Εισαγωγή στη Βιοπληροφορική

2024-25

Project 1

|  |  |
| --- | --- |
| Δημητρακόπουλος Χρήστος | Μπερτσεκάς Παρασκευάς - Σωτήριος |
| chris.dimitrako@ac.upatras.gr | up1093445@ac.upatras.gr |
| 1079500 | 1093445 |

Περιεχόμενα

[1. Ερώτημα 1 3](#_Toc200399775)

[1.1 Εργαλεία πρόβλεψης πρωτεϊνικών δομών 3](#_Toc200399776)

[1.1.1 AlphaFold3 3](#_Toc200399777)

[1.1.2 ESMFold2 3](#_Toc200399778)

[1.2 Αναζήτηση BlastP 4](#_Toc200399779)

[1.3 Πολλαπλή στοίχιση 5](#_Toc200399780)

[1.3.1 Default παράμετροι 5](#_Toc200399781)

[1.3.2 Προσαρμοσμένοι παράμετροι 5](#_Toc200399782)

[1.3.3 Τελική σύγκριση 6](#_Toc200399783)

[1.3.4 Συμπέρασμα 6](#_Toc200399784)

[2. Ερώτημα 2 7](#_Toc200399785)

[3. Ερώτημα 3 8](#_Toc200399786)

[3.1 Μέγιστη κοινή υπακολουθία SARS-CoV-2 & MERS-CoV 8](#_Toc200399787)

[3.2 Σύγκριση δομής πρωτεϊνών με Dali 8](#_Toc200399788)

[4. Ερώτημα 4 9](#_Toc200399789)

[4.1 Πρώτη επιλογή 9](#_Toc200399790)

[5. Ερώτημα 5 10](#_Toc200399791)

[5.1 Ολική στοίχιση 10](#_Toc200399792)

[5.2 Τοπική στοίχιση 10](#_Toc200399793)

[6. Ερώτημα 6 12](#_Toc200399794)

[6.1 Άσκηση 1 12](#_Toc200399795)

[6.2 Άσκηση 2 12](#_Toc200399796)

# Ερώτημα 1

## Εργαλεία πρόβλεψης πρωτεϊνικών δομών

### AlphaFold3

Το AlphaFold3 είναι εργαλείο που προβλέπει την τρισδιάστατη δομή πρωτεϊνών μόνο με βάση την πρωτοταγή αλληλουχία αμινοξέων. Ανήκει στην κατηγορία των μοντέλων βαθιάς μάθησης και χρησιμοποιεί εξελιγμένες τεχνικές τεχνητής νοημοσύνης για να υπολογίσει με υψηλή ακρίβεια τη χωρική διάταξη μιας πρωτεΐνης. Σε αντίθεση με απλούστερες μεθόδους για πρόβλεψη δομής πρωτεϊνών, το AlphaFold3 μπορεί να προβλέψει τη δομή πολλαπλών τύπων βιομορίων, συμπεριλαμβανομένων DNA, RNA, protein complexes κ.α. Αυτό το καθιστά καίριο για τη μελέτη μοριακών αλληλεπιδράσεων και την κατανόηση πολύπλοκων βιολογικών μηχανισμών.

Χρησιμοποιείται είτε μέσω της πλατφόρμας alphafoldserver.com είτε μέσω του ColabFold, εργαλείο που βασίζεται στο Google Colab. Ο χρήστης δίνει ως είσοδο μια ή περισσότερες αλληλουχίες σε μορφή FASTA και λαμβάνει ως έξοδο μοντέλα δομής σε μορφή Protein Data Bank (PDB) καθώς και ενδείξεις αξιοπιστίας όπως το pLDDT score. Για βασική web χρήση δεν απαιτούνται ειδικές γνώσεις ή μεγάλη εμπειρία, αρκεί ο χρήστης να γνωρίζει ως ένα βαθμό το αντικείμενο προς μελέτη.

Το AlphaFold3 είναι κατάλληλο για έρευνα όπου η *ακρίβεια* και η *λεπτομέρεια* είναι κρίσιμες, όπως στον σχεδιασμό φαρμάκων ή στη μελέτη λειτουργικών περιοχών σε πρωτεΐνες.

### ESMFold2

Το ESMFold2 είναι και αυτό ένα μοντέλο πρόβλεψης πρωτεϊνικών δομών, το οποίο ανέπτυξε η Meta AI. Βασίζεται σε transformer αρχιτεκτονική και αξιοποιεί τεχνικές self-suprvised learning για να προβλέψει τη δομή μιας πρωτεΐνης απευθείας από την αλληλουχία, χωρίς να χρησιμοποιεί MSA (multiple sequence alignment) ή templates. Το βασικότερο πλεονέκτημα έναντι άλλων εργαλείων είναι η ταχύτητα, αφού δεν χρησιμοποιεί την χρονοβόρα διαδικασία πολλαπλής στοίχισης. Αυτό καθιστά το ESMFold2 κατάλληλο για μαζική ανάλυση πρωτεϊνών, όπως σε μεταγονιδιωματικά δεδομένα, όπου πρέπει να αναλυθούν χιλιάδες (ή και περισσότερες) ακολουθίες χωρίς ομόλογα.

Το ESMFold2 είναι διαθέσιμο στο GitHub και μπορεί να εγκατασταθεί τοπικά, ενώ παράλληλα υπάρχει και online demo λειτουργίας στα πλαίσια του ESM Metagenomic Atlas. Η εγκατάστασή του προϋποθέτει μερικές βασικές γνώσεις Python και Linux.

Συνοπτικά, είναι ένα μοντέλο που εστιάζει στην *ταχύτητα,* σε μαζικές αναλύσεις πολλών πρωτεϊνών, με σκοπό τη γενίκευση της έρευνας. Δεν εστιάζει στη λεπτομερή μοντελοποίηση κάθ επρωτεινής ξεχωριστά, αλλά λειτουργεί ως εργαλείο αρχική διαλογής ώστε να εντοπιστούν αυτές που αξίζουν περεταίρω ανάλυση και σε επόμενο βήμα να αναλυθούν με πιο ακριβή μοντέλα, όπως το AlphaFold3.

## Αναζήτηση BlastP[[1]](#footnote-1)

A screenshot of a computer program

AI-generated content may be incorrect.

Εικόνα : Αναζήτηση

A screenshot of a computer

AI-generated content may be incorrect.

Εικόνα : Αποτελέσματα

## Πολλαπλή στοίχιση

Ένας από τους βασικότερους τρόπους αξιολόγησης του αποτελέσματος μιας πολλαπλής στοίχισης είναι η ανάλυση των συντηρημένων περιοχών. Αυτές οι περιοχές, σε αντίθεση με τις N-terminal και C-terminal περιοχές οι οποίες εμφανίζουν διαφοροποιήσεις, είναι κατά βάση ίδιες αφού αντιπροσωπεύουν δομικές και λειτουργικές περιοχές της πρωτεΐνης. Στο παράδειγμα της άσκησής μας, έχουμε επιλέξει 10 συγγενείς πρωτεΐνες για στοίχιση άρα αναμένουμε εκτενείς συντηρημένες περιοχές. Άρα, ως πρώτο κριτήριο ορίζουμε το κατά πόσο κάθε εργαλείο μπορεί να διατηρήσει αυτές τις περιοχές ευθυγραμμισμένες με το δυνατόν λιγότερα κενά. Τα εργαλεία από το EBI μας δίνουν την δυνατότητα να εμφανίσουμε τα αποτελέσματα μέσω του MView, το οποίο μας δίνει αυτό το ποσοστό «υψηλής συντήρησης» ως *consensus.* Επιπλέον, μας δίνει το *pid* το οποίο δείχνει ποσοστό ομοιότητας με την reference αλληλουχία. Αυτό σημαίνει ότι όσο πιο υψηλό είναι τόσο πιο εύκολο και πιο αξιόπιστη θα είναι η στοίχιση.

### Default παράμετροι[[2]](#footnote-2)

Με βάση αυτά τα κριτήρια, καταλαβαίνουμε ότι και στις 3 περιπτώσεις των εργαλείων, τα αποτελέσματα θα είναι αρκετά όμοια, όμως μπορούμε να δούμε ότι υπάρχουν μικρές διαφορές στην τοποθέτηση κενών αλλά και στη διατήρηση των συντηρημένων περιοχών. Πιο συγκεκριμένα:

* MAFFT: Η στοίχιση που δίνει έχει τη μεγαλύτερη συνοχή στις συντηρημένες περιοχές με ελάχιστα κενά και μεγάλα blocks υψηλού consensus.
* MUSCLE: Ενώ συνολικά δίνει καλή στοίχιση εισάγει περισσότερα κενά σε ενδιάμεσες περιοχές, κάτι που οδηγεί σε πιο σύνθετη στοίχιση σε σχέση με το MAFFT.
* T-Coffee: Παρόλο που είναι το πιο ευαίσθητο από τα 3 εργαλεία, εισάγει περισσότερα κενά ακόμα και εντός περιοχών υψηλής συντήρησης, κάτι που υποδηλώνει ότι ακόμα και σε σημεία υψηλής ομοιότητας η στοίχιση δεν θα είναι καλή.

Τέλος, παρατηρώντας το pid επιβεβαιώνεται ότι οι περισσότερες αλληλουχίες έχουν υψηλή ομοιότητα με την reference αλληλουχία (πάνω από 90%). Αυτό σημαίνει ότι οι μικρές αποκλείσεις των εργαλείων δεν επηρεάζουν ουσιαστικά την βιολογική αξιοπιστία.

### Προσαρμοσμένοι παράμετροι[[3]](#footnote-3)

Οι παράμετροι – ρυθμίσεις που παρέχονται αφορούν κυρίως το MAFFT μιας και είναι το μόνο το οποίο μπορούμε να αλλάξουμε gap penalty, max iterations κ.α. Γι’ αυτό θα ασχοληθούμε μόνο με την παραμετροποίηση του MAFFT, και συγκεκριμένα θα θέσουμε gap open penalty 1.8, gap extension penalty 0.2, max iterate 100 και θα ενεργοποιήσουμε FFT σε local pair.

Με αυτές τις παραμέτρους μπορούμε να δούμε αισθητή βελτίωση στην ποιότητα της στοίχισης. Οι συντηρημένες περιοχές είναι μεγαλύτερες και παραμένουν ευθυγραμμισμένες χωρίς την προσθήκη κενών.

### Τελική σύγκριση

* MAFFT (default)

Η στοίχιση είναι σε γενικές γραμμές καλή με ικανοποιητική διατήρηση των συντηρημένων περιοχών με λίγα κενά. Ωστόσο υπάρχουν ενδιάμεσες περιοχές που εισάγονται κενά που διασπούν μικρά blocks με υψηλό consensus, κάτι που δείχνει ελαττωμένη ακρίβεια στην στοίχιση.

* MAFFT (modified)

Με την ενεργοποίηση του FFT σε local pair και όλες τις υπόλοιπες αλλαγές που αναφέραμε προηγουμένως, βλέπουμε ως αποτέλεσμα τη βέλτιστη δυνατή στοίχιση. Οι συντηρημένες περιοχές είναι πλήρως στοιχισμένες με ελάχιστα κενά, και τα consensus blocks είναι μεγαλύτερα και συνεχή.

* MUSCLE (default)

Η στοίχιση είναι αποδεκτή με υψηλά pid και καλό επίπεδο στοίχισης, όμως εισάγονται περισσότερα κενά σε ενδιάμεσες περιοχές σε σχέση με το MAFFT (και στις 2 περιπτώσεις του). Μερικά consensus blocks διασπώνται χωρίς προφανή λόγο, με αποτέλεσμα η ποιότητα της στοίχισης να πέφτει.

* T-Coffee (default)

Είναι το πιο επιθετικό στην εισαγωγή κενών. Ακόμα και σε περιοχές «υψηλής συντήρησης» εμφανίζονται περιττά κενά τα οποία μειώνουν την στοίχιση δομικά κρίσιμων περιοχών. Η συνολική του ποιότητα είναι μειωμένη, κάτι που αποδεικνύει πως δεν είναι ιδανικό εργαλείο για περιπτώσεις στοίχισης με υψηλή ομοιότητα, όπως έχουμε.

### Συμπέρασμα

Η modified στοίχιση του MAFFT παρέχει την πιο συνεπή και ακριβή στοίχιση. Η χρήση των default παραμέτρων είναι επαρκής για μια αρχική αξιολόγηση των εργαλείων, αλλά ανάλογα το ζητούμενο και τις αλληλουχίες ο MAFFT μπορεί να αποδώσει καλύτερα.

# Ερώτημα 2

# Ερώτημα 3

## Μέγιστη κοινή υπακολουθία SARS-CoV-2 & MERS-CoV

## Σύγκριση δομής πρωτεϊνών με Dali

# Ερώτημα 4

## Πρώτη επιλογή

# Ερώτημα 5

Καθαρά για εμφανισιακούς σκοπούς χρησιμοποιήθηκαν διαθέσιμα online εργαλεία για την οπτικοποίηση των δύο πινάκων που ζητούνται. Για την ολική στοίχιση χρησιμοποιήθηκε το [Global Alignment App](https://bioboot.github.io/bimm143_W20/class-material/nw/) και για την τοπική στοίχιση ένα [εργαλείο εκπαίδευσης του αλγορίθμου Smith-Waterman](https://rna.informatik.uni-freiburg.de/Teaching/index.jsp?toolName=Smith-Waterman).

## Ολική στοίχιση

Στον παρακάτω πίνακα φαίνεται το αποτέλεσμα της ολικής στοίχισης. Η πρώτη γραμμή και στήλη του πίνακα αρχικοποιούνται με τιμές κόστους στοίχισης με κενό, υποθέτοντας ότι μέχρι εκείνο το σημείο προσθέτουμε διαρκώς κενά. Για κάθε επόμενο κελί υπολογίζονται οι τιμές για στοίχιση, ασυμφωνία και στοίχιση με κενό και αποθηκεύεται η μέγιστη από αυτές. Κάθε φορά που συμπληρώνεται μια τιμή, σημειώνεται και το κελί προέλευσης ώστε αργότερα να εκτελεστεί το Traceback (κόκκινο μονοπάτι) το οποίο ξεκινά από το τελευταίο κελί του πίνακα (κάτω δεξιά) μέχρι να φτάσει στην αρχή του πίνακα (πάνω αριστερά). Το μονοπάτι που θα δημιουργηθεί είναι η βέλτιστη στοίχιση.

A screenshot of a game

AI-generated content may be incorrect.

Εικόνα : Πίνακας δυναμικού προγραμματισμού ολικής στοίχισης

## Τοπική στοίχιση

Στον παρακάτω πίνακα φαίνεται το αποτέλεσμα της τοπικής στοίχισης. Η βασική διαφορά της τοπικής στοίχισης από την ολική είναι ότι δεν προσπαθεί να ταιριάξει ολόκληρες τις ακολουθίες αλλά μόνο το ή τα κομμάτια που ταιριάζουν καλύτερα. Πρακτικά αυτό σημαίνει ότι μας ενδιαφέρουν κομμάτια της ακολουθίας και όχι ολόκληρη η ακολουθία. Για αυτόν τον λόγο η πρώτη γραμμή η πρώτη στήλη του πίνακαα αρχικοποιούνται με 0, και επιτρέπεται επίσης να εισαχθεί μηδενική τιμή στα ενδιάμεσα κελιά. Όπως και στην ολική στοίχιση κάθε κελί υπολογίζεται με βάση τις πιθανές κινήσεις αλλά το Traceback ξεκινά από τη μεγαλύτερη τιμή του πίνακα και τελειώνει μόλις συναντήσει μηδενική τιμή. Ακριβώς για αυτόν τον λόγο μπορεί να υπάρχουν πολλές βέλτιστες εναλλακτικές στοιχίσεις, όπως φαίνεται δεξιά του πίνακα.

A screenshot of a computer

AI-generated content may be incorrect.

Εικόνα : Πίνακας δυναμικού προγραμματισμού τοπικής στοίχισης

# Ερώτημα 6

## Άσκηση 1

## Άσκηση 2

1. Αναλυτικά τα 10 αποτελέσματα της αναζήτησης θα τα βρείτε στο [GitHub](https://github.com/xrhstosdim1/ceid_bioInformatics/tree/main/Project%201) RES\_blastp\_NP-000137s [↑](#footnote-ref-1)
2. Αναλυτικά τα αποτελέσματα της στοίχισης και με τα 3 εργαλεία θα τα βρείτε στο GitHub. [↑](#footnote-ref-2)
3. Αναλυτικά τα προσαρμοσμένα αποτελέσματα της MAFFT στοίχισης θα τα βρείτε στο GitHub. [↑](#footnote-ref-3)