Generalization of the Simplified Case for Strain Diversity Problem

Stanley Gan, Guo Liang

April 28, 2017

1 Simple illustration and description of the problem

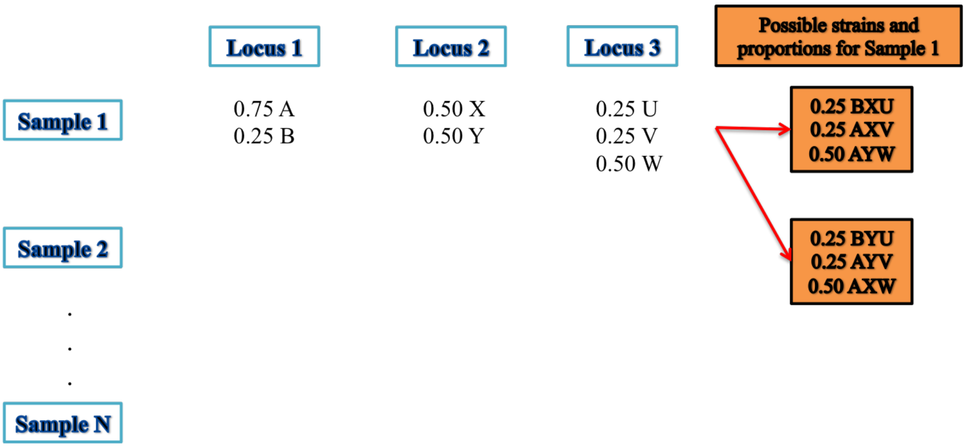

Figure 1: Illustration of the problem

We are given the data about the genotypes observed at all loci for N samples and their respective proportions. For example in Figure 1, we observed genotype of type A and B with proportion 0.75 and 0.25 respectively at locus 1 for sample 1. Given the genotypes observed at all loci for N samples, we produce different strains based on combinations of genotypes at all loci while preserving proportions. For example in Figure 1, we have shown 2 different examples which are J={0.25BXU, 0.25AXV, 0.5AYW} and K={0.25BYU, 0.25AYV, 0.5AXW}. In J, proportion of B is 0.25(from BXU), proportion of A is 0.25(from AXV) + 0.5(AYW)=0.75, ... and so on. The proportions are preserved. Our problem and goal are as follows:

Given a library of known strain types (600+ for Borrelia case), we are trying to explain what we see(genotypes) using as few new strain types as possible. From a mathematical perspective, we want to minimize the number of new strains introduced to our existing library.

For example in a simple case, if our library contains {BXU, AXV, ... }, we will choose J instead of K as we have to introduce at least 2 new strain types if we choose K. Definitely, there are other criteria to consider such as the proportions.

2 Problem Formulation using Mixed Integer Linear Programming

The idea in this program is to formulate the problem rigorously and minimize the number of new strains, the proportions of the new strains using 0/1 weights indicator. Besides, this program also captures the possible errors between the true proportion of a variant and its observed proportion, in which these errors may happen due to sampling in the lab. These errors will also be included into the objective function.

Known Parameters

- Number of loci: 8
- Number of samples: 30
- Set of different genotypes observed on sample i at locus j, $G_{ij} = \{g_{ij}^{(1)}, g_{ij}^{(2)}, ...\}$.
- $P_{ij} = \{p_{ij}^{(1)}, p_{ij}^{(2)}, ...\}$ where $p_{ij}^{(k)}$ corresponds to the proportion of genotype $g_{ij}^{(k)}$ in G_{ij}. (Note: $\displaystyle\sum_{k=1}^{|P_{ij}|} p_{ij}^{(k)} = 1$, ∀1≤i≤30, ∀1≤j≤8)
- Reference = Ω where $|\Omega|$=683
- Set of possible different combinations of the genotypes we observed at all loci for sample i, $V_i = \{V_i^{(1)}, V_i^{(2)}, ..., V_i^{(H_i)}\}$, where $H_i = \prod_{j=1}^{L} |G_{ij}|$
- Unique strain types S_i among all sample i.e. unique elements in $\{V_i^{(k)}\}$ for all i, k.
- A representation of the strain type, $V_i^{(k)} = \begin{bmatrix} a_{i1,1}^{(k)} & a_{i2,1}^{(k)} & \cdots & a_{iL,1}^{(k)} \\ a_{i1,2}^{(k)} & \cdots & \cdots & a_{iL,2}^{(k)} \\ \vdots & \ddots & & \vdots \\ a_{i1,|N_{i1}|}^{(k)} & & \ddots & \vdots \\ \vdots & & & \vdots \\ a_{i1,R_i}^{(k)} & a_{i2,R_i}^{(k)} & \cdots & a_{iL,R_i}^{(k)} \end{bmatrix}$, i-th sample k-th combination, $\forall 1 \leq i \leq 30$. $a_{ij}^{(k)} = \{0,1\}$, $R_i = \max_j |G_{ij}|$. For those j such that $|G_{ij}| < R_i$, $a_{ij}^{(k)} = 0$ for $k = |G_{i(j+1)}|, .., R_i$
- For the example in the problem description, if $V_1 = \{V_1^{(1)} = \text{BXU}, V_1^{(2)} = \text{AXV}, V_1^{(3)} = \text{AYW}\}$, the matrix representation is as follows: $V_1^{(1)} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, $V_1^{(2)} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$, $V_1^{(3)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
- Weight for each unique strain type S_i, w_i where w_i=1 iff S_i is a new strain type, otherwise w_i=0
- Weight for the proportion $\pi_i^{(k)}$(will be explained later), $c_i^{(k)}$ where $c_i^{(k)}$=1 iff $\pi_i^{(k)}$ corresponds to a new strain type, otherwise $c_i^{(k)}$=0

Decision Variables

- Indicator variable a_i where a_i=1 iff S_i is chosen to explain the samples, otherwise a_i=0
- Proportion of the combination $V_i^{(k)}$, $\pi_i^{(k)}$
- $E_{ij} = \{e_{ij}^{(1)}, e_{ij}^{(2)}, ...\}$ where $e_{ij}^{(k)}$ corresponds to the error of the observed proportion $p_{ij}^{(k)}$ of $g_{ij}^{(k)}$ from its true proportion.
- For convenience, let $\Phi_i = \begin{bmatrix} p_{i1}{}^{(1)} & p_{i2}{}^{(1)} & \cdots & p_{iL}{}^{(1)} \\ p_{i1}{}^{(2)} & \cdots & \cdots & p_{iL}{}^{(2)} \\ \vdots & \ddots & & \vdots \\ p_{i1}{}^{(|N_{i1}|)} & & \ddots & \vdots \\ \vdots & & & \vdots \\ p_{i1}{}^{(R_i)} & p_{i2}{}^{(R_i)} & \cdots & p_{iL}{}^{(R_i)} \end{bmatrix}$, $\forall 1 \leq i \leq 30$. For those j such that $|G_{ij}| < R_i$, $p_{ij}^{(k)} = 0$ for $k = |G_{i(j+1)}|, .., R_i$

Constraints

- $p_{ij}^{(k)} \in [0,1]$, $e_{ij}^{(k)} \in [-p_{ij}^{(k)}, 1 - p_{ij}^{(k)}]$ $\forall i, j, k$ and $\sum_{k=1}^{|P_{ij}|} (p_{ij}^{(k)} + e_{ij}^{(k)}) = 1$, $\forall 1 \leq i \leq 30$, $\forall 1 \leq j \leq 8$
- $a_i \in \{0, 1\}$ $\forall i$
- $\pi_i^{(k)} \in [0,1]$ and $\sum_{k=1}^{H_i} \pi_i^{(k)} = 1$ for all i
- $a_i \geq \pi_j^{(k)}$, where $V_j^{(k)} = S_i$
- $0 \leq a_i - \frac{1}{|\{\pi_j^{(k)} : V_j^{(k)} = S_i\}|} \cdot \sum_{V_j^{(k)} = S_i} \pi_j^{(k)} \leq 1 - \epsilon_i$
- $e_{ij}^{(k)} \leq T_i^{(k)} - p_{ij}^{(k)}$ and $e_{ij}^{(k)} \leq p_{ij}^{(k)} - T_i^{(k)}$ where $T_i^{(k)} = \sum_{(i,k): g_{ij}^{(k)} \in V_i^{(k)}} \pi_i^{(k)}$
- For all sample i where $1 \leq \ i \leq 30$,

$$\sum_{k=1}^{H_i} \pi_i^{(k)} \cdot V_i^{(k)} = \Phi_i$$

Objective Function

$$min \quad \sum_i w_i \cdot a_i + \sum_{j,k} c_j^{(k)} \cdot \pi_j^{(k)} + \sum_{p,q,r} e_{pq}{}^{(r)}$$