

Ερώτημα 1 (ii)

COBALT

Το COBALT χρησιμοποιεί προοδευτική πολλαπλή ευθυγράμμιση για τον συνδυασμό διαζευκτικών περιορισμών από διάφορες πηγές σε μια πολλαπλή ευθυγράμμιση. Χρησιμοποιώντας μια μέθοδο βαθμολόγησης, λαμβάνει υπόψη του πληροφορίες από διάφορες πηγές και τις ενσωματώνει σε μια ενιαία ευθυγράμμιση. Ξεκινά με την ευθυγράμμιση των πιο σχετικών ακολουθιών και στη συνέχεια προσθέτει προοδευτικά τις λιγότερο σχετικές. Οι διαζευκτικοί περιορισμοί που χρησιμοποιούνται μπορεί να προέρχονται από διάφορες πηγές, όπως η φυλογενετική ανάλυση, η δομική βιολογία ή η βιοχημική πληροφορία. Το αποτέλεσμα είναι μια πολλαπλή ευθυγράμμιση που αντικατοπτρίζει την πληροφορία από όλες αυτές τις πηγές, παρέχοντας ένα πλαίσιο για την κατανόηση της εξέλιξης και της λειτουργίας των πρωτεϊνών.

Χρησιμοποιώντας αρχικά το RPS-BLAST, ψάχνει στο CDD για παρόμοιους τομείς πρωτεϊνών με τις εισόδους. Όταν ένας τομέας ταιριάζει σε πολλές εισόδους τότε αποκτούμε μεγάλη πληροφορία. Μέσω του CDD δημιουργούμε και προσωρινά προφίλ στις εισόδους με αποδοτικό και φθηνό τρόπο. Τέλος σε επόμενα στάδια χρησιμοποιούμε PROSITE μοτίβα για την ανάλυσή μας (καθώς είναι μικρότερα).

Το COBALT είναι ένα ισχυρό εργαλείο για την ανάλυση των ακολουθιών πρωτεϊνών και την πρόβλεψη της δομής και της λειτουργίας τους.

Γραμμένο σε c,c++,perl.

Για την βαθμολόγηση πολλαπλών πρωτεϊνών χρησιμοποιεί έναν συνδυασμό από την καταγραφή των λογαριθμικών αναλογιών (log-odds) και τη βασισμένη στην εντροπία βαθμολόγηση. Αυτή η μέθοδος βαθμολογεί μια ακολουθία βάσει της πιθανότητας εμφάνισης ενός συγκεκριμένου αμινοξέου σε σχέση με την πιθανότητα εμφάνισης αυτού.

Βήματα αλγορίθμου :

- Εύρεση ευθυγραμμίσεων για τη δημιουργία περιορισμών.- RPS-BLAST,- CDD,- PROSITE,- PHI-BLAST
- Εύρεση πρόχειρων προφίλ και συνεπών περιορισμών.
- Δημιουργία Δένδρου-οδηγού.-συνεπείς περιορισμοί,- BLOSUM62
- Δημιουργία πολλαπλής ευθυγράμμισης χρησιμοποιώντας το τρέχον σύνολο περιορισμών και το δέντρο.- Needleman-Wunsch,- (Edgar, 20004a),- (Edgar and Sjölander, 2004; Wang and Dunbrack, 2004)
- Δημιουργία διαχωρισμών και επανευθυγράμμιση.

- Εκτέλεση (προαιρετικής) εκλεπτυσμένης διαδικασίας με τον καθορισμό ενός νέου συνόλου περιορισμών και επανάληψη από το Βήμα 4 όσο το πλήθος των περιορισμών συνεχίζει να αυξάνεται.

Το COBALT είναι ένα εργαλείο με τις εξής επιλογές:

Gap Penalties Πρόστιμο κενών (Διαφορετικό για το πρώτο κενό που θα βρεθεί)

Opening
Extension

End-Gap Penalties Πρόστιμο κλείσιμου κενών (Διαφορετικό για το πρώτο κενό που θα βρεθεί)

Opening
Extension

Constraint Parameters

RPS blast Εδώ έχουμε μια επιλογή να απενεργοποιήσουμε το RPS blast. Αυτό συνελεί στο να μην χρησιμοποιηθεί αναζήτηση για συντηρημένα πεδία και άρα ο αλγόριθμος τελειώνει πιο γρήγορα. Όταν έχουμε πρωτεΐνες ίδια οικογένειας αποεπιλέγουμε το rps blast. Όταν έχουμε πρωτεΐνες διαφορετικών οικογενειών το επιλέγουμε γιατί θα μας δωθούν αποτελέσματα κακής ποιότητας.

Use RPS BLAST to guide alignment

Constraint E-value Καθορίζει την μέση αναμενόμενη τιμή σφαλμάτων στα αμινοξέα. Αν έχουμε πολύ σημαντικές ακολουθίες, βάζουμε χαμηλό δείκτη. “Μόνο οι αντιστοιχίες με Evalue χαμηλότερο από το κατώφλι θα χρησιμοποιηθούν”

Conserved columns Αυτή η επιλογή επανεξετάζει τις στήλες που είχαν επιτυχία μετά το πέρας του αλγορίθμου για τυχόν βελτιώσεις. Αν το αποεπιλέξουμε θα γλιτώσουμε χρόνο αλλά θα χάσουμε ποιότητα.

Find Conserved Columns and Recompute Alignment

Query Clustering Parameters

Query Clustering Parameters Η επιλογή “Use query clusters” στο Cobalt είναι μια προαιρετική ρύθμιση που μπορεί να βοηθήσει στη μείωση του χρόνου υπολογισμού¹. Όταν αυτή η επιλογή είναι ενεργοποιημένη, το Cobalt χρησιμοποιεί

συστάδες από παρόμοιες ακολουθίες για να μειώσει τον αριθμό των απαραίτητων συγκρίσεων.

Η ιδέα πίσω από τη χρήση συστάδων είναι ότι οι περιορισμοί δεν συμβάλλουν στην ευθυγράμμιση πολύ παρόμοιων ακολουθιών. Αυτό σημαίνει ότι, αν έχετε πολλές ακολουθίες που είναι πολύ παρόμοιες μεταξύ τους, το Cobalt μπορεί να τις θεωρήσει ως μια συστάδα και να τις επεξεργαστεί ως μια ενιαία οντότητα, αντί να τις επεξεργαστεί ξεχωριστά. Αυτό μπορεί να επιταχύνει σημαντικά τη διαδικασία ευθυγράμμισης, ειδικά για μεγάλα σύνολα δεδομένων. Ωστόσο, πρέπει να σημειωθεί ότι η χρήση συστάδων μπορεί να έχει επιπτώσεις στην ακρίβεια της τελικής ευθυγράμμισης, εξαρτάται δε από τα δεδομένα και τους στόχους της ανάλυσης.

Use query clusters

Word Size Η επιλογή “Word Size” στο Cobalt αναφέρεται στο μέγεθος των λέξεων που χρησιμοποιούνται για τη σύγκριση των ακολουθιών. Συγκεκριμένα, ορίζει τον αριθμό των συνεχόμενων χαρακτήρων (ή “λέξεων”) που πρέπει να ταιριάζουν μεταξύ δύο ακολουθιών για να θεωρηθούν ως αντιστοιχία.

Για παράδειγμα, αν ορίσετε το “Word Size” σε 3, τότε το Cobalt θα ψάχνει για ακολουθίες 3 συνεχόμενων χαρακτήρων που ταιριάζουν μεταξύ των δύο ακολουθιών. Αν βρεθούν τέτοιες αντιστοιχίες, τότε οι ακολουθίες θεωρούνται ότι έχουν κάποια ομοιότητα.

Η επιλογή αυτή είναι σημαντική για την απόδοση και την ακρίβεια της ανάλυσης. Μια μεγαλύτερη τιμή για το “Word Size” μπορεί να οδηγήσει σε πιο ακριβείς αντιστοιχίες, αλλά μπορεί επίσης να αυξήσει τον χρόνο εκτέλεσης της ανάλυσης. Αντίθετα, μια μικρότερη τιμή μπορεί να οδηγήσει σε πιο γρήγορη ανάλυση, αλλά μπορεί να μην εντοπίζει όλες τις πιθανές αντιστοιχίες. Η κατάλληλη τιμή εξαρτάται από τα δεδομένα και τους στόχους της ανάλυσης.

Max Cluster distance Η επιλογή “Max Cluster distance” στο Cobalt αναφέρεται στη μέγιστη επιτρεπόμενη απόσταση μεταξύ δύο ακολουθιών σε μια συστάδα. Αυτό το όριο αποτρέπει το Cobalt από το να δημιουργεί συστάδες από άσχετες ακολουθίες. Η απόσταση μεταξύ δύο ακολουθιών υπολογίζεται ως το ποσοστό των λέξεων που εμφανίζονται και στις δύο ακολουθίες σε σχέση με τον αριθμό όλων των λέξεων στη μεγαλύτερη ακολουθία.

Οι τιμές που παραθέτετε (0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95) είναι πιθανά ρυθμίσεις για αυτήν την επιλογή. Μια χαμηλότερη τιμή θα οδηγήσει σε πιο στενές συστάδες, ενώ μια υψηλότερη τιμή θα επιτρέψει μεγαλύτερη απόσταση μεταξύ των ακολουθιών σε μια συστάδα, πιθανώς δημιουργώντας πιο ευρείες συστάδες. Η κατάλληλη τιμή εξαρτάται από τα δεδομένα και τους στόχους της ανάλυσης.

0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95

Alphabet Αυτή η παράμετρος αναφέρεται στο αλφάβητο που χρησιμοποιείται για τη δημιουργία αναπαραστάσεων αριθμού k-mer των ακολουθιών. Οι διαθέσιμες επιλογές είναι SE-B15 και SE-V10, που είναι αλφάβητα 15 και 10 γραμμάτων. Το "Regular" αναφέρεται στο κανονικό αλφάβητο 20 αμινοξέων

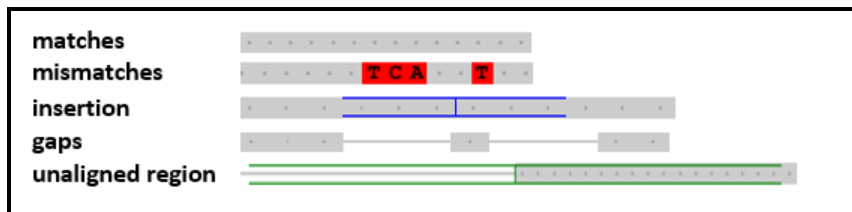
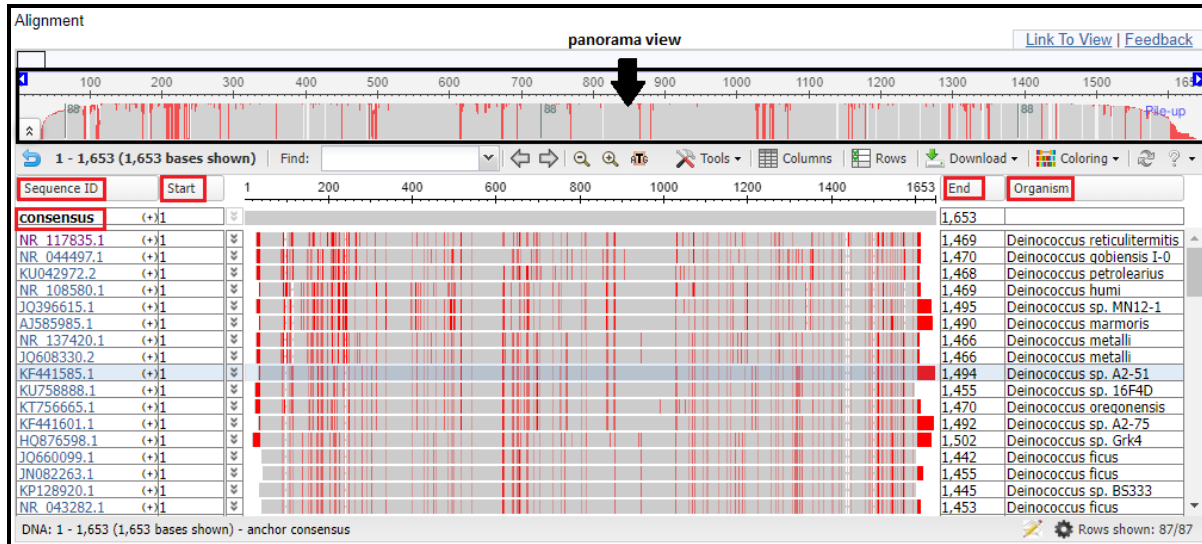
SE-B15
SE-V10
Regular

Πηγή: COBALT: constraint-based alignment tool for multiple protein sequences
Jason S. Papadopoulos and Richa Agarwala
[## NCBI Multiple Sequence Alignment Viewer](https://watermark.silverchair.com/bioinformatics_23_9_1073.pdf?token=AQECAHi208BE49Ooan9kKhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAA3kwggN1BgkqhkiG9w0BBwagggNmMIIDYglBADCCA1sGCSqGSIb3DQEHAeBglghkgBZQMEAS4wEQQML65iZrBU8t8wrr_6AgEQglIDLKGqXCH6PYRFCFnIm5B1vQb6MzJqNBCfPAEQH2KtYmBlWiyu00YGIUU3L6-fURopfaaCyTp_R_xTPSnNLQ4r260BKmxwVEhjUSPGa_Ze6ccnfPpfW2G0HNNH5t1_bBiqHFcK_vAdGYFcHoSGfoVPUwaawFZcs4YCoIZ_7A4LZpbo62KvTanZ6L_PYGMBXcBddZjiN_uTaKlzz7uze9-N-lZsof72-q8grvhjWzslhC7fpM21JNKRT8X58SoPOrpRwxk5VH2rycWS3AvvHRAqseOHxZBklhEd9Gjz_h2eqRxM7BDVnMFrYoB4oYgpkU8lBUYLSABmY_UWdHqWaTqHjd9aZF8m74zYGBmh50lRb2XQ6rWpEn8pNVEKDvBqglxmdrKTtnlr2gRIQpPsLqP0jLd0klumnFKIQ8Sh21ofTMX9LIQ3ihK5ru8g1glvHhLDmH4Gmhd70DzN_RDhZl02fObp9OqrD70fDkVs-nwq5xmrxWJTKDf2z2Jk7yAj7irEjthulGmvcdXrRtRUS-tgj_-xcylIrU1W4TbjLVY2gkGmWo51nllTGdYPxrMEbYAREbiA6AKz1rtnNdXffcLOJYcvYTQz6FQoYNRZcvpQK_x9rdNb2J4_ba7k1B1Wer06g5os6KPv7znLghn1dqu4poiahimUQxL6LxC LmUfGCJZBY_NxtTJZNClXh7x9XfQPlxZCCZkmFuM7wOh3R493cPoyPnaob2HBvHL_aKq1wsiLh8gRbCd4H3bqj_Fr_G93aU2eK0nUDQw0ktZlGmfr7b1AAXzbzsBmPqlG0pdK7a7c kC4Y_Z6zpthwnCcN89UflvM9eZPUJ0cMkiF_Ue210vAlecPK_a2WwbQzqJMnCW7EdvAE N9wFFPYB3iehMuDK82a6Vhlu0aO6SqtK5YgMysdmqqan9WpVTK-QqvxDa16f1JMT0bul03u7gEd5v6uT7RE9x-Nuy5t2ofuc904GwYb2ZLpPLMzoyim6tGFLPJ6gzuZPvg5S7BA27p8ZBW5cgl4Lb-Xtxe7ML9nzBBFZ4PAW0pfJByJ3u0QslPk6Rx0ZKMOQt-uDv5xWr8y</p></div><div data-bbox=)

Ένα χρήσιμο εργαλείο για την αναπαράσταση ευθυγραμμίσεων με γραφικά, οι οποίες έχουν γίνει με εργαλεία όπως το Cobalt.

Χρησιμοποιεί χρώματα, (τα οποία μπορούν να οριστούν από το χρήστη) για να δείχνει τις διαφορές στις ευθυγραμμίσεις. προεπιλογή είναι τα κενά με γκρι, οι αστοχίες με κόκκινο και οι εισαγωγές με μπλε. Αν τα αρχεία μας είναι πολύ μεγάλα μπορούμε να κάνουμε ζουμ σε διάφορες περιοχές και να δούμε πιο αναλυτικά αποτελέσματα. Υπάρχουν πολλές επιλογές χρωματισμού. Μια από αυτές είναι να χρωματίσουμε τα αμινοξέα μέσω της υδροπάθειάς τους, πράγμα που θα ήταν χρήσιμο στη μελέτη που

έγινε για την δημοσίευση με τις υδροπάθειες των μορίων πο χρησιμοποιήθηκαν για να εξετάσουν την δομή ενός μορίου στη 2ρη εργασία βιοπληροφορικής.



Αν στοχεύουμε σε συγκεκριμένη ακολουθία μπορούμε να την θέσουμε ως "άγκυρα" και να δούμε όλες τις άλλες σε σχέση προς αυτήν.

Βλέπουμε ότι οι στήλες προσδιορισμού δεξιά και αριστερά έχουν τα ID και το όνομα του οργανισμού. Αυτές αλλάζουν για την διευκόλυνσή μας με ότι πεδία θέλουμε, πχ με την χώρα που προήλθε κάτι (πολύ χρήσιμο για ιούς). Αφού γίνει αυτό υπάρχουν και επιλογές ταξινόμησης για ότι μας ενδιαφέρει, πχ best alignments at the top. Συνήθεις επιλογές είναι **Sequence ID** , **Organism** , **Gene** , **Date** , **Country** , **Host** , **Source** , **(%) Identity**, **(%) Coverage**, **Mismatches** relative to an anchor

Επιπλέον μπορούμε να ζητάμε αναλυτική περιγραφή σε διάφορα σημεία που θέλουμε παραπάνω πληροφορίες. Γίνεται να συνδεθούμε και κατευθείαν με γνωστές βάσεις οργανισμών όπως GenBank και να βρούμε το αρχείο μας με ένα κλικ. Αν θέλουμε να εντοπίσουμε μια συγκεκριμένη ακολουθία αμινοξέων απλά την πληκρολογούμε.

- Clustal Omega: Είναι ένα νέο πρόγραμμα πολλαπλής ευθυγράμμισης ακολουθιών που χρησιμοποιεί οδηγούς δέντρων και τεχνικές HMM profile-

profile για τη δημιουργία ευθυγραμμίσεων μεταξύ τριών ή περισσότερων ακολουθιών. Είναι κατάλληλο για μεσαίες έως μεγάλες ευθυγραμμίσεις.
Επιλογές: OUTPUT FORMAT, DEALIGN INPUT, MBED-LIKE CLUSTERING, MBED-LIKE CLUSTERING, COMBINED ITERATIONS ,MAX GUIDE TREE ,MAX HMM ITERATIONS ,ORDER ,DISTANCE MATRIX ,OUTPUT GUIDE TREE

- EMBOSS Cons: Το EMBOSS Cons δημιουργεί μια ακολουθία συναίνεσης από μια πολλαπλή ευθυγράμμιση πρωτεϊνών ή νουκλεοτιδίων. Επιλογές: PLURALITY ,SETCASE ,IDENTITY ,NAME ,MATRIX
- Kalign: Είναι ένα πολύ γρήγορο εργαλείο MSA που επικεντρώνεται σε τοπικές περιοχές. Είναι κατάλληλο για μεγάλες ευθυγραμμίσεις. Επιλογές: OUTPUT FORMAT ,MACSIM,GAP OPEN PENALTY ,GAP EXTENSION PENALTY ,TERMINAL GAP PENALTIES
- MAFFT: Είναι ένα εργαλείο MSA που χρησιμοποιεί γρήγορους μετασχηματισμούς Fourier. Είναι κατάλληλο για μεσαίες έως μεγάλες ευθυγραμμίσεις. Επιλογές: OUTPUT FORMAT ,MATRIX (PROTEIN ONLY) ,GAP OPEN PENALTY ,GAP EXTENSION PENALTY ,ORDER ,TREE REBUILDING NUMBER ,GUIDE TREE OUTPUT ,MAX ITERATE ,PERFORM FFTS
- MUSCLE: Είναι ένα ακριβές εργαλείο MSA, ιδιαίτερα καλό με πρωτεΐνες. Είναι κατάλληλο για μεσαίες ευθυγραμμίσεις. Επιλογές: OUTPUT FORMAT ,OUTPUT TREE
- MView: Μετατρέπει ένα αποτέλεσμα αναζήτησης ομοιότητας ακολουθίας σε μια πολλαπλή ευθυγράμμιση ακολουθιών ή μεταμορφώνει μια πολλαπλή ευθυγράμμιση ακολουθιών χρησιμοποιώντας το πρόγραμμα MView. Επιλογές: INPUT FORMAT ,OUTPUT FORMAT ,HTML MARKUP ,CSS ,PCID ,ALIGNMENT ,RULER ,ALIGNMENT WIDTH ,COLORING ,COLOR MAP ,CONSENSUS ,CONCOLOURING ,GROUPMAP ,CONCOLORMAP ,CONGROUPMAP ,CONGAPS

- **T-Coffee:** Είναι ένα εργαλείο MSA βασισμένο στη συνέπεια που προσπαθεί να αντιμετωπίσει τις παγίδες των προοδευτικών μεθόδων ευθυγράμμισης. Είναι κατάλληλο για μικρές ευθυγραμμίσεις. Επιλογές: OUTPUT FORMAT ,MATRIX , ORDER
 - **WebPRANK:** Το EBI έχει ένα νέο πρόγραμμα πολλαπλής ευθυγράμμισης ακολουθιών που είναι ευαισθητοποιημένο στη φυλογένεση και χρησιμοποιεί πληροφορίες εξέλιξης για να βοηθήσει στην τοποθέτηση εισαγωγών και διαγραφών – Αυτό το εργαλείο ξεχωρίζει από τα παραπάνω καθώς έχει ενσωματωμένα μοντέλα δομής των ακολουθιών, που μπορούν να βοηθήσουν στην κατανόηση της λειτουργίας των πρωτεϊνών και στην ακριβέστερη ευθυγράμμιση τους.
1. **OUTPUT FORMAT:** Καθορίζει τη μορφή των αποτελεσμάτων που επιστρέφονται από το πρόγραμμα.
 2. **DEALIGN INPUT:** Αφαιρεί τις αντιστοιχίσεις από τις εισαγόμενες ακολουθίες.
 3. **MBED-LIKE CLUSTERING:** Χρησιμοποιεί μια τεχνική ομαδοποίησης για να βελτιώσει την απόδοση της διαδικασίας ευθυγράμμισης.
 4. **COMBINED ITERATIONS:** Συνδυάζει πολλαπλές επαναλήψεις για να βελτιώσει την ακρίβεια της ευθυγράμμισης.
 5. **MAX GUIDE TREE:** Καθορίζει τον μέγιστο αριθμό των δέντρων οδηγών που θα χρησιμοποιηθούν.
 6. **MAX HMM ITERATIONS:** Καθορίζει τον μέγιστο αριθμό των επαναλήψεων HMM.
 7. **ORDER:** Καθορίζει τη σειρά των ακολουθιών στην ευθυγράμμιση.
 8. **DISTANCE MATRIX:** Υπολογίζει μια μήτρα αποστάσεων μεταξύ των ακολουθιών.
 9. **OUTPUT GUIDE TREE:** Επιστρέφει το δέντρο οδηγό που χρησιμοποιήθηκε για την ευθυγράμμιση.
 10. **PLURALITY:** Καθορίζει τον αριθμό των ακολουθιών που πρέπει να συμφωνούν για να σχηματίσουν μια συναίνεση.
 11. **SETCASE:** Καθορίζει την περίπτωση των ακολουθιών εξόδου.
 12. **IDENTITY:** Υπολογίζει την ταυτότητα μεταξύ των ακολουθιών.
 13. **NAME:** Καθορίζει το όνομα της ευθυγράμμισης.
 14. **MATRIX:** Καθορίζει τη μήτρα αντικατάστασης που θα χρησιμοποιηθεί.
 15. **MACSIM:** Επιτρέπει την ανάλυση των αποτελεσμάτων με το εργαλείο MACSIM.
 16. **GAP OPEN PENALTY:** Καθορίζει την ποινή για το άνοιγμα ενός κενού.
 17. **GAP EXTENSION PENALTY:** Καθορίζει την ποινή για την επέκταση ενός κενού.

18. **TERMINAL GAP PENALTIES:** Καθορίζει τις ποινές για τα κενά στα άκρα των ακολουθιών.
19. **MATRIX (PROTEIN ONLY):** Καθορίζει τη μήτρα αντικατάστασης πρωτεϊνών που θα χρησιμοποιηθεί.
20. **TREE REBUILDING NUMBER:** Καθορίζει τον αριθμό των επαναδομήσεων δέντρου που θα πραγματοποιηθούν.
21. **GUIDE TREE OUTPUT:** Επιστρέφει το δέντρο οδηγό που χρησιμοποιήθηκε για την ευθυγράμμιση.
22. **MAX ITERATE:** Καθορίζει τον μέγιστο αριθμό επαναλήψεων που θα πραγματοποιηθούν.
23. **PERFORM FFTS:** Καθορίζει αν θα πραγματοποιηθούν γρήγοροι μετασχηματισμοί Fourier.
24. **OUTPUT TREE:** Επιστρέφει το δέντρο που παράγεται από την ευθυγράμμιση.
25. **INPUT FORMAT:** Καθορίζει τη μορφή των εισαγόμενων ακολουθιών.
26. **HTML MARKUP:** Καθορίζει αν θα χρησιμοποιηθεί HTML για τη διαμόρφωση των αποτελεσμάτων.
27. **CSS:** Καθορίζει το CSS που θα χρησιμοποιηθεί για τη διαμόρφωση των αποτελεσμάτων.
28. **PCID:** Υπολογίζει την ταυτότητα των ακολουθιών.
29. **ALIGNMENT:** Καθορίζει τη μορφή της ευθυγράμμισης.
30. **RULER:** Καθορίζει αν θα εμφανιστεί ένας χάρακας για την ευθυγράμμιση.
31. **ALIGNMENT WIDTH:** Καθορίζει το πλάτος της ευθυγράμμισης.
32. **COLOR MAP:** Καθορίζει τον χάρτη χρωμάτων που θα χρησιμοποιηθεί για την ευθυγράμμιση.
33. **CONSENSUS:** Υπολογίζει μια ακολουθία συναίνεσης από την ευθυγράμμιση.
34. **CONCOLOURING:** Καθορίζει τον τρόπο χρωματισμού της ακολουθίας συναίνεσης.
35. **GROUPMAP:** Καθορίζει τον χάρτη ομάδας που θα χρησιμοποιηθεί για την ευθυγράμμιση.
36. **CONCOLORMAP:** Καθορίζει τον χάρτη χρωμάτων που θα χρησιμοποιηθεί για την ακολουθία συναίνεσης.
37. **CONGROUPMAP:** Καθορίζει τον χάρτη ομάδας που θα χρησιμοποιηθεί για την ακολουθία συναίνεσης.
38. **CONGAPS:** Καθορίζει αν θα εμφανιστούν κενά στην ακολουθία συναίνεσης.
39. **COLORING:** Καθορίζει τα χρώματα.