Title: Machine Learning Reveals Unexpected Environmental Imprints in Microbial Genomes

Abstract: Genomes are traditionally viewed as records of ancestry, not of the environments organisms inhabit. We present computational evidence that environmental conditions can leave detectable, genome-wide sequence signatures. Our alignment-free machine-learning framework uncovers these signals using compact “genome proxies” constructed by stitching distributed sequence segments, enabling efficient large-scale analysis while preserving key statistical structure. Using Chaos Game Representations of DNA sequences and both supervised and unsupervised learning on hundreds of microbial genomes, we systematically optimize feature design and proxy length to balance accuracy and computational cost.

The resulting computational analysis reveals pervasive patterns that reflect shared environments rather than evolutionary relatedness, identifying multiple pairs of distantly related organisms with strikingly similar genomic signatures.

These findings demonstrate how scalable machine learning can expose unexpected environmental imprints in genomic data, providing practical tools for large-scale comparative analysis and data-driven discovery.