SNP

SINGLE-NUCLEOTIDE POLYMORPHISM - SITE IN HUMAN GENOME WHERE TWO NUCLEOTIDES COMMONLY OCCUR

HAPLOTYPE SPECIFIES NUCLEOTIDES AT n SELECTED POLYMORPHIC SITES ON A CHROMOSOME. THE PATERNAL AND MATERNAL CHROMOSOMES EACH SPECIFY A HAPLOTYPE

GENOTYPE SPECIFIES PAIR OF NUCLEOTIDES AT A SITE, BUT NOT THEIR ASSORTMENT BETWEEN HAPLOTYPES

HAPLOTYPES

A C G & T
T C C G C

GENOTYPE

{A/T} C {C/G} G {C/T}

PHASE PROBLEM

GIVEN THE GENOTYPES OF A POPULATION OF INDIVIDUALS, COMPUTE THEIR HAPLOTYPES

GENOTYPE SPECIFIES PAIR OF NUCLEOTIDES AT A SITE, BUT NOT THEIR ASSORTMENT BETWEEN HAPLOTYPES

HAPLOTYPES

A C G & T
T C C G C

GENOTYPE

{A/T} C {C/G} G {C/T}

PHASE PROBLEM

GIVEN THE GENOTYPES OF A POPULATION OF INDIVIDUALS, COMPUTE THEIR HAPLOTYPES

BASED ON OBSERVATION THAT GENOMIC DNA CAN BE PARTITIONED INTO LONG BLOCKS WHERE RECOMBINATION HAS BEEN RARE AND THE NUMBER OF DISTINCT HAPLOTYPES IS SMALL.

ASSUME THAT, WITHIN A BLOCK, HAPLOTYPES ARE GENERATED BY A PERFECT PHYLOGENY

HAP ESKIN, HALPERIN

MATHEMATICAL MODEL

- m SITES
- AT EACH SITE, THE TWO NUCLEOTIDES DENOTED BY 0 AND 1

HAPLOTYPE $h \in \{0,1\}^m$

GENOTYPE $g \in \{0,1,2\}^m$

2 MEANS THAT BOTH NUCLEOTIDES OCCUR

THE PAIR $(h^{(1)}, h^{(2)})$ OF HAPLOTYPES IS CONSISTENT WITH GENOTYPE g IF $\forall j$

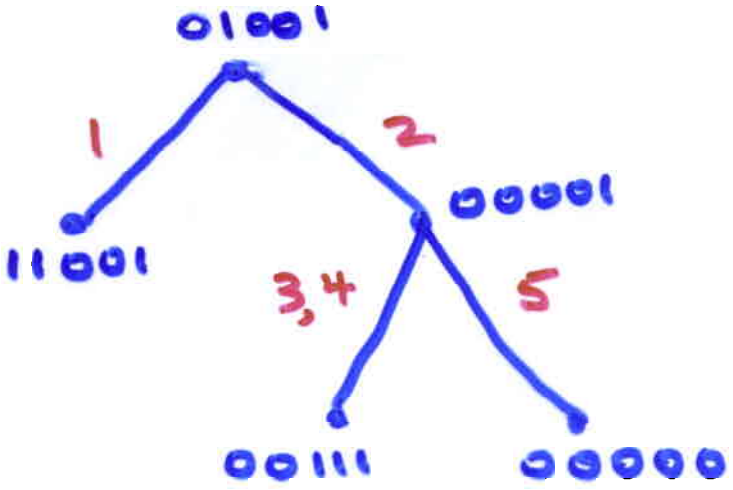
$$g_j = 0 \Rightarrow h_j^{(1)} = h_j^{(2)} = 0$$

$$g_j = 1 \Rightarrow h_j^{(1)} = h_j^{(2)} = 1$$

$$g_j = 2 \Rightarrow h_j^{(1)} \neq h_j^{(2)}$$

PERFECT PHYLOGENY

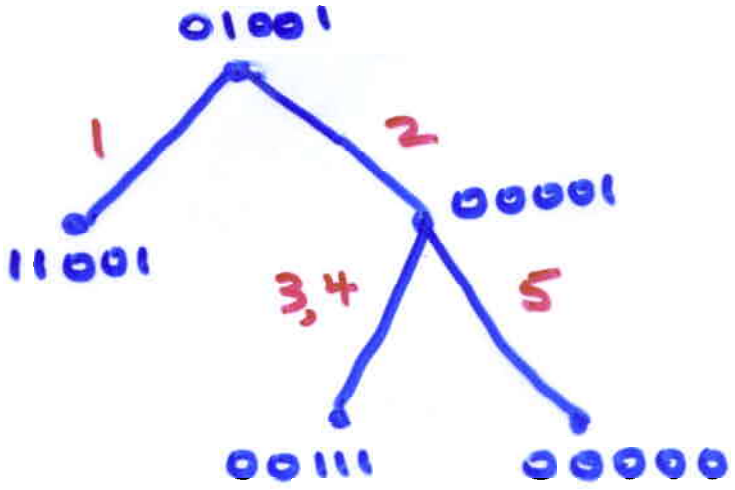
NO SITE HAS MUTATED MORE THAN ONCE



GUSFIELD POSTULATES THAT, WITHIN EACH BLOCK, THE HAPLOTYPES HAVE EVOLVED WITHOUT RECOMBINATION ACCORDING TO A PERFECT PHYLOGENY

PERFECT PHYLOGENY

NO SITE HAS MUTATED MORE THAN ONCE



GUSFIELD POSTULATES THAT, WITHIN EACH BLOCK, THE HAPLOTYPES HAVE EVOLVED WITHOUT RECOMBINATION ACCORDING TO A PERFECT PHYLOGENY

$n \times m$ HAPLOTYPE MATRIX H

$H = (h_{ij})$ $h_{ij} \in \{0,1\}$

ROW = HAPLOTYPE

- H IS REALIZABLE IF ITS ROWS BELONG TO A PERFECT PHYLOGENY

DUO 2×2 SUBMATRIX

e.g. $\begin{matrix} & s & t \\ i & 0 & 1 \end{matrix}$

H IS REALIZABLE IFF, IN EACH PAIR OF COLUMNS, AT MOST THREE DISTINCT DUOS OCCUR

(BUNEMAN)

$n \times m$ GENOTYPE MATRIX G

$G = (g_{ij})$ $\quad g_{ij} \in \{0,1,2\}$

ROW = GENOTYPE

G IS REALIZABLE IF EACH GENOTYPE CAN BE REPLACED BY A CONSISTENT PAIR OF HAPLOTYPES SO THAT THE RESULTING HAPLOTYPE MATRIX IS REALIZABLE

G → H

$$\begin{bmatrix} 2000 \\ 2202 \\ 2220 \\ 0122 \end{bmatrix}$$

$$\begin{bmatrix} 0000 \\ 1000 \\ 1000 \\ 0101 \\ 1000 \\ 0110 \\ 0101 \\ 0110 \end{bmatrix}$$

GUSFIELD (2002) SHOWED THAT REALIZABILITY OF G CAN BE TESTED IN SLIGHTLY SUPERLINEAR TIME BY REDUCTION TO TUTTE'S TREE REALIZATION PROBLEM: GIVEN A COLLECTION $\{E_k\}$ OF SUBSETS OF A FINITE SET E, IS THERE A TREE WITH EDGE SET E IN WHICH EACH E_k IS THE EDGE SET OF A PATH?

PRACTICAL $O(m^2n)$ ALGOR'S

BAFNA, GUSFIELD, LANCIA, YOOSEPH JCB 2003

ESKIN, HALPERIN, KARP RECOMB 2003

SIMPLE GENOTYPE REALIZATION ALGORITHM

$G \rightarrow H$

DUOS IN G INDUCE DUOS IN H

[0,2] INDUCES [0,1]

[0,2] " [0,0], [0,1]

[2,1] " [0,1], [1,1]

[2,2] CAN BE RESOLVED EITHER EQUALLY [0,1], [1,0] OR UNEQUALLY [0,0], [1,1]

IN A GIVEN PAIR j, k OF COLUMNS ALL OCCURRENCES OF [2,2] MUST BE RESOLVED THE SAME WAY

$$L(j,k) = \begin{cases} 0 & \text{IF } [2,2] \text{ DUOS RESOLVED EQUALLY} \\ 1 & \text{IF RESOLVED UNEQUALLY} \end{cases}$$

CONSTRAINTS

- IF $\overset{j\ k}{[0,1]}$ AND $\overset{j\ k}{[1,0]}$ ARE INDUCED THEN $L(j,k) = 1$
- IF $\overset{j\ k}{[0\ 0]}$ AND $\overset{j\ k}{[1\ 1]}$ ARE INDUCED THEN $L(j,k) = 0$

IF SOME GENOTYPE CONTAINS 2's IN COL's j, k, l THEN

- $L(j,k) + L(j,l) + L(k,l) \equiv 0 \mod 2$

REALIZABILITY CAN BE TESTED BY SOLVING LINEAR EQUATIONS OVER GF[2]

PARTIAL HAPLOTYPE MATRIX

$\hat{H} = (\hat{h}_{ij})$

$\hat{h}_{ij} \in \{0, 1, x\}$

x INDICATES MISSING DATA

$\hat{H}$ IS REALIZABLE IF ONE CAN OBTAIN A REALIZABLE HAPLOTYPE MATRIX H BY REPLACING EACH x IN $\hat{H}$ BY 0 OR 1

- REALIZABILITY OF $\hat{H}$ IS NP-COMPLETE, BUT DECIDABLE IN LINEAR TIME IF ONE ROW OF H IS GIVEN

PE'ER, PUPKO, SHAMIR, SHARAN
SICOMP (TO APPEAR)