Predicting EC numbers using OrthoMCL and sequence features

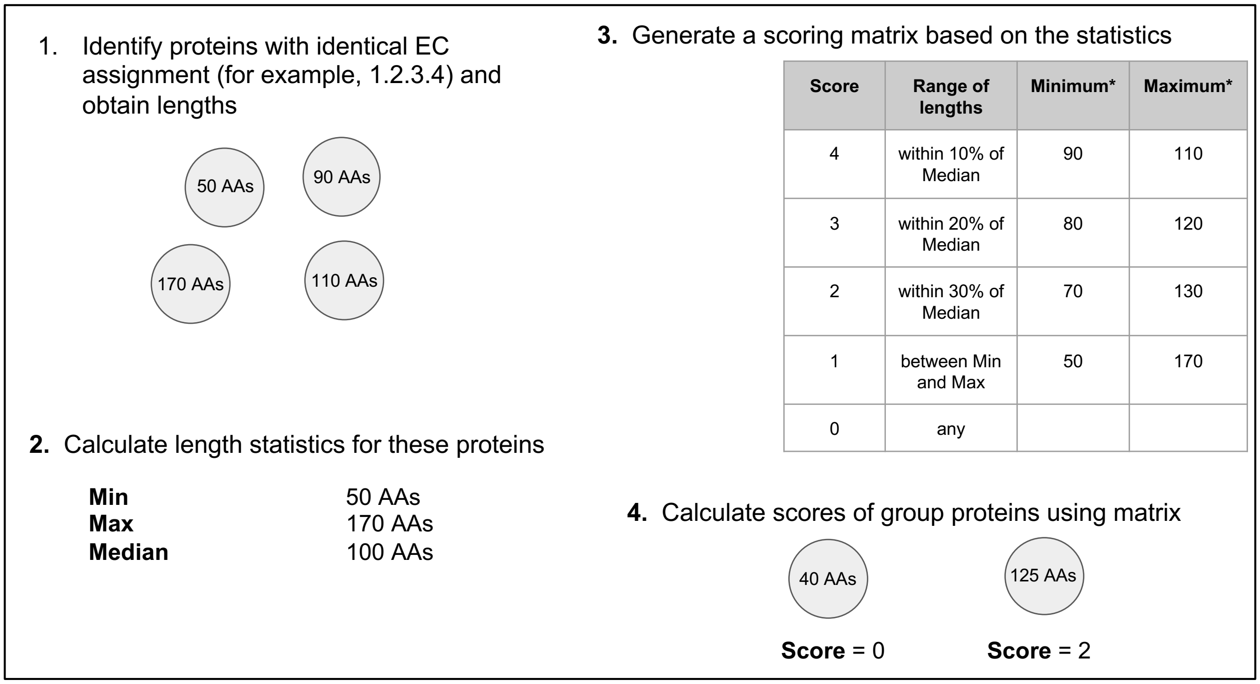
1. Obtain all of the proteins from an OrthoMCL group
2. Identify the set of EC numbers in the group and expand the set to include parent EC numbers. For example, if the group contains these EC numbers: 1.2.3.4, 1.2.3.5, 5.6.7.8, then also consider: 1.2.3.-
3. Determine statistics (median, min and max) for the set of proteins that share a given EC. Evaluate three features of each protein:
    a. Protein length
    b. BLAST score (for all pairwise comparisons)
    c. Domain architecture
4. Assess whether each protein in the OrthoMCL group (regardless of whether they are assigned an EC number or not) should be assigned an EC number by comparing the protein to proteins already assigned that EC number. A score of 0-4 is given for the three features (length, BLAST, and domains). **See below for the details of scoring.**
5. Remove parent EC numbers (like 1.2.3.-) if they have the same or lower score than the complete EC number. For example, the initial set of EC numbers consists of 1.2.3.4 (score 444) and 1.2.3.- (score 434), while the final set only includes 1.2.3.4 (score 444).
6. Calculate other statistics for each EC prediction:
    a. Number of proteins in group previously annotated with that EC number (i.e. not inferred by OrthoMCL)
    b. Total number of proteins in OrthoMCL group
    c. Number of genera in group previously annotated with that EC number (i.e. not inferred by OrthoMCL)
    d. Total number of genera in OrthoMCL group
7. Keep predictions that have a score of 1 for each feature (length, BLAST, and domains). Our testing has shown that 93.9% of 56,003 previously assigned EC numbers were predicted when using this threshold.

Score details

For each method, refer to the accompanying figure below the text description.

Length:

1. Identify proteins with identical EC assignment (for example, 1.2.3.4) and obtain lengths.
2. Calculate length statistics for these proteins: calculate the median length, the min length, and the max length. For example, four proteins are assigned EC 1.2.3.4 and have lengths 50, 90, 110, and 170 amino acids. EC median is 100, EC min is 50 and EC max is 170.
3. Generate a scoring matrix based on the statistics. See the figure below.
4. Calculate scores of all proteins in that OrthoMCL group, regardless of EC assignment, using the matrix.



1. Identify proteins with identical EC assignment (for example, 1.2.3.4) and obtain lengths

2. Calculate length statistics for these proteins

Min	50 AAs
Max	170 AAs
Median	100 AAs

3. Generate a scoring matrix based on the statistics

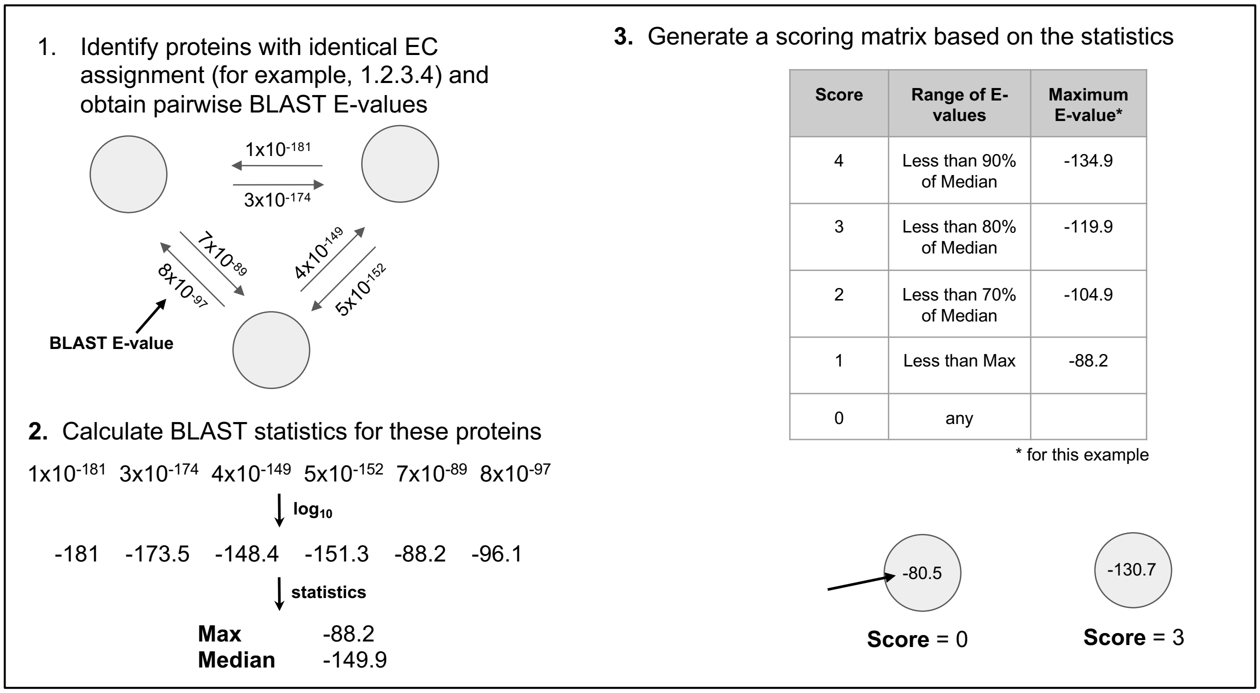
Score	Range of lengths	Minimum*	Maximum*
4	within 10% of Median	90	110
3	within 20% of Median	80	120
2	within 30% of Median	70	130
1	between Min and Max	50	170
0	any		

4. Calculate scores of group proteins using matrix

Score = 0 **Score** = 2

BLAST:

1. Identify proteins with identical EC assignment (for example, 1.2.3.4) and obtain pairwise BLAST E-values. For example, consider six BLAST comparisons with E-values of $1x10^{-181}$, $3x10^{-174}$, $4x10^{-149}$, $5x10^{-152}$, $7x10^{-89}$, and $8x10^{-97}$.
2. Calculate BLAST statistics for these proteins. Note that the lower the E-value, the better the BLAST score; $1x10^{-181}$ is the best E-value while 1 is worst. Using $\log_{10}$ transformed E-values, calculate the max and the median.
3. Generate a scoring matrix based on the statistics. See the figure below.
4. Calculate scores of all proteins in that OrthoMCL group, regardless of EC assignment, using the matrix.



1. Identify proteins with identical EC assignment (for example, 1.2.3.4) and obtain pairwise BLAST E-values

2. Calculate BLAST statistics for these proteins

$1x10^{-181}$ $3x10^{-174}$ $4x10^{-149}$ $5x10^{-152}$ $7x10^{-89}$ $8x10^{-97}$

$\downarrow$ **$\log_{10}$**

-181 -173.5 -148.4 -151.3 -88.2 -96.1

$\downarrow$ **statistics**

Max -88.2
Median -149.9

3. Generate a scoring matrix based on the statistics

Score	Range of E-values	Maximum E-value*
4	Less than 90% of Median	-134.9
3	Less than 80% of Median	-119.9
2	Less than 70% of Median	-104.9
1	Less than Max	-88.2
0	any	

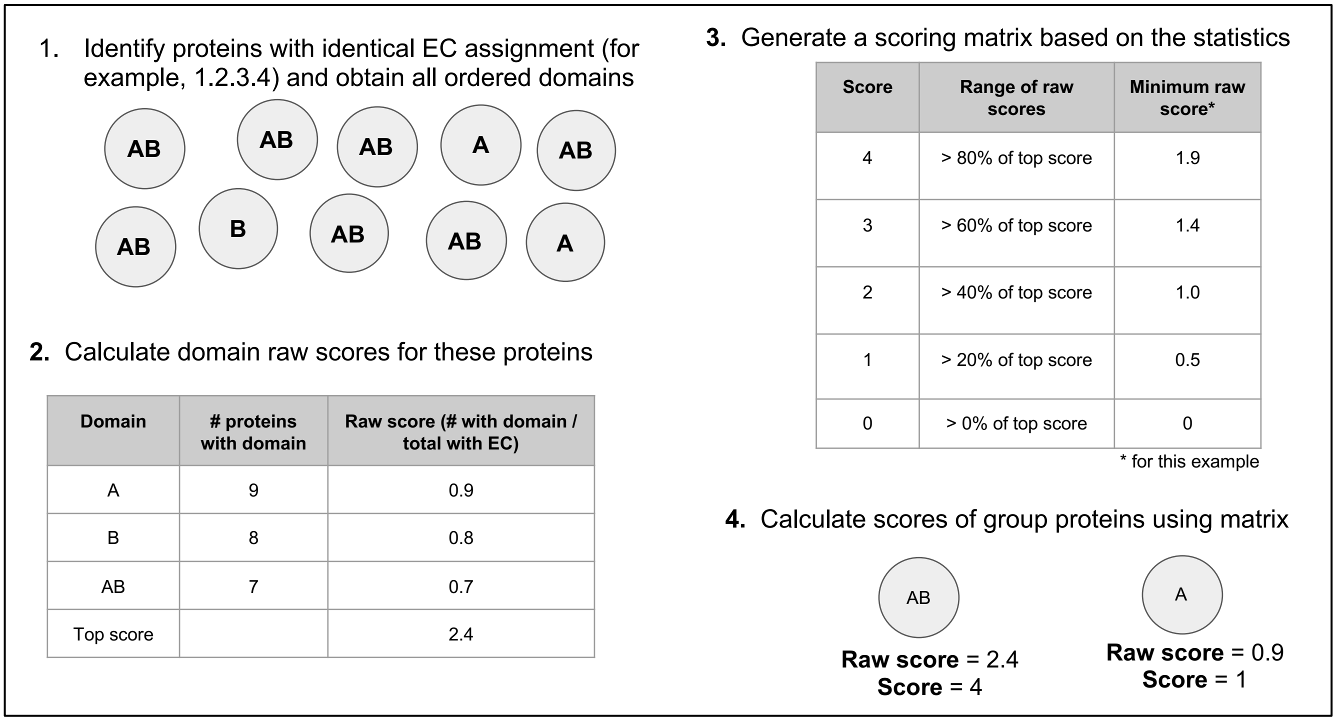
* for this example

Score = 0

Score = 3

Domain architecture:

1. Identify proteins with identical EC assignment (for example, 1.2.3.4) and obtain all ordered domains
2. Calculate domain raw scores for these proteins. First, determine the number of proteins containing each string (e.g., A in 10 proteins, B in 8 proteins, AB in 6 proteins). Then, divide the domain occurrence by the total number of proteins assigned this specific EC number to get a score for that domain (e.g., A: 9/10 = 0.9, B: 8/10 = 0.8, AB: 7/10 = 0.7). The theoretical top score that a protein can have is the sum of all domain scores (e.g., 0.9 + 0.8 + 0.7 = 2.4).
3. Generate a scoring matrix based on the statistics. See the figure below.
4. Calculate scores of all proteins in that OrthoMCL group, regardless of EC assignment, using the matrix. For each protein, determine a score by summing its domain scores (e.g., protein has BA, so only A and B = 0.9 + 0.8 = 1.7; note, AB is not present because it is the reverse order).



1. Identify proteins with identical EC assignment (for example, 1.2.3.4) and obtain all ordered domains

2. Calculate domain raw scores for these proteins

Domain	# proteins with domain	Raw score (# with domain / total with EC)
A	9	0.9
B	8	0.8
AB	7	0.7
Top score		2.4

3. Generate a scoring matrix based on the statistics

Score	Range of raw scores	Minimum raw score*
4	> 80% of top score	1.9
3	> 60% of top score	1.4
2	> 40% of top score	1.0
1	> 20% of top score	0.5
0	> 0% of top score	0

* for this example

4. Calculate scores of group proteins using matrix

Raw score = 2.4
Score = 4

Raw score = 0.9
Score = 1