1 EC-calling methods

Through using contrastive learning, we essentially establish a ranking model, where each EC cluster center is represented by the average of all the enzyme ids that share this EC number. Therefore for any query enzyme, the correct set of EC number for it can be ranked based on Euclidean distances between all the EC cluster centers and itself. However, because an enzyme could potentially have multiple EC number, there need to be algorithms to 'call' which of the top-ranked EC numbers are considered correct for the query enzyme. Simply calling the top-1, top-2 or top-k ranked EC numbers for all query enzymes would not be suitable, since while most enzymes have only one function (and one EC number), there exists promiscuous enzymes that have multiple EC numbers. Calling the top-1 EC numbers would neglect enzyme promiscuity, but calling more than that, the precision of prediction drastically drops. Here we propose two EC-calling algorithms to determine, with these distance ranking information, what is the correct cutoff for each query enzyme: 1) **Random-nk**, and 2) **Max-Separation**.

Random-nk algorithm is based on statistical significance, where for each EC number, n-thousand randomly chosen enzymes from the training set are ranked based on their distances with the EC cluster center. An EC number is only called for the query enzyme if the query enzyme are statistically close to the EC cluster center. For example, if the p-value is chosen to be 0.001, then the target EC number will be called if the query enzyme is closer to the EC cluster center than 99.9% of all randomly chosen enzymes. By choosing different p-value, users can tune the precision and recall for the prediction.

Unlike Random-nk, **Max-Separation** algorithm has no tunable parameter and will give a single prediction. Max-Separation call the set of EC numbers with distances close query enzyme but far from all other EC numbers, just like any person looking at the distance data would call. For example, here are the top-10 ranked EC numbers for a enzyme:

$$[5.6, 5.7, 11.9, 12.22, 12.23, 12.4, 12.5, 12.6, 12.6, 12.7]$$

and it would be obvious that EC numbers with distances 5.6 and 5.7 should be called. We prove mathematically that Max-Separation makes the same decision. Empirically, Max-Separation results in both better precision and better recall than using any possible p-value for Random-nk.

2 Problem formulation for EC-calling:

Let $S = s_0, s_1, ..., s_{n-1}$ be a sorted sequence of distances s_i ($s_i \leq s_{i+1}$) where s_i represents the distance between the query enzyme sequence embedding and the cluster center embedding with EC number EC_i. The goal is to call the correct set of EC numbers $\{EC_i\}$ for the query sequence (for example, 2.4.1.376 and 2.4.2.63 for enzyme Q8NBL1).

2.1 Random-*n*k

For Random-*n*k algorithm, we pick *n*k (for example, 20k when $n = 20$) randomly chosen enzyme embeddings (after a forward pass thorough the trained model) from the training set. With a chosen p-value (for example, $p = 0.001$), these *n*k embeddings are used to determine statistically whether a EC number should be called for a particular query enzyme.

Instead of picking the *n*k enzyme embeddings uniformly, we weight the probability of picking an enzyme with EC number EC_i with $1/|EC_i|$, the inverse of the number of enzymes in that EC class. We record the picked *n*k enzyme embeddings' Euclidean distances with a particular EC cluster center. In this way, we obtain a distance matrix between all EC cluster centers in the training set and all *n*k enzyme embeddings.

When a particular query enzyme needs call the set of EC numbers EC_i, the algorithm starts with EC number EC_0 with the smallest distance s_0, then s_0 is compared with the *n*k embeddings' distances with EC_0' cluster center. If s_0 is among the top $p \times n \times 1000$ in these *n*k distances, then EC_0 is called for the query enzyme.

2.2 Max-separation

For Max-separation algorithm, we assume there is a background noise distance γ, such that all distances $s_{i'}$ from **incorrect** set of EC numbers for the query sequence are close to γ by ε, and all distances s_i from the **correct** set $\{EC_i\}$ for the query sequence are far from γ by δ:

$$|s_{i'} - \gamma| \leq \varepsilon, \;\; |s_i - \gamma| \leq \delta, \text{ and } \varepsilon \ll \delta \tag{1}$$

A example of how $\{EC_i\}$ can be selected follows: the first 10 smallest sequences for query q is:

$$[5.6, 6.1, 7.5, 12.22, 12.23, 12.4, 12.5, 12.6, 12.6, 12.7],$$

and the background noise is 12.88. It is apparent that EC numbers corresponding to distances 5.6, 6.1 and 7.5 are the correct set of ECs for q. We assume that the sequence $S = s_0, s_1, ..., s_{n-1}$ is long enough that at most 50% of the distances are coming from the correct ECs, that is, $|\{EC_i\}| \leq n/2$. We want to find $\mathring{\imath}$ subject to the following requirements: 1) $|\varepsilon - \delta|$ is **maximized**, 2) the $|\{EC_i\}|$ is **minimized**.

We come up with the following algorithm for calling $\{EC_i\}$ using Max-Separation:

Proof:
To prove that this algorithm indeed finds the correct set of EC numbers $\{EC_i\}$, we first assume that the maximum separation index $\mathring{\imath}$ produced by the algorithm indeed finds the separating index such that $\{EC_i\} = \{EC_0, ..., EC_{\mathring{\imath}}\}$ and $\{EC_{\mathring{\imath}+1}, ..., EC_{n-1}\} \in \{EC_i\}$.

Firstly, since we only have access to the first n number of distances, and the first ranked EC number EC_0 corresponding to s_0 will always be called, $\hat{\gamma}$ from line 3 in the algorithm is a proper estimate of the true γ.

Algorithm 1 Max-Separation

1: **function** MAXSEP(S)
2: **Require** S is the sequence of distances $s_0, s_1, ..., s_{n-1}$ in sorted order
3: **Let** background noise distance $\hat{\gamma} = \text{mean}(s_1 + s_2 + ... + s_{n-1})$
4: **Let** noise separation distances $D = d_0, ..., d_{n-1} = |s_0 - \hat{\gamma}|, ..., |s_{n-1} - \hat{\gamma}|$
5: **Let** slope of separation curve $G = g_0, ..., g_{n-1} = |d_1 - d_0|, ..., |d_{n-1} - d_{n-2}|$
6: **Initialize** maximum separation index $\mathbb{i} \leftarrow 0$
7: **Let** mean slope $\overline{g} = \text{mean}(G)$
8: **Let** maximum separation index $\mathbb{i} \leftarrow i$ be the first i that satisfies $g_i > \overline{g}$.
9: **Return** the correct set of EC numbers for query $\{EC_i\} = \{EC_0, ..., EC_{\mathbb{i}}\}$

The noise separation distances D is the sequence of approximate differences between s_i and γ, and we want to find $\mathbb{i}$ such that similar to statement (1):

$$|s_i - \hat{\gamma}| \leq \delta \text{ given } i \leq \mathbb{i}, \quad \left|s_j - \hat{\gamma}\right| \leq \varepsilon \text{ given } j > \mathbb{i}, \quad \text{and } \varepsilon \ll \delta \tag{2}$$

The first constraint, maximizing $|\varepsilon - \delta|$, requires that the called EC number EC_i to have its distance s_i be distinct from the background noise γ. The second constraint, minimizing $|\{EC_i\}|$, requires that as few EC number as possible to be called, since most enzymes only have one EC number.

By assumption, $\mathbb{i} \leq n/2$, which implies that $EC_{n/2+1}, ..., EC_{n-1}$ will never be called. By computing the G in line 4 of the algorithm, we are computing the slope of the curve plotted by the sequence D. The larger the slope, the larger the difference $|\varepsilon - \delta|$. The maximum separation index $\mathbb{i}$ is the first index i with a large slope (by comparing with the average slope). This is required because if later large slope is used, $|\{EC_i\}|$ is no longer minimized. Since both requirements are satisfied, Max-Separation indeed finds the separating index that calls the correct set of EC numbers.