Students:

- Julian Buetzer
- Giacomo Frison
- Cindy Guerrero

Viterbi Algorithm: Manual Solution for CG-rich Genome Regions

Problem Setup: Case 1

Sequence: GGACTGAA (8 nucleotides)

Hidden States: - **R** (CG-rich): emits C and G with higher probability - **P** (CG-poor): emits T and A with higher probability

Initial Probabilities: - _R = 0.5 - _P = 0.5

Transition Probabilities: - a_RR = 0.5 (R → R) - a_RP = 0.5 (R → P) - a_PR = 0.6 (P → R) - a_PP = 0.4 (P → P)

Emission Probabilities:

Nucleotide	State R	State P
T	0.2	0.3
C	0.3	0.2
A	0.2	0.3
G	0.3	0.2

Viterbi Algorithm Steps

(Exercise 1) Use a Viterbi algorithm to define the most likely sequence of hidden states for sequence GGACTGAA.

The Viterbi algorithm finds the most likely sequence of hidden states using dynamic programming.

Notation

- **P[t][s]**: Maximum probability of ending at state s at time t
- **Q[t][s]**: Previous state that led to state s at time t (for backtracking)

Step 0: Initialization (t=0, observation: G)

P[0][R] = _R × b_R(G) = 0.5 × 0.3 = **0.15**

P[0][P] = _P × b_P(G) = 0.5 × 0.2 = **0.10**

Q[0][R] = None (initial state)
Q[0][P] = None (initial state)

Step 1: t=1 (observation: G)

For state R: - From R: P[0][R] × a_RR × b_R(G) = 0.15 × 0.5 × 0.3 = 0.0225 - From P: P[0][P] × a_PR × b_R(G) = 0.10 × 0.6 × 0.3 = 0.0180 - **P[1][R]** = max(0.0225, 0.0180) = **0.0225** - **Q[1][R]** = R

For state P: - From R: P[0][R] × a_RP × b_P(G) = 0.15 × 0.5 × 0.2 = 0.0150 - From P: P[0][P] × a_PP × b_P(G) = 0.10 × 0.4 × 0.2 = 0.0080 - **P[1][P]** = max(0.0150, 0.0080) = **0.0150** - **Q[1][P]** = R

Step 2: t=2 (observation: A)

For state R: - From R: P[1][R] × a_RR × b_R(A) = 0.0225 × 0.5 × 0.2 = 0.00225 - From P: P[1][P] × a_PR × b_R(A) = 0.0150 × 0.6 × 0.2 = 0.00180 - **P[2][R]** = max(0.00225, 0.00180) = **0.00225** - **Q[2][R]** = R

For state P: - From R: P[1][R] × a_RP × b_P(A) = 0.0225 × 0.5 × 0.3 = 0.003375 - From P: P[1][P] × a_PP × b_P(A) = 0.0150 × 0.4 × 0.3 = 0.00180 - **P[2][P]** = max(0.003375, 0.00180) = **0.003375** - **Q[2][P]** = R

Step 3: t=3 (observation: C)

For state R: - From R: P[2][R] × a_RR × b_R(C) = 0.00225 × 0.5 × 0.3 = 0.0003375 - From P: P[2][P] × a_PR × b_R(C) = 0.003375 × 0.6 × 0.3 = 0.0006075 - **P[3][R]** = max(0.0003375, 0.0006075) = **0.0006075** - **Q[3][R]** = P

For state P: - From R: P[2][R] × a_RP × b_P(C) = 0.00225 × 0.5 × 0.2 = 0.000225 - From P: P[2][P] × a_PP × b_P(C) = 0.003375 × 0.4 × 0.2 = 0.00027 - **P[3][P]** = max(0.000225, 0.00027) = **0.00027** - **Q[3][P]** = P

Step 4: t=4 (observation: T)

For state R: - From R: P[3][R] × a_RR × b_R(T) = 0.0006075 × 0.5 × 0.2 = 0.00006075 - From P: P[3][P] × a_PR × b_R(T) = 0.00027 × 0.6 × 0.2 = 0.0000324 - **P[4][R]** = max(0.00006075, 0.0000324) = **0.00006075** - **Q[4][R]** = R

For state P: - From R: P[3][R] × a_RP × b_P(T) = 0.0006075 × 0.5 × 0.3 = 0.000091125 - From P: P[3][P] × a_PP × b_P(T) = 0.00027 × 0.4 × 0.3 =

0.0000324 - **P[4][P]** = max(0.000091125, 0.0000324) = **0.000091125** - **Q[4][P]** = R

Step 5: t=5 (observation: G)

For state R: - From R: P[4][R] × a_RR × b_R(G) = 0.00006075 × 0.5 × 0.3 = 0.0000091125 - From P: P[4][P] × a_PR × b_R(G) = 0.000091125 × 0.6 × 0.3 = 0.0000164025 - **P[5][R]** = max(0.0000091125, 0.0000164025) = **0.0000164025** - **Q[5][R]** = P

For state P: - From R: P[4][R] × a_RP × b_P(G) = 0.00006075 × 0.5 × 0.2 = 0.000006075 - From P: P[4][P] × a_PP × b_P(G) = 0.000091125 × 0.4 × 0.2 = 0.00000729 - **P[5][P]** = max(0.000006075, 0.00000729) = **0.00000729** - **Q[5][P]** = P

Step 6: t=6 (observation: A)

For state R: - From R: P[5][R] × a_RR × b_R(A) = 0.0000164025 × 0.5 × 0.2 = 0.00000164025 - From P: P[5][P] × a_PR × b_R(A) = 0.00000729 × 0.6 × 0.2 = 0.0000008748 - **P[6][R]** = max(0.00000164025, 0.0000008748) = **0.00000164025** - **Q[6][R]** = R

For state P: - From R: P[5][R] × a_RP × b_P(A) = 0.0000164025 × 0.5 × 0.3 = 0.000002460375 - From P: P[5][P] × a_PP × b_P(A) = 0.00000729 × 0.4 × 0.3 = 0.0000008748 - **P[6][P]** = max(0.000002460375, 0.0000008748) = **0.000002460375** - **Q[6][P]** = R

Step 7: t=7 (observation: A)

For state R: - From R: P[6][R] × a_RR × b_R(A) = 0.00000164025 × 0.5 × 0.2 = 0.000000164025 - From P: P[6][P] × a_PR × b_R(A) = 0.000002460375 × 0.6 × 0.2 = 0.000000295245 - **P[7][R]** = max(0.000000164025, 0.000000295245) = **0.000000295245** - **Q[7][R]** = P

For state P: - From R: P[6][R] × a_RP × b_P(A) = 0.00000164025 × 0.5 × 0.3 = 0.0000002460375 - From P: P[6][P] × a_PP × b_P(A) = 0.000002460375 × 0.4 × 0.3 = 0.000000295245 - **P[7][P]** = max(0.0000002460375, 0.000000295245) = **0.000000295245** - **Q[7][P]** = P

Backtracking to Find the Most Likely Path

At the final time step (t=7): - P[7][R] = 0.000000295245 - P[7][P] = 0.000000295245

Both states have the same probability! We can choose either, but typically we choose the one with the higher index or use a tie-breaking rule. At this point, we can trace back from P (since it's the last state in our Q matrix, or A matrix according to the other book, Lecture 5, part II, slide 24):

Path reconstruction: - t=7: **P** (final state) - t=6: Q[7][P] = **P** - t=5: Q[6][P] = **R** - t=4: Q[5][R] = **P** - t=3: Q[4][P] = **R** - t=2: Q[3][R] = **P** - t=1: Q[2][P] = **R** - t=0: Q[1][R] = **R**

Most likely sequence of hidden states: R-R-P-R-P-R-P-P

Sequence alignment:

```
Observation:  G  G  A  C  T  G  A  A
Hidden State: R  R  P  R  P  R  P  P
```

Summary

The Viterbi algorithm has determined that the most likely sequence of hidden states for "GGACTGAA" is **R-R-P-R-P-R-P-P**, where: - **R** = CG-rich region - **P** = CG-poor region

This makes biological sense because: - The sequence starts with GG (two Gs), which favors the CG-rich state - The middle contains A, C, T which are more likely in CG-poor regions - The algorithm balances emission probabilities with transition probabilities to find the optimal path

Probability P(X) of the sequence GGCA calculated through forward and backward algorithm:

For states, initial and transition probabilities and emission probabilities see above

Forward altgorithm for sequence GGCA:

initialization (t=1, G) a1(R) = R x P(G|R) = 0.5 x 0.3 = 0.15 a1(P) = P x P(G|P) = 0.5 x 0.2 = 0.10

recursion (t=2, G) a2(R) = [a1(R)(a_PR) + a1(P)(a_PP)] x P(G|R) = [(0.15 x 0.6) + (0.10 x 0.4)] x 0.3 = **0.0039** a2(R) = [a1(R)(a_PP) + a1(P)(a_PR)] x P(G|R) = [(0.15 x 0.4) + (0.10 x 0.6)] x 0.2 = **0.024**

recursion (t=3 C) a3(R) = [a2(R)(a_PR) + a2(P)(a_PP)] x P(C|R) = [(0.0039 x 0.6) + (0.024 x 0.4)] x 0.3 = **0.0099** a3(P) = [a2(R)(a_PP) + a2(P)(a_PR)] x P(C|P) = [(0.0039 x 0.4) + (0.024 x 0.6)] x 0.2 = **0.006**

recursion (t=4, A) a4(R) = [a3(R)(a_PR) + a3(P)(a_PP)] x P(A|R) = [(0.0099 x 0.6) + (0.006 x 0.4)] x 0.2 = **0.001668** a4(P) = [a3(R)(a_PR) + a3(P)(a_PP)] x P(A|P) = [(0.0099 x 0.4) + (0.006 x 0.6)] x 0.2 = **0.001512**

termination P(GGCA) a4(R) + a4(P) = 0.001668 + 0.001512 = **0.00318**

Backward algorithm

Initialization (t=4, A) b4(R) = 1 ; b4(P) = 1

Step t=3, C to A b3(R) = [0.6 x ((P(A|R) x b4(R)) + (0.4 X (P(A|P) x b4(P))] = [(0.6 X 0.2 x 1) + (0.4 x 0.2 x 1)] = 0.12 + 0.08 = 0.20 b3(P) = [0.6 x ((P(A|P) x b4(P) + (0.4 X (P(A|R) x b4(R))] = [(0.6 X 0.2 x 1) + (0.4 x 0.2 x 1)] = 0.12 + 0.08 = 0.20

Step t=2, G to C b2(R) = [0.6 x ((P(C|R) x b3(R)) + (0.4 X (P(C|P) x b3(P))] = [(0.6 X 0.3 x 0.20) + (0.4 x 0.2 x 0.20)] = 0.036 + 0.016 = 0.052 b2(P) = [0.6 x ((P(C|P) x b3(P)) + (0.4 X (P(C|R) x b3(R))] = [(0.6 X 0.2 x 0.20) + (0.4 x 0.3 x 0.20)] = 0.024 + 0.024 = 0.048

Step t=1, G to G b1(R) = [0.6 x ((P(G|R) x b3(R)) + (0.4 X (P(C|P) x b3(P))] = [(0.6 X 0.3 x 0.0052) + (0.4 x 0.2 x 0.0048)] = 0.00936 + 0.00384 = 0.0132 b1(P) = [0.6 x ((P(C|P) x b3(P)) + (0.4 X (P(C|R) x b3(R))] = [(0.6 X 0.3 x 0.0048) + (0.4 x 0.2 x 0.0052)] = 0.00576 + 0.00624 = 0.012

P sequence P(s) x P(G|R-P) x b1(R-P) = (0.5 x 0.3 x 0.0132) + (0.5 x 0.2 x 0.012) = 0.00198 + 0.0012 = **0.00318**

Scaling variables

s = 1 / (at(S))

Initilization t=1, G

a1(R) + a1(P) = 0.15 + 0.10 = 0.25 s1 = 1/0.25 = 4 as1(R) = 0.6 as1(P) = 0.4

t=2, G as2(R) = [a1(R)(a_PR) + a1(P)(a_PP)] x P(G|R) = **0.0039** as2(P) = [a1(P)(a_PR) + a1(R)(a_PP)] x P(G|R) = **0.024** s2 = 1/(0.0039 + 0.0024) = 1/0.0063 = 15.87 as2 (R) = 0.0039 x 15.87 = 0.619 as2 (P) = 0.024 x 15.87 = 0.381

t=3, C as3(R) = [a2(R)(a_PR) + a2(P)(a_PP)] x P(G|R) = **0.0099** as3(P) = [a2(P)(a_PR) + a2(R)(a_PP)] x P(G|R) = **0.006** s3 = 1/(0.0099 + 0.006) = 1/0.0159 = 62.9 as3 (R) = 0.0099 x 62.9 = 0.622 as3 (P) = 0.006 x 62.9 = 0.378

t=4, A as4(R) = [a3(R)(a_PR) + a3(P)(a_PP)] x P(G|R) = **0.001668** as4(P) = [a3(P)(a_PR) + a3(R)(a_PP)] x P(G|R) = **0.001512** s4 = 1/(0.001668 + 0.001512) = 1/0.00318 = 314.5 as4 (R) = 0.001668 x 314.5 = 0.525 as4 (P) = 0.001512 x 314.5 = 0.475

P(GGCA) $1 \,/\, (\mathrm{s1} + \mathrm{s2} + \mathrm{s3} + \mathrm{s4}) = 1 \,/\, (4 + 15.87 + 62.9 + 314.5) =$ **0.00318**