[2]p()*1/2-^ [1]p1-^ TJ
3 10000pt

0.00.5em

0.0.00.5em

0.0.0.00.5em

0.0.0.0.00.5em

0.0.0.0.0.00.5em

efe-bgc

Davide Rossotto <drossotto@crimson.ua.edu>

Jun 24, 2025

Contents

This documentation describes the *efe-bgc* pipeline for Exponential Family Embeddings of Biosynthetic Gene Clusters (BGCs). It includes usage instructions, a formal technical specification, and implementation details.

Chapter 1

Contents

1.1 EFE: Exponential Family Embedding Pipeline

1.1.1 Overview

This module performs probabilistic modeling of Biosynthetic Gene Clusters (BGCs) using Exponential Family Embeddings (EFE). It enables inference of novel or rare BGCs by embedding co-occurrence patterns of biosynthetic domains.

1.1.2 Core Concepts

- **Input matrix**: Wide-format binary matrix (rows = BGCs, columns = PFAM domains).
- **Long-form data**: Triplet format (BGC_ID, Domain_ID, Count) for EFE training.
- **Context vectors**: Probabilistic embeddings of BGCs in a latent feature space.
- **Target embeddings**: Learnable vectors for biosynthetic features (e.g., PFAM domains).
- **Novelty scoring**: Negative log-likelihood of new BGCs under a reference GMM.

1.1.3 Workflow Summary

1. Preprocessing - Convert wide-form matrix to long-form (process load-for-efe) - Create index maps for BGCs and domains
2. Training - Train EFE model using long-form data (train train-efe-model)
3. Inference - Apply trained model to new BGCs (process infer-embeddings)
4. Extraction - Save embedding matrices (process extract-embeddings)
5. Scoring - Calculate GMM-based novelty (calculate calculate-gmm-novelty) - Compute embedding-space dissimilarity (calculate calculate-dissimilarity)

1.1.4 Input/Output Formats

- **Input TSV matrix**: BGC-feature matrix with binary or integer values.
- **Index maps**: JSON files mapping BGC/domain names to indices.
- **Embeddings**: TSV files with BGC ID and embedding vector columns.
- **Trained model**: PyTorch *.pt* file with serialized model state.