Identifying Variants and Calculating Their Proportions Once Identified

SFU

April 28, 2017

1 Simplified case

The reads here are mapped with $\leq k$ number of mismatches, assuming k=3. The input is a matrix M of reads vs variants at a locus, where the entry of the matrix m_{ij}=1 if read i is mapped to variant j with ≤ 3 mismatches, otherwise = n/a. For example:

$$\begin{array}{cc} \begin{array}{ccccc} var_1 & var_2 & var_3 & var_4 & var_5 \end{array} & \\ \begin{pmatrix} 1 & 1 & 1 & 1 & na \\ na & 1 & na & 1 & 1 \\ na & na & 1 & na & 1 \\ na & 1 & na & 1 & na \\ 1 & na & na & na & 1 \end{pmatrix} & \begin{array}{c} read_1 \\ read_2 \\ read_3 \\ read_4 \\ read_5 \end{array} \end{array}$$

The idea here is analogous to the set cover problem, but with some relaxation, i.e. to find minimum number of variants needed to cover most of the reads.

1.1 Known Parameters

- The set of all reads, $R = \{r_1, r_2, ...\}$
- The set of reads covered by variant j, $S_j = \{r_i : m_{ij} = 1\}$
- Total variants = n

1.2 Decision Variables

- $x_j = 1$ if S_j is chosen, otherwise 0
- $y_k = 1$ if r_k is used, otherwise 0

1.3 Constraints

- $x_j = \{0,1\}$
- $y_k = \{0,1\}$
- We require most of the reads, *i.e.* at least $(1-\alpha)|R|$ reads to be covered. Hence,

$$\sum_{r_i \in R} y_i \geq (1-\alpha)|R|$$

- If a read r_i is used, then there must be some variant, S_j covering it. Hence,

$$\sum_{j : r_i \in S_j} x_j \geq y_i$$

1.4 Objective Function

$$min \sum_{j=1}^{n} x_j$$

2 Weighted version

The reads here are mapped with $\leq$ k number of mismatches, assuming k=3. The input is a matrix M of reads vs variants at a locus, where the entry of the matrix m_{ij}=k if read i is mapped to variant j with k mismatches, otherwise = n/a. For example:

$$\begin{array}{cc} \begin{array}{ccccc} var_1 & var_2 & var_3 & var_4 & var_5 \end{array} & \\ \begin{pmatrix} 1 & 0 & 2 & 1 & na \\ na & 3 & na & 1 & 0 \\ na & na & 3 & na & 2 \\ na & 1 & na & 2 & na \\ 1 & na & na & na & 1 \end{pmatrix} & \begin{array}{c} read_1 \\ read_2 \\ read_3 \\ read_4 \\ read_5 \end{array} \end{array}$$

The idea here is to find a set of variants(maybe limit to at most 10 variants?) which cover most of the reads with minimum number of mismatches.

2.1 Known Parameters

- The set of all reads, R= $\{r_1, r_2, ...\}$
- The set of reads covered by variant j, S_j = $\{r_i : m_{ij} \neq na\}$
- Total variants = n
- For each read i, A_i=set of distinct number of mismatches i.e. A_i= $\{k : m_{ij} = k\}$. For the example above, A_1 = {0,1,2}

2.2 Decision Variables

- x_j = 1 if S_j is chosen, otherwise 0
- y_{ik} = 1 if read i with k mismatches is chosen, otherwise = 0

2.3 Constraints

- x_j = {0,1}
- y_{ik} = {0,1}
- We require most of the reads, *i.e.* at least $(1-\alpha)|R|$ reads to be covered. Hence,

$$\sum_{r_i \in R} \sum_{k \in A_i} y_{ik} \geq (1-\alpha)|R|$$

- If a read i with k number of mismatches is used i.e. y_{ik}=1, then there must be some variant j covering it with k number of mismatches. Hence,

$$\sum_{\{j: r_i \in S_j, m_{ij}=k\}} x_j \geq y_{ik}$$

- Only allow a read to be mapped with a unique number of mismatches(The read cannot map with 1 mismatch and 2 mismatches at the same time, for example). Hence, for each read i

$$\sum_{k: k \in A_i} y_{ik} \leq 1$$

2.4 Objective Function

$$min \sum_{r_i \in R} \sum_{k \in A_i} k \cdot y_{ik} + \sum_{j=1}^{n} x_j$$

3 Proportions

3.1 Calculating Proportions

We are interested in computing

$$P(var_i \mid read_j) \tag{1}$$

That is we are interested in computing the probability of a variant given a read. Bayes' rule states that

$$P(var_i \mid read_j) = \frac{P(read_j \mid var_i)\, P(var_i)}{P(read_j)} \tag{2}$$

Thus we do not need to directly compute equation (1) we can compute

$$P(read_j \mid var_i) \tag{3}$$

and multiply it by a proportionality constant to get

$$P(var_i \mid read_j) = P(read_j \mid var_i)\, k_j \tag{4}$$

where

$$k_j = \frac{P(var_i)}{P(read_j)} \tag{5}$$

By summing over all variants and equating to 1, we can solve for (5) i.e.

$$\sum_i k_j\, P(read_j \mid var_i) = 1 \tag{6}$$

To compute (3) we can appeal to the binomial distribution since it is given that $\frac{1}{100}$ of mismatches within a mapping is due to sequencing errors. Thus we can use the number of mismatches which bowtie reports for a mapped read and the Binomial Distribution to compute a probability distribution over (3). Thus we have that

$$P(read_j \mid var_i) = \binom{m_j}{l_j} (\frac{1}{100})^{l_j} (\frac{99}{100})^{m_j - l_j} \tag{7}$$

where m_j is the length of the $read_j$, and l_j is number of mismatches for the mapping between $read_j$ and $variant_i$. Plugging (7) into (1), we get that

$$\sum_j k_j \binom{m}{l_j} (\frac{1}{100})^{l} (\frac{99}{100})^{m - l_j} = 1 \tag{8}$$

Thus we are now able to compute k_j for all $read_j$. Now we are fully equipped to compute the proportions for each variant in a gene. The proportion of a $variant_i$ for a gene G will be

$$\frac{\sum_j P(var_i \mid read_j)}{\sum_h \sum_i P(var_h \mid read_j)} \tag{9}$$

We just sum up over all reads that maps to a particular variant for a gene G and divide by the sum over all variants for that gene and multiply by 100 to get proportions in percentages.