Integration of Heterogeneous DNA Data Sources for Constructing a Balanc

Overview:

In this study, a DNA sequence dataset was created to simulate classification tasks in disease risk assessment based on genomic patterns. The dataset was constructed by integrating synthetic sequences with reference to patterns from reputable databases. Below, the contribution of different datasets is highlighted.

Data Sources Used:

1. NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/)

GenBank is one of the largest and most comprehensive nucleotide sequence databases. For this project, patterns from several disease-related nucleotide sequences were studied. This helped in understanding the nucleotide composition, sequence motifs, and typical sequence lengths used in various disease research datasets.

2. Ensembl Genome Browser (https://www.ensembl.org/)

Ensembl provided insights into genomic annotations and species-specific sequence variations. By analyzing DNA sequences of multiple species, realistic variations were introduced in the synthetic dataset to simulate biological diversity and mutation patterns.

3. Synthetic DNA Generator Script (Custom)

To ensure balanced classes and controlled randomness, custom Python scripts were used to generate sequences of exactly 100 nucleotides, following A, C, G, T distributions. This allowed the creation of balanced classes (e.g., Disease_A, Disease_B) while preserving realistic features such as k-mer distributions, inspired by the studied patterns from GenBank and Ensembl.

Datasets Created:

- sequences.csv: Raw sequences (synthetic)
- parameters.csv: Parameters for sequence mutation and variations
- dna_dataset.csv: Final labeled dataset used for model training

Significance:

The combination of these datasets ensured that the final dna_dataset.csv was:

- Balanced across classes.
- Mimicking real-world patterns (from GenBank and Ensembl).
- Controlled and reproducible for machine learning algorithm benchmarking.

References:

- Benson, D. A., et al. (2018). GenBank. Nucleic Acids Research, 46(D1), D41-D47.

https://doi.org/10.1093/nar/gkx1094

- Yates, A. D., et al. (2020). Ensembl 2020. Nucleic Acids Research, 48(D1), D682-D688. https://doi.org/10.1093/nar/gkz966