

---

## Tmmpred 0.2.0

Pål Puntervoll - NORCE



2024-04-08

## Tmmpred - prediction of bacterial-type trimethylamine monooxygenase sequences

Tmmpred is a Python-module based on PyHMMER that offers prediction of bacterial-type trimethylamine monooxygenase sequences.

Bacterial trimethylamine monooxygenases (Tmms) belong to a well-defined subfamily of flavin-containing monooxygenase (FMOs)<sup>1</sup>. This group also contain eukaryotic Tmms, that are more closely related to the bacterial Tmms than to any other FMO.

### Construction of profile HMMs

As a first step, a sequence similarity network (SSN) was built from reference eukaryotic, bacterial, and archaeal proteomes (UniProt release 2023\_02) using the online EFI-EST tool:

1. All eukaryotic reference proteomes with BUSCO<sup>2</sup> > 99% were downloaded in TSV format (132 proteomes: data/euk\_ref\_proteomes\_busco\_99.tsv).
2. All bacterial reference proteomes with BUSCO > 99% were downloaded in TSV format (3067 proteomes; data/bac\_ref\_proteomes\_busco\_99.tsv).
3. All archaeal reference proteomes with BUSCO > 99% were downloaded in TSV format (55 proteomes; data/arc\_ref\_proteomes\_busco\_99.tsv).
4. All reference proteome sequences annotated as Flavin monooxygenase-like sequences (InterPro IPR020946) were downloaded from UniProt release 2023\_02 (1235 eukaryotic, 2659 bacterial, and 1 archaeal sequences; data/ref\_proteomes\_busco\_99.fasta).
5. The sequences were submitted to EFI-EST (with the parse headers option) to generate an SSN. The SSN was finalised with score threshold 55 (corresponding to a sequence identity of 30%) and only including sequences with lengths 360-680 (3,724 sequences).
6. The resulting SSN was divided into three subnetworks (using Cytoscape):
  1. Eukaryotic sequences (excluding plants), which contained 9 clusters with > 20 sequences, named E1-E9.
  2. Plant sequences, which contained 3 clusters named V1-V3.
  3. Bacterial sequences, which contained 8 clusters with > 20 sequences, named B1-B8.
7. Sequences were downloaded for each cluster (data/initial\_cluster\_sequences/).

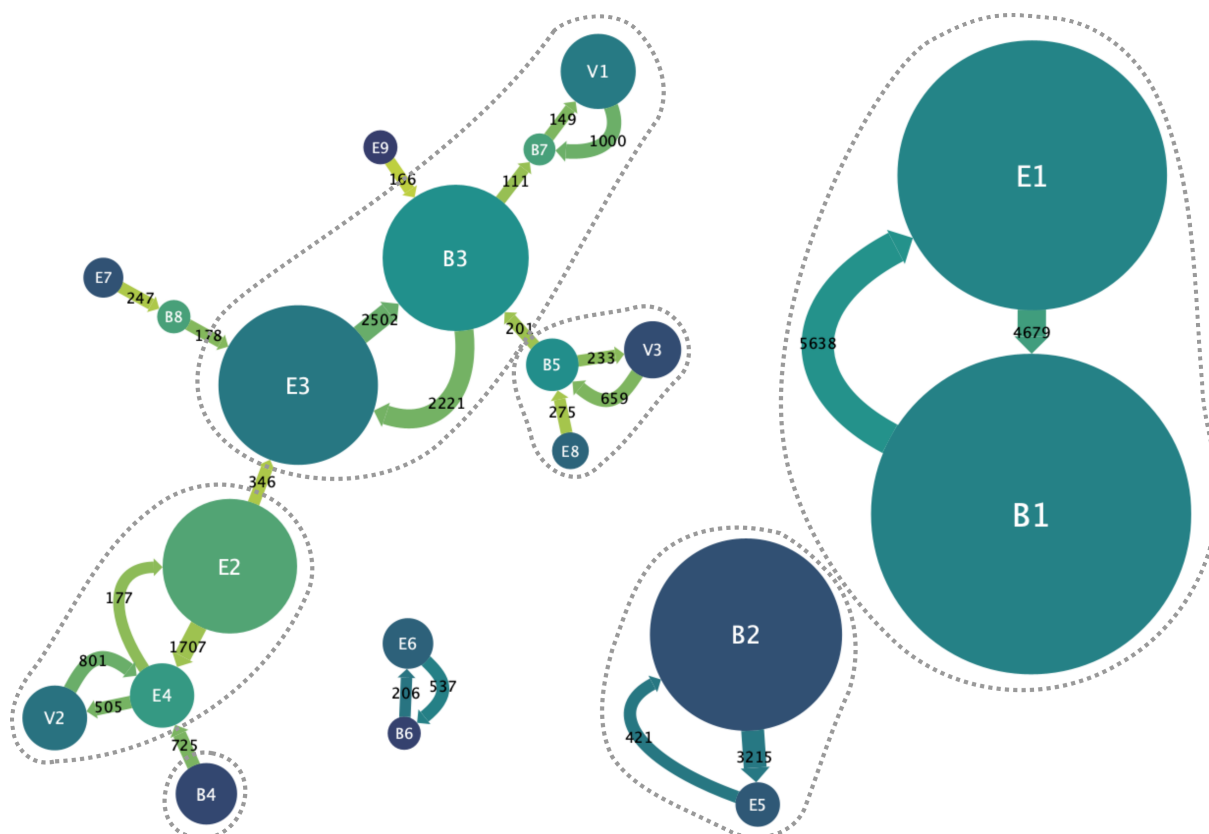
---

<sup>1</sup>Goris, Marianne, Pål Puntervoll, David Rojo, Julie Claussen, Øivind Larsen, Antonio Garcia-Moyano, David Almendral, Coral Barbas, Manuel Ferrer, and Gro Elin Kjæreng Bjerga. 2020. 'Use of Flavin-Containing Monooxygenases for Conversion of Trimethylamine in Salmon Protein Hydrolysates'. *Applied and Environmental Microbiology* 86 (24): e02105-20. <https://doi.org/10.1128/AEM.02105-20>.

<sup>2</sup><https://www.sib.swiss/about/news/10131-gauging-the-completeness-of-genomics-data-with-busco>.

For each SSN cluster, a representative set of sequences were generated using CD-HIT and an identity threshold of 70%. For each representative set of cluster sequences, MAFFT (-auto) was used to generate multiple sequence alignments (MSAs) and sequences with spurious deletions and insertions were removed to generate high quality MSAs (data/initial\_msas). These were in turn used to generate profile HMMs using HMMER (data/initial\_hmms).

The set of profile HMMs (compiled in data/initial/hmms/ref\_prot\_fmo-like.hmm) were used to search (using HMMER) a subset of the 61,993 sequences annotated as FMO-like sequences (InterPro: IPR020946) in UniProt release 2023\_04: only sequences with lengths 360-680 were included (54,695 sequences), and CD-HIT was used to cluster sequences  $\geq 90\%$  identical (resulting in a set of 27,569 sequences; data/ipr020946\_range\_360-680\_nr90.fasta). A new network was constructed where each node represents one SSN cluster (from the initial analysis) and contains the sequences which scored best against the profile HMM for that cluster (Figure 1). The directed edges represent the number of sequences for a particular SSN cluster, that scored second best against another SSN cluster. Tmm sequences are found in B4.

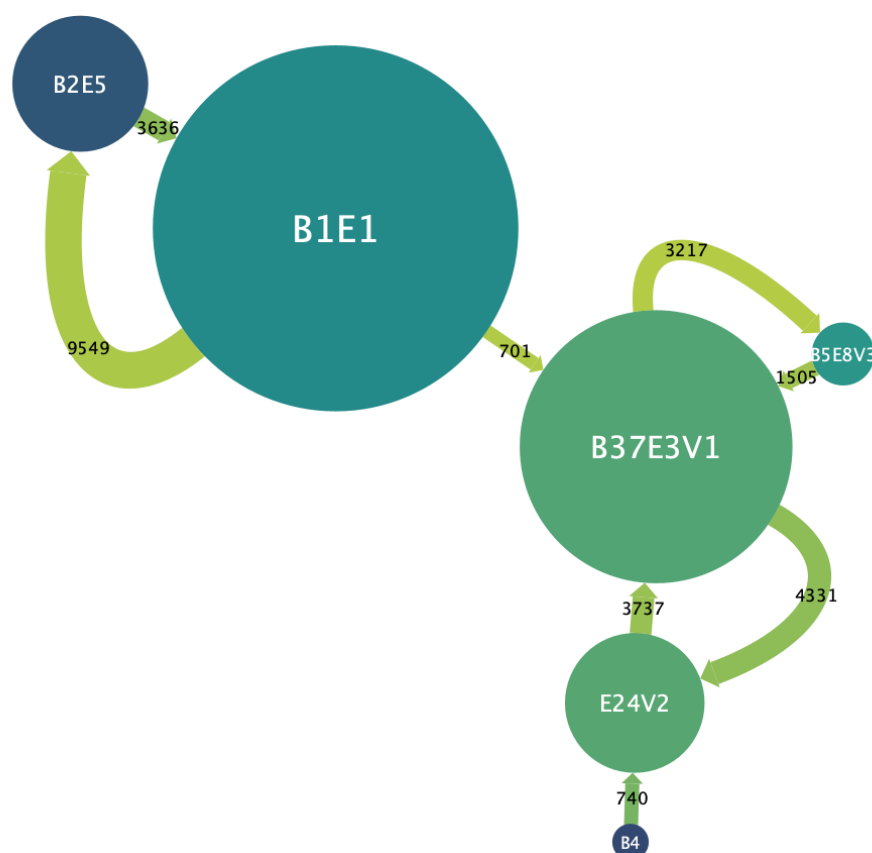


A set of second generation of profile HMMs was generated by merging sequences of selected nodes, as indicated by the dotted lines in Figure 1. The new superclusters are:

1. **BVMO1 (B1E1)** - containing B1 and E1 sequences.
2. **BVMO2 (B2E5)** - containing B2 and E5 sequences
3. **FM01 (B37E3V1)** - containing B3, B7, E3, and V1 sequences
4. **Tmm (B4)** - containing B4 sequences
5. **FM03 (B5E8V3)** - containing B5, E8, and V3 sequences
6. **FMNO2 (E24V2)** - containing E2, E4, and V2 sequences

Sequences from the following clusters were not included in the construction of second generation profile HMMs: B6, B8, E6, E7, and E9. The new supercluster sequences (data/second\_generation\_cluster\_sequences) were used to build MSAs (data/second\_generation\_msas) and profile HMMs (data/second\_generation\_hmms) as described above.

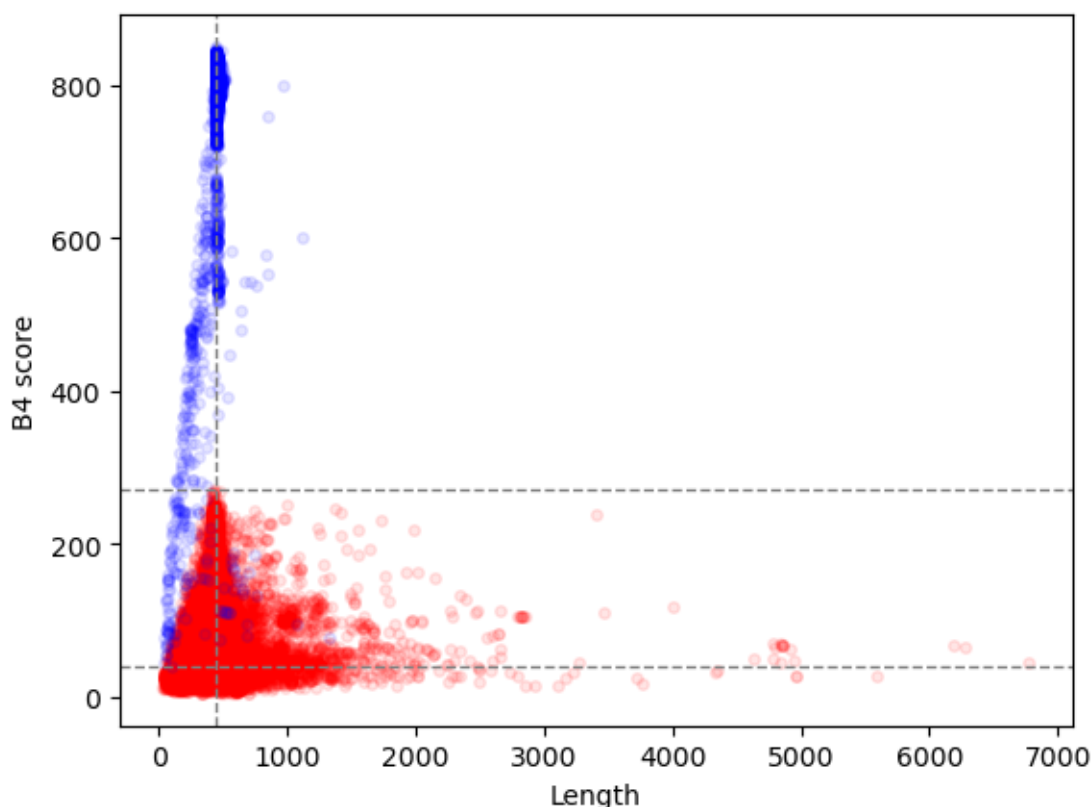
The second generation set of profile HMMs (compiled in data/second\_generation\_hmms/ref\_prot\_fmo-like\_02.hmm) were used to search the subset of FMO-like (InterPro: IPR020946) UniProt sequences (as described above; 27,569 sequences; data/ipr020946\_range\_360-680\_nr90.fasta). A network was constructed where each node represents a supercluster and contains the sequences which scored best against the profile HMM for that supercluster (Figure 2).



**Figure 2:** Network representing best hits (nodes) and second best hits (edges) of the second generation supercluster profile HMMs (node names) searched against IPR020946 UniProt sequences with sequence identity  $\leq 90\%$  and lengths 360-680. Figure visualised as in Figure 1. Only edges with sequence numbers  $> 100$  are shown.

Figures 1 and 2 illustrate that the sequences scoring best against the B4 profile HMM on average score much higher against that profile (756) than the second best scoring profile HMM E24V2 (302). Tmmpred

takes advantage of this when predicting Tmms and separating Tmms from other FMOs or FMO-like sequences. The second generation set of profile HMMs (Figure 1) was used to scan all 61,993 UniProt sequences annotated with InterPro FMO-like family (IPR020946) and a plot of sequence length vs score against the Tmm B4 profile is shown in Figure 3.

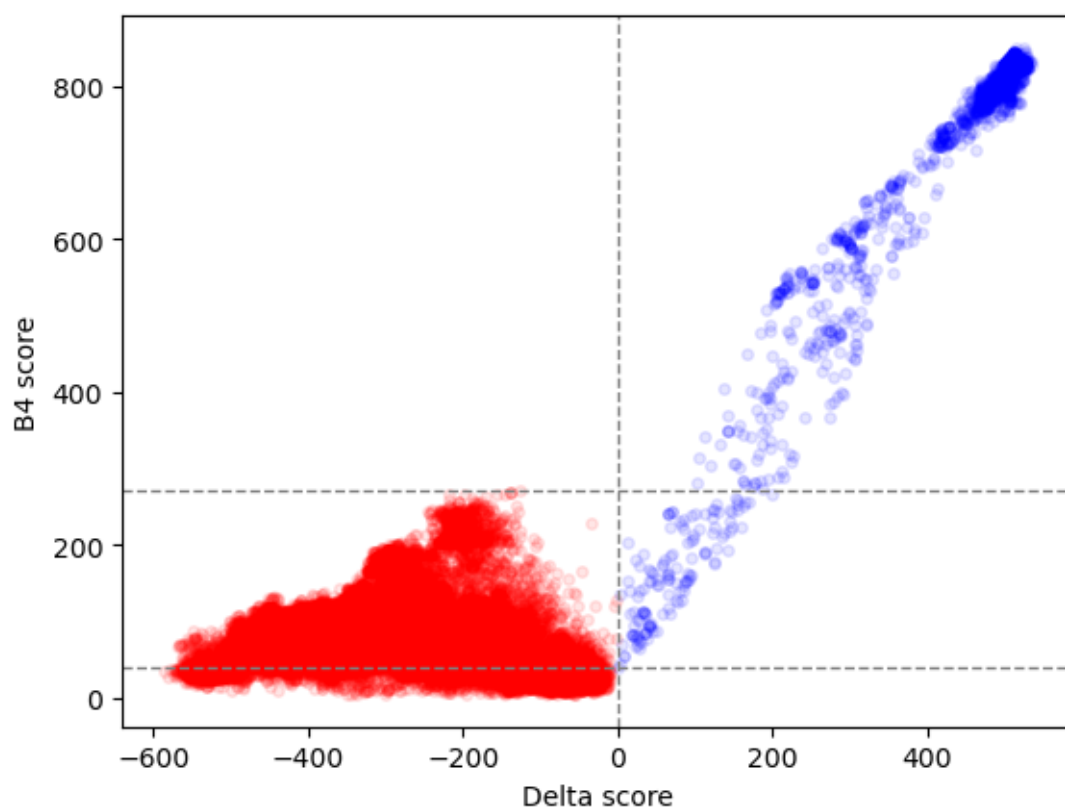


**Figure 3:** All UniProt sequences annotated with InterPro FMO-like family (IPR020946) were scanned against the profile HMMs depicted in Figure 2. The horizontal axis shows sequence length, and the vertical axis shows the score against the B4 profile. Sequences scoring better against the B4 profile than against any other profile are shown in blue (considered as true Tmm candidates), and those scoring best against any of the other profiles are shown in red (considered as false Tmm candidates). The vertical dotted line shows the median sequence length (452), and the two horizontal lines show lowest score of a true Tmm (score 40; noise cutoff) and the highest score of a false Tmm (score 272; trusted cutoff).

Figure 3 shows that 2,195 sequences has the B4 as the best scoring profile and a score above 272, which is higher than best-scoring sequence that scores better against one of the other FMO-like profiles (and considered not to belong to the Tmm subfamily). This set of sequences contain all Tmms that have been shown experimentally to oxidise TMA to TMAO. In addition, there are 136 sequences that score

between 272 (trusted cutoff) and 40 (noise cutoff). Many of these sequences are likely to be fragments and lack one or more of the cofactor binding motifs.

Tmmpred computes a so-called *delta score*, which is the B4 score subtracted by the score of the highest scoring other FMO-like profile. Figure 4 shows how the delta score can be used to separate likely Tmm candidates from non-Tmm FMO-like sequences.



**Figure 4:** The *delta score* separates sequences that are most likely to belong to the Tmm subfamily from those that most likely belong to a different FMO-like family. The delta score is calculated by subtracting the the score of the highest scoring other FMO-like profiles from the B4 score. Colors and dotted lines as in Figure 3.

## Tmmpred

### Tmmpred usage

```
1 usage: tmmpred [-h] [-q] [-d] [-c CUTOFF] [-n] [--html] [-v]
   sequence_file
2
```

```
3 Predict Tmm sequences (FMO subfamily with trimethylamine monooxidase
   activity).
4
5 positional arguments:
6   sequence_file      FASTA-formated protein sequences.
7
8 options:
9   -h, --help          show this help message and exit
10  -q, --quick          Search with Tmm HMM profile only and use
                        trusted score cutoff: 272.0.
11  -d, --deep          Search with Tmm HMM profile using noise score
                        cutoff 40.0 and filter using all FMO-like HMM profiles.
12  -c CUTOFF, --cutoff CUTOFF
                        Score cutoff [float, default=40.0 (noise score
                        cutoff)].
13
14  -n, --nofilter       Do not filter using other FMO-like HMM profiles
                        .
15  --html              Format results as HTML.
16  -v, --verbose        Show details about running tmmpred.
```

The default mode of Tmmpred is to first scan the input sequences against the Tmm (B4) profile, and use the noise cutoff (40.0) to filter out false hits. Next, the remaining input sequences are scanned against the closest neighbour profile FMO2 (E24V2) (Figure 1), and sequences with negative delta scores are filtered out.

The quick mode (**-q**) only scans sequences against the Tmm (B4) profile and uses the trusted cutoff (272.0) to filter out false hits.

The deep mode (**-d**) is similar to the default mode, but in the second step it scans the sequences that passed the first step against all other FMO-like profiles. Some Tmm predictions reported by the default mode may be filtered out by the deep mode.

The user can override the default cutoff settings using the **-c** flag and override filtering using the **-n** flag. When using the latter option, all sequences that score above the cutoff against Tmm (B4) will be reported, also those with negative delta scores.

### Tmm result format

Tmmpred reports the following for each hit sequence (separated by tabs):

- **Query** - sequence identifier
- **Length** - length of query sequence
- **Coverage** - the length of the sequence fragment that matches the profile HMM divided by the profile HMM length (given as percentage)
- **Score** - the Tmm (B4) profile HMM score

- **Delta score** - the Tmm (B4) profile HMM score subtracted by the the score of the highest scoring other FMO-like profile
- **D name** - Name of the highest scoring other FMO-like profile
- **FAD motif** - (consensus sequence of FAD motif as represented in the profile HMM) subsequence matching the FAD motif (with sequence numbers indicated)
- **NAD motif** - (consensus sequence of NAD motif as represented in the profile HMM) subsequence matching the NAD motif (with sequence numbers indicated)

The quick mode does not report *Delta score* or *D name*.