Final Project Report

1. Project Topic and Context

The objective of this project was to compare two approaches for identifying the origin of replication (oriC) in bacterial genomes:

- A **machine learning-based method**, using k-mer frequency features and logistic regression.
- The **cumulative GC skew method**, a classical approach discussed in class, which identifies oriC based on asymmetries in nucleotide composition around the replication origin.

This comparison aimed to evaluate whether machine learning could provide a viable or improved alternative to traditional sequence analysis techniques.

2. Initial Plan

The initial plan involved:

- Extracting oriC sequences from the DoriC database and preparing them as positive training examples.
- Sampling non-oriC sequences from a representative bacterial genome as negative examples.
- Constructing a feature matrix from k-mer frequencies (with k = 3,4,5).
- Training and testing a logistic regression classifier.
- Comparing the locations and accuracy of machine learning predictions to the output of cumulative GC skew analysis.

3. Challenges Encountered

Several challenges arose during the project:

- **New to Machine Learning**: Since this was the first experience using ML, understanding how to structure data, choose models, and interpret results was initially overwhelming.

- **Class Imbalance**: The dataset had many more oriC examples than non-oriC sequences, which biased the model and hurt generalization. This was addressed by carefully sampling a balanced number of negative examples.

4. Methodology and Reasoning

- **Data Source**: OriC sequences were downloaded from the DoriC database and truncated to 188 bp since 188 was the average length of the OriC sequence. Also to combat against data imbalance, I picked the top 35 most frequently appearing bacterial genomes, then randomly generated 1000 188 sequence lengths from each of the 35.
- **Feature Engineering**: I converted each DNA window into **k-mer frequency vectors** — essentially counting short DNA patterns (k = 3, 4, 5) to use as model input.
- **Modeling**: A logistic regression model was trained on balanced positive and negative datasets. This choice was made for simplicity and interpretability.
- **Baseline Comparison**: A comparison of Oric prediction Method was drawn shows the probabilities of Oric region in a genome
- **Evaluation**: The two methods were compared in terms of predicted oriC positions, precision/recall (for ML), and qualitative agreement with the GC skew peak.

5. Results

I compared performance using k =3,4,5
Metrics like **accuracy, precision, recall, and F1 score** all improved as k increased, with the best results atk = 5
This made sense since longer k-mers capture more specific patterns.

But Then I Tested It on Real DNA:

I ran the model on an actual bacterial genome and compared its predictions to calSkew():

- At **k = 3**, the top prediction was **only 52,000 bp** away from the skew result.

- At **k = 4**, it got worse — **943,000 bp** off.

- At **k = 5**, it drifted dramatically, **over 3 million bp** from the skew prediction.

Despite performing well on training data, higher k values **failed to generalize** on real DNA.

Potential reasons for this higher k performs worse on real data

- Real DNA sequences (like those from GenBank) are **longer, more diverse**, and contain **noisy or unknown patterns**.
- **Higher k-mers are too specific** — they may not generalize to unseen DNA because those exact patterns may not occur again.
- This leads to **overfitting**: the model was too tuned to training patterns and can't recognize different valid oriC signatures.
- Also, as k increases, the number of possible k-mers grows exponentially (e.g., $4^3 = 64$, but $4^5 = 1024$). Many of those may **never appear** in test data, making predictions unstable.

Conclusions

Both the machine learning and GC skew methods proved effective in approximating the oriC region. In testing, the machine learning model using **k = 3** for k-mer frequencies yielded predictions that aligned more closely with the oriC position identified by the calSkew() function introduced in class. While the GC skew method remains a simple and biologically grounded approach, it is less adaptable in cases where GC asymmetry is weak or ambiguous. In contrast, machine learning offers greater flexibility and, with sufficient training data and proper validation, has the potential to complement or even surpass classical sequence analysis techniques.

7. Future Directions

- **Sliding Window Genome Scanning:** Apply the trained ML model across the entire genome using a sliding window to generate oriC probability profiles and compare prediction peaks to GC skew minima.

- **Cross-Species Validation:** Test the generalization capability of the model by applying it to bacterial species not included in the training data.

- **More Complex Models:** Investigate the use of convolutional neural networks or ensemble learning techniques to capture broader sequence dependencies and improve classification accuracy.

- **Integration with Motif Discovery:** Combine ML outputs with known oriC motifs (e.g., DnaA boxes) to enhance prediction precision and biological interpretability.