Estimate haplotypes for multiallelic present-absent loci

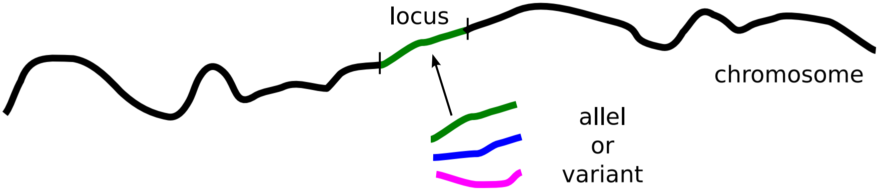
Robert Nowak

Warsaw University of Technology
Electronics and Computer Science Department
Electronics Systems Institute

Jun 16, 2008

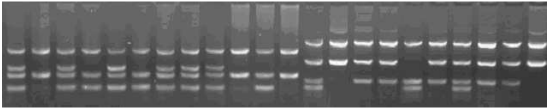
introduction

- locus - fixed position in chromosome
- allele - variant of gene in locus
- haplotype - alleles at multiple loci transmitted together

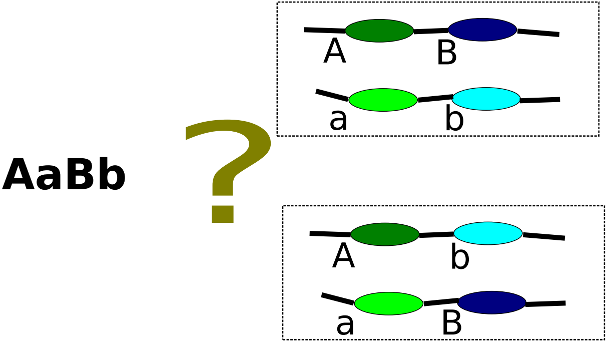


haplotype analyze

Popular typing methods: amplification, selected PCR primers



- lack of phase information

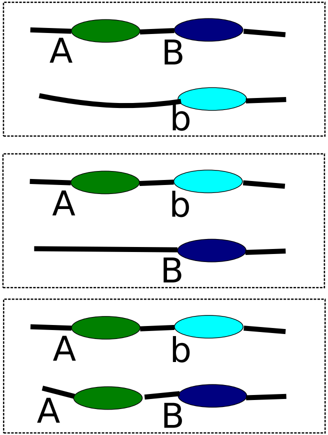


null variants in haplotype analyze

present - absent polymorphisms

- ambiguous to the heterozygous status
- important in medicine

ABb ?



probabilistic analyse of population data

Sample: n individuals, where G genotypes observed:

$$S = (n_1, n_2, ..., n_G), \text{ where } \sum_{j=0}^{G} n_j = n \tag{1}$$

haplotype frequency estimation h_i (maximum likelihood approach):

$$\underset{h_1,h_2,...,h_H}{\arg\max}\ P(S\ |h_1, h_2, ..., h_H) = \underset{h_1,h_2,...,h_H}{\arg\max} \prod_{j=1}^{G} (\sum_{i=0}^{r_j} z_{mn})^{n_j} \tag{2}$$

$$\text{where } z_{mn} = \begin{cases} h_m^2 & \text{for } m = n \\ 2\ h_m h_n & \text{for } m \neq n \end{cases}$$

haplotype estimation based on population analyzes

- no need multi-generation families

Problem not new:

- `Arlequin` (http://cmpg.unibe.ch/software/arlequin3/)
- `PHASE` (http://stephenslab.uchicago.edu/software.html)
- `Haplo-IHP` (http://www.soph.uab.edu/Statgenetics/)

does not considered multiallelic loci with null variants

computational difficulty

R = number haplotype pairs

$$R = \frac{1}{2}H \ * (H+1), \text{ where } H = \prod_{i=1}^{k} l_i \tag{3}$$

G = number genotypes:

$$G \ = \prod_{i=1}^{k} \frac{(l_i - \delta_i)(l_i + 1 \ - \delta_i) + 2\delta_i}{2}, \delta_i = \begin{cases} 1 & \text{loci with null variants} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

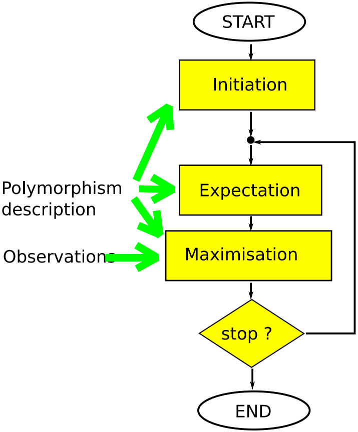
Number haplotype pairs for given genotype j:

$$r_j = \begin{cases} 2^{s_j - 1} * 3^{t_j} & \text{for } s_j > 0 \\ \frac{3^{t_j} + 1}{2} & \text{for } s_j = 0 \end{cases} \tag{5}$$

algorithm

Algorithm EM:

- iteration:
  - expectation value of unknown parameters (step E)
  - maximisation the goal function (step M)
- stop when no changes in adjoining steps
- local optimization algorithm (multiple starting points)
- fast convergence



implementation NullHap application

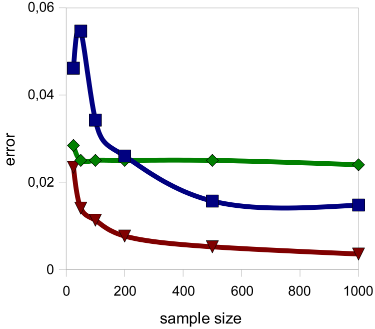
- portable
- efficiency (C++, boost)
- open source (LGPL)
  `http://nullhap.sourceforge.net`
- binaries: Windows 2000/XP/Vista, Linux (Debian)

Testing:

- generated data sets
- real data sets:
  - HLA (100 individuals, 2 multi-allelic loci, no null variants, source: Arlequin)
  - KIR (200 individuals, 10 biallelic loci with null variants, source: Haplo-IHP)

population size effect

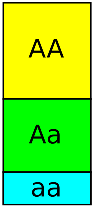
- generate population of *n* individuals with given haplotype frequency
- estimate haplotype frequency by application



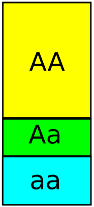
$$error = \frac{1}{N}\sum_{i=1}^{N}|x - x^*|$$

- 2 loci, 6 haplotypes
- 3 loci, 24 haplotypes, $P_{max} = 0.3$
- 3 loci, 40 haplotypes, $P_{max} = 0.025$

departure of Hardy-Weinberg proportion

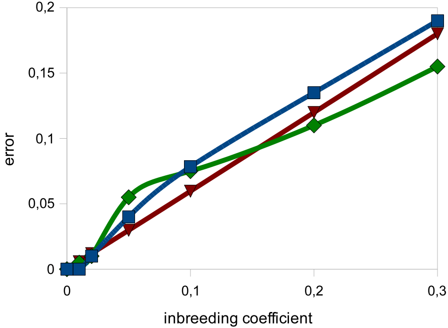


$$P(AA) = P_A^2$$
$$P(Aa) = P_A P_a$$
$$P(aa) = P_a^2$$



$$P(AA) = P_A^2(1-f) + P_a f$$
$$P(Aa) = P_A P_a (1-f)$$
$$P(aa) = P_a^2(1-f) + P_A f$$

f - inbreeding coefficient



$$error = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{x-x^*}{x}\right|$$

- 2 loci, 6 haplotypes
- 3 loci, 24 haplotypes
- 3 loci, 40 haplotypes

summary

The type of analyzed loci

program name	biallelic	multiallelic	null variants
Arlequin	+	+	-
PHASE	+	+	-
Haplo-IHP	+	-	+
NullHap	+	+	+

Only the NullHap can handle multiallelic loci with null variants.

Acknowledgement

- dr hab. Rafał Płoski

r.m.nowak@elka.pw.edu.pl

Thank you