A mathematical model for the analysis of LdMNPV isolate compositions

Michael Rihlmann

As part of the publication

Amplicon-based sequence analyses of single nucleotide polymorphisms reveal the genetic structure of baclovirus forest populations

in *Virus Evolution*

1 Model description

We consider the following problem: Let I be a LdMNPV-isolate with n_{pos} SNP-positions and B be a set of pure genotypes $V^i, i = 1, ..., n_{\text{pure}}$. Find a frequency distribution of the pure genotypes that are present in the isolate I such that it can be explained as a linear combination of the pure genotypes as much as possible.

A given isolate I is modelled as $n_{\text{pos}} \times 4$-matrix of real number in $[0,1]$ representing the relative frequencies that each of the four nucleotide A, C, G, T was found with at this position:

$$\begin{pmatrix} s_{1A} & s_{1C} & s_{1G} & s_{1T} \\ \vdots & \vdots & \vdots & \vdots \\ s_{nA} & s_{nC} & s_{nG} & s_{nT} \end{pmatrix} \in [0,1]^{n_{\text{pos}} \times 4}$$

Example 1.1. Consider the following example isolate with two SNP-positions:

	A	C	G	T
1	2	1	1	0
2	0	0	0	4

It is represented by the 2 × 4-matrix

$$I = \begin{pmatrix} 0,5 & 0,25 & 0,25 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \in [0,1]^{2\times 4}.$$

◁

The n_{pure} pure genotypes are modelled analogously as matrices $V^i \in \{0,1\}^{n_{\text{pos}}\times 4}$ for $i = 1, .., n_{\text{pure}}$. Hence, for all $i = 1, ..., n_{\text{pure}}$ each row of V^i contains exactly one entry equal to 1.

Example 1.2. The pure genotypes *AT* and *GT* are modelled as

$$V^1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \in \{0,1\}^{2\times 4}$$

and

$$V^2 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

◁

We introduce the optimization variables $x_i \in [0,1]$ for $i = 1, ..., n_{\text{pure}}$. They give the portion of the pure genotype V^i that is contained within the sample I. In an ideal world, we could write I as linear combination of the pure genotypes V^i with the variables x_i as coefficients. Probably, the isolates also contain yet unknown pure genotypes and there are some numerical inaccuracies, so we introduce an error matrix $F \in \mathbb{R}^{n_{\text{pos}}\times 4}$.

$$I = \sum_{i=1}^{n_{\text{pure}}} x_i V^i + F \tag{1}$$

We assume that an optimal estimation of the real distribution values is achieved if the sum of all absolute values of the entries of F is minimal.

$$\min \sum_{j=1}^{n_{\text{pos}}} \sum_{X\in\{A,C,G,T\}} |F_{jX}| \tag{2}$$

Example 1.3. Consider again the isolate I from Example 1.1 and the pure genotypes *AT* and *GT* from Example 1.2. We want to write I as a "good" linear combination of the pure genotypes V^1 and V^2. Thus, we are looking for real numbers x_1, x_2 in $[0,1]$ with

$$\begin{aligned} & I = x_1 \cdot V^1 + x_2 \cdot V^2 + F \\ \Leftrightarrow & \begin{pmatrix} 0.5 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = x_1 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + x_2 \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} F_{1A} & F_{1C} & F_{1G} & F_{1T} \\ F_{2A} & F_{2C} & F_{2G} & F_{2T} \end{pmatrix} \end{aligned}$$

where $|F_{1A}| + |F_{1C}| + |F_{1G}| + |F_{1T}| + |F_{2A}| + |F_{2C}| + |F_{2G}| + |F_{2T}|$ should be minimal. A feasible solution is

$$x_1 = 0.5 \qquad x_2 = 0.25 \qquad F = \begin{pmatrix} 0 & 0.25 & 0 & 0 \\ 0 & 0 & 0 & 0.25 \end{pmatrix}.$$

This is indeed an optimal solution for this instance. ◁

We reformulate (1):

$$\begin{aligned}
&\sum_{i=1}^{n_{\text{pure}}} x_i V^i + F = I \\
\Leftrightarrow &\sum_{i=1}^{n_{\text{pure}}} x_i V^i_{jX} + F_{jX} = I_{jX} \qquad \text{for all } (j, X) \in \{1, ..., n_{\text{pos}}\} \times \{A, C, G, T\}
\end{aligned} \tag{3}$$

In equation (3), I_{jX} denotes the entry of I at position (j, X). The second line is a component-wise reformulation. As already discussed, we must have

$$0 \leq x_i \leq 1 \tag{4}$$

for all $i = 1, ..., n_{\text{pure}}$. This results in the following formulation of a *Linear Program (LP)*:

$$\begin{aligned}
\min \sum_{j=1}^{n_{\text{pos}}} \sum_{X \in \{A,C,G,T\}} &|F_{jX}| \\
\sum_{i=1}^{n_{\text{pure}}} x_i V^i_{jX} + F_{jX} = I_{jX} &\qquad \text{for all } (j, X) \in \{1, ..., n_{\text{pos}}\} \times \{A, C, G, T\} \\
x_i \geq 0 &\qquad \text{for all } i = 1, ..., n_{\text{pure}} \\
x_i \leq 1 &\qquad \text{for all } i = 1, ..., n_{\text{pure}}
\end{aligned}$$

We want to find a matrix F and values x_i fulfilling the constraints (such tupel (F, x) are called feasible) and minimizing the objective function (2). There is always (at least) one feasible solution for *(LP)*, namely $x_i = 0$ for all $i = 1, ..., n_{\text{pure}}$ and $F = I$. Thus, the program is always feasible. The objective function is bounded by 0 from below and hence by the fundamental theorem of linear programming there exists always an optimal solution to *(LP)*. In a next step, the values $|F_{jX}|$ in the objective function are replaced by new variables in order to achieve a truly linear objective function. This is a standard approach that can be found in any literature on linear programming.

2 Implementation

We implemented a framework to read and process the input data for this problem, solve the linear program with Gurobi and write the results.

2.1 Datasets

Each sub-directory in `datasets/` represents an independent dataset. Within such a dataset, two files must be present to run the code:

- basis.txt: Containing the pure genotypes as strings consisting of the characters A, C, G, and T. In front of each string is the id of the pure genotype followed by a colon.
- input.csv: This file contains (multiple) samples I. The first column gives the name of the sample as string. For each sample, there are n_{pos} rows representing the SNP-positions each containing the absolute values the nucleotide A, C, G, T where found there.

The dataset original contains the samples that were analysed for the above mentioned publication. The dataset minimal_example contains the sample from Example 1.1.

To select a dataset, set the variable dataset at the beginning of the scrip src/main.py to the according name of the dataset. The code can be executed by running the script src/main.py.

The following output files are produced within the selected dataset sub-directory:

- analysis.csv Each row corresponds to one sample. Following the sample name, the computed values of the variables $x_1, ..., x_{n_{\text{pure}}}$ are shown. The next column contains the entry of the corresponding error matrix F with the largest absolute value. The last column contains a comma-separated list of all SNP-positions that have an entry in F for any of the four nucleotide with an absolute value larger than $\varepsilon = 0.1$ and the value itself. The choice of ε can be adapted in the script src/main.py.
- errormatrix.csv This file contains the error matrices F for all samples, following the style of input.csv.
- interpolation.csv Follows the style of input.csv. It contains the interpolations for all samples, i.e. the linear combinations

$$\sum_{i=1}^{n_{\text{pure}}} x_i V^i$$

  ignoring the error matrix.