2.

Στο συγκεκριμένο ερώτημα μας ζητήθηκε να υλοποιήσουμε ένα Hidden Markov Model το οποία θα αποκωδικοποιεί την πιο πιθανή ακολουθία δύο καταστάσεων, α και β, δοθείσης μιας ακολουθίας από πουρίνες και πυριμιδίνες. Στην περίπτωση μας η αλληλουχία η οποία δόθηκε ήταν η “GGCT”.

Με βάση το σχήμα το οποίο δόθηκε καταλήξαμε στις παρακάτω πιθανότητες.

* Πιθανότητες μετάβασης(Transition Probabilities):

α β

α 0.9 0.1

β 0.1 0.9

* Πιθανότητες εκπομπής(Emission Probabilities):

A G T C

α 0.4 0.4 0.1 0.1

β 0.2 0.2 0.3 0.3

Προκειμένου ωστόσο να υλοποιήσουμε το HMM είναι αναγκαίο να γνωρίζουμε και την πιθανότητα εμφάνισης καθενός εκ των δύο καταστάσεων α και β(a priori probabilities). Επομένως έχουμε τα παρακάτω στοιχεία.

Η πιθανότητα να εμφανιστεί η κατάσταση «α» προκύπτει από τις πιθανότητες μετάβασης από α->α και β->α. Επομένως η πιθανότητα εμφάνισης της κατάστασης «α» εκφράζεται από την σχέση *α = 0.9α+0.1β* (1)

Ακολουθώντας την ίδια λογική, για την κατάσταση «β» προκύπτει η σχέση:

*β = 0.1α+0.9β* (2)

Τελικά, λύνοντας το σύστημα των σχέσεων (1),(2) προκύπτουν οι πιθανότητες εμφάνισης των α, β τις οποίες αναζητούσαμε.

Στο Hidden Markov Model κατά βάση γίνεται εφαρμογή του θεωρήματος Bayes. Kατά τον υπολογισμό της πιθανότητας για την πρώτη παρατήρηση( ένα εκ των AGCT ) να είναι α ή β χρειαζόμαστε μονάχα τις πιθανότητες εκπομπής και των a priori πιθανοτήτων. Αυτό συμβαίνει διότι δεν υφίσταται προηγούμενη κατάσταση επομένως οι πιθανότητες μετάβασης δεν χρησιμοποιούνται. Σε αυτό το σημείο αξίζει να σημειωθεί πως οι παραπάνω πιθανότητες προστίθενται και δεν πολλαπλασιάζονται αφού χρησιμοποιούμε λογαριθμικές βαθμολογίες πιθανοτήτων , όπως υποδείχθηκε στην εκφώνηση.

Στη συνέχεια για την αποκωδικοποίηση των επόμενων παρατηρήσεων, ο υπολογισμός των πιθανοτήτων καθενός εκ των δύο καταστάσεων προκύπτει προσθέτοντας την πιθανότητα της προηγούμενης κατάστασης μαζί με τις αντίστοιχες πιθανότητες μετάβασης και εκπομπής. Ξανά, λόγο της λογαριθμικής βαθμολογίας οι πιθανότητες προστίθενται και δεν πολλαπλασιάζονται.

Για την υλοποίηση της παραπάνω λειτουργίας έγινε χρήση των βασικών αρχών του αλγορίθμου Viterbi. Επομένως στην περίπτωση μας, κάθε φορά που υπολογίζονται οι πιθανότητες για την αποκωδικοποίηση μιας σειράς παρατηρήσεων(sequence), τα αποτελέσματα αποθηκεύονται. Αυτό έχει ως αποτέλεσμα με κάθε νέα προσθήκη μιας παρατήρησης στην προηγούμενη σειρά παρατηρήσεων, να υπολογίζουμε μονάχα τόσες πιθανότητες όσο και οι καταστάσεις μας. Αφού στην περίπτωση μας οι καταστάσεις είναι δύο, σε κάθε του βήμα του ο αλγόριθμος θα κάνει υπολογισμούς για μονάχα δύο καταστάσεις. Αν δεν χρησιμοποιήσουμε την παραπάνω λογική, το σύνολο τον πράξεων το οποίο θα έπρεπε να γίνουν π.χ. για την σειρά «GGCT» θα ήταν ίσο με Σ*2n*, όπου 1<=n<=4, ενώ τώρα γίνονται μονάχα 2\*n πράξεις.