Scaling MIST's Specificity

Ben Polacco, Krogan Lab.

2022-11-16

Specificity is main MIST component, but does not scale with study size

The MIST score, with range between 0 and 1, is a weighted average of Specificity, Reproducibility and Abundance. We focus here on Specificity because it is the highest-weighted score but is sensitive to experiment size in ways that Reproducibility and Abundance are not.

Specificity Defined

For a given prey, the specificity (S_i) of a bait/prey interaction is simply the abundance of the i-th bait, B_i, divided by the sum of the abundance for that prey across all baits in the dataset. This calculation is done entirely within a prey. Once the signal is normalized between runs, there is no calculation that depends on other preys, so to simplify the notation we consider only a single prey.

$S_i = \frac{B_i}{\sum B}$

Abundance is averaged (mean) across replicates, can be spectral counts or intensity, and is typically normalized by MIST for prey sequence length and total signal in each APMS run. But those details are unimportant to this discussion.

Specificity and Experiment Size

One issue with this score is the denominator entirely depends on how many other baits there are. The more baits to sum together, the lower your Specificity. To begin to rework specificity so that it can scale we separate $\sum B$ into the abundance for bait of interest, B_i, and for all the other baits $\sum_{k!=i} B_k$

$S_i = \frac{B_i}{B_i + \sum_{k!=i} B_k}$

Thus you can see that as the number of other baits with non-zero abundance increase, it changes the specificity . Doing some algebra (in Note 1 below), this equation can be reworked to relate specificity to the ratio of B_i versus all other baits.

$\frac{S_i}{1-S_i} = \frac{B_i}{\sum_{k!=i} B_k}$

To be able to scale this to different sizes of experiments, we can rewrite the sum of abundance in other runs in terms of the average abundance in other runs using definition of mean: $\overline{O_{!i}} = \frac{\sum_{k!=i} B_k}{N-1}$. Here, $\overline{O_{!i}}$ is the average abundance in all other baits, and N is the total number of baits in the study. Then we can rewrite the right-hand side using this definition of average.

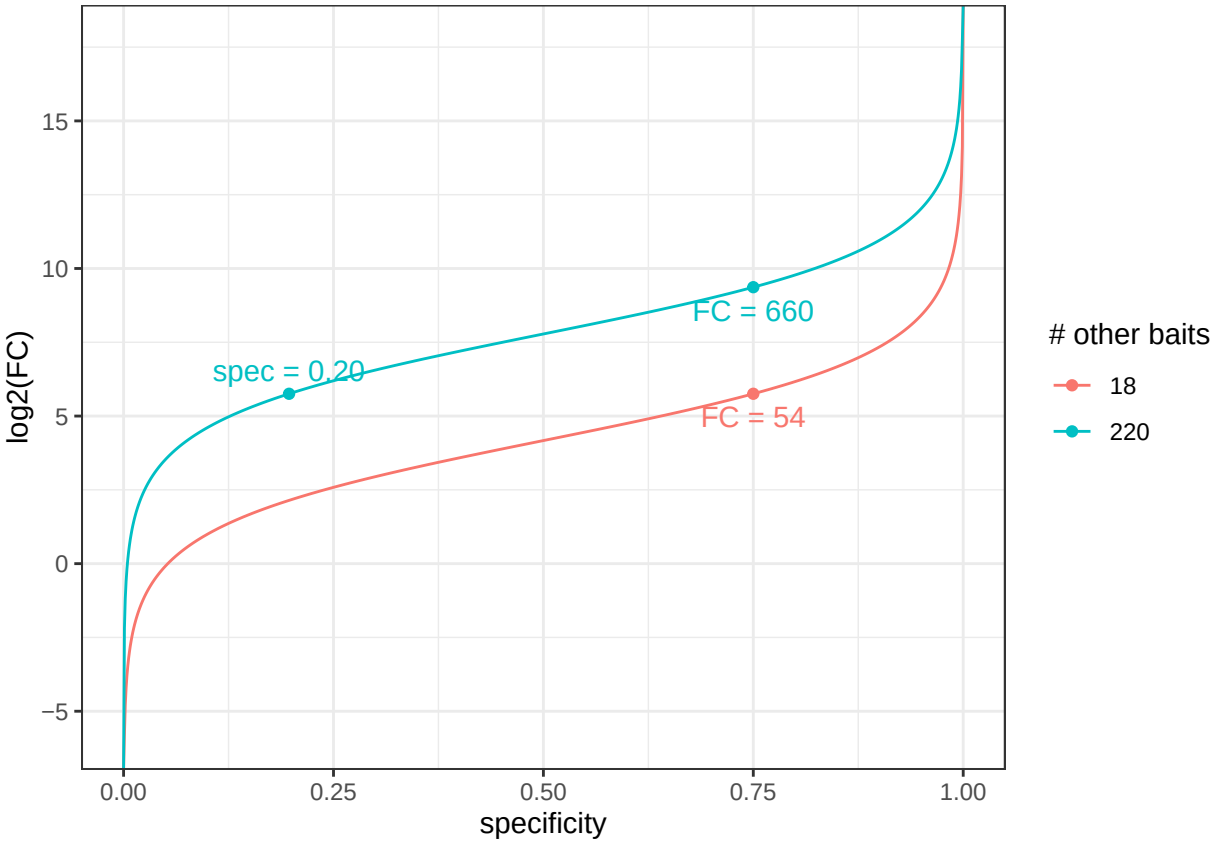
$\frac{S_i}{1-S_i} = \frac{B_i}{(N-1) \cdot \overline{O_{!i}}}$

Thus, for any experiment:

$\frac{S}{1-S} \cdot (N-1) = \frac{B_i}{\overline{O_{!i}}}$

Specificity is a transformed fold change!

This means the specificity value is simply a transformed fold change (and vice versa) where we use the average of all other baits as the background. The exact relationship depends on the number of other baits in the experiment. For HIV, there might be 18 other baits, and for Vaccinia, 220. Thus the transformations look like this:



It takes a very different fold change to produce the same specificity score of 0.75 in these two different experiments!

Specificity can be used alone

(section in progress) The contribution of abundance is trivially small, with a weight of 0.01 and typical values much lower. Reproducibility is more important, but we find that other filters we typically use do a good job of requiring reproducibility, such as SAINT and requiring no missing values. Specificity, once adapted to be insensitive to experiment size, can be used as the primary score from MIST.

Note 1: The algebra

$S_i = \frac{B_i}{B_i + \sum_{k!=i} B_k}$

Simplify the symbols:

$$
\begin{aligned}
S &= \frac{B}{B+O} \\
S(B+O) &= B \\
SB + SO &= B \\
SO &= B - SB \\
SO &= B(1-S) \\
\frac{S}{1-S} &= \frac{B}{O}
\end{aligned}
$$

Unsimplify:

$\frac{S_i}{1-S_i} = \frac{B_i}{\sum_{k!=i} B_k}$