**Πανεπιστήμιο Πειραιώς**

Σχολή Πληροφορικής

Τμήμα Πληροφορικής

Εικόνα που περιέχει κείμενο, clipart

Περιγραφή που δημιουργήθηκε αυτόματα

Ιούλιος 2022

# Βιοπληροφορική

Ομάδα εκπόνησης εργασίας

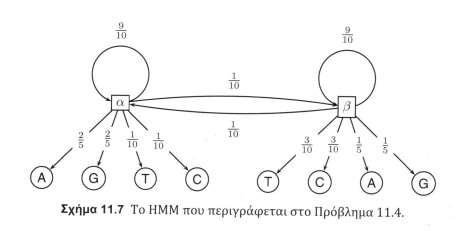
|  |
| --- |
| Μιχάλης Βασιλακάκης, Π19026 |
| Γεωργιάδης Ευθύμιος Πάτροκλος, Π19031 |
| Ιωάννης Ιάσων Μποϊδάνης, Π19217 |
| Ιωάννης Σταθόπουλος, Π19240 |

Θέμα 2

Άσκηση 11.4

Για την επίλυση του προβλήματος θα χρειαστεί να αξιοποιήσουμε τον αλγόριθμο Viterbi. Oι τρεις βασικοί πίνακες που απαιτούνται για την εκτέλεση του αλγορίθμου είναι ο πίνακας πιθανότητας αρχικής κατάστασης Π, ο πίνακας πιθανότητας μετάβασης από κατάσταση σε κατάσταση Α και o πίνακας εκπομπής συμβόλων πιθανών καταστάσεων Β. Μέσω του σχήματος Hidden Markov Model της άσκησης, θα μπορέσουμε να εξάγουμε τις πληροφορίες για τους παραπάνω πίνακες.

Το σχήμα:



Γνωρίζουμε ήδη τα εξής από την εκφώνηση:

* Υπάρχουν 2 καταστάσεις: α, β
* Και οι 2 καταστάσεις περιέχουν τις παρατηρήσεις Α, G, T, C

Από το σχήμα μπορούμε να εξάγουμε τα εξής δεδομένα:

* Πίνακας πιθανότητας αρχικής κατάστασης Π: [0,5 ; 0,5]
* Πίνακας πιθανότητας μετάβασης από κατάσταση σε κατάσταση Α: [0,9 0,1 ; 0,1 0,9] (αναλυτικά από κατάσταση σε κατάσταση: [p(α/α), p(β/α) ; p(α/β) p(β/β)])

Με λογαριθμική βαθμολογία: log2([0,9 0,1 ; 0,1 0,9]) [2]

* Πίνακας εκπομπής συμβόλων πιθανών καταστάσεων Β:

[0,4 0,4 0,1 0,1 ; 0,2 0,2 0,3 0,3]

([p(A/α), p(G/α), p(T/α), p(C/α) ; p(A/β), p(G/β), p(T/β), p(C/β)])

Με λογαριθμική βαθμολογία: log2([0,4 0,4 0,1 0,1 ; 0,2 0,2 0,3 0,3]) [2]

Χρησιμοποιώντας αυτούς τους 3 πίνακες, τον πίνακα καταστάσεων, τον πίνακα παρατηρήσεων και την ακολουθία παρατηρήσεων που μας δίνεται “GGCT”, μπορούμε να εφαρμόσουμε τον αλγόριθμο Viterbi για να αποκωδικοποιήσουμε την πιο πιθανή ακολουθία καταστάσεων α, β για την ακολουθία “GGCT”.

Αλγόριθμος Viterbi

Είσοδος: [1]

1. To διάστημα καταστάσεων (α, β)
2. Το διάστημα παρατηρήσεων (A, T, C, G)
3. Πίνακας πιθανότητας αρχικής κατάστασης Π
4. Πίνακας πιθανότητας μετάβασης από κατάσταση σε κατάσταση Α
5. Πίνακας εκπομπής συμβόλων πιθανών καταστάσεων Β
6. Η ακολουθία παρατηρήσεων “GGCT”

Έξοδος:

Η πιο πιθανή ακολουθία καταστάσεων της ακολουθίας παρατηρήσεων

Έστω Ν, το πλήθος των καταστάσεων (Ν=2) και Μ το πλήθος των παρατηρήσεων (Μ=4). Αρχικοποιούμε 3 πίνακες Τ1, Τ2, k ως πίνακες ΝxΜ διαστάσεων. Τι αντιπροσωπεύουν αυτοί οι πίνακες όμως;

* Τ1: To κάθε στοιχείο (i, j) του πίνακα Τ1 αντιπροσωπεύει την πιθανότητα της αντίστοιχης ακολουθίας παρατηρήσεων μέχρι στιγμής
* Τ2: To κάθε στοιχείο (i, j) του πίνακα Τ2 αντιπροσωπεύει την τελευταία κατάσταση της πιο πιθανής ακολουθίας παρατηρήσεων μέχρι στιγμής
* k: To κάθε στοιχείο (i, j) του πίνακα k αντιπροσωπεύει την πιθανότητα της προηγούμενης κατάσταση της τωρινής κατάστασης του αλγορίθμου

Αρχικά θα βρούμε τις πιθανότητες των αρχικών καταστάσεων δεδομένου της πρώτης παρατήρησης, δηλαδή ποια είναι η πιο πιθανή αρχική κατάσταση δεδομένου της παρατήρησης “G”. Για i = 1 έως Ν=2 και j=1, εισάγουμε στον πίνακα Τ1 το γινόμενο του Π(i, 1) και B(i, S1),

όπου S1 είναι η πρώτη παρατήρηση της ακολουθίας παρατηρήσεων.

Έπειτα, θα χρησιμοποιήσουμε έναν τριπλό βρόγχο for για την εύρεση των υπόλοιπων καταστάσεων δεδομένου της αντίστοιχης παρατήρησης. Για j=2 έως Μ=4, εκτελούμε τον επόμενο βρόγχο: Για i=1 έως N=2, εκτελούμε τον 3ο βρόγχο: Για q=1 έως N=2, υπολογίζουμε το k(q, j) ως το γινόμενο του Τ1(q, j-1), του Α(q, i) και του Β(i, j). Ουσιαστικά, υπολογίζουμε την πιθανότητα της προηγούμενης κατάστασης της j-στης κατάστασης.

Αφού βγούμε από τον 3ο βρόγχο, υπολογίζουμε την πιο πιθανή προηγούμενη κατάσταση της j-στης κατάστασης max, με την χρήση της μεθόδου argmax() [3]. H μεταβλητή max θα αντιπροσωπεύεται με φυσικό αριθμό αντί του ονόματος της κατάστασης (δηλαδή με 1 αντί για α και 2 αντί για β). Έπειτα υπολογίζουμε το στοιχείο T1(i, j) ως το γινόμενο του Τ1(max, j-1) με το Α(max, i) και το Β(i, j) και εξισώνουμε το στοιχείο T2(i,j) με το max. Τέλος και των 2 υπόλοιπων βρόγχων.

Υπολογίζουμε την πιο πιθανή προηγούμενη κατάστασης της τελευταίας κατάστασης, την μεταβλητή k, για τις 2 περιπτώσεις όπου η τελευταία κατάσταση είναι η α ή η β. Ύστερα, εκτελούμε την διαδικασία αναδρομής για την εύρεση της πιο πιθανής ακολουθίας καταστάσεων της δεδομένης ακολουθίας παρατηρήσεων.

Για o=M έως 1, με βήμα -1, η ο-οστή πιο πιθανή κατάσταση της ακολουθίας καταστάσεων της δεδομένης ακολουθίας παρατηρήσεων είναι ίση με την αντίστοιχη κατάσταση που αντιπροσωπεύει η μεταβλητή k. Έπειτα, εξισώνουμε την μεταβλητή k με το στοιχείο Τ2(k, o) του πίνακα Τ2 και επαναλαμβάνουμε μέχρι το τέλους του βρόχου. Επιστρέφουμε την ακολουθία καταστάσεων της δεδομένης ακολουθίας παρατηρήσεων και λήγουμε τον αλγόριθμο.

Βιβλιογραφία

[1] <https://en.wikipedia.org/wiki/Viterbi_algorithm>

[2] <https://www.mathworks.com/help/matlab/ref/log.html>

[3] <https://www.mathworks.com/matlabcentral/answers/408966-is-there-a-function-for-argmax>

# O πηγαίος κώδικας (source code) αυτού του θέματος βρίσκεται στα αρχεία main\_ex2.m και viterbi.m

Θέμα 3

Άσκηση 6.12

Για τα βιολογικά δεδομένα έχουμε επιλέξει τα ισοένζυμα της φωσφορυλάσης του γλυκογόνου του ήπατος (PYGL glycosen phosphorylase L [Homo Sapiens]) και του εγκεφάλου (PYGB glycosen phosphorylase B [Homo Sapiens]).

Κανόνες του παιχνιδιού:

Δύο παίκτες παίζουν το αντίστοιχο παιχνίδι με 2 χρωμοσώματα μήκους n και m νουκλεοτίδια αντίστοιχα. Σε κάθε γύρο του παιχνιδιού, ένας παίκτης μπορεί να καταστρέψει ένα από τα χρωμοσώματα και να διαχωρίσει το άλλο σε 2 μη κενά τμήματα. Ο παίκτης που διαγράφει το τελευταίο νουκλεοτίδιο κερδίζει.

*Προσοχή:* Για να κερδίσει ένας παίκτης πρέπει να ολοκληρώσει τον γύρο του. Δηλαδή υποχρεούται να διαγράψει ένα από τα χρωμοσώματα και να διαχωρίσει το άλλο σε 2 μη κενά τμήματα. Εάν δεν μπορεί να εκτελέσει και τις δύο παραπάνω ενέργειες, τότε χάνει. Παραδείγματος χάρη, όταν ένας παίκτης διαγράφει το ένα χρωμόσωμα και το άλλο χρωμόσωμα έχει μόνο ένα νουκλεοτίδιο, τότε δεν μπορεί να το διαχωρίσει σε 2 μη κενά ακέραια μέρη και δεν μπορεί να ολοκληρώσει τον γύρο του. Επομένως, παρόλο που έχει τεχνικά διαγράψει το τελευταίο χρωμόσωμα, χάνει επειδή δεν μπορεί να εκπληρώσει τον γύρο του.

Ως αποτέλεσμα, η νικηφόρα συνθήκη για έναν παίκτη είναι τα 2 χρωμοσώματα να έχουν μόνο ένα νουκλεοτίδιο ακριβώς όταν τελειώσει τον γύρο του. Άρα, όταν τελειώσει τον προηγούμενο του γύρο, είναι θεμιτό να έχει χωρίσει το δεύτερο χρωμόσωμα σε 2 χρωμοσώματα μήκους 1 και 3 αντίστοιχα. Αντιθέτως, δεν είναι θεμιτό να χωρίσει το δεύτερο χρωμόσωμα σε τμήματα μήκους 2 και n-2 (ή m-2), καθώς ο άλλος παίκτης μπορεί να διαγράψει το μεγαλύτερο χρωμόσωμα και να χωρίσει το άλλο σε 1 και 1 νουκλεοτίδιο, κερδίζοντας του το παιχνίδι.

Η καλύτερη λοιπόν τακτική είναι ο παίκτης να προσπαθεί να διαγράψει ένα χρωμόσωμα και να διαχωρίσει το άλλο σε 2 χρωμοσώματα μονού μήκους. Έτσι, ο επόμενος παίκτης μπορεί να διαχωρίσει μόνο κάποιο μονό χρωμόσωμα σε ένα ζυγό και ένα μονό. Ύστερα, ο παίκτης μπορεί να επιλέξει να διαγράψει τον μονό και να χωρίσει το ζυγό σε αναγκαστικά 2 χρωμοσώματα μονού μήκους πάλι. Με αυτή την τακτική, ο παίκτης αυξάνει τις πιθανότητες του είτε να φτάσει να χωρίσει κάποια στιγμή τα χρωμοσώματα σε χρωμοσώματα μήκους 1 και 3 αντίστοιχα ή το χρωμόσωμα ζυγού μεγέθους του άλλου παίκτη να είναι ίσο με 2, νικώντας έτσι το παιχνίδι όπως αναφερθήκαμε πιο πάνω.

Πρόγραμμα:

Με την μέθοδο getgenbank(), λαμβάνουμε τις 2 ακολουθίες των ενζύμων της φωσφορυλάσης του γλυκογόνου του ήπατος και του εγκεφάλου, από την βάση δεδομένων του NCBI. Θα χρειαστούμε 3 ορίσματα για την κάθε ακολουθία: [1]

* Accession number: Μια αλφαριθμητική ακολουθία που προσδιορίζει το ακριβές ένζυμο (π.χ ΝC\_000014)
* PartialSeq: Ένα εύρος τιμών που προσδιορίζει από ποιο νουκλεοτίδιο ξεκινά το κομμάτι της αλυσίδας του ενζύμου μας και σε ποιο τελειώνει, δηλαδή την θέση του 1ου και την θέση του τελευταίου νουκλεοτιδίου του τμήματος του γονιδίου που επιθυμούμε.
* SequenceOnly: Θέτοντας αυτό το όρισμα ως ‘True’, επιβεβαιώνουμε ότι επιθυμούμε μόνο το τμήμα της αλυσίδας του ενζύμου (δηλαδή την ακολουθία των νουκλεοτιδίων) και δεν επιθυμούμε άλλες πληροφορίες για το επιλεγμένο γονίδιο.

Παράδειγμα:

Brain = getgenbank('NC\_000020', 'PARTIALSEQ', [25248070, 25298012],'SequenceOnly','True');

Ως Ν και Μ θέτουμε τα μήκη των τμημάτων των ενζύμων της φωσφορυλάσης του γλυκογόνου του ήπατος και του εγκεφάλου αντιστοίχως. Θέτουμε τον τρέχοντα παίκτη ως τον παίκτη 1, τυπώνουμε στην οθόνη τους κανόνες και εκκινούμε το παιχνίδι.

Σε κάθε γύρο, το παιχνίδι συνεχίζεται εάν τα μήκη Ν και Μ είναι μεγαλύτερα από 0 (δηλαδή και τα 2 χρωμοσώματα έχουν νουκλεοτίδια). Στην περίπτωση που τα χρωμοσώματα έχουν μήκη Ν=1 και Μ=1, τότε το παιχνίδι τελειώνει και έχουμε ως νικητή τον παίκτη του τελευταίου γύρου. Σε κάθε γύρο, τυπώνουμε στην οθόνη τα 2 χρωμοσώματα και ελέγχουμε εάν το παιχνίδι έχει νικητή.

Εάν δεν έχει τελειώσει το παιχνίδι, παραθέτουμε στον αντίστοιχο παίκτη του γύρου να καταστρέψει ένα από τα 2 χρωμοσώματα μήκος Ν και Μ αντίστοιχα, με την μέθοδο chromosome\_destroy(). Αφού επιλέξει ποιο χρωμόσωμα επιθυμεί να διαγραφεί, ελέγχουμε ποιο χρωμόσωμα καταστράφηκε και ποιο θα διαχωριστεί. Άμα το μη διαγραμμένο χρωμόσωμα έχει μήκος 1, δεν μπορεί να διαχωριστεί, και ο τρέχων παίκτης χάνει το παιχνίδι επειδή είναι ανίκανος να ολοκληρώσει τον γύρο του.

Με την μέθοδο chromosome\_split(), διεκπεραιώνουμε την 2η ενέργεια του παίκτη, τον διαχωρισμό του 2ου χρωμοσώματος. Εξηγώντας στον χρήστη πως δουλεύει η διαδικασία διαχωρισμού, τον ρωτάμε να γράψει το μήκος του πρώτου από τα δύο τμήματα, του χρωμοσώματος που χωρίζουμε στα δύο, θεωρώντας πως το μήκος της αλυσίδας είναι ίσο με n. Ελέγχουμε εάν η είσοδος του χρήστη είναι έγκυρη. Στην περίπτωση που δεν είναι, επαναλαμβάνουμε τα παραπάνω βήματα της μεθόδου chromosome\_split(). Στρογγυλοποιώντας το μήκος που δίνει ο χρήστης, υπολογίζουμε τα μήκη των 2 νέων χρωμοσωμάτων και τα επιστρέφουμε στον κύριο αλγόριθμο. Θέτουμε τα νέα μήκη ως Ν και Μ.

Πριν το τέλος του γύρου, υπολογίζουμε τις ακριβείς ακολουθίες νουκλεοτιδίων των 2 νέων χρωμοσωμάτων, χρησιμοποιώντας το αντίστοιχο μη διαγραμμένο χρωμόσωμα και το μήκος του πρώτου νέου χρωμοσώματος (το νέο μήκος Ν). Τέλος του γύρου.

Βιβλιογραφία

[1] <https://www.mathworks.com/help/bioinfo/ref/getgenbank.html>

# O πηγαίος κώδικας (source code) αυτού του θέματος βρίσκεται στο αρχείο ex3.m