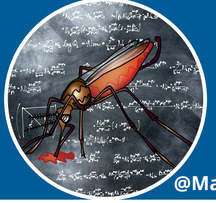

738

@MalariaMath

Estimating Multiplicity of Infection, allele Frequencies and Prevalences from incomplete data

Meraj Hashemi, Kristan A. Schneider

Faculty of Applied Computer- and Biosciences, University of Applied Sciences Mittweida

2020 ANNUAL MEETING NOVEMBER 15–19 VIRTUAL MEETING

HOCHSCHULE MITTWEIDA University of Applied Sciences

Abstract

Multiplicity of infection (MOI) refers to the number of super-infections due to the occurrence of multiple infectious contacts. Accurate estimates of MOI based on SNP or microsatellite data are highly desirable, as it is revealing clinical, genetic and epidemiological purposes. Due to limitations and difficulties of molecular methods for sequencing SNPs or calling STRs, incomplete information or undetected alleles are common. Although, estimating MOI and allele (or lineage) frequencies/prevalences is recognized to be fundamental in malaria genetic studies, unobserved and incomplete genetic/molecular information is not properly accounted by current methods. This potentially biases results and reduces confidence of estimates. Here, we develop a statistical model to estimate MOI and allele frequencies and prevalences from molecular data containing incomplete information. The model assumes that alleles/lineages present within an infection will be detected in a blood sample independently of each other, potentially leading to samples with empty information. We apply the expectation-maximization (EM) algorithm to derive maximum-likelihood estimates (MLE) of the MOI distribution, allele-frequency spectrum and prevalences. The distinct feature of this method is that it incorporates patient blood samples with completely missing information. The method has desirable analytical (asymptotic) properties. Furthermore, as shown by a systematic numerical study, the method performs well for realistic sample sizes. As an example, the method is applied to a data set previously collected in Asembo Bay, Kenia. An implementation of the method to estimate allele frequency spectra at a single SNP or microsatellite locus alongside MOI in R is provided.

1. Background - MOI

- ⇝ presence of different parasite haplotypes in infections due to multiple infectious contacts (**multiplicity of infection – MOI**; **Fig. 1**).
- **The importance of MOI:**
  - mediates the amount of recombination;
  - associats with disease severity;
  - to accurately monitor transmission intensity and prevalence (of drug-resistance associated mutations);
  - to determine connection between frequency and prevalence of drug resistance;
  - to track changes in seasonal malaria.
- **Freqency and prevalence are estimated from:**
  - ⇝ alleles/haplotypes determined by SNP barcodes and neutral markers (microsatellites);
  - to monitor drug-resistant mutations/haplotypes (characterized by point mutations at several codons);
  - for further population-genetic analysis, e.g., genetic diversity, heterozygosity, linkage disequilibria.
- MOI and frequencies are estimated from molecular data of malaria positive blood samples. Molecular assays are not perfect (inaccurate lab protocols, biochemical failure) ⇝ **incomplete data**.

2. Statistical Model

- n pathogen haplotypes: $A_1, \ldots, A_n$, with frequencies $p_1, \ldots, p_n$;
- natural census point for frequencies: salivary glands of mosquito;
- infective events are rare and independent;
- at each infective event only one parasite haplotype (allele) infects the host ⇝ another haplotype is added (super-infection);
- a haplotype identifies with an allele;
- consider only disease-positive samples;
- ⇝ MOI follows a **Poisson distribution** truncated at zero:

$$P(\text{MOI} = m) = \frac{1}{e^{\lambda} - 1} \frac{\lambda^m}{m!}, \quad m > 0;$$

- m_k is the number of super-infections by allele A_k ⇝ $m = \sum_{k=1}^{n} m_k$;
- given MOI, number of super-infections with alleles $A_1, \ldots, A_n$ are multinomially distributed $\sim$ Mult$(m, p_1, \ldots, p_n)$.

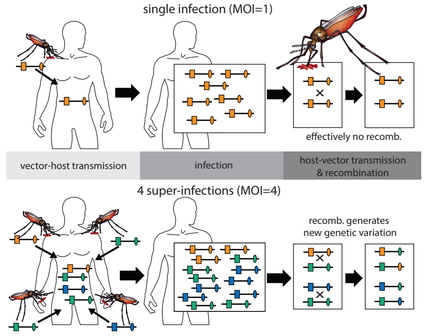

Fig. 1: Single infection and super-infection with MOI = 4.

3. Data collection

- For a given infection, m_k's are unobservable;
- only the **absence**/**presence** of alleles in a blood sample is observed (**Fig. 2**);
- y_k indicates the absence/presence of allele A_k;
- the observed data from a blood sample is denoted by $\boldsymbol{y} = (y_1, \ldots, y_n)$;
- the probability of observing $\boldsymbol{y}$:

$$Q_{\boldsymbol{y}} := P(\boldsymbol{y}) = \frac{1}{e^{\lambda} - 1} \prod_{k=1}^{n} (e^{\lambda p_k} - 1)^{y_k};$$

- the prevalence of allele A_k:

$$P(y_k = 1) = \frac{1 - e^{-\lambda p_k}}{1 - e^{-\lambda}};$$

- a molecular dataset consists of N independent observations $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(N)}$.

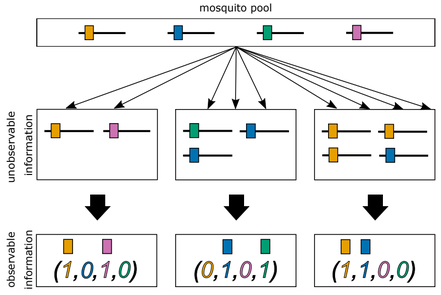

Fig. 2: The relation between unobserved and observed data.

4. Accounting for incomplete data

- An allele can be present in an infection but remain unobserved;
- detection of an allele is independnt of the number of super-infections with that allele;
- the observed data accounting for incomplete data is $\boldsymbol{x} = (x_1, \ldots, x_n)$ (**Fig. 3**);
- $x_k \leq y_k$ for all k ⇝ $\boldsymbol{x} \preceq \boldsymbol{y}$;
- detection failure occurs with probability ε, i.e., $P(x_k = 0 | y_k = 1) = \varepsilon$;
- the probability of observing $\boldsymbol{x}$ (non-empty record):

$$\tilde{Q}_{\boldsymbol{x}} := Q_{\boldsymbol{x}} (1 - \varepsilon)^{|\boldsymbol{x}|} \prod_{k=1}^{n} \left(\varepsilon (e^{\lambda p_k} - 1) + 1 \right)^{1 - x_k};$$

- the probability of observing an **empty record**:

$$\tilde{Q}_{\boldsymbol{0}} = \frac{1}{e^{\lambda} - 1} \left(\prod_{k=1}^{n} \left(\varepsilon (e^{\lambda p_k} - 1) + 1 \right) - 1 \right).$$

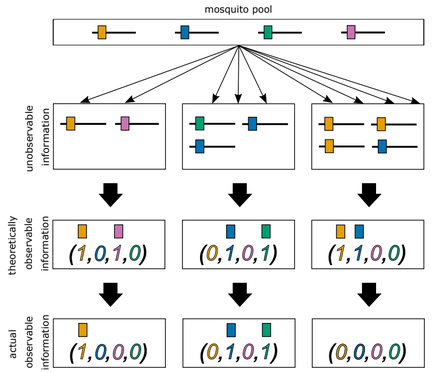

Fig 3. The relation between actual and theoretical observations.

5. The Maximum-Likelihood Estimate (MLE) - Prevalence

Likelihood function too complicated:

⇝ **EM Algorithm:**

- Start from initial values $\lambda^{(0)}, \boldsymbol{p}^{(0)}$ and $\varepsilon^{(0)}$.
- The next iteration is derived as

$$p_k^{(t+1)} = \frac{U_k^{(t)}}{\sum_{k=1}^{n} U_k^{(t)}} \quad \text{and} \quad \varepsilon^{(t+1)} = \frac{1}{1 + \frac{1}{V^{(t)}} \sum_{k=1}^{n} N_k}.$$

- $\lambda^{(t+1)}$ is derived iteratively by a Newton method

$$\lambda_{s+1} = \lambda_s - \frac{\lambda_s - \frac{W^{(t)}}{N}(1 - e^{-\lambda_s})}{1 - \frac{W^{(t)}}{N} e^{-\lambda_s}}.$$

- $U_k^{(t)}, V^{(t)}$ and $W^{(t)}$ are functions of $\lambda^{(t)}, \boldsymbol{p}^{(t)}$ and $\varepsilon^{(t)}$.
- Repeat until convergence.

- For a molecular dataset:
  - N_k := "number of samples infected by allele A_k";
  - N_+ := "number of samples with at least one detected allele";
- the observed prevalence of allele A_k: $\frac{N_k}{N_+}$;
- the true prevalence of allele A_k:

$$P(x_k = 1) = \frac{(1 - \varepsilon)(1 - e^{-\lambda p_k})}{1 - \prod_{k=1}^{n} \left(\varepsilon (1 - e^{-\lambda p_k}) + e^{-\lambda p_k} \right)};$$

- **Prevalence**: the probability of allele's presence in an infection (within the population of disease-positive individuals), whereas **frequency**: the relative abundance of the allele in parasite population.

6. Conclusions

- The MLE derived by the EM algorithm exists & is unique except in pathological cases;
- the incomplete-data model can also be applied to multi-loci models;
- prevalence is mediated by MOI and the probability of detection failure;
- empty records are informative of allele frequencies and MOI ⇝ preserved in data processing.

Contact

Meraj Hashemi, MSc.
University of Applied Sciences Mittweida
Technikumplatz 17, 09648, Mittweida, Germany
e-mail: mhashemi@hs-mittweida.de
p: +49 3727 58-1036

Prof. Dr. Kristan A. Schneider
University of Applied Sciences Mittweida
Technikumplatz 17, 09648, Mittweida, Germany
e-mail: kristan.schneider@hs-mittweida.de
p: +49 3727 58-1057

Acknowledgements

Hanns Seidel Stiftung DAAD DFG

This research was supported by grants from the DAAD ("Mathematics against malaria within the AIMS network", project-ID 57417782), DFG ("Ökologisch nachhaltige Wertschöpfungsketten in in der Landwirtschaft durch Optimierung des Insektizid-Gebrauchs aufgrund von automatisiertem Schädlings-Monitoring"), the SMWK ("Vorlaufforschung Technologieentwicklung 4.0") and a scholarship from the Hans-Seidel-Stiftung (HSS).

STAATSMINISTERIUM FÜR WISSENSCHAFT UND KUNST | Freistaat SACHSEN