Assignment 3: Hidden Markov Models and Prokaryotic Gene Prediction

Due Date: 10/2
Topics Covered: HMMs (Forward & Viterbi algorithms), probabilistic sequence modeling, Glimmer for gene prediction

---

Reading Assignments

- Review: Durbin et al., Chapter 4 (HMMs), including dishonest casino example (p. 54)
- Review: Glimmer paper and Chapter 2 (gene prediction principles)
- Prepare: Chapters 5 and 6 (future material)

Problems

1. Simulate a Hidden Markov Model (8 pts)

Task:

- Implement the "dishonest casino" HMM (Durbin Fig. 3.5).
- Start in the Fair state: `Pr(F) = 1`, `Pr(L) = 0`.
- Transition probability: 0.05 between states.
- Emission: Fair emits 1–6 uniformly; Loaded favors 6.
- Generate a random sequence of 300 dice rolls.

Deliverables:

- Source code (`casino_simulator.py` or similar).
- Instructions to run the code in your `report.txt`.
- Submission of the actual 300-roll instance is optional.

---

2. Load Benchmark Data (provided)

Task:

- Download two sequences from the course website (`casino.benchmark1.txt`, `casino.benchmark2.txt`).

---

3. Forward Algorithm – Compute Probabilities (10 pts)

Task:

- Implement the forward algorithm.

- Compute and report the probability of each benchmark sequence given the HMM.

Deliverables:

- Source code with instructions.
- Report of `log P(sequence | model)` for each benchmark.

---

4. Viterbi Decoding – Most Likely State Path (10 pts)

Task:

- Implement the Viterbi algorithm.
- Compute the most likely state sequence (`F`, `L`) for each benchmark.
- Save results in `viterbi.1.txt` and `viterbi.2.txt`, one label per roll.

Deliverables:

- Source code.
- Two labeled output files.

---

5. Genome Retrieval (0 pts)

Task:

- Download the following two Bacillus anthracis strains from GenBank:
  - Ames ancestor (NC_007530) – virulent
  - Ames (NC_003997) – lab strain

Note: These genomes will be used in both Homework #3 and #4.

---

6. GLIMMER Gene Prediction (8 pts total)

Task:

- Visit the GLIMMER website.
- Review documentation and install Glimmer3.
- Run gene prediction on each downloaded genome from Step 5 above.

Deliverables:

- Gene prediction results for both genomes.
- Submit only the output gene calls (e.g., `.predict` or `.coords` files).
- Include basic summary statistics in your `report.txt` (e.g., number of predicted genes).