ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ

Απαλλακτικής Εργασία

*Ακαδημαϊκό Έτος 2019 - 2020*

# 

Δημήτρης Ματσαγγάνης, Π17068

Αλέξανδρος Σκαρπέλος, Π17122

Περιεχόμενα

[Άσκηση 6.13 3](#_Toc45296524)

[1. Εκφώνηση 3](#_Toc45296525)

[2. Κεντρική Ιδέα Υλοποίησης 3](#_Toc45296526)

[2.1. Κεντρικό Μενού 4](#_Toc45296527)

[2.2. Εύρεση ακολουθιών νουκλεοτιδίων 4](#_Toc45296528)

[2.3. Κύριο Μέρος 4](#_Toc45296529)

[2.4. Συναρτήσεις Υλοποίησης 5](#_Toc45296530)

[2.5. Παραδείγματα Στρατηγικής 6](#_Toc45296531)

[3. Αποτελέσματα 7](#_Toc45296532)

[Άσκηση 6.22 8](#_Toc45296533)

[1.Εκφώνηση 8](#_Toc45296534)

[2. Κεντρική Ιδέα Υλοποίησης 8](#_Toc45296535)

[Άσκηση 6.23 8](#_Toc45296536)

[1.Εκφώνηση 8](#_Toc45296537)

[2. Κεντρική Ιδέα Υλοποίησης 8](#_Toc45296538)

[Άσκηση 7.2 8](#_Toc45296539)

[1.Εκφώνηση 8](#_Toc45296540)

[2. Κεντρική Ιδέα Υλοποίησης 8](#_Toc45296541)

[Άσκηση 11.4 9](#_Toc45296542)

[1.Εκφώνηση 9](#_Toc45296543)

[2. Κεντρική Ιδέα Υλοποίησης 9](#_Toc45296544)

[2.1. Δηλώσεις Μεταβλητών 10](#_Toc45296545)

[2.2. Αλγόριθμος Viterbi 10](#_Toc45296546)

[2.3. Κύριο Μέρος 10](#_Toc45296547)

[2.4. Τελικό Στάδιο 11](#_Toc45296548)

[3. Αποτελέσματα 11](#_Toc45296549)

[Απαραίτητα Εργαλεία 13](#_Toc45296550)

# Άσκηση 6.13

## 1. Εκφώνηση

Δυο παίχτες παίζουν το παρακάτω παιχνίδι με δύο χρωμοσώματα που έχουν μήκος n και m νουκλεοτίδια αντίστοιχα.

Σε κάθε γύρο του παιχνιδιού, ένας παίχτης μπορεί να αφαιρέσει έναν τυχαίο αριθμό νουκλεοτιδίων από τη μία αλληλουχία ή τον ίδιο (αλλά και πάλι τυχαίο) αριθμό νουκλεοτιδίων και από τις δύο αλληλουχίες.

Ο παίχτης που αφαιρεί το τελευταίο νουκλεοτίδιο κερδίζει.

Ποιος θα κερδίσει; Περιγράψτε την νικηφόρα στρατηγική για όλες τις τιμές n και m.

## 2. Κεντρική Ιδέα Υλοποίησης

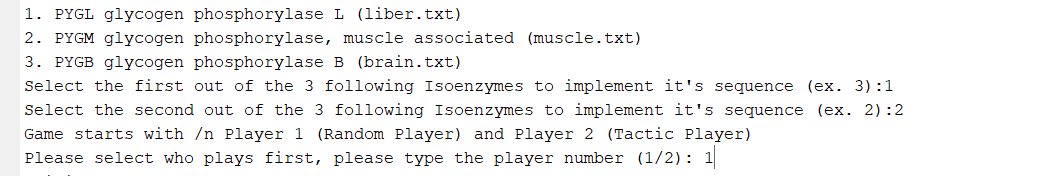
Το πρόγραμμα μας αποτελείται από ένα αρχείο το οποίο πρέπει και να εκτελέσουμε ώστε να δούμε τα αποτελέσματα.

Σε αυτήν την άσκηση προσπαθούμε να ανακαλύψουμε ποιος χρήστης θα κερδίσει για αυτό εκτελώντας το πρόγραμμα υπάρχει κεντρικό μενού στο οποίο επιλέγουμε ποιος παίχτης θα ξεκινήσει αλλά και ποια αρχεία θα χρησιμοποιηθούν.

### 2.1. Κεντρικό Μενού

Υλοποιήσαμε την άσκηση αυτή και προσπαθήσαμε να την κάνουμε πιο ευχάριστη για τον χρήστη με την ύπαρξη κεντρικού μενού μέσω του οποίου μπορεί να επιλεγεί ποιος από τους δύο παίχτες θα ξεκινήσει καθώς και ποια δύο αρχεία θα επιλεγούν (από τα τρία διαθέσιμα: liver, brain, muscle).

Ακολουθεί ενδεικτική εικόνα:



Εικόνα Αποτελεσμάτων 1

### 2.2. Εύρεση ακολουθιών νουκλεοτιδίων

Αρχικά, κατεβάζουμε από την ιστοσελίδα της NCBI (όπως μας το συστήσατε) όπου θα κατεβάσουμε και την ακολουθία FASTA και θα τη μετατρέψουμε σε αρχείο txt.

Έπειτα στο Matlab χρησιμοποιώντας την εντολή *fastaread* θα δημιουργηθεί ένα αντικείμενο με τον header και την ακολουθία από το αρχείο και εμείς θα επιλέξουμε την ακολουθία για να υλοποιήσουμε την υπόθεση της άσκησης.

### 2.3. Κύριο Μέρος

Έπειτα από το κεντρικό μενού αναγράφονται οι αλληλουχίες των νουκλεοτιδίων που επιλέξαμε (τα δύο από τα τρία αρχεία που επιλέγει ο χρήστης παραπάνω).

Στην συνέχεια κάνει κίνηση δηλαδή αφαιρεί νουκλεοτίδια ο παίχτης που επιλέξαμε εμείς στο κεντρικό μενού.

Για αυτόν τον λόγο δημιουργήσαμε συναρτήσεις αναφορικά με τους παίχτες (τις Player1, Player2 και την losing\_position).

### 2.4. Συναρτήσεις Υλοποίησης

Για την επιλογή του τυχαίου αριθμού, ο παίκτης1 θα επιλέξει τυχαία αναμεσά στις κινήσεις move1 & move2, για να αφαιρέσει οποιανδήποτε τυχαίο αριθμό από ολόκληρη/ρες την ακολουθία/ες.

Στο τέλος της συνάρτησης, επιστρέφονται οι νέες ακολουθίες μετά την αφαίρεση νουκλεοτιδίων.

Για την συνάρτηση Player2, όπου σε αυτή την περίπτωση ο παίκτης δεν παίζει τυχαία, ακολουθεί η ίδια λογική στην περίπτωση της αφαίρεσης ενός (όχι τυχαίου) αριθμού.

Στην περίπτωση όπου ο παίκτης1 παίζει τυχαία, εδώ δημιουργήσαμε τη συνάρτηση losePosi όπου ελέγχει αν η θέση που βρίσκεται ο παίκτης2 είναι μειονεκτική (losing position).

Αν οι ακολουθίες έχουν το ίδιο μήκος, τότε επιστρέφει τιμή 0. Αν η ακολουθία m ισούται με 0 και η n είναι μεγαλύτερη από το 0 πάλι θα επιστραφεί η τιμή 0. Αυτό γίνεται διότι από δω και πέρα θα πρέπει να κάνει κινήσεις που αφορούν τη μια ακολουθία, όχι και τις 2.

Αντιστρόφως ανάλογα, αν η n ισούται με 0 και η m μεγαλύτερη του μηδενός, η επιστρεφόμενη τιμή θα είναι 0. Τελικά επιστρέφει 1 μόνο όταν οι ακολουθίες είναι διάφορες του μηδενός και όχι ίσες.

### 2.5. Παραδείγματα Στρατηγικής

Σε αυτή την ενότητα θα παρουσιαστεί ένα σύντομο παράδειγμα που θα αναδεικνύει τη σειρά αφαίρεσης νουκλεοτιδίων την οποία ακολούθησε ο κάθε παίκτης για να νικήσει, καθώς και τον νικητή στη κάθε περίπτωση.

Για παράδειγμα: Η στρατηγική παιξίματος κάθε φορά έχει να κάνει με το ποιος παίκτης θα φτάσει πρώτος στα 3 νουκλεοτίδια, αν φτάσει ο παίκτης1 πρώτος στα 3 νουκλεοτίδια και είναι η σειρά του παίκτης2, τότε όποιο αριθμό νουκλεοτιδίων αφαιρέσει, ο παίκτης1 θα κερδίσει. Υποθετικά βγάζει 2 νουκλεοτίδια ο παίκτης2. Τώρα ο παίκτης1 μπορεί να αφαιρέσει το 1 νουκλεοτίδιο και να κερδίσει. Αν από την άλλη ο παίκτης2 αφαιρέσει 1 νουκλεοτίδιο, τότε ο παίκτης2 πάλι μπορεί να κερδίσει αφαιρώντας και τα 2 νουκλεοτίδια.

Επεκτείνοντας τη στρατηγική που προαναφέρθηκε για n αριθμό νουκλεοτιδίων έχουμε τα ακόλουθα.

Αν πάρουμε για παράδειγμα 10 νουκλεοτίδια, σημασία έχει ποιος θα φτάσει πρώτος στα 9 νουκλεοτίδια, για να μπορεί να οδηγεί το παιχνίδι μέχρι το 3 όπου και είναι ο απώτερος σκοπός για να κερδίσει.

Ο παίκτης1 παίζει και αφαιρεί 1, άρα έχουμε 9 νουκλεοτίδια και είναι η σειρά του παίκτης2. Αν και εκείνος αφαιρέσει 2 τότε θα πάμε στα 7.

Η επόμενη κίνηση του παίκτη1 είναι να αφαιρέσει 1 νουκλεοτίδιο για να πάει στα 6.

Αν ο παίκτης2 αν αφαιρέσει 2, θα πάμε στα 4 νουκλεοτίδια και ο παίκτης1 θα αφαιρέσει 1 νουκλεοτίδιο οπότε θα είναι και αυτός που θα κερδίσει γιατί έφτασε πρώτος στα 3.

Τώρα αν πάλι ο παίκτης2 αφαιρέσει 1, θα πάμε στα 5 νουκλεοτίδια και ο παίκτης1 θα αφαιρέσει 2 αυτή τη φορά για να φτάσει πρώτος στα 3 και θα κερδίσει το παιχνίδι.

Παρατηρούμε δηλαδή πως η στρατηγική επεκτείνεται στους περιττούς αριθμούς κάθε φορά και σημασία έχει ποιος θα φτάσει στο 9, μετά στο 7, 5 και τελικά στο 3.

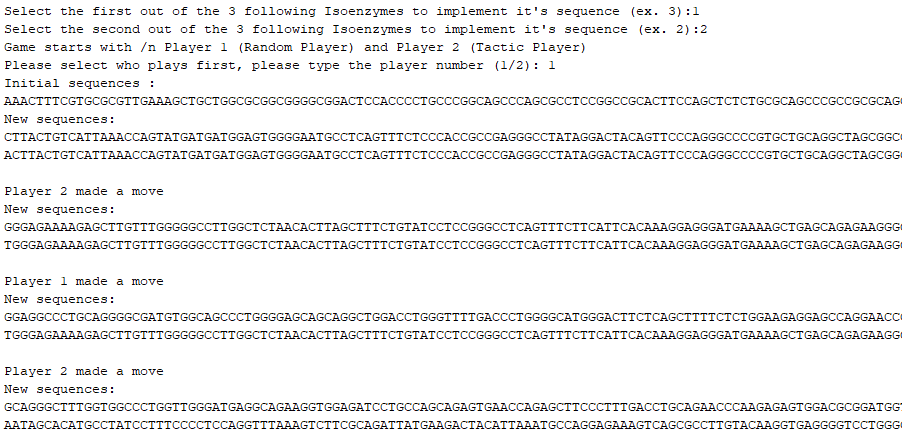
Αν γίνει κάποιο “λάθος” και φτάσει ο παίκτης2 πρώτος στο 7 ή στο 9 (όπως στο παραπάνω παράδειγμα) τότε ο παίκτης1 θα πρέπει να κάνει κίνηση η οποία δεν θα επιτρέπει στον παίκτης2 να φτάσει πρώτος στον ζητούμενο περιττό αριθμό, 3.

Στην περίπτωση όπου ο αριθμός των νουκλεοτιδίων είναι περιττός πχ: n=11 τότε ο παίκτης ο οποίος παίζει πρώτος, έχει από την αρχή το προβάδισμα διότι ο αριθμός είναι ήδη περιττός. Άρα θα πρέπει να κρατήσει αυτή τη λογική και να στέλνει στον αντίπαλο παίκτη άρτιους.

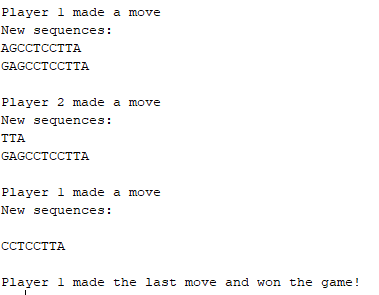
Οποιοσδήποτε και να είναι ο αριθμός του n, αρκεί κάθε φορά ο παίκτης που θα επιχειρήσει την νικηφόρα στρατηγική να έχει κατά νου ότι πρέπει να φτάσει πρώτος στους περιττούς αριθμούς και να στέλνει στον αντίπαλο τους άρτιους.

### 3. Αποτελέσματα

Ακολουθούν τα αποτελέσματα της παραπάνω υλοποίησης (ενδεικτικά από την αρχή και το τέλος των αποτελεσμάτων):



Εικόνα Αποτελεσμάτων 2



Εικόνα Αποτελεσμάτων 3

# Άσκηση 6.22

## 1.Εκφώνηση

## 2. Κεντρική Ιδέα Υλοποίησης

# Άσκηση 6.23

## 1.Εκφώνηση

Διατυπώστε έναν αλγόριθμο που υπολογίζει τη βέλτιστη στοίχιση προσαρμογής. Εξηγήστε πώς συμπληρώνεται η πρώτη γραμμή και η πρώτη στήλη του πίνακα δυναμικού προγραμματισμού και γράψτε μια σχέση επανάληψης για τη συμπλήρωση του υπόλοιπου πίνακα. Παρουσιάστε μια μέθοδο που βρίσκει την καλύτερη στοίχιση συμπληρωθεί ο πίνακας. Ο αλγόριθμος θα πρέπει να εκτελείται σε χρόνο O(nm).

Επεξήγηση: Γενικότερα οι δύο βασικές μέθοδοι στοίχισης είναι δύο: η καθολική και η τοπική. Υπάρχει και μια μέση κατάσταση, η οποία είναι η ημικαθολική και μοιάζει στην προσαρμοστική αλλά δεν υπολογίζονται τα ‘κοστοβόρα’ κενά που βρίσκονται στην άκρη. Με βάση της ημικαθολικής στοίχισης θα υλοποιηθεί η άσκηση.

## 2. Κεντρική Ιδέα Υλοποίησης

Για την υλοποίηση της άσκησης θα πρέπει στην ουσία να υλοποιήσουμε τον αλγόριθμο Needleman–Wunsch με κάποιες αλλαγές ώστε να κάνει την στοίχιση χωρίς όμως να αφαιρεί στοιχεία της δεύτερης ακολουθίας από το τελικό αποτέλεσμα.

## 2.1 Εύρεση ακολουθιών νουκλεοτιδίων

Αρχικά, κατεβάζουμε από την ιστοσελίδα της NCBI (όπως μας το συστήσατε) όπου θα κατεβάσουμε και την ακολουθία FASTA και θα τη μετατρέψουμε σε αρχείο txt.

Έπειτα στο Matlab χρησιμοποιώντας την εντολή *fastaread* θα δημιουργηθεί ένα αντικείμενο με τον header και την ακολουθία από το αρχείο και εμείς θα επιλέξουμε την ακολουθία για να υλοποιήσουμε την υπόθεση της άσκησης.

## 2.2

## 2.3 Κύριο Μέρος-Υλοποίηση της κεντρικής ιδέας- Εύρεση καλύτερης αντιστοίχισης

Αρχικά θα βρούμε τη καλύτερη βαθμολογία σε όλη την τελευταία στήλη. Αυτό θα το κάνουμε ώστε να ξεκινήσουμε την διαδικασία του backtracking από την αντιστοιχία που δίνει τη μέγιστη συνολική βαθμολογία μεταξύ του τελευταίου γράμματος της δεύτερης ακολουθίας (ακολουθίας προσαρμογής) με το εκάστοτε γράμμα της πρώτης ακολουθίας. Άρα ξεκινώντας από αυτή τη θέση στη μήτρα F κάνουμε backtracking εξασφαλίζοντας ότι η ακολουθία προσαρμογής θα μείνει όπως 24 είναι και απλά θα έχουμε πια την πιο όμοια υπο-ακολουθία της πρώτης ακολουθίας για να τη συγκρίνουμε. Τέλος στη διαδικασία του backtracking, ελέγχουμε κάθε φορά αν με βάση τους κανόνες σύγκρισης των κελιών που αναφέραμε πριν, τα κελία έχουν τις αντίστοιχες τιμές και ανάλογα συγκρίνουμε αν τα γράμματα των ακολουθιών που συναντάμε στο «μονοπάτι» είναι ίδια η διαφορετικά, και ανάλογα αν κάνουμε αντικατάσταση, προσθήκη ή αφαίρεση προσθέτουμε τους κατάλληλους χαρακτήρες ή σύμβολα στους αντίστοιχους πίνακες Alignment που αναφέραμε πριν. Σημαντικό είναι να αναφέρουμε ότι λόγω της φύσης του αλγόριθμου Needleman-Wunch μπορεί να υπάρχουν παραπάνω από μία βέλτιστες στοιχίσεις ακολουθιών προσαρμογής επειδή στο backtracking γίνεται διακλάδωση σε διαφορετικά μονοπάτια αν υπάρχει ισότητα στις τιμές των γειτονικών κελιών.

## 3.Αποτελέσματα

# Άσκηση 7.2

## 1.Εκφώνηση

## 2. Κεντρική Ιδέα Υλοποίησης

# Άσκηση 11.4

## 1.Εκφώνηση

Στο σχήμα 11.7 φαίνεται ένα HMM με δύο καταστάσεις α και β. Όταν το HMM βρίσκεται στην κατάσταση α, έχει την μεγαλύτερη πιθανότητα να εκπέμψει πουρίνες (A και G). Όταν βρίσκεται στην κατάσταση β, έχει μεγαλύτερη πιθανότητα να εκπέμψει πυριμιδίνες (C και T). Αποκωδικοποιήστε την πιο πιθανή ακολουθία των καταστάσεων (α/β) για την αλληλουχία GGCT. Χρησιμοποιήστε λογαριθμικές βαθμολογίες για κανονικές βαθμολογίες πιθανοτήτων.

## 2. Κεντρική Ιδέα Υλοποίησης

Το πρόγραμμα μας αποτελείται από ένα μοναδικό αρχείο το οποίο πρέπει και να εκτελέσουμε ώστε να δούμε τα αποτελέσματα.

Σε αυτήν την άσκηση προσπαθούμε να ανακαλύψουμε την ‘καλύτερη’ διαδρομή (αυτή με το μεγαλύτερο σκορ σαν σύνολο) για την κωδικοποίηση της ακολουθίας νουκλεοτιδίων “GGCT”.

### 2.1. Δηλώσεις Μεταβλητών

Θεωρούμε ισοπίθανες αρχικές πιθανότητες να ξεκινάει είτε από την κατάσταση Α είτε από την B (0.5 για κάθε μια) εφόσον το πρόγραμμα δεν μας υποδεικνύει κάτι διαφορετικό.

Έπειτα δημιουργήσαμε έναν πίνακα που αναφέρεται στην πιθανότητα να πηγαίνει από την μία κατάσταση σε άλλη ή ίδια να πηγαίνει στον εαυτό της.

Η ακολουθία GGCT μετατρέπετε σε αριθμούς μέσω της εντολής nt2int για να μπορεί να γίνει αντιληπτή. Επομένως μετατρέπεται σε ακολουθία αριθμών (3,3,2,4).

### 2.2. Αλγόριθμος Viterbi

Η υλοποίηση του αλγόριθμου Viterbi σε αυτήν την άσκηση έγινε ως εξής: Αρχικά, μετατρέπουμε όλες τις τιμές των πινάκων πιθανοτήτων σε λογαριθμικές τιμές για να μην υπάρχει πρόβλημα - underflow με τις κανονικές πιθανότητες που θα υπολογίσουμε (προτείνεται από την άσκηση).

Μετά ξεκινάμε βάζοντας στην αρχή του πίνακα Viterbi τις αρχικές πιθανότητες των καταστάσεων μας (50% - 50% για κάθε κατάσταση) και συνεχίζουμε για κάθε σύμβολο της ακολουθίας τον υπολογισμό της πιθανότητας να εκπεμφθεί δοσμένης της τωρινής κατάστασης και της πιθανότητας της προηγούμενης κάθε φορά.

Αφού υπολογίσουμε ολόκληρο το πίνακα Viterbi και βρούμε την μέγιστη πιθανότητα στο τέλος ανάμεσα στις δύο καταστάσεις ξεκινάμε τη διαδικασία του backtracking.

### 2.3. Κύριο Μέρος

Αρχικά, δημιουργούμε έναν πίνακα δύο διαστάσεων ο οποίος αποτελείται μόνο από μηδενικά. Έπειτα προσπαθούμε μέσω εμφολευμένων loops να ακολουθήσουμε όλες τις πιθανές διαδρομές που προκύπτουν από το διάγραμμα που κατασκευάζουμε στο χαρτί.

Πιο συγκεκριμένα, κοιτάμε κάθε φορά ποιο στοιχείο εξετάζουμε (πρώτο loop) και στην συνέχεια ελέγχουμε σε ποια κατάσταση βρισκόμαστε και έπειτα τα αποθηκεύουμε στον πίνακα που υπήρχαν τα μηδενικά ο οποίος χρησιμεύει σαν ένας λευκός πίνακας πάνω στον οποίο γράφουμε τις αντλούμενες πληροφορίες.

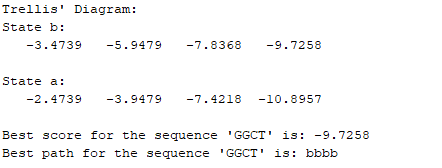
### 2.4. Τελικό Στάδιο

Στο τελικό στάδιο της εφαρμογής, υπολογίζεται το συνολικό σκορ των δύο διαδρομών (στο συγκεκριμένο πρόβλημα, σε κάποιο άλλο θα μπορούσε να έχει παραπάνω καταστάσεις άρα και παραπάνω διαδρομές), συγκρίνονται και αποφασίζεται ποια διαδρομή είναι η καλύτερη.

### 3. Αποτελέσματα

Όπως φαίνεται και στην παραπάνω φωτογραφία συγκρίνονται στο τέλος οι τιμές του πίνακα (το 1ο και το 2ο στοιχείο της 5ης στήλης) και επιλέγεται η ο μεγαλύτερος ο αριθμός δηλαδή η διαδρομή με το καλύτερο σκορ.

Άρα επιλέγουμε την διαδρομή ‘bbbb’ αφού είναι αυτή που μας δίνει το καλύτερο αποτέλεσμα με βάση τον αλγόριθμο Viterbi.



Εικόνα Αποτελεσμάτων 4

Συνεπώς, και όπως προαναφέρθηκε βλέπουμε ότι για την ακολουθία GGCT η καλύτερη ακολουθία καταστάσεων που δίνει μέγιστη πιθανότητα (λογαριθμική) -9.7258 είναι η ΒΒΒΒ, δηλαδή μόνο η δεύτερη κατάσταση.

Ενδεικτικά, εάν το λύναμε το πρόβλημα χωρίς να άγουμε τις πιθανότητες στη λογαριθμική κλίμακα, τότε το αποτέλεσμα του διαγράμματος Trellis, θα ήταν το παρακάτω:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Κατάσταση Β** | 0,1 --> | 0,018 --> | 0,00486 --> | 0,0013122 |
| **Κατάσταση Α** | 0,2 --> | 0,072 --> | 0,00648 --> | 0,0005832 |
|  | **G** | **G** | **C** | **T** |

# Απαραίτητα Εργαλεία

Η εργασία μας υλοποιήθηκε στο προγραμματιστικό-εκπαιδευτικό εργαλείο Matlab.

Απαραίτητο εργαλείο-βιβλιοθήκη αποτελεί το Bionformantics tool που πρέπει να είναι και αυτό εγκατεστημένο για την υλοποίηση της. Κάνουμε ιδιαίτερη αναφορά σε αυτήν την βιβλιοθήκη επειδή ο χρήστης θα πρέπει να φροντίσει να την επιλέξει κατά την εγκατάσταση του Matlab μέσω installer ή σε μεταγενέστερη τροποποίησης της εγκατάστασης.