Lecture 15: January 31, 2002

Lecturer: Ron Shamir *Scribe: Tamir Tuller and Yair Sade*

15.1 Preface

In most organisms, genetic investigation can be carried out simply by breeding a pair of parents, and examining the offspring. Of course, parent selection for experimental breeding is not practical in humans. Therefore, humans genetics involves retrospective analysis of inheritance, and statistical analysis of hereditary pattern. Genetics analysis in human, will mostly serve as a low resolution method for localizing genes, but it is the only method available if the only prior information about the gene is some hypothesis about its function, or some presumed phenotype that results from a particular DNA sequence (*allele*) of that gene.

15.1.1 Diploids and Haploids

Humans and other mammals are *diploid* organisms, which means they contain two copies of each homologous chromosome in contrast to *haploids*, which contains only single copy of each chromosome. Sperm and egg cells are haploids called *gametes*, and created by a particular cell division process called *meiosis*. When a sperm and an egg combine, a diploid cell is created. The offspring inherits one copy of the genome from each parent.

15.1.2 Biology terms

- A *locus* is a site along the chromosome.

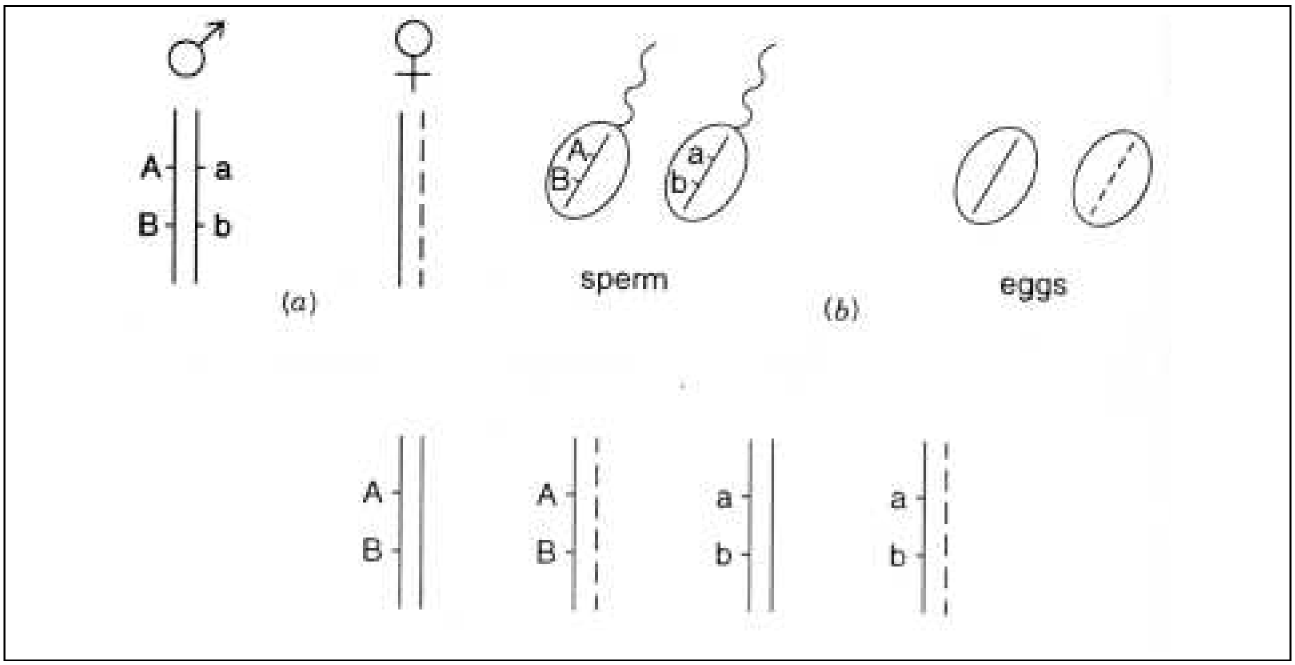

Figure 15.1: Source:[1] Basic scheme for chromosome segregation in the genetics of diploid organisms. (a) Parental genome. (b) Possible gametes. (c) Possible offspring (all equally probable).

- An *allele* is a particular DNA sequence variant at the locus of interest.
- When two homologous chromosomes in a diploid cell have the same allele at the locus of interest, the cell is said to be *homozygote* for that locus. Otherwise it is called *heterozygote*.
- The set of observable inherited characteristics of an individual (what we see) is called *phenotype*.
- The genetic constitution, which is the particular set of alleles inherited by the individual (what it inherits) is its *genotype*.
- Whenever a single allele determines the phenotype (even if present in one homologous chromosome), this allele is called *dominant*. Dominant disease alleles are rare.
- In contrast *recessive* allele is one that produces a particular phenotype only if it is present in both homologous chromosomes.
- The chance of an allele to produce the expected phenotype is called *penetrance*.
- The configuration of particular alleles on homologous chromosomes is called *phase*. For example, in figure 15.1 the phase is `AB/ab`.

15.1.3 Mendel's Laws

First Law: Segregation

Two alleles segregate from each other in gametes. Each gamete takes each allele with probability of 0.5.

Second Law: Independent Assortment

Genes of different traits are inherited independently.

Third Law:"The Third Law"

In heterozygote, one allele will dominate.

Mendel published the above laws in 1865. These laws were the basis of modern genetics. Today we know the Third Law fails on codominant, non bi-allelic traits. For example, let `A` be the dominant allele and `a` recessive one. In non bi-allelic traits, `Aa` and `AA` would be different, in contrast to regular traits.

15.2 describes basic mendelian pedigree patterns.

Figure 15.2: A sample pedigree. The squares describe the males and the circles the females. The shaded shapes describes a phenotype of hearing loss.

15.1.4 Linkage

Observable genomic traits are called *Markers*. Markers must have more than one form in population and each form is called *allele*. Markers on different chromosomes have 50% chance to be passed to offspring together. *Linkage* is the tendency of two markers on the same chromosome to be coinherited. Linkage and genetic analysis can help find a connection between known observable markers to a phenotypes such as a lethal disease.

15.1.5 Hardy – Weinberg equilibrium

Given random pedigrees, we would like to know the probability of a dominant trait eventually taking over. For simplicity, we shall make the following assumptions:

- No emigration (a close system).
- No cross generation breeding.
- No mutations.
- No natural selection that favors any trait.

Given two traits `A`,`a` and genotype probabilities $P_0(AA) = u \ P_0(Aa) = v \ P_0(aa) = w$, the probabilities for each possible first generation genotype are given in the table 15.1

Parents	Probability	Genotype
AA,AA	u^2	AA
AA,Aa	$2uv$	$\frac{1}{2}$(AA) + $\frac{1}{2}$(Aa)
AA,aa	$2uw$	Aa
Aa,Aa	v^2	$\frac{1}{4}$(AA) + $\frac{1}{2}$(Aa) + $\frac{1}{4}$(aa)

Aa,aa	$2vw$	$\frac{1}{2}$(Aa) + $\frac{1}{2}$(aa)
aa,aa	w^2	aa

Table 15.1: Probabilities for first generation genotypes

The corollaries are:

$$P_1(AA) = u^2 + uv + v^2/4$$
$$P_1(Aa) = 2uw + vw + uv + v^2/2$$
$$P_1(aa) = w^2 + wv + v^2/4$$

Let $p = u + v/2, q = 1 - p = w + v/2$, then probabilities can be written as:

$$P_1(AA) = p^2$$
$$P_1(Aa) = 2pq$$
$$P_1(aa) = q^2$$

Probabilities for possible genotypes in generation 2:

$$P_2(AA) = u_1^2 + u_1v_1 + v_1^2/4 = p^4 + 2p^3q + p^2q^2 = p^2$$
$$P_2(Aa) = 2u_1w_1 + v_1w_1 + u_1v_1 + v_1^2/2 = 2pq$$
$$P_2(aa) = w_1^2 + w_1v_1 + v_1^2/2 = q^2$$

Equilibrium is achieved after one generation. This equilibrium is called *Hardy-Weinberg equilibrium*. Note that the assumptions made are not realistic, especially the one that ignoring natural selection, although natural selection is slow.

15.2 Recombination

During meiosis, homologous chromosomes pair up and align to each other, physically attached at several random locations called *Holiday junctions*. Each junction can inflict a swap of chromosome segments. This phenomenon is called *meiotic recombination* (see figure 15.3).

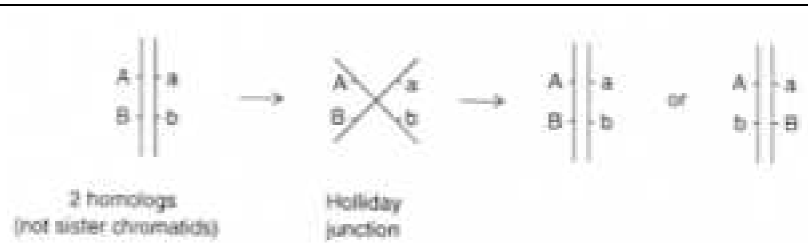

Figure 15.3: Source: [1] Meiotic recombination in a single pair of homolouges chromosomes. Formations and resolutions of a Holiday junction.

Meiotic recombination occurs at least once for each pair of homologous chromosomes. Recombination is important for evolution since it enables the expression of alleles that reside beside a dominant lethal mutation. Meiotic recombination in human occurs within a sequence of 1MB with probability of 1%. In different organisms, this probability varies. That low probability of recombination, and the small size of human families make genetic analysis more difficult in human. The closer two markers are, the larger the chances of coinheritance, since the chance of recombination occurring between them is lower. The probability that two markers were on the same haploid before meiosis, but on different ones after it called *recombination fraction*. Recombination fraction is denoted by θ.

In Figure 15.4, `A` is a marker for a disease allele `D`. Squares represent the males and circles the females. Shaded shapes represent disease affected offsprings. We first compute the recombination fraction without prior knowledge of the parental (first generation))genotypes and phenotypes. The likelihood of the recombination fraction assuming phase `DA/da` for the father is the probability of selecting the 3 affected offsprings out of the 10 offsprings that

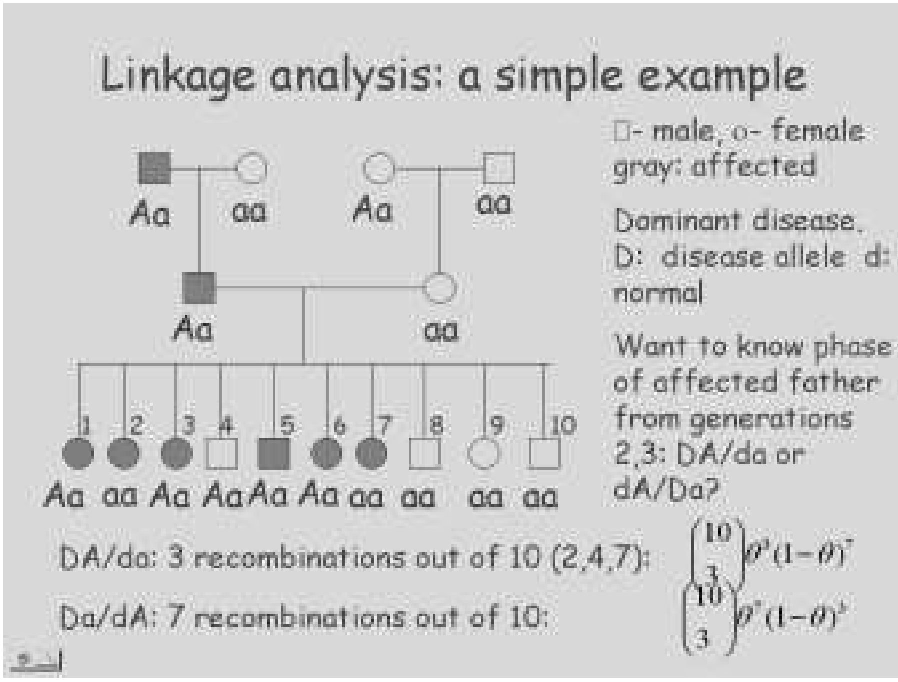

Figure 15.4: Recombination Fraction Calculation

match `DA/da` (Offsprings number 2,4,7):

$$L(\theta) = \frac{1}{2}\begin{pmatrix} 10 \\ 3 \end{pmatrix}[\theta^7(1-\theta)^3 + \theta^3(1-\theta)^7]$$

The value of θ maximizing $L(\theta)$ is $\theta = 0.318$.

Considering the first generation of the family we see that the affected father must be `DA/da` because the dominant disease was inherited from the `Aa` father. The likelihood becomes therefore,

$$L(\theta) = \begin{pmatrix} 10 \\ 3 \end{pmatrix}[\theta^3(1-\theta)^7]$$

And the value for optimal $L(\theta)$ is $\theta = 0.3$.

15.3 Pedigree Analysis

The following section is based on [5]. Define $L(\theta) = \Pr(Data|\theta)$ as the likelihood of θ given the data .The problem we study is the following :

Problem 15.1 (Maximum likelihood problem)

INPUT: Given two alleles in two loci A,a and B,b.

GOAL: Estimate θ the recombination fraction , maximizing $L(\theta)$.

15.3.1 The Bayesian Approach

In this section we show how to solve the maximum likelihood problem using Bayesian Approach .Let the prior distribution for θ be :

$$h(\theta) = \begin{cases} 1/2 & \text{with probability } 21/22 \\ uniform[0,1/2] \equiv u(\theta) & \text{with probability } 1/22 \end{cases} \tag{15.1}$$

Explanation:the probability that the loci of the two alleles are on two different chromosomes is 21/22 (human have 22 pairs of chromosomes ignoring the sex chromosomes).The probability for recombination in this case is 1/2 . If the loci of the two alleles lie on the same chromosome , the the recombination fraction is uniform distributed between 0 and 1/2 .

Given a pedigree F we calculate $p(\theta|F)$ as follow , according to Bayes theorem :

$p(F,\Theta) = p(F|\Theta) \cdot p(\Theta) \equiv L(F|\Theta) \cdot h(\Theta)$

$p(F) = \int p(F,\Theta)d\Theta = \int\limits_{0<\theta<1/2} L(F|\theta) \cdot u(\theta) + L(F|1/2) \cdot h(\theta = 1/2)$

$$p(\theta|F) = \frac{L(F|\theta)\cdot h(\theta)}{\int\limits_{0<\theta<1/2} L(F|\theta)\cdot u(\theta) + L(F|1/2)\cdot h(\theta=1/2)} \tag{15.2}$$

We next calculate the distribution of θ in the range : $0 < \theta < 1/2$, i.e when the two alleles are on the same chromosome.The prior density function in this range , based on (15.1) is :

$$h(\theta) = \left\{ 1/11 \quad 0 < \theta < 1/2 \right.$$

Denote $L(\theta) \equiv L(F|\theta)$, $L^*(\theta) \equiv \frac{L(\theta)}{L(\theta=1/2)}$, equation (15.2) becomes :

$$p(\theta|F) = \frac{L(\theta)\cdot h(\theta)}{\frac{1}{11}\int L(\theta)d\theta + L(1/2)\cdot\frac{21}{22}} = \frac{22L^*(\theta)h(\theta)}{2\int L^*(t)dt + 21}$$

Denoting $\Lambda \equiv 2\int\limits_0^{1/2} L^*(t)dt$,we derive the following formula for $p(\theta < 1/2|F)$:

$$p(\theta < 1/2|F) = \int\limits_0^{1/2} p(\theta|F)d\theta = \frac{22\int\limits_1^{1/2} L^*(t)\cdot\frac{1}{11}dt}{\Lambda + 21} = \frac{\Lambda}{\Lambda+21}$$

Bayesian inference : examples

Now we demonstrate the use of the relations and formulas we developed .

Example 1 : In this example we want to calculate the probability that $\theta < 1/2$ (the two alleles are on the same chromosome), when we observe 1 recombination out of 12 meioses:

$$L(\theta) = 12\theta(1-\theta)^{11}$$

$$\Lambda = 2\int_0^{1/2} L^*(t)dt = 2 \cdot \frac{1}{12\cdot(1/2)} \int_0^{12} 12t(1-t)^{11}dt =$$

$$= 2^{13} \cdot 0.0064 = 52.043$$

$$p(\theta < 1/2) = \frac{52.043}{52.043+21} = 0.714$$

Example 2: In this example we calculate how many meioses without recombination we need to observe if we want $p(\theta < 1/2) \approx 0.95$:

Let n denote the number of meioses , we get : $L(\theta) = (1-\theta)^n$

$$\Lambda = 2^{n+1} \int_0^{1/2} (1-\theta)^n d\theta = 2^{n+1} \frac{1-2^{-(n+1)}}{n+1} \approx \frac{2^{n+1}}{n+1}$$

$$\Rightarrow p(\Theta < 1/2|F) = \frac{2^{n+1}/(n+1)}{2^{n+1}/(n+1)+21} = 0.95$$

In this case : $n = 11 \Rightarrow p(\Theta < 1/2|F) = 0.942$, $n = 12 \Rightarrow p(\Theta < 1/2|F) = 0.968$. Hence We need $\geq$11 meioses even for $p(\Theta < 1/2|F) \approx 0.95$.

15.3.2 Likelihood Ratio Inference

Suppose we wish to test the following two hypotheses :

$H_0 : \theta = 1/2$, two alleles are not on the same chromosome.

$H_1 : \theta < 1/2$, two alleles are on the same chromosome.

Define the likelihood ratio $\lambda = \frac{L(\theta=1/2)}{\max_\theta L(\theta)}$ and let $Q = -2\ln\lambda$.

The distribution of Q is : $1/2 \cdot \chi^2 + 1/2\,\{0\}$ Where χ^2 denotes chi-square density , and $\{0\}$ denotes delta function at the point 0 which is due to the delta in equation (15.1) . Hence $p\{-2\ln\lambda > x_0\} = 1/2 \cdot p\{\chi^2 > x_0\}$.

Define $LOD(\log odds) = \log_{10}\frac{1}{\lambda} = -0.4343\ln\lambda = 0.2171 \cdot Q = \frac{Q}{4.6}$.

Typically need $LOD \geq 3$ for rejecting H_0: $1/2 \cdot pr\{\chi^2 \geq 4.6 \cdot 3\} \approx 0.0001$. We need about 20 markers per chromosome , total 440 . And by Bonferroni inequality , bound false alarm rate by $440 \cdot 10^{-4} = 0.044$.

Alternatively , we can get the same results by using the following relations :

Let $Z(\theta) = \log_{10} \frac{L(\theta)}{L(\theta=1/2)}$,we need $Z_{\max} = \max_\theta Z(\theta) = LOD$. For independent families $F_1, .., F_n$: $Z(\theta) = \sum_{i=1}^{n} Z(\theta_i)$,and in our case we have $n = 440$.

15.4 Inference of Haplotypes

15.4.1 SNP and Haplotypes

The follow section is based on [4].
SNP , Single Nucleotide Polymorphism ,is a single nucleotide sites where at least 2 (out of 4) different nucleotides occur in a large percentage of the population. There is a hope that *SNP* screens will help identify genetic elements of complex diseases. In *diploids* , as mentioned before, there are 2 copies of each chromosomes. Those copies are not completely identical, each copy of it called *haplotype and the combination of them is called* Genotype.

In complex diseases (diseases that are affected by more than single gene), *haplotype* data is more informative than *Genotype* data. Obtaining the *haplotype* out of the *genotype* data by experiments is hard to be done and expensive, therefore we would like to have an algorithm that inputs *genotypes* (i.e. *haplotype* pairs) from individuals and infer the population's *haplotypes*.

15.4.2 Clark's method

Given n vectors each of length m where each value is either $0, 1, 2$. Each position in the vector is associated with locus in the chromosome. The state of the chromosome can be either 0 or 1. The associated vector position is 0 or 1 if the chromosome site has that state in both copies, and the value of 2 of both states are present (in *heterozygous* site).Values of $0, 1$ called "*resolved*" and values of 2 called "*ambiguous*". A vector, which all its values are resolved is called a *resolved* vector and otherwise called *ambiguous*. Given two non-identical *ambiguous* vectors N and R, the *conflation* of the two vectors creates a an *ambiguous* vector A, with 0 (respectively 1) values of both N,R had 0 in that index, and 2 in each site the values differs.

The idea of Clark's method [2] is to find few *haplotypes* to explain data justified by population genetics arguments and by simulations.

Let S be the current collection of vectors. Find $a \in S$ ambiguous with h heterozygous sites and $r \in S$ resolved vector that equals to one of the 2^h potential resolutions of a.

The method consists of the following steps :

1. Define $\widehat{a}$ as r_i for all resolved i, and as $\bar{r}_i$ for each unresolved position i.
2. $S = S \setminus \{a\} \cup \{\widehat{a}\}$
3. Goto 1 unless all the vectors are resolved .

when we stop the resolved vectors are the wanted *haplotypes*.

Example for step 1: if $a = 0212$ and $r = 0110$ we will choose :

$\widehat{a} = 0011$. Clark's method presents the following problem:

Problem 15.2 (Maximum Resolution Problem (MR))

INPUT: A Set of given vectors.

QUESTION: What is the maximum number of ambiguous vectors that can be resolved by applying Clark's inference rule?

The MR problem is known to be NP hard [3], and in fact Max-SNP complete [3].

15.4.3 Graph-Theoretic view of the MR problem

Given the input of n vectors of length m, we will create the following directed graph G containing a set of nodes $N(A)$, for each ambiguous node, and a set of nodes, I, containing one node for each resolved vector in the input. In details, a node r in I for each resolved

$r \in S$; a set of nodes for each unresolved $a \in S$, where $N(a)$ is all possible resolutions of a (2^{h_a}); and an edge $v \to v'$ iff :

1. v' in some $N(A)$ (not in I).

2. Applying Clark's rule to a and v gives v'.

Figure 15.5 demonstrate such a graph.

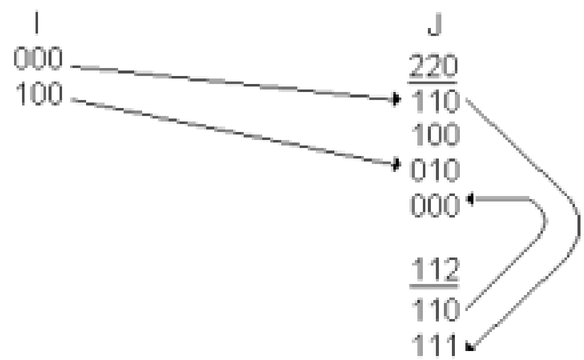

Figure 15.5: MR Problem - Graph example.

Maximum Resolution on Graphs (MRG).

The following example show that the order of choices of a,r affects the number of resolved ambiguous vectors:

S: 000 100 200 112

first possibility : $\begin{matrix} 000 \\ 220 \end{matrix} \to 110, \begin{matrix} 112 \\ 110 \end{matrix} \to 111, \to 000, 100, 110, 111.$

second possibility : $\begin{matrix} 100 \\ 220 \end{matrix}$ $\to$ 010,112 can not be resolved by 010,000,100.

This gives rise to the following problem:

Problem 15.3 Maximum Resolution on Graphs (MRG)

INPUT: A graph G from the input vectors.

PROBLEM: Find the largest number of nodes reachable by vertex-disjoint directed trees rooted at nodes in I, and for each ambiguous `a` at most one node in $N(a)$ is reached.

Unfortunately MRG problem is NP hard, even the input to the problem is exponential ($N(a)$ consists of 2^{h_a} possible resolutions).

15.4.4 Integer Programming Formulation of MRG

Since solving MRG takes exponential time, we would like to show more realistic methods to solve the MRG problem. Define $\overline{G}$ as subgraph of G in which nodes not reachable from I are removed along with the incident edges.

$\overline{G}$ can be built using mathematical integer programming in $O(mn^2 + \left|E(\overline{G})\right|)$ time. Even though $\overline{G}$ edges number is still exponential, practice shows that it is much more efficient than the naive approach (of building $\overline{G}$ from G.

15.4.5 Network Flow Formulation

Following [4] we shall formalize the MRG problem as a network flow problem with non-linear constraints. Define $\overline{G}$ edges with capacity of ∞. Add source node and sink node to $\overline{G}$. Define directed edges entering I node from source node with capacity ∞. Define all edges from each node not in I to the sink node with capacity 1.

Clearly, a feasible solution of MRG that reaches q nodes (and hence resolve q ambiguous vectors) defines a source-sink flow with value of q exactly. However the converse does not hold, since we have not excluded the possibility of reaching more than one node from any set $N(A)$. To solve that we use linear programming formulation of the flow problem described above.

Let x_e denote the variable flow on edge e. Then for every pair of edges e, e' entering nodes in same set N(A), we add a non-linear constraint $x_e \cdot x_{e'} \leq 1$ to the network flow formulation. An integer solution to that mathematic formulation exactly solves the MRG problem, but solving formulation with non-linear constraints is hard to solve in practice.

15.4.6 An Alternative Formulation

We create a binary variable y_v for each node in $\overline{G}$. We'll add inequalities to insure that the variable set to 1 corresponding to nodes that are reached in a solution of the MRG problem, hence corresponding to haplotypes that are inferred by application of Clark's inference rule. For each set $N(A)$ we add the inequality:

$$\sum_{v \in N(A)} y_v \leq 1$$

These inequalities assure that at most one node in $N(A)$ will be reached. For each node $w \notin I$ let $P(w)$ be the set of nodes with an edge directed into w, which means node v is in $P(w)$ iff there is a directed edge $v \rightarrow w$ in G.

For each node $w \notin I$ we add inequality

$$y_w \leq \sum_{v \in P(w)} y_v$$

These inequalities are called "predecessor" constraints. These constraints assure that a node w not in I cannot be reached without reaching at least one node acting as a predecessor of w, i.e., leading to w. The objective function is: $Maximize \sum_v y_v$

It is simple to see that the optimal solution to the MRG problem defines a feasible solution to this integer programming formulation. However if $\overline{G}$ contains directed cycles, the integer programming formulation is not guaranteed to solve the MRG problem. But usually the input date does not contain directed cycles, and if there are ones, and there are ways of obtaining solution to MRG by modifying solution found with directed cycles. Linear programming is a NP-hard problem but with relaxation of the problem, good results can be found in reasonable time.

Bibliography

[1] C. R. Cantor and C. L. Smith. *Genomics.* John Wiley and Sons, Inc., 1999.

[2] A. Clark. Inference of haplotypes from pcr-amplified samples of diploid populations. *Mol. Biol.*, 7:111–122, 2000.

[3] D. Gusfield. Inference of haplotypes from peramplified samples of diploid pupulations: Complexity and algorithms. *UC Davis, Department of Computer Science technical report*, techreport cse-99-6, 1999.

[4] D. Gusfield. A practical algorithm for optimal inference of haplotypes from diploid populations. *American Association of Artificial Intelegence*, 2000.

[5] M. C. K. Yang. *Introduction to Statistical Methods in Modern Genetics.* Gordon and Breach Science Publishers, 2000.