987)
    - Inheritance vector
    - Hidden Markov Model
    - Good for small pedigree and large number of markers
    - Implemented in the GENHUNTER program
- Our algorithm
  - "Probability inheritance" algorithm (Sugaya and Shibata,2005)
    - Good for large pedigree and large number of markers

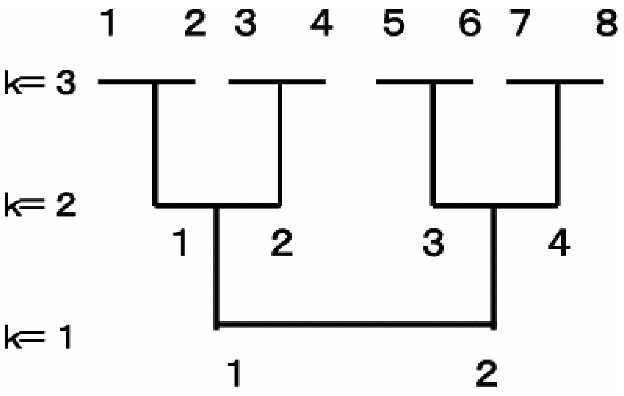
Evaluation of the likelihood

- Existing Algorithm
  - Elston-Stewart algorithm(1971)
    - backward algorithm
    - Good for large pedigree and small number of markers
    - Implemented in the LINKAGE program
  - Lander-Green algorithm(1987)
    - Inheritance vector
    - Hidden Markov Model
    - Good for small pedigree and large number of markers
    - Implemented in the GENHUNTER program
- Our algorithm
  - "Probability inheritance" algorithm (Sugaya and Shibata,2005)
    - Good for large pedigree and large number of markers

Phenotype to haplotype



phenotype



haplotype

- $A_{k,l}$: a (affected) or n (not affected)

$$(H_{k,l},\ H_{k,l+1}) \in \mathscr{H}^2 = \mathscr{H}_a^2 \cup \mathscr{H}_n^2,\ \ \mathscr{H}_a^2 \cap \mathscr{H}_n^2 = \emptyset$$

$$A_{k,l} = a \ \Leftrightarrow\ (H_{k,l},\ H_{k,l+1}) \in \mathscr{H}_a^2$$
$$A_{k,l} = n \ \Leftrightarrow\ (H_{k,l},\ H_{k,l+1}) \in \mathscr{H}_n^2$$

(k, l)

k : level

l : odd $\Rightarrow$ paternal origin

even $\Rightarrow$ maternal origin

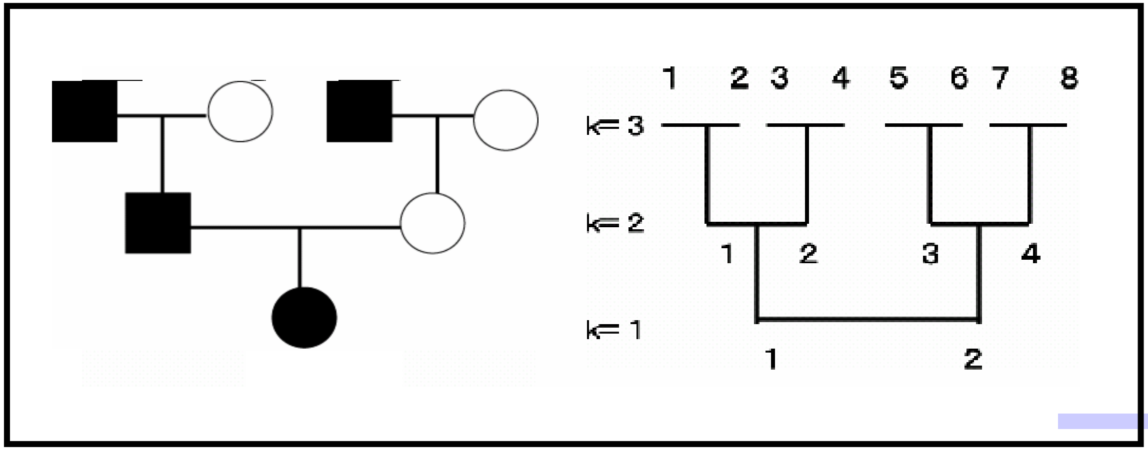
- $H_{k,l}$: Haplotype $(H_{k,l} \in \mathscr{H})$

- $\mathscr{A}_{k,l}$: A set of random variables that expresses affection status of an individual (k, l) and his ancestors

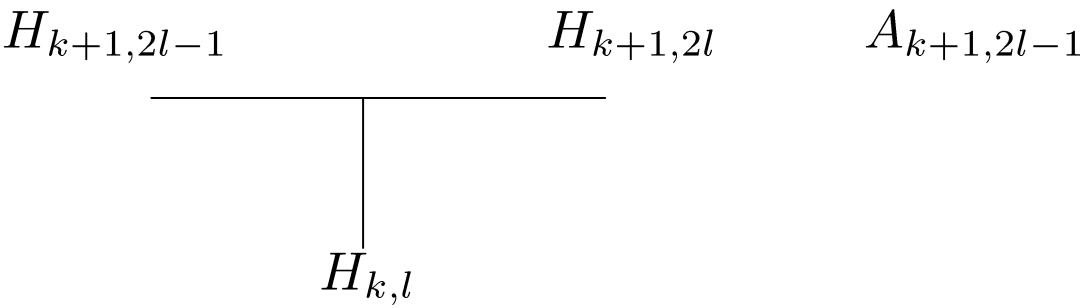
Probability of phenotypes in a pedigree

$$P(\mathscr{A}_{1,1}) = \sum_{\mathscr{H}^2_{A_{1,1}}} P((H_{1,1}, H_{1,2}) = (h_{1,1}, h_{1,2}), \mathscr{A}_{2,1}, \mathscr{A}_{2,3})$$

$$= \sum_{\mathscr{H}^2_{A_{1,1}}} \{\underline{P(H_{1,1} = h_{1,1}, \mathscr{A}_{2,1})}\ P(H_{1,2} = h_{1,2}, \mathscr{A}_{2,3})\}$$



Probability inheritance



$H_{k+1,2l-1}$ $H_{k+1,2l}$ $A_{k+1,2l-1}$

$H_{k,l}$

$$P(H_{k,l} = h_{k,l}, \mathscr{A}_{k+1,2l-1})$$

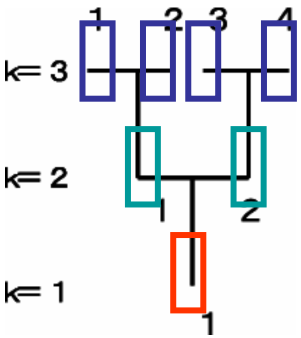
$$= \sum_{\mathscr{H}^2_{A_{k+1,2l-1}}} P(H_{k,l} = h_{k,l} \mid H_{k+1,2l-1} = h_{k+1,2l-1}, H_{k+1,2l} = h_{k+1,2l})$$

$$\cdot \underline{P(H_{k+1,2l-1} = h_{k+1,2l-1}, \mathscr{A}_{k+2,2(2l-1)-1})}$$

$$\cdot \underline{P(H_{k+1,2l} = h_{k+1,2l}, \mathscr{A}_{k+2,2(2l)-1})}$$

Probability inheritance in three generations

$$
\begin{aligned}
&P(H_{1,1} = h_1,\ \mathscr{A}_{2,1}) \\
&= \sum_{\mathscr{H}^2_{A_{2,1}}} P(H_{1,1} = h_1 \mid H_{2,1} = h_{2,1}, H_{2,2} = h_{2,2}) \\
&\quad \times \left\{ \left\{ \sum_{\mathscr{H}^2_{A_{3,1}}} P(H_{2,1} = h_{2,1} \mid H_{3,1} = h_{3,1}, H_{3,2} = h_{3,2}) \right. \right. \\
&\qquad\qquad \left. \times P(H_{3,1} = h_{3,1}) P(H_{3,2} = h_{3,2}) \right\} \\
&\quad \times \left\{ \sum_{\mathscr{H}^2_{A_{3,3}}} P(H_{2,2} = h_{2,2} \mid H_{3,3} = h_{3,3}, H_{3,4} = h_{3,4}) \right. \\
&\qquad\qquad \left. \left. \times P(H_{3,3} = h_{3,3}) P(H_{3,4} = h_{3,4}) \right\} \right\}
\end{aligned}
$$



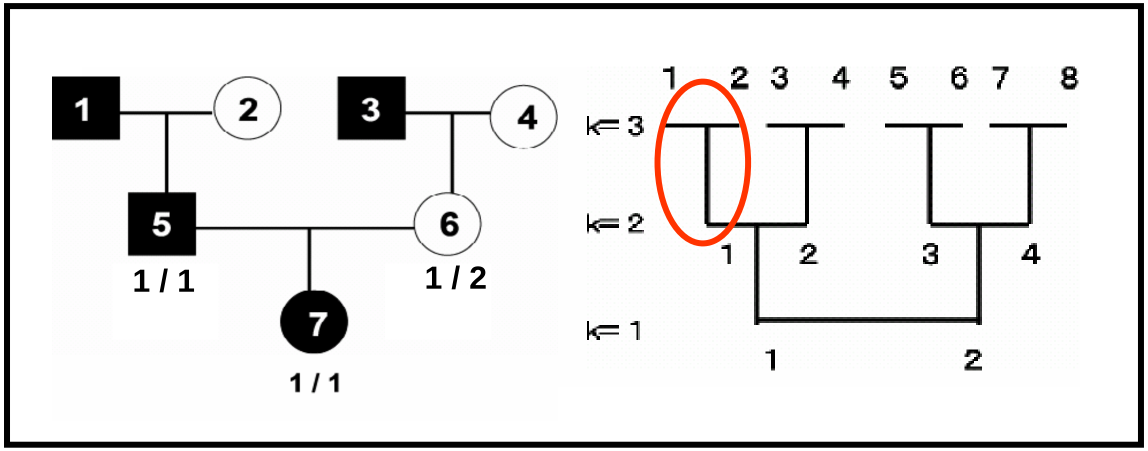
Example of “Probability inheritance”

maker allele : 1,2

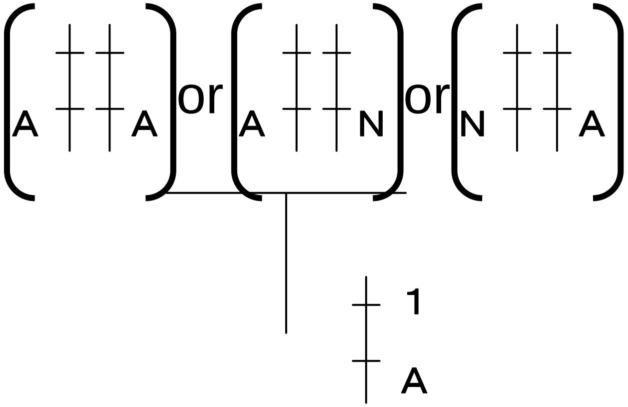
disease allele : A,N

$$A_{k,l} = a \Leftrightarrow ((*,A),(*,A))or((*,A),(*,N))$$

$$A_{k,l} = n \Leftrightarrow ((*,N),(*,N))$$

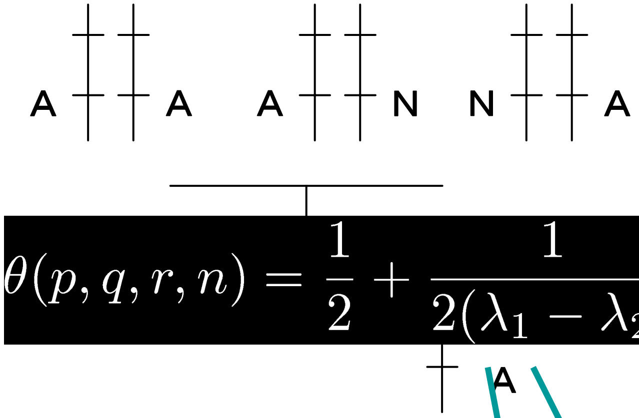


$$H_{2,1} = (1,A),\ \ A_{3,1} = a$$



Probability inheritance + Markov modelling of recombination fraction

$$\underline{H_{1,1} = (1, A),\ A_{1,1} = a}$$



$$P\left(H_{k,l} = (1, A)\right) = P_{1A}(k,l)$$
$$P\left(H_{k,l} = (*, A)\right) = P_{\bullet A}(k,l)$$
$$P\left(H_{k,l} = (1, *)\right) = P_{1\bullet}(k,l)$$

$$\theta(p,q,r,n) = \frac{1}{2} + \frac{1}{2(\lambda_1 - \lambda_2)}\left\{(2r-1)(\lambda_1^n - \lambda_2^n) + \lambda_1\lambda_2(\lambda_1^{n-1} - \lambda_2^{n-1})\right\}$$

$$\boxed{\begin{aligned} &P\left(H_{1,1} = (1, A),\ A_{2,1} = a\right) \\ &\qquad = \frac{1}{2}\left(1 - \underline{\theta}\right)\left\{P_{1A}(2,1) + P_{1A}(2,2)\right\} \\ &\qquad\quad + \frac{1}{2}\,\underline{\theta}\left\{P_{\bullet A}(2,1)P_{1\bullet}(2,2) + P_{1\bullet}(2,1)P_{\bullet A}(2,2)\right\} \end{aligned}}$$

Summary

- Markov Modelling of recombination fraction

$$\theta(p,q,r,n) = \frac{1}{2} + \frac{1}{2(\lambda_1 - \lambda_2)}\left\{(2r-1)(\lambda_1^n - \lambda_2^n) + \lambda_1\lambda_2(\lambda_1^{n-1} - \lambda_2^{n-1})\right\}$$

- “Probability inheritance” algorithm
  - Likelihood can be evaluated by a simple recursive formula
  - Good for large pedigree and large number of markers

Markov modelling of recombination fraction
+ Probability inheritance

References

- 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
- 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.

- Jurg Ott. Analysis of Human Genetic Linkage .

Thank you for your listening.