Integration of Heterogeneous DNA Data Sources for Constructing a Balanc

Overview:

This documentation describes how multiple data sources were integrated to create the final DNA sequence dataset (dna_dataset.csv), used for machine learning-based disease risk classification.

Three primary data sources were leveraged:

- NCBI GenBank: Provided real-world disease-related DNA sequences.
- Ensembl Genome Browser: Offered annotated genomic variations.
- Synthetic DNA Generator: Ensured class balance and controlled sequence variations.

Data Sources Used and Integration Process:

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

Real DNA sequences related to known diseases were retrieved and analyzed. These sequences offered authentic patterns such as repetitive motifs, GC content, and known mutation signatures.

These patterns were extracted and served as reference motifs.

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

Genomic annotations and variations across different species were utilized. These sequences helped simulate realistic variations and mutation hotspots. Patterns such as SNPs (Single Nucleotide Polymorphisms) and structural variations were observed and introduced in the synthetic sequences.

3. Synthetic DNA Generator Script (Custom)

Using Python-based custom scripts, the final sequences were generated by combining insights from the above sources. The generator applied motifs from GenBank and variations from Ensembl while ensuring the sequence length (100 nucleotides) and balanced class distribution.

This facilitated controlled dataset creation ensuring:

- Class labels were balanced (Disease_A, Disease_B, etc.).
- Patterns from real datasets were realistically represented.
- Sequences included synthetic noise and diversity inspired by real-world data.

Datasets Created:

- sequences.csv: Contains raw DNA sequences generated with embedded real-world motifs.
- parameters.csv: Contains mutation rates and sequence diversity configurations used in the generator.
- dna_dataset.csv: Final dataset with labeled sequences ready for ML tasks.

Key Contribution of Each Dataset to dna_dataset.csv:

1. GenBank sequences provided disease-specific nucleotide patterns.

2. Ensembl data introduced realistic variations such as SNPs.

3. Synthetic generator combined these patterns while balancing the dataset and ensuring quality control.

Significance of Integration:

The resulting dna_dataset.csv combines biological realism with machine learning readiness, allowing robust benchmarking of classification algorithms while ensuring traceability to reference data.

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