The microbial contamination index and the leave-one-out cross-validation model

Initially, the top ranking genes $g_1, \ldots, g_n$ were selected using L-scores that defined them as possible marker genes for microbial infection. The positive integer $n$ was determined to satisfy certain optimization conditions which are previously discussed in materials and methods section. Using the marker genes $g_1, \ldots, g_n$, we defined the microbial contamination index $MCI(s)$ for a given sample $s$, as follows:

$$MCI(s) = MCI(s; g_1, \ldots, g_n) = \frac{1}{n} \sum_{i=1}^{n} \log_{-odds}(g_i, s).$$

To obtain a suitable value for $n$ and to calculate the power of our model, we used the following leave-one-out method for cross-validation:

1) For $j$ in $1, \ldots, K$, the score $L$ was redefined omitting the sample $s_j$ as follows:

$$L^{(-j)}(g) = \frac{1}{K-1} \sum_{k \neq j} \log_{-odds}(g, s_k)$$

2) For each positive integer $n$, define the cross-validation score $CV(n)$ was defined by

$$CV(n) = \min_{j=1}^{K} \frac{1}{n} \sum_{j=1}^{n} \log_{-odds}(g_j^{(-j)}, s_j)$$

where $g_1^{(-j)}, \ldots, g_n^{(-j)}$ are the top ranked genes as determined by $L^{(-j)}$.

3) We let $n^*$ be the number which maximizes the cross-validation score $CV(n)$.

4) The criteria $CV(n^*) > 1.0$ or $CV(n^*) > m + 3s$ was used to validate our prediction model, where $m$ and $n$ are mean and standard deviation of $MCI(s)$ respectively. The random $MCI(s)$ is calculated using the same formula used for the index $MCI(s)$ replacing $g_1, \ldots, g_n^*$ by randomly selected genes.