The Universal Language of Life – Leveraging Deep Transfer Learning to Model the Biogeosphere

ADRIENNE HOARFROST[1] AND YANA BROMBERG[2]

[1] Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Drive, New Brunswick, NJ: adrienne.hoarfrost@rutgers.edu

[2] Department of Biochemistry and Microbiology, Rutgers University, 76 Lipman Drive, New Brunswick, NJ: yanab@sebs.rutgers.edu

The interactions between the biosphere and geosphere are highly complex, and simple models have often proven insufficient to capture the complexity of biogeochemical systems. Advanced deep learning approaches are promising, but require large datasets to be effective over more traditional statistical methods. While data resources measuring biology and its environmental context in conjunction have grown rapidly, this multidisciplinary growth has lagged behind the explosion of publicly available biological sequencing datasets in recent years. We are using this massive, unlabeled pool of sequencing data to train a deep learning model to learn contextually-relevant representations of biological sequences. Our model encodes functionally and evolutionarily relevant features underlying biological systems as a whole, with more functionally and taxonomically similar sequences having more similar feature representations than more distant sequences. These feature representations, or 'embeddings', are then able to be transferred to a new model, which is fine tuned to predict a task of interest for which labeled training data is limited. In particular, we transfer the knowledge of biological organization encoded in these deeply-learned features in order to predict the environmental context of microbial communities, using relatively small datasets from marine and sedimentary environments for which biological and environmental data have both been measured. We are also able to accurately predict the functional annotation, taxonomic identity, and correct reading frame translation of short-read DNA sequences alone, demonstrating the broad utility of our pretrained embeddings. The downstream prediction tasks for which these pretrained embeddings are useful are not limited to any particular domain, with potential applications spanning questions important to microbial ecology, function, biogeochemistry, and evolution. Our approach and model may be useful in the future to improve biogeochemical components of Earth system models, and to generate hypothetical biological information associated with putative environmental conditions of Earth's past or future, or of conditions characterizing potentially habitable exoplanets.