Desired Attributes of a Protein Function Description Model

Meet Barot

February 24, 2022

Attribute 1

Annotation coherence.

Given a sequence set that the model is assigning scores of function descriptions:

Descriptions of GO terms that annotate the entire sequence set should be scored higher than terms that do not annotate the entire sequence set.

Let D_S be the GO term descriptions associated with sequence set S.

$$P(d \in D_S|S) > P(d \notin D_S|S)$$

Attribute 2

Specificity preference.

Among terms that do annotate the whole set, the model should score more specific terms higher than less specific terms.

Let $t(d)$ be the depth of a GO term description d, and δ is an arbitrary depth.

$$P(d \in D_S, t(d) \geq \delta|S) > P(d \in D_S, t(d) < \delta|S)$$

Attribute 3

Branch equality.

Among the most specific terms that annotate the whole set, descriptions from all three branches of GO should be scored equally.

Let B_i be the ith GO branch, and D_{S,B_i} be the descriptions associated with sequence set S that are in branch B_i.

$$P(d \in D_{S,B_i}, t(d) = \delta|S) = P(d \in D_{S,B_j}, t(d) = \delta|S)$$

Attribute 4

Annotation robustness.

Any set of sequences that have the same exact set of GO descriptions in common should produce the same scores for those GO descriptions.

Let S_i and S_j be different sequence sets such that $D_{S_i} = D_{S_j}$.

$$P(d \in D_{S_i}|S_i) = P(d \in D_{S_j}|S_j)$$