

In the realm of genomics, accurate DNA sequence classification is essential for understanding evolutionary relationships, taxonomic distinctions, and microbial adaptations to different environments. This study uses ML-DSP, a groundbreaking method that combines supervised Machine Learning with Digital Signal Processing to create a versatile, alignment-free tool for classifying DNA sequences at various taxonomic levels. ML-DSP's innovation lies in its unique feature vector, which leverages the Pairwise Pearson Correlation Coefficient (PCC) between the magnitude spectra of the Discrete Fourier Transform (DFT) of a numerical representation of DNA sequences and those of other sequences in the training set. This approach not only enhances speed but also maintains an impressive average classification accuracy of over 97%. The methodology of ML-DSP involves three key components: DNA numerical representations, DFT analysis, and the computation of PCC for pairwise distances. Supervised Machine Learning classifiers are employed to classify new DNA sequences based on these distance measurements. A 10-fold cross-validation technique is used to evaluate classifier performance, ensuring the reliability of the results. Additionally, Classical Multidimensional Scaling (MDS) is used for visualization, creating a 3D representation of sequence relationships known as Molecular Distance Maps (MoDMap). To demonstrate ML-DSP's applicability, an extensive analysis of extremophilic microbial genomes was conducted. This dataset, comprising 693 high-quality genome assemblies of microorganisms adapted to extreme temperatures and pH levels, was categorized into two distinct groups: Temperature and pH. This categorization allowed for detailed taxonomic and environmental classifications. The results of the study showcase ML-DSP's proficiency in both supervised and unsupervised learning, exhibiting high accuracy in taxonomic classifications across various k-mer values. Furthermore, the method's ability to identify environmental components in genomic signatures of extremophiles suggests its potential as a valuable tool for understanding how microorganisms adapt to extreme conditions. In summary, ML-DSP represents a significant advancement in DNA sequence classification, offering a robust and efficient approach for researchers studying genetic diversity and microbial adaptation in various environments.