A014 · Fungal Gene Essentiality Prediction with Genomic Language Models

Presenter: Chen Liao — Dartmouth College

Authors: Chen Liao, Hannah G. Thomas, Robert A. Cramer

Session: Day 2 · Fri Oct 2, 2026 · 2:15–4:15 PM

Predicting phenotype from genome sequence remains a major challenge in fungal biology, especially for non-model species and clinical isolates with limited functional data. Here, we asked whether genomic language models (GLMs), which learn biologically meaningful representations from DNA sequence, can predict fungal gene essentiality, a phenotype relevant to fundamental biological processes and antifungal target discovery. Using Candida albicans as a primary benchmark, we found that GLM-derived DNA embeddings contain essentiality-related signals, and that model performance was not limited by downstream classifier head capacity, but by the biological information captured in the DNA representation itself. To overcome this bottleneck, we integrated GLM embeddings with two complementary genome-derived features: ortholog-based essentiality and predicted protein-protein interaction information. This multimodal integration model achieved robust, consistent performance across Candida albicans, Saccharomyces cerevisiae, and Schizosaccharomyces pombe, both within species and in cross-species transfer settings. Overall, our results show that GLM-based multimodal models can generalize across diverse fungi and provide a practical framework for prioritizing candidate essential genes from genome sequence alone.

Fungal Gene Essentiality Prediction with Genomic Language Models

Chen Liao[1,*], Hannah G. Thomas[1], Robert A. Cramer[1]

[1]Department of Microbiology and Immunology, Dartmouth College, Hanover, New Hampshire, USA

Background. Genome-wide essentiality profiles reveal genes required for fungal viability and provide a rational starting point for antifungal target discovery. However, such profiles are available for only a limited number of model species because experimental screens are laborious and resource intensive. Although machine learning models can, in principle, predict gene function from sequence, they typically require large amount of species-specific, labeled training data and may not transfer well to genetically divergent fungi. Genomic language models (GLMs), which are pretrained on large and diverse collections of DNA sequences, offer a potential solution by learning transferable sequence representations (i.e., embeddings) without requiring task-specific labels during pretraining. In this work, we tested whether GLM embeddings can capture gene essentiality signals and support prediction both within and across fungal species.

Methods. Using *Candida albicans* (*C. albicans*) as the primary benchmark, we evaluated zero-shot mutation scoring and supervised classifiers trained on frozen Evo2-7B-base embeddings. Each gene was represented by layer-26 embeddings derived from an 8,192-nucleotide genomic window and mean-pooled across its open reading frame. We first compared a lightweight multilayer perceptron (MLP) trained on Evo2 embeddings with published language-model-based approaches, then varied MLP capacity and evaluated alternative classifier families to determine whether performance was constrained by the downstream prediction head. We also assessed the extent to which Evo2 embeddings could recover 13 manually curated biological features. Our final model integrated Evo2 embeddings with ortholog-based essentiality information inferred using OrthoFinder and node2vec representations of flashPPI-predicted protein-protein interaction (PPI) networks. Modality-specific representations and their pairwise interactions were combined through bilinear fusion (**Fig. 1A**). The final multimodal model architecture was evaluated by five-fold cross-validation in *C. albicans*, *Saccharomyces cerevisiae* (*S. cerevisiae*), and *Schizosaccharomyces pombe* (*S. pombe*), and cross-species transfer was assessed by training on two species while holding out the third. Finally, we applied integrated gradients and TF-MoDISco to identify nucleotide-level sequence patterns associated with model predictions.

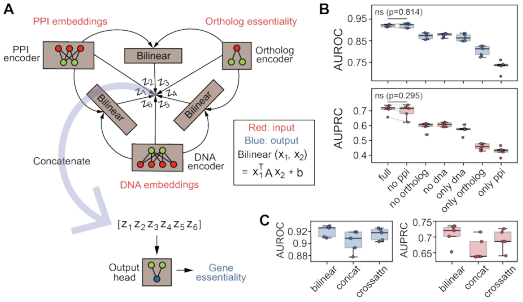

Figure 1. Multimodal architecture and ablation analysis. (A) Bilinear integration of Evo2 DNA, ortholog-essentiality, and predicted PPI-network embeddings. (B) Ablation identifies DNA and ortholog information as the major contributors. (C) Bilinear fusion outperforms concatenation and cross-attention.

Results. Zero-shot Evo2 scores distinguished essential from non-essential *C. albicans* genes across mutation and context settings (AUROC 0.597-0.645; AUPRC 0.249-0.271), with the ~8kb genomic context performing best. Supervised Evo2 embeddings increased performance to mean AUROC 0.864 and AUPRC 0.582, exceeding public baseline models PIC (0.862/0.560) and Bingo (0.850/0.533). Expanding the first hidden layer from 64 to 4,096 neurons or switching among linear, tree-based, and nearest-neighbor classifiers produced no meaningful performance gain. In contrast, a simple MLP based on 13 manually curated features reached AUROC 0.921 and AUPRC 0.676, indicating that model performance was limited not by classifier head capacity, but by the biological information captured in the DNA representation itself. The full multimodal model achieved AUROC 0.921 and AUPRC 0.709 in *C. albicans*. As shown in **Fig. 1B**, removing DNA embeddings and ortholog essentiality reduced performance to 0.877/0.606 and 0.871/0.593, respectively, while removing predicted PPI features had little effect in this species (0.920/0.700). Bilinear fusion also outperformed simple concatenation (0.904/0.662) and cross-attention (0.915/0.690), as demonstrated in **Fig. 1C**. The same framework generalized to *S. cerevisiae* (0.889/0.751) and *S. pombe* (0.856/0.738). When a target species was completely excluded from training, cross-species models still achieved AUROC above 0.82 and AUPRC above 0.57. Attribution-derived motifs were enriched among essential genes and genes encoding core RNA-protein machinery, although most sequence patterns were species-specific rather than universally conserved.

Impacts. This study develops a broadly transferable framework for predicting fungal gene essentiality directly from genome sequence, enabling functional inference in non-model fungal species and clinical isolates that lack functional genomics data. More broadly, it illustrates the value of combining DNA sequence embeddings with complementary, high-level biological features to improve and optimize sequence-to-function models.