ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΒΙΟΧΗΜΕΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΤΣΙΜΠΙΔΗΣ ΜΙΧΑΗΛ

ΛΑΡΙΣΑ 2016

T-RECs: Ένα εργαλείο για γρήγορο και μεγάλης κλίμακας εντοπισμό ανασυνδυασμών ανάμεσα σε διαφορετικές εξελικτικές ομάδες ιικών γονιδιωμάτων

T-RECs: Rapid and large-scale detection of recombination events among different evolutionary lineages of viral genomes.

T-RECs: Ένα εργαλείο για γρήγορο και μεγάλης κλίμακας εντοπισμό ανασυνδυασμών ανάμεσα σε διαφορετικές εξελικτικές ομάδες ιικών γονιδιωμάτων

Όνομα: Τσιμπίδης Μιχαήλ

Επιβλέπων Καθηγητής: Γρηγόριος Αμούτζιας

Εργαστήριο Βιοπληροφορικής του τμήματος
Βιοχημείας & Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ

Αμούτζιας Γρηγόριος (Επιβλέπων)

Επίκουρος Καθηγητής Βιοπληροφορικής στη Γενωμική
Τμήμα Βιοχημείας & Βιοτεχνολογίας

Μαρκουλάτος Παναγιώτης

Καθηγητής Εφαρμοσμένης Μικροβιολογίας με έμφαση στη Βιοτεχνολογία
Τμήμα Βιοχημείας & Βιοτεχνολογίας

Μόσιαλος Δημήτριος

Επίκουρος Καθηγητής Βιοτεχνολογίας Μικροβίων
Τμήμα Βιοχημείας & Βιοτεχνολογίας

Αρχικά θα ήθελα να ευχαριστήσω τον κύριο Αμούτζια Γρηγόριο, Επίκουρο Καθηγητή Βιοπληροφορικής στη Γενωμική, του τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας, για τη στήριξη και τη βοήθεια που μου παρείχε καθ' όλο το διάστημα εκπόνησης της διπλωματικής μου εργασίας.

Ακόμη, θα ήθελα να ευχαριστήσω τον κύριο Μαρκουλάτο Παναγιώτη, Καθηγητή Εφαρμοσμένης Μικροβιολογίας με έμφαση στη Βιοτεχνολογία, του τμήματος Βιοχημείας και βιοτεχνολογίας και μέλος της τριμελούς επιτροπής καθώς και τον κύριο Μόσιαλο Δημήτριο, Επίκουρο Καθηγητή Βιοτεχνολογίας Μικροβίων, του τμήματος Βιοχημείας και Βιοτεχνολογίας και μέλος της τριμελούς επιτροπής για τις συμβουλές που μου έδωσαν κατά την εκπόνηση της διπλωματικής μου εργασίας.

Περιεχόμενα

Περίληψη

Στις μέρες μας υπάρχουν πολλά υπολογιστικά εργαλεία τα οποία είναι σε θέση να εντοπίζουν γεγονότα ανασυνδυασμών σε ιικά γονιδιώματα. Το πρόβλημα με τα προγράμματα αυτά έγκειται στο γεγονός ότι δεν είναι κατασκευασμένα ώστε να μπορούν να ανταπεξέλθουν στον εκθετικά αυξανόμενο όγκο δεδομένων. Έτσι, είναι απαραίτητη η ύπαρξη ενός υπολογιστικού εργαλείου το οποίο θα είναι σε θέση να αναλύσει εκατοντάδες ή ακόμη και χιλιάδες γονιδιώματα ή τμήματα ακολουθιών και να εντοπίσει πιθανά γεγονότα ανασυνδυασμού τα οποία θα μπορούν να αναλυθούν στη συνέχεια με πιο ευαίσθητες και εξειδικευμένες μεθόδους. Το πρόγραμμα T-RECs (**T**ool for **REC**ombination**S**) αποτελεί ένα εργαλείο με γραφικό περιβάλλον για λειτουργικά συστήματα Windows το οποίο χρησιμοποιεί στοίχιση κατά ζεύγη με συρόμενα παράθυρα ώστε να εντοπίσει πρόσφατα γεγονότα ανασυνδυασμού μεταξύ διαφορετικών εξελικτικών ομάδων. Μέσα από το πρόγραμμα είναι δυνατή η δημιουργία διαγραμμάτων ομοιότητας τα οποία ενσωματώνουν γραφικά τα αποτελέσματα της στοίχισης κατά ζεύγη καθώς και οποιονδήποτε διαθέσιμο σχολιασμό (annotation) των υπό μελέτη γονιδιωμάτων. Τα αποτελέσματα της ανάλυσης 555 γονιδιωμάτων Norovirus με το T-RECs έδειξαν πως είναι πολύ αποτελεσματικό.

Το T-RECs διατίθεται δωρεάν και η λήψη του μπορεί να γίνει ακολουθώντας τον παρακάτω σύνδεσμο: http://bioinf.bio.uth.gr/t-recs.html

1 Εισαγωγή

1.1 Ο ανασυνδυασμός

Ο γενετικός ανασυνδυασμός μαζί με τις σημειακές μεταλλάξεις αποτελούν τους βασικούς μηχανισμούς οι οποίοι παίζουν σημαντικό ρόλο στην εξέλιξη των ιών [1, 2]. Ο ανασυνδυασμός μπορεί να λειτουργήσει ως μηχανισμός που θα συνδυάσει ιδιότητες από διαφορετικά γονιδιώματα δημιουργώντας έτσι ένα νέο. Άμεση συνέπεια της παραπάνω διαδικασίας είναι η πιθανή ικανότητα των νέων στελεχών να ξεφεύγουν από το ανοσοποιητικό σύστημα αλλά και να εμφανίζουν ανθεκτικότητα σε φάρμακα. Για το λόγο αυτό είναι πολύ σημαντικό τα γεγονότα ανασυνδυασμού να εντοπίζονται και να συσχετίζονται με νέες ιδιότητες ή νέες επιδημίες.

Τα πλεονεκτήματα του ανασυνδυασμού χωρίζονται σε δύο μεγάλες τάξεις [3]:

1) Ο ανασυνδυασμός συμβάλλει στη δημιουργία και τη διάδοση χαρακτηριστικών που προσδίδουν κάποιο πλεονέκτημα

2) Ο ανασυνδυασμός επιτρέπει την αφαίρεση επιβλαβών μεταλλάξεων.

Σημαντικά ευρήματα υποδεικνύουν πως ορισμένοι RNA ιοί επωφελούνται από την αναδιάταξη του γονιδιώματος που οφείλεται σε γεγονότα ανασυνδυασμού. Ένα πλήθος πειραματικών μελετών έδειξε πως υπάρχει η δυνατότητα ανασυνδυασμού σε αδύναμα ή ακόμη και μη αντιγραφόμενα στελέχη που μπορεί να οδηγήσει στον σχηματισμό νέων, βιώσιμων, και με εξαιρετική προσαρμοστικότητα ιών. Τα παραδείγματα περιλαμβάνουν τις λειτουργικές χίμαιρες που σχηματίστηκαν μεταξύ RNAs και DI RNAs *tombusviruses* [4] και ανασυνδυασμένους, μολυσματικούς ιούς που προέκυψαν από διαφορετικούς συνδυασμούς μεταλλαγμένων RNAs από *Sindbis virus* [5, 6].

Επιδιόρθωση γονιδιωμάτων μέσω του ανασυνδυασμού έχει παρατηρηθεί επίσης και σε ιούς των φυτών [7, 8, 9, 10].

Πρόσφατα παρατηρήθηκαν ανασυνδυασμοί σε RNA ιούς για τους οποίους πληθώρα εμβολίων ήδη χρησιμοποιούνται ή δοκιμάζονται [11, 12, 13]. Το δυναμικό που έχει ένας ανασυνδυασμός να παράγει νέα παθογόνα υβριδικά είδη και οι πιθανές επιπτώσεις τέτοιων ανεξέλεγκτων περιπτώσεων ανασυνδυασμού, πρέπει να ληφθούν ιδιαίτερα υπ’ όψιν ειδικά σε περιπτώσεις όπου πληθώρα σχετικών εμβολίων χρησιμοποιούνται για να αντιμετωπιστούν αυτοί οι RNA ιοί [31].

Είναι πλέον βέβαιο πως ο ανασυνδυασμός και η αυξημένη παθογένεια συνδέονται άμεσα. Κατά τη διάρκεια της αντιγραφής, η πολυμεράση μπορεί να μεταπηδήσει από το εκμαγείο που αντιγράφει σε ένα άλλο παράγοντας ένα νέο, ανασυνδυασμένο ιό. Η διαδικασία αυτή είναι υπεύθυνη για πολλές περιπτώσεις

δημιουργίας των παθογόνων ειδών γνωστών και ως *feline infectious peritonitis viruses* (FIPVs) μέσω αλλαγής των ασυμπτωματικών *feline enteric coronaviruses* [14].

1.2 Μέθοδοι για τον εντοπισμό ανασυνδυασμών

Η μοριακή επανάσταση που ξεκίνησε με την ανάπτυξη της μεθόδου της PCR έχει παίξει πολύ σημαντικό ρόλο στη μελέτη των ιικών ανασυνδυασμών. Αναλύσεις ακολουθιών αλλά και φυλογενετικές τεχνικές έχουν αποδειχτεί τα τελευταία χρόνια εξαιρετικά αποτελεσματικές μέθοδοι για τον εντοπισμό και το χαρακτηρισμό ανασυνδυασμών ανάμεσα σε ιούς τόσο στη φύση [11, 12, 13, 15, 16, 17, 18, 19] όσο και στο εργαστήριο [6, 9, 20, 21, 22, 23]. Αξίζει να σημειωθεί πως οι παραπάνω μέθοδοι παίζουν ρόλο όχι μόνο στην ανακάλυψη πληροφοριών σχετικά με γεγονότα ανασυνδυασμού που μπορεί να έχουν προκύψει πολύ καιρό πριν ή είναι εξαιρετικά σπάνιοι [24, 25], αλλά και στην αναζήτηση λεπτομερειών που έχουν να κάνουν με τον ίδιο το μηχανισμό [20, 26].

Πληθώρα προγραμμάτων και εργαλείων Βιοπληροφορικής έχουν αναπτυχθεί τα τελευταία 20 χρόνια που είναι σε θέση να εντοπίσουν γεγονότα ανασυνδυασμού [27, 28, 29]. Τα προγράμματα αυτά βασίζονται σε τέσσερις μεγάλες κατηγορίες μεθόδων:

1) Σύγκριση ακολουθιών κατά ζεύγη
2) Φυλογενετική
3) Πληθυσμιακή γενετική
4) Μοτίβα περιοχών

Αρκετές εφαρμογές με γραφικό περιβάλλον συμπεριλαμβανομένων των «bootscanning» [28], «PhylPro» [25], «TOPAL» [30] και «DIVERT» [15] χρησιμοποιούν τη μέθοδο του «συρόμενου παραθύρου» (sliding window) για να εντοπίσουν διαφορές στις υπό μελέτη ακολουθίες που μπορεί να υποδεικνύουν γεγονός ανασυνδυασμού. Παρ' όλα αυτά, με την αλληλούχιση νέας γενιάς (Next generation sequencing - NGS) παράγεται συνεχώς πληθώρα νέων δεδομένων παρέχοντας αρκετό υλικό για ανάλυση. Το πρόβλημα όμως που δημιουργείται είναι ότι τα διαθέσιμα προγράμματα δεν είναι βελτιστοποιημένα ώστε να μπορούν να χειρίζονται με ευκολία έναν τόσο μεγάλο όγκο δεδομένων.

<u>Bootscanning</u>

Το Bootscanning κάνει μια φυλογενετική προσέγγιση η οποία ξεκινά με τη δημιουργία ενός φυλογενετικού δέντρου από ένα μικρό κομμάτι του «παραθύρου» στο ένα άκρο της στοίχισης και το αξιολογεί με βάση το Bootstrap. Στη συνέχεια το παράθυρο μεταφέρεται στο επόμενο σημείο της στοίχισης και ένα νέο φυλογενετικό δέντρο δημιουργείται και αξιολογείται. Η διαδικασία αυτή

συνεχίζεται για ολόκληρη τη στοίχιση. Έτσι το πρόγραμμα είναι σε θέση να εντοπίσει πιθανές περιοχές όπου έχει γίνει ανασυνδυασμός [31].

PhylPro και TOPAL

Τα προγράμματα PhylPro και TOPAL χρησιμοποιούν ζεύγη «συρόμενων παραθύρων» κατά μήκος της ακολουθίας. Το κάθε πρόγραμμα χρησιμοποιεί διαφορετικές παραμέτρους για τη φυλογένεση αλλά και στα δύο η φυλογενετική πληροφορία που περιέχεται σε ένα «παράθυρο» συγκρίνεται με εκείνη του γειτονικού «παραθύρου». Όταν δεν υπάρχει ανασυνδυασμός όλα τα «παράθυρα» αναμένεται να παράγουν παρόμοιο μοτίβο. Σε περίπτωση που έχει συμβεί ανασυνδυασμός, κάποια από τα ζεύγη παραθύρων αναμένεται να προσδώσουν διαφορετικό σήμα και μάλιστα η διαφορά τους θα πρέπει να είναι μεγαλύτερη όταν «σαρώνουν» ένα σημείο όπου ξεκινά ο ανασυνδυασμός [31].

DIVERT

Αποτελεί την απλούστερη μέθοδο «συρόμενου παραθύρου» (συχνά και την πιο αποτελεσματική), η οποία παράγει ένα γράφημα σύγκρισης γενετικών αποστάσεων μεταξύ επιλεγμένων ακολουθιών και ακολουθιών σύγκρισης, το οποίο μπορεί να απεικονίσει ομοιότητες και διαφορές μεταξύ των ακολουθιών που ενδεχομένως να οδηγήσουν σε γεγονότα ανασυνδυασμού [31].

Τα διαγράμματα αυτά (Diversity plots) έχουν χρησιμοποιηθεί με μεγάλη αποτελεσματικότητα στην εύρεση ανασυνδυασμένων στελεχών του ιού της ανοσοανεπάρκειας των ανθρώπων (HIV) και του ιού της ανοσοανεπάρκειας των πιθήκων (SIV) [15, 32].

RDP4

Η δύναμη του προγράμματος RDP4 έγκειται στην ταυτόχρονη χρήση ενός εύρους μεθόδων εντοπισμού ανασυνδυασμών τόσο για την εύρεση όσο και για τον χαρακτηρισμό των γεγονότων ανασυνδυασμού που προκύπτουν μέσα από μια στοίχιση ακολουθιών. Το RDP4 περιλαμβάνει τις μεθόδους BOOTSCANning [28, 33], GENECONV [34], Maximum Chi Square [35, 36], CHIMAERA [36], Sister Scanning [37], 3SEQ [38], VisRD [39] καθώς και την μέθοδο BURT. Ένα μειονέκτημα του προγράμματος ωστόσο είναι η αδυναμία στοίχισης ακολουθιών η οποία θα πρέπει να πραγματοποιηθεί από τον χρήστη με τη βοήθεια άλλων προγραμμάτων.

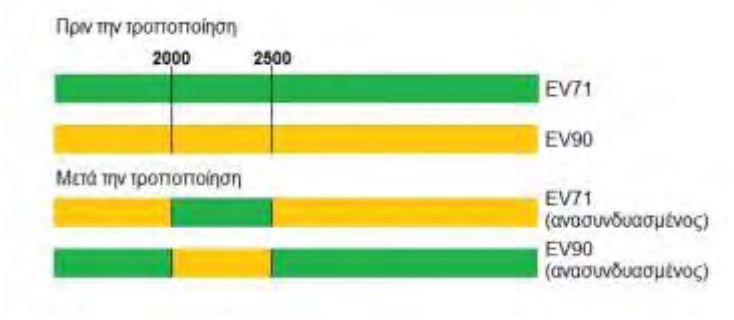
Παρ’ όλες τις δυνατότητες που προσφέρουν, οι μέθοδοι που χρησιμοποιούνται σήμερα δεν είναι διαμορφωμένες για την ανάλυση ολόκληρων γονιδιωμάτων από εκατοντάδες ή ακόμη και χιλιάδες ιούς με ένα απλό βήμα. Επίσης πολλά από τα προγράμματα αυτά δεν παρέχουν ένα γραφικό περιβάλλον φιλικό προς τον χρήστη. Οι συνεχείς βελτιώσεις στο χώρο των τεχνολογιών αλληλούχισης αναμένεται να

διογκώσουν το πρόβλημα αυτό καθώς ήδη πλατφόρμες αλληλούχισης όπως η Illumina και η Ion Proton επιτρέπουν την άμεση αλληλούχιση ολόκληρων ιικών γονιδιωμάτων [40].

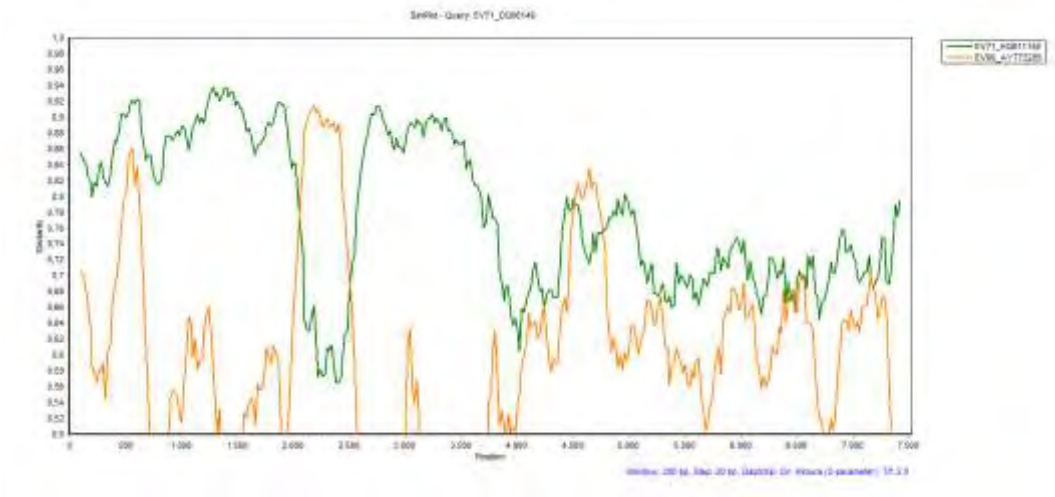
1.3 Διαγράμματα ομοιότητας (Similarity Plots)

Τα διαγράμματα ομοιότητας (Similarity Plots) αποτελούν ένα πολύ σημαντικό υπολογιστικό εργαλείο για τη σύγκριση ακολουθιών σε μια πολλαπλή στοίχιση. Η λειτουργία τους βασίζεται στη μέθοδο «συρόμενων παραθύρων» όπου συγκρίνεται η κάθε ακολουθία της στοίχισης με την ακολουθία επερώτησης, ώστε να αξιολογηθεί η μεταξύ τους ομοιότητα και να εντοπιστούν τυχόν ανασυνδυασμοί. Το αποτέλεσμα της διαδικασίας αυτής είναι ένα γράφημα όπου φαίνεται το ποσοστό ομοιότητας για όλες τις θέσεις της κάθε ακολουθίας που εξετάζεται ως προς την ακολουθία επερώτησης. Ένα από τα πιο γνωστά προγράμματα που παράγουν τέτοια διαγράμματα είναι το SimPlot.

Στη συνέχεια ακολουθεί ένα παράδειγμα διαγράμματος ομοιότητας:



Εικόνα 1: Προσομοίωση ανασυνδυασμού μεταξύ δύο γονιδιωμάτων (EV71 και EV90) στις θέσεις 2000 έως 2500



Εικόνα 2: Αποτελέσματα SimPlot της παραπάνω προσομοίωσης ανασυνδυασμού

1.4 Εργαλείο BLAST (Basic Local Alignment Search Tool)

Η εύρεση ομόλογων ακολουθιών από μια βάση δεδομένων είναι ένα πολύ σημαντικό εργαλείο Βιοπληροφορικής. Η διαδικασία αυτή επιτυγχάνεται με αλγόριθμους τοπικής στοίχισης κατά ζεύγη. Το BLAST αποτελεί τον πιο διαδεδομένο αλγόριθμο αυτού του είδους. Με τη χρήση του BLAST είναι επίσης δυνατός ο εντοπισμός ομοιοτήτων και οι διαφορών μεταξύ δύο ομόλογων ακολουθιών.

Ο καθορισμός των παραμέτρων με τις οποίες γίνεται κάθε αναζήτηση μέσω του BLAST είναι πολύ σημαντικός καθώς ακόμη και μικρές αλλαγές μπορεί να επηρεάσουν τα αποτελέσματα που θα επιστρέψει ο αλγόριθμος. Το *e-value* αποτελεί τη σημαντικότερη παράμετρο. Πρόκειται για μια τιμή η οποία περιγράφει την πιθανότητα ενός αποτελέσματος να εμφανιστεί ως ομόλογο καθαρά από τύχη.

Το εργαλείο BLAST παρέχει τη δυνατότητα πραγματοποίησης διαφόρων μορφών στοίχισης. Στον πίνακα 1 φαίνονται τα προγράμματα που μπορούν να χρησιμοποιηθούν μέσα από το BLAST.

Πρόγραμμα	Βάση δεδομένων	Ακολουθία επερώτησης
BLASTN	Νουκλεοτιδική	Νουκλεοτιδική
BLASTP	Πρωτεϊνική	Πρωτεϊνική
BLASTX	Πρωτεϊνική	Νουκλεοτιδική μεταφρασμένη σε πρωτεϊνική
TBLASTN	Νουκλεοτιδική μεταφρασμένη σε πρωτεϊνική	Πρωτεϊνική
TBLASTX	Νουκλεοτιδική μεταφρασμένη σε πρωτεϊνική	Νουκλεοτιδική μεταφρασμένη σε πρωτεϊνική

Πίνακας 1: Προγράμματα του BLAST

1.5 Πολλαπλή στοίχιση (Multiple Sequence Alignment – MSA)

Ως πολλαπλή στοίχιση, ορίζουμε τη στοίχιση όπου ορισμένες ακολουθίες μιας οικογένειας (από τρεις και πάνω) στοιχίζονται με αποτέλεσμα τον εντοπισμό συντηρημένων και μη περιοχών μεταξύ των ακολουθιών αυτών. Χρησιμοποιείται για τη δημιουργία profiles/motifs που χαρακτηρίζουν μια επικράτεια (domain), για την ανίχνευση συντηρημένων DNA-binding sites σε προαγωγείς γονιδίων, στην πρόβλεψη δευτεροταγούς και τριτοταγούς δομής πρωτεϊνών, στο σχεδιασμό εκφυλισμένων εκκινητών PCR και ειδικότερα στη φυλογενετική ανάλυση.

Η πολλαπλή στοίχιση μπορεί να πραγματοποιηθεί με διαφορετικές μεθόδους οι οποίες εφαρμόζονται από εξειδικευμένα προγράμματα. Οι μέθοδοι αυτοί μπορεί να είναι ο *Δυναμικός προγραμματισμός* ή κάποια *Ευρετική μέθοδος* (Προοδευτική

στοίχιση, Στοίχιση με διαδοχικές βελτιώσεις, Στοίχιση βασισμένη σε Blocks). Ορισμένα αρκετά διαδεδομένα προγράμματα που πραγματοποιούν πολλαπλή στοίχιση είναι το MUSCLE, το ClustalW και το T-coffee τα οποία βασίζονται στην προοδευτική στοίχιση.

Κανένα πρόγραμμα δεν παράγει τη βέλτιστη στοίχιση και για το λόγο αυτό θα πρέπει πάντα να πραγματοποιείται βελτίωση της στοίχισης χειροκίνητα. Δύο γνωστά προγράμματα που χρησιμοποιούνται για το σκοπό αυτό είναι το Seaview και το Bioedit.

2 Υλικά και μέθοδοι

2.1 T-RECs

Το λογισμικό που περιγράφεται στην παρούσα εργασία ονομάζεται T-RECs (**T**ool for **REC**ombination**s**) και βασίζεται στην ευρετική μέθοδο τοπικής στοίχισης κατά ζεύγη του BLASTN με χρήση «συρόμενων παραθύρων». Πρόκειται για ένα εργαλείο το οποίο μπορεί να φιλτράρει ακολουθίες μέσα από ένα μεγάλο αριθμό ακολουθιών και να εντοπίζει πιθανά γεγονότα ανασυνδυασμού σε αυτές. Σημαντικά πλεονεκτήματα του προγράμματος έναντι των προγραμμάτων που ήδη χρησιμοποιούνται για εντοπισμό ανασυνδυασμών είναι τα εξής: 1) Το T-RECs είναι σε θέση να αναλύσει εκατοντάδες, ακόμη και χιλιάδες ακολουθίες ταυτόχρονα, 2) Παρέχει τη δυνατότητα παραγωγής διαγραμμάτων ομοιότητας στα οποία ο χρήστης μπορεί να εισάγει σχολιασμό και 3) Το πρόγραμμα παρέχει ένα γραφικό περιβάλλον φιλικό προς τον χρήστη και είναι διαθέσιμο για Microsoft Windows 7, 8, 8.1 και 10. Η λήψη του προγράμματος μπορεί να γίνει ακολουθώντας τον παρακάτω σύνδεσμο: *http://bioinf.bio.uth.gr/t-recs.html*

Για τη σωστή λειτουργία του προγράμματος είναι απαραίτητη η παράλληλη εγκατάσταση των προγραμμάτων Muscle, BLASTN και UCLUST η οποία γίνεται αυτόματα για τα δύο πρώτα, ενώ το UCLUST πρέπει να εγκατασταθεί ξεχωριστά από τον χρήστη.

2.2 Microsoft Visual Studio 2010

Για την ανάπτυξη του T-RECs χρησιμοποιήθηκε η γλώσσα προγραμματισμού της Microsoft, Visual Basic 2010 ενώ η κατασκευή του προγράμματος έγινε στο περιβάλλον του Microsoft Visual Studio 2010.

2.3 UCLUST – USEARCH

Όπως ήδη έχει αναφερθεί, το T-RECs χρειάζεται το UCLUST για να είναι πλήρως λειτουργικό. Πρόκειται για ένα πρόγραμμα που ενσωματώνει τον αλγόριθμο UCLUST ο οποίος έχει τη δυνατότητα να δημιουργεί ομάδες (clusters) ακολουθιών με βάση την ομοιότητά τους. Το UCLUST ανήκει στη βιβλιοθήκη προγραμμάτων του USEARCH που δημιουργήθηκε από τον Robert Edgar και η λήψη του μπορεί να γίνει από τον παρακάτω σύνδεσμο: *http://drive5.com/usearch/*

2.4 BLASTN

Ένα ακόμη πρόγραμμα που χρειάζεται το T-RECs για να λειτουργήσει είναι το BLASTN. Ο βασικός μηχανισμός του T-RECs για τον εντοπισμό ανασυνδυασμών στηρίζεται στο BLASTN το οποίο λαμβάνει νουκλεοτιδικές ακολουθίες επερώτησης και τις συγκρίνει με τις ακολουθίες που υπάρχουν στην προκαθορισμένη από τον

χρήστη βάση δεδομένων. Το BLASTN εγκαθίσταται αυτόματα παράλληλα με την εγκατάσταση του T-RECs.

2.5 makeblastdb

Μαζί με το BLASTN εγκαθίσταται αυτόματα και ένα ακόμη πρόγραμμα που ονομάζεται makeblastdb. Το πρόγραμμα αυτό είναι υπεύθυνο για τη σωστή λειτουργία του T-RECs όσον αφορά τη μετατροπή της προκαθορισμένης βάσης δεδομένων σε μορφή που μπορεί να χρησιμοποιηθεί από το BLASTN.

Τόσο το makeblastdb όσο και το BLASTN περιλαμβάνονται στη βιβλιοθήκη προγραμμάτων του BLAST+ και μπορούν να ληφθούν μέσω του παρακάτω συνδέσμου:
https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download

2.6 Muscle

Το τελευταίο πρόγραμμα που είναι απαραίτητο είναι το Muscle. Το πρόγραμμα αυτό πραγματοποιεί την πολλαπλή στοίχιση που χρειάζεται το T-RECs προκειμένου να παράγει τα διαγράμματα ομοιότητας. Δημιουργός του Muscle είναι ο Robert Edgar και η λήψη του προγράμματος μπορεί να γίνει από τον παρακάτω σύνδεσμο: *http://drive5.com/muscle/*

2.7 Αξιολόγηση του T-RECs

Για την αξιολόγηση του T-RECs χρησιμοποιήθηκαν δεδομένα από την πτυχιακή του κ. Γιώργου Μπαχούμη:

Αρχικά έγινε λήψη 555 ολόκληρων γονιδιωμάτων Norovirus για τα οποία πραγματοποιήθηκε φυλογενετική ανάλυση. Με τη βοήθεια perl scripts έγινε η εξαγωγή πληροφοριών, όπως αυτές ήταν καταγεγραμμένες από την Genbank, για τις ληφθείσες ακολουθίες. Οι πληροφορίες αυτές είχαν να κάνουν τόσο με την εξελικτική ομάδα στην οποία ανήκε κάθε ακολουθία όσο και με τα όρια των τριών ORFs και των οκτώ πεπτιδίων του γονιδιώματος. Το ORF1 κωδικοποιεί για ένα πολυπεπτίδιο που παράγει 6 επιμέρους πεπτίδια (p48, NTPase, p22, VPG, 3C, RNApol), ενώ τα ORF2 και 3 κωδικοποιούν για δύο καψιδιακές περιοχές (VP1 και VP2 αντίστοιχα). Η γενοτύπηση για τους Norovirus γίνεται με βάση την περιοχή VP1. Έτσι, όταν ορισμένες ακολουθίες δεν ήταν ξεκάθαρο σε ποια εξελικτική ομάδα ανήκουν, η ταυτοποίησή τους έγινε με φυλογενετική ανάλυση στο ORF2.

Οι ακολουθίες αυτές αναλύθηκαν από το πρόγραμμα RDP4 για τον εντοπισμό πιθανών γεγονότων ανασυνδυασμού. Στη συνέχεια, οι ίδιες ακολουθίες αναλύθηκαν με το T-RECs ώστε να ελεγχθεί η λειτουργία του. Για τη χρήση τους στο RDP4, οι στοιχίσεις των επιμέρους πεπτιδίων ενώθηκαν ώστε να σχηματιστεί μία ενιαία στοίχιση. Το RDP4 εντόπισε 11 διαοροτυπικά ανασυνδυασμένες ακολουθίες όπου το μέγεθος του δότη ήταν μεγαλύτερο από 200 nt. Ύστερα από έλεγχο

αποκαλύφθηκε πως 7 από αυτές τις 11 περιπτώσεις ήταν ξεκάθαρα γεγονότα ανασυνδυασμού, 1 περίπτωση δεν ήταν τόσο ξεκάθαρη ενώ 3 από αυτές τις περιπτώσεις αφορούσαν παλαιά γεγονότα. Κατά την ανάλυση των ίδιων 555 ολόκληρων γονιδιωμάτων με το T-RECs, εντοπίστηκαν οι 7 ξεκάθαρες περιπτώσεις που είχαν βρεθεί και από το RDP4. Από τα τρία παλαιά γεγονότα που εντοπίστηκαν από το RDP4, το T-RECs απέτυχε στον εντοπισμό των δύο ενώ κατάφερε να εντοπίσει το ένα. Επίσης, ένα πιθανό γεγονός που εντοπίστηκε από το T-RECs ήταν ψευδώς θετικό και τέλος υπήρχε ένα ακόμη γεγονός που εντοπίστηκε επίσης από το RDP4 όπου το κομμάτι του δευτερεύων δότη ήταν 160 nt. Όλες οι παραπάνω περιπτώσεις εξετάσθηκαν με φυλογενετικά δέντρα και διαγράμματα ομοιότητας ώστε να επαληθευτεί η εγκυρότητά τους.

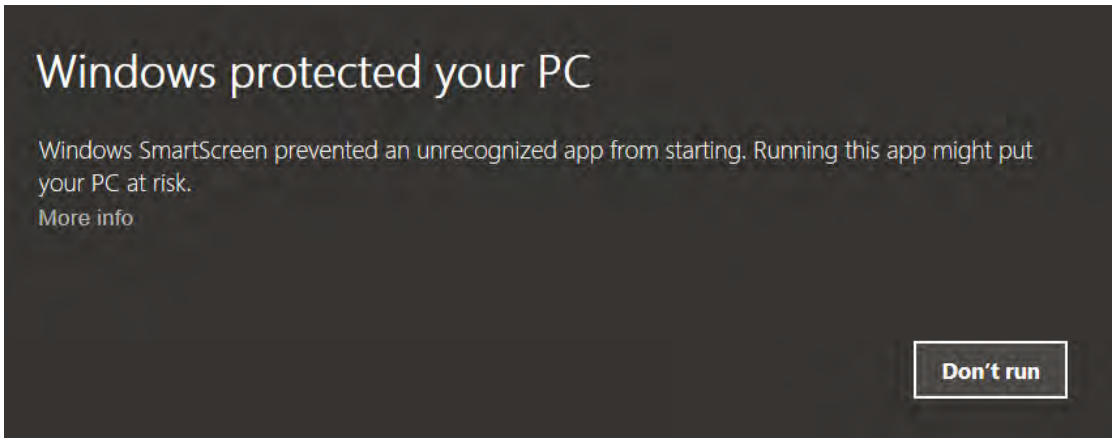
3 Αποτελέσματα

3.1 Λήψη και εγκατάσταση του T-RECs

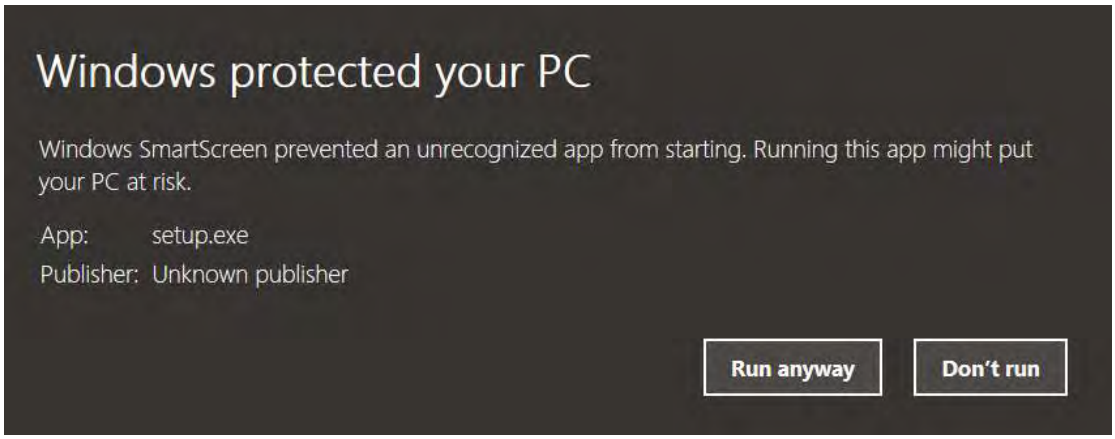
Το T-RECs διατίθεται δωρεάν από την ιστοσελίδα του εργαστηρίου Βιοπληροφορικής του τμήματος Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας και η λήψη του μπορεί να γίνει ακολουθώντας το σύνδεσμο *http://bioinf.bio.uth.gr/t-recs.html* . Από εκεί θα πρέπει να επιλεχθεί το Download T-RECs software ώστε να ξεκινήσει η λήψη. Το αρχείο λήψης θα έχει το όνομα *T-RECs_software.zip* του οποίου η κατάληξη (.zip) υποδηλώνει πως πρόκειται για ένα συμπιεσμένο αρχείο. Θα πρέπει λοιπόν να γίνει εξαγωγή/αποσυμπίεση των περιεχομένων του σε οποιοδήποτε σημείο του δίσκου επιθυμεί ο χρήστης.

Για την εγκατάσταση του προγράμματος, αφού γίνει η εξαγωγή των περιεχομένων του φακέλου *T-RECs_software.zip* θα πρέπει να εκτελεστεί το αρχείο *setup.exe* το οποίο βρίσκεται στο φάκελο *T-RECs*. Στο σημείο αυτό μπορεί να εμφανιστούν τα εξής 2 προβλήματα:

1) Το πρόγραμμα SmartScreen των Windows ενδέχεται να μην αναγνωρίζει το T-RECs ως έγκυρο πρόγραμμα εμφανίζοντας το μήνυμα της εικόνας 3. Για την αντιμετώπιση του προβλήματος αυτού θα πρέπει να επιλεχθεί το «More info» και έπειτα το «Run anyway» όπως φαίνεται στην εικόνα 4.

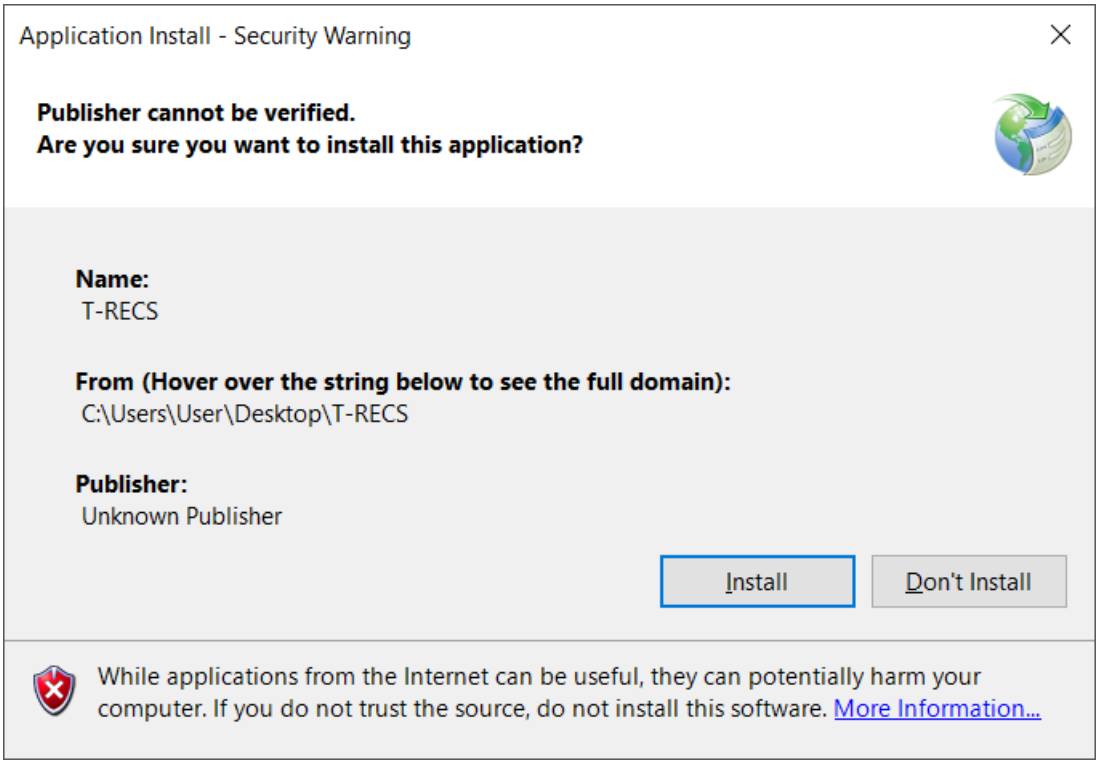


Εικόνα 3: Το πρόγραμμα SmartScreen των Windows δεν αναγνωρίζει το T-RECs ως έγκυρο πρόγραμμα. (Εικόνα από Windows 10)



Εικόνα 4: Για να εκτελεστεί το αρχείο setup.exe θα πρέπει να επιλεχθεί το «More info» και στη συνέχεια το «Run anyway» (Εικόνα από Windows 10)

Στη συνέχεια , για την έναρξη της εγκατάστασης του προγράμματος θα πρέπει να επιλεχθεί το κουμπί με τίτλο «*Install» όπως φαίνεται στην εικόνα 5.*

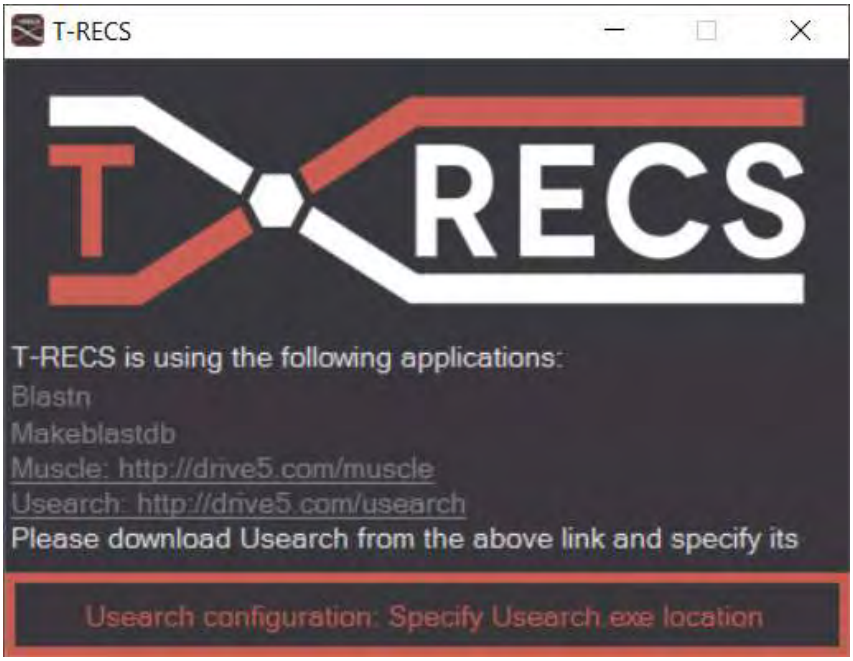


Εικόνα 5: Για την έναρξη της εγκατάστασης του T-RECs θα πρέπει να επιλεχθεί το «Install».

2) Προγράμματα Anti-virus όπως είναι το Avast, ενδεχομένως να θεωρήσουν πως το T-RECs αποτελεί πιθανό κίνδυνο για τον υπολογιστή του χρήστη και να προβούν σε εκτεταμένο έλεγχο του προγράμματος αποτρέποντας ταυτόχρονα την εκτέλεσή του από τον χρήστη. Σε αυτό το σημείο θα πρέπει να αναφερθεί πως άλλα προγράμματα προστασίας από κακόβουλο λογισμικό όπως είναι το AVG ή το Windows Defender το οποίο είναι ενσωματωμένο στο λειτουργικό σύστημα Windows, δεν εντόπισαν κάποιου είδους απειλή.

3.2 Πρώτη εκτέλεση - εγκατάσταση του UCLUST

Όπως έχει ήδη αναφερθεί, για τη σωστή λειτουργία του T-RECs, είναι απαραίτητη η λήψη και η ενσωμάτωση του προγράμματος UCLUST που ανήκει στη βιβλιοθήκη προγραμμάτων του USEARCH. Κατά την πρώτη εκτέλεση του T-RECs, ο χρήστης ενημερώνεται για την ανάγκη της εγκατάστασης του UCLUST και του παρέχεται ένας ηλεκτρονικός σύνδεσμος μέσω του οποίου μπορεί να πραγματοποιήσει τη λήψη του (*http://drive5.com/usearch/*) όπως φαίνεται στην εικόνα 6. Θα πρέπει να σημειωθεί πως για τη λήψη του προγράμματος Usearch, ο χρήστης θα πρέπει να δηλώσει διεύθυνση ηλεκτρονικού ταχυδρομείου (e-mail). Ύστερα από τη λήψη του προγράμματος, αυτό θα πρέπει να μετονομαστεί (για λόγους ασφαλείας και συνεργασίας με το T-RECs) σε «usearch.exe».



Εικόνα 6: Για την λήψη του Usearch θα πρέπει να επιλεγεί ο σύνδεσμος http://drive5.com/usearch και έπειτα να ορισθεί από το χρήστη η θέση όπου έγινε η λήψη επιλέγοντας το κουμπί «Usearch configuration: Specify Usearch.exe location».

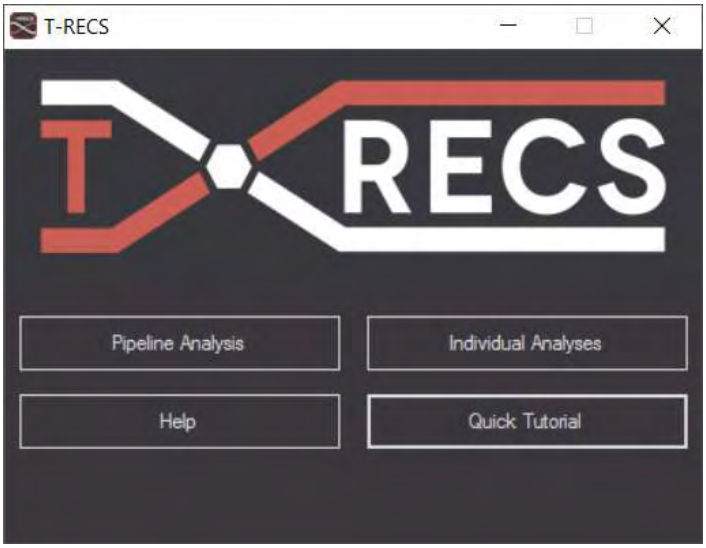
Κατόπιν, θα πρέπει να επιλεγεί το κουμπί «Usearch configuration: Specify Usearch.exe location» μέσω του οποίου ο χρήστης θα μπορέσει να υποδείξει το αρχείο Usearch.exe που έλαβε από τον παραπάνω σύνδεσμο. Μόλις ολοκληρωθεί και αυτό το βήμα, το T-RECs θα είναι πλέον έτοιμο να λειτουργήσει.

3.3 Αρχική φόρμα του T-RECs

Η αρχική φόρμα του T-RECs όπως φαίνεται στην εικόνα 7, παρέχει τέσσερις επιλογές:

- Pipeline Analysis
- Individual Analyses
- Help
- Quick Tutorial

Συνοπτικά, η πρώτη επιλογή αφορά μια αυτοματοποιημένη ανάλυση που συνδυάζει τις δυνατότητες του T-RECs με απλά βήματα, η δεύτερη αποτελεί ένα εργαλείο μέσα από το οποίο μπορούν να γίνουν επιμέρους αναλύσεις με λεπτομερείς ρυθμίσεις ως προς τις παραμέτρους που θα χρησιμοποιηθούν. Η επιλογή «Help» οργανώνει ολόκληρη τη λειτουργία του T-RECs σε θεματικές ενότητες, περιγράφοντας κάθε επιλογή που έχει ο χρήστης αλλά και κάθε στοιχείο του γραφικού περιβάλλοντος που προσφέρει το πρόγραμμα σε κάθε σημείο της ανάλυσης. Τέλος, η επιλογή «Quick Tutorial» αποτελεί ένα μικρό οδηγό εκμάθησης της βασικής ανάλυσης που προσφέρει το T-RECs με έμφαση σε μικρές λεπτομέρειες που μπορεί να φανούν χρήσιμες σε αναλύσεις του είδους αυτού.



Εικόνα 7: Η αρχική φόρμα του T-RECs προσφέρει 2 επιλογές για ανάλυση, 1 επιλογή για γρήγορη εκμάθηση της χρήσης της βασικής μεθόδου ανάλυσης του T-RECs και 1 επιλογή βοήθειας η οποία παρέχει πληροφορίες για κάθε επιλογή του προγράμματος.

3.4 Ο μηχανισμός εντοπισμού πιθανών γεγονότων ανασυνδυασμού του T-RECs

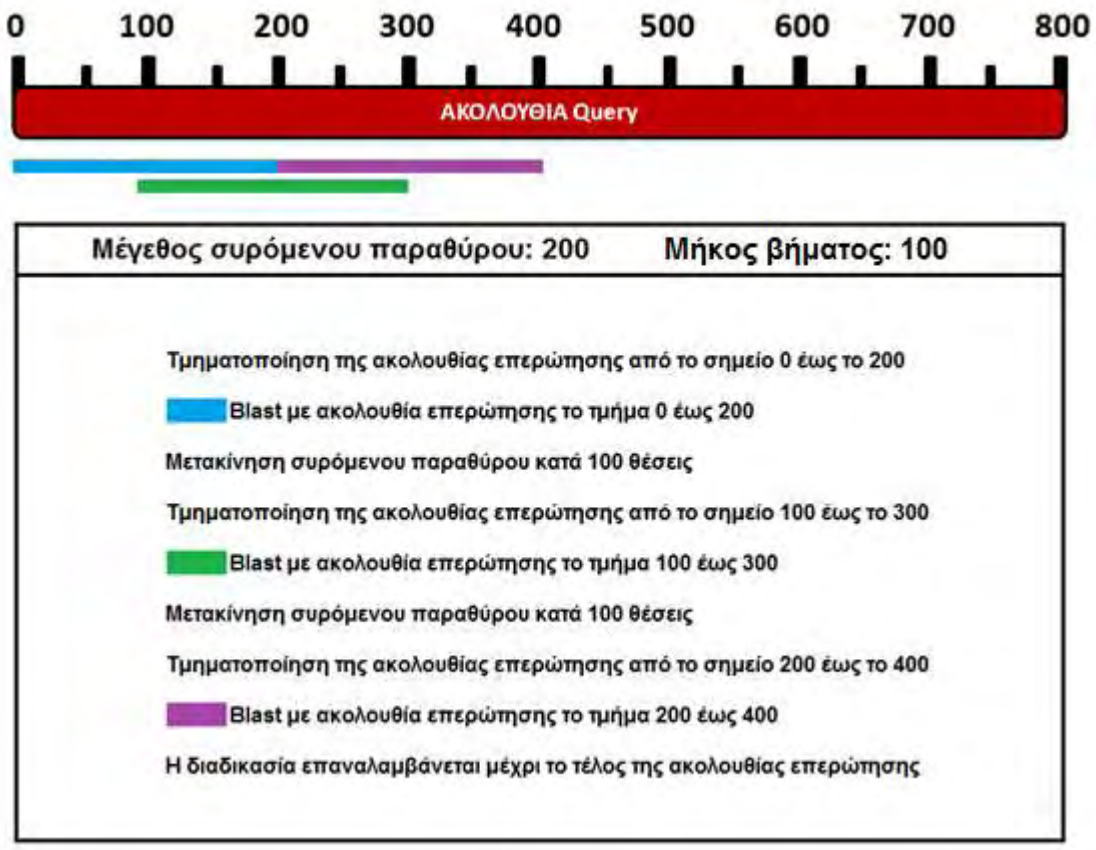
Πριν την εκτενή ανάλυση των επιλογών του T-RECs θα πρέπει να αναφερθεί και να αναλυθεί ο μηχανισμός με τον οποίο το πρόγραμμα είναι σε θέση να αναζητά και να εντοπίζει πιθανά γεγονότα ανασυνδυασμού. Το T-RECs μπορεί να πραγματοποιεί Blast τοπικά στον υπολογιστή ενώ ο χρήστης θα πρέπει να παρέχει ο ίδιος τη βάση δεδομένων καθώς και την ακολουθία (ή ακολουθίες) επερώτησης. Τόσο οι ακολουθίες της βάσης δεδομένων όσο και η ακολουθία/ες επερώτησης θα πρέπει να βρίσκονται σε δύο αρχεία (database _file και query _file αντίστοιχα) σε μορφή FASTA και η μία κάτω από την άλλη με την ακόλουθη δομή:

>accession_number1
Αλληλουχία1

>accession_number2
Αλληλουχία2

Εκτενέστερη περιγραφή για τα αρχεία εισόδου αλλά και εξόδου του T-RECs γίνεται στο Παράρτημα 1.

Βασική λειτουργία του T-RECs αποτελεί ο έλεγχος για ανασυνδυασμούς μεταξύ μιας ακολουθίας επερώτησης και μιας ακολουθίας από αυτές της βάσης δεδομένων. Ο έλεγχος αυτός πραγματοποιείται με τμηματικό Blast της ακολουθίας επερώτησης: η ακολουθία αυτή «κόβεται» από το T-RECs σε μικρότερα τμήματα, των οποίων το μέγεθος μπορεί να καθοριστεί από τον χρήστη και στη συνέχεια με κάθε ένα από αυτά τα τμήματα γίνεται Blast επάνω στις ακολουθίες της βάσης δεδομένων. Επίσης το πρόγραμμα είναι ικανό να χωρίζει την ακολουθία επερώτησης σε μικρότερα τμήματα συγκεκριμένου μεγέθους βάσει ενός συρόμενου παραθύρου. Η βασική αυτή λειτουργία παρουσιάζεται σχηματικά στην εικόνα 8.



Εικόνα 8: Συνοπτική παρουσίασης τρόπου λειτουργίας του T-RECs. Στο συγκεκριμένο παράδειγμα το μέγεθος των επιμέρους τμημάτων που προκύπτουν από τον «τεμαχισμό» της ακολουθίας επερώτησης (μέγεθος συρόμενου παραθύρου) είναι 200 ενώ το μήκος βήματος είναι 100 θέσεις.

Θα πρέπει να σημειωθεί ότι σε πολλές περιπτώσεις ίσως υπάρξει πρόβλημα με το τελευταίο τμήμα που θα πρέπει να αποσπαστεί από την ακολουθία επερώτησης ώστε να γίνει το Blast με αυτό. Το πρόβλημα είναι ότι το μέγεθος του τμήματος αυτού ενδεχομένως να είναι αρκετά μικρότερο από το μέγεθος των τμημάτων που έχει οριστεί από τον χρήστη. Στην περίπτωση αυτή ο αλγόριθμος παίρνει ένα κομμάτι της ακολουθίας με το προκαθορισμένο μέγεθος ξεκινώντας όμως από το τέλος της.

Αφού πραγματοποιηθούν τα Blast για όλα τα τμήματα κάθε μιας από τις ακολουθίες που βρίσκονται στο αρχείο με τις ακολουθίες επερώτησης ξεκινά μια σειρά από φιλτραρίσματα των αποτελεσμάτων του Blast ώστε να αποκαλυφθούν τυχόν γεγονότα ανασυνδυασμού. Αρχικά από όλα τα αποτελέσματα αφαιρείται η πρώτη γραμμή καθώς αυτή αντιπροσωπεύει την ομολογία μεταξύ της ίδιας της ακολουθίας επερώτησης η οποία συμπεριλαμβάνεται στη βάση δεδομένων. Η ουσιαστική πορεία του φιλτραρίσματος είναι η ακόλουθη:

Ως αποτέλεσμα πιθανού ανασυνδυασμού θεωρείται το αποτέλεσμα του Blast που ικανοποιεί τις παρακάτω συνθήκες:

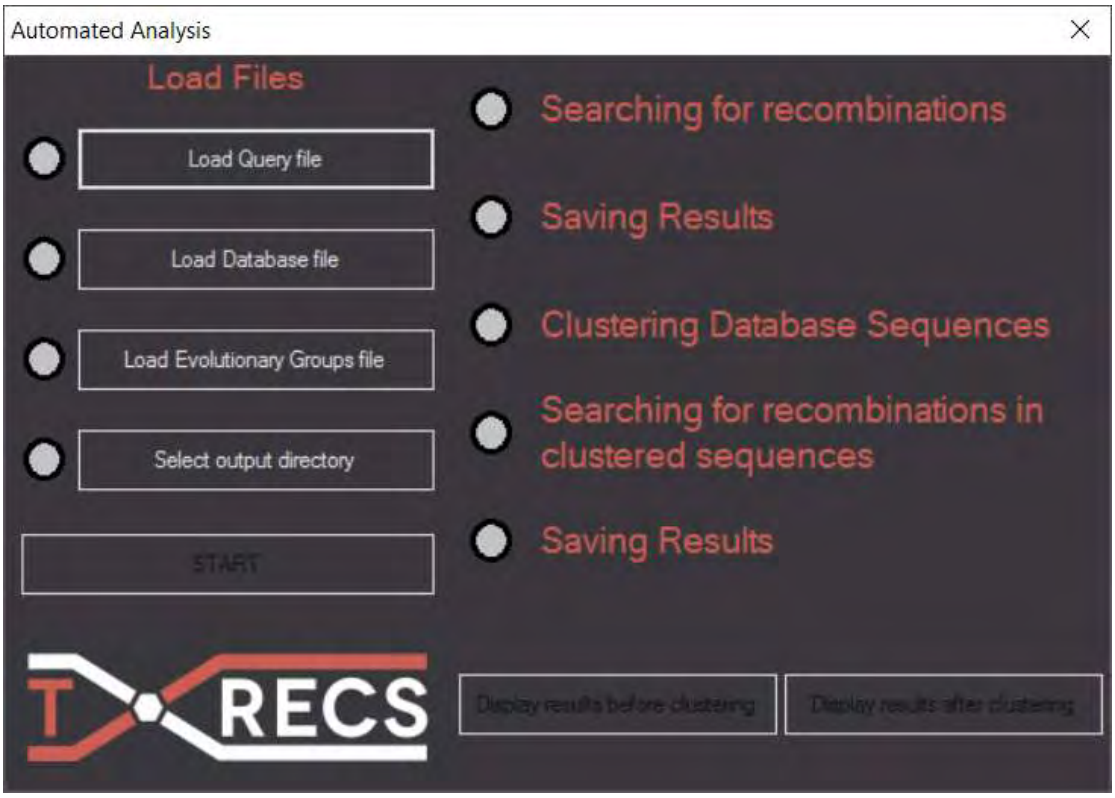
- Το ελεγχόμενο αποτέλεσμα θα πρέπει να έχει το μεγαλύτερο bitscore
- Το όνομα της ομάδας στην οποία ανήκει το ελεγχόμενο αποτέλεσμα θα πρέπει να είναι διαφορετικό από αυτό της ακολουθίας επερώτησης
- Η διαφορά του ποσοστού ταύτισης μεταξύ του ελεγχόμενου αποτελέσματος και του αμέσως καλύτερου αποτελέσματος που έχει ίδιο όνομα ομάδας με αυτό της ακολουθίας επερώτησης να ξεπερνάει ένα ποσοστό που ορίζεται από τον χρήστη

Με τον τρόπο αυτό εξασφαλίζονται οι καλύτερες δυνατές προϋποθέσεις για την ύπαρξη τυχόν περιπτώσεων ανασυνδυασμού.

Για τη σωστή λειτουργία του βασικού μηχανισμού εντοπισμού ανασυνδυασμών του T-RECs είναι αναγκαία η ύπαρξη του αρχείου απλού κειμένου το οποίο θα περιέχει μία στήλη με τους accession numbers κάθε ακολουθίας που χρησιμοποιείται τόσο στο αρχείο με τις ακολουθίες επερώτησης όσο και σε αυτό με τις ακολουθίες της βάσης δεδομένων, καθώς και μία στήλη με τους οροτύπους που αντιστοιχούν σε καθέναν από αυτούς.

3.5 Pipeline Analysis

Η επιλογή αυτή δίνει στο χρήστη τη δυνατότητα να εκτελέσει δύο ειδών σαρώσεις για πιθανά γεγονότα ανασυνδυασμού, αυτόματα και μέσω ελαχίστων βημάτων χρησιμοποιώντας τις προκαθορισμένες τιμές παραμέτρων του T-RECs. Η πρώτη σάρωση που λαμβάνει χώρα αναζητά ανασυνδυασμούς ανάμεσα σε ακολουθίες επερώτησης και ακολουθίες της βάσης δεδομένων. Έπειτα οι ακολουθίες που περιέχονται στο αρχείο της βάσης δεδομένων ομαδοποιούνται και κρατούνται μόνο εκείνες οι ακολουθίες οι οποίες αποτελούν τον αντιπρόσωπο της κάθε ομάδας (cluster). Η ομαδοποίηση γίνεται με τη βοήθεια του προγράμματος UCLUST. Η φόρμα της «Pipeline Analysis» φαίνεται στην εικόνα 9.



Εικόνα 9: Η κεντρική φόρμα της «Pipeline Analysis».

Στο αριστερό μέρος της εικόνας 9 βρίσκονται οι τέσσερις παράμετροι που θα πρέπει να οριστούν από τον χρήστη:

***Load Query file**:* Ο χρήστης θα πρέπει, μετά το πάτημα του σχετικού κουμπιού, να επιλέξει το αρχείο με τις ακολουθίες επερώτησης.

***Load Database file**:* Ο χρήστης θα πρέπει, μετά το πάτημα του σχετικού κουμπιού, να επιλέξει το αρχείο με τις ακολουθίες της βάσης δεδομένων.

***Load Evolutionary Groups file**:* Ο χρήστης θα πρέπει, μετά το πάτημα του σχετικού κουμπιού, να επιλέξει το αρχείο με τις αντιστοιχίσεις του κάθε accession number με το αντίστοιχο όνομα της ομάδας στην οποία ανήκει η εκάστοτε ακολουθία.

Η δομή που θα πρέπει να έχει το κάθε ένα από τα παραπάνω αρχεία περιγράφεται στο Παράρτημα 1.

***Select output directory**:* Η καθεμιά από τις δύο σαρώσεις θα παράγει από ένα αρχείο κειμένου που θα περιέχει τα αποτελέσματά της. Τα αρχεία αυτά αποθηκεύονται και μπορούν να φορτωθούν οποιαδήποτε στιγμή στο T-RECs χωρίς να πρέπει να εκτελεστούν εκ νέου οι δύο σαρώσεις. Για το λόγο αυτό, το πάτημα

του σχετικού κουμπιού ορίζει το σημείο στο οποίο θα αποθηκευτούν τα αρχεία αυτά.

Όταν οι παραπάνω παράμετροι οριστούν από τον χρήστη, το κουμπί «START» θα ενεργοποιηθεί και ο χρήστης θα είναι σε θέση να το επιλέξει ώστε να ξεκινήσει η ανάλυση. Από τη στιγμή που θα ξεκινήσει να τρέχει η ανάλυση, η δεξιά πλευρά της φόρμας της «Pipeline Analysis» θα είναι υπεύθυνη για την ενημέρωση του χρήστη όσον αφορά το στάδιο στο οποίο βρίσκεται κάθε στιγμή η ανάλυση:

Searching for recombinations*:* Αναζήτηση ανασυνδυασμών μεταξύ των ακολουθιών του αρχείου ακολουθιών επερώτησης και αυτών του αρχείου της βάσης δεδομένων.

Saving Results*:* Η πρώτη σάρωση ολοκληρώθηκε και αποθηκεύονται τα αποτελέσματά της.

Clustering Database Sequences*:* Γίνεται ομαδοποίηση των ακολουθιών της βάσης δεδομένων και δημιουργείται ένα νέο, προσωρινό, αρχείο βάσης δεδομένων που περιέχει μόνο τις ακολουθίες των αντιπροσώπων της κάθε ομάδας.

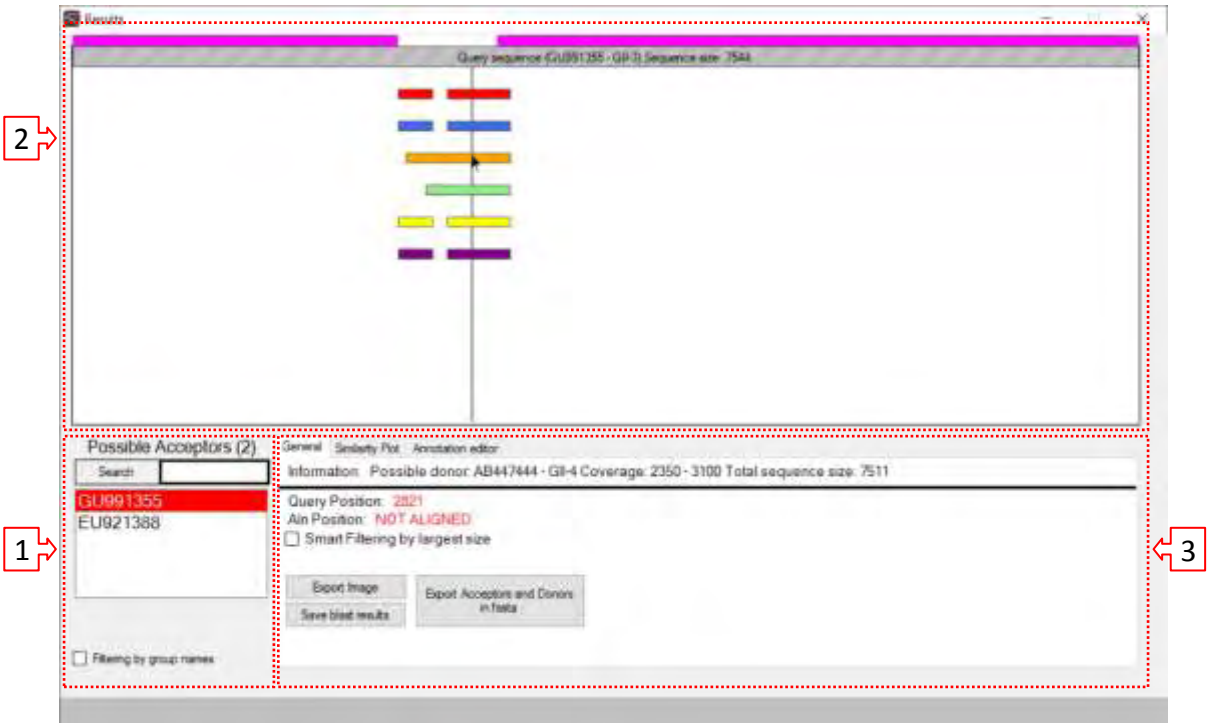
Searching for recombinations in clustered sequences*: Α*ναζήτηση ανασυνδυασμών μεταξύ των ακολουθιών του αρχείου ακολουθιών επερώτησης και αυτών του νέου αρχείου της βάσης δεδομένων το οποίο περιέχει μόνο τους αντιπροσώπους κάθε ομάδας που προέκυψε ύστερα από την ομαδοποίηση.

Saving Results*:* Η δεύτερη σάρωση ολοκληρώθηκε και αποθηκεύονται τα αποτελέσματά της.

Μετά την ολοκλήρωση και του τελευταίου βήματος θα ενεργοποιηθούν τα δύο κουμπιά στο δεξιό κάτω μέρος της φόρμας της «Pipeline Analysis» με τίτλο «Display results before clustering» και «Display results after clustering» αντίστοιχα. Πατώντας ένα από τα δύο αυτά κουμπιά εμφανίζεται η φόρμα αποτελεσμάτων που απεικονίζει γραφικά τα αποτελέσματα ύστερα από την επιλεγμένη ανάλυση.

3.6 Φόρμα αποτελεσμάτων

Η φόρμα αποτελεσμάτων (Results form) αποτελεί το εργαλείο με τη βοήθεια του οποίου ο χρήστης είναι σε θέση να προβάλλει γραφικά αλλά και να επεξεργαστεί τα αποτελέσματα ύστερα από την ολοκλήρωση της ανάλυσης που έχει εκτελέσει. Κατά την εμφάνιση της φόρμας αποτελεσμάτων εμφανίζεται και ένα μήνυμα που ενημερώνει τον χρήστη για τον αριθμό των ακολουθιών επερώτησης οι οποίες εντοπίστηκαν από το T-RECs ως πιθανώς ανασυνδυασμένες. Στην εικόνα 10 απεικονίζεται ένα παράδειγμα της φόρμας αποτελεσμάτων.



Εικόνα 10: Η φόρμα αποτελεσμάτων. 1) Περιοχή επιλογής γεγονότος. 2) Περιοχή απεικόνισης. 3) Περιοχή επεξεργασίας.

Για την ευκολότερη κατανόηση της φόρμας αποτελεσμάτων η εικόνα 10 χωρίζεται σε 3 μέρη:

1) Την Περιοχή επιλογής γεγονότος
2) Την Περιοχή απεικόνισης
3) Την περιοχή επεξεργασίας

Περιοχή επιλογής γεγονότος: Βρίσκεται στο κάτω αριστερό μέρος της φόρμας αποτελεσμάτων. Στην περιοχή αυτή βρίσκεται μια λίστα που περιέχει τους accession numbers των ακολουθιών επερώτησης που έχουν εντοπιστεί ως πιθανά γεγονότα ανασυνδυασμού. Στην κορυφή της περιοχής επιλογής γεγονότος βρίσκεται η ένδειξη «Possible Acceptors (X)» η οποία ενημερώνει τον χρήστη για τον αριθμό των ακολουθιών που περιέχονται στη λίστα. Στην περίπτωση αυτού του παραδείγματος ο αριθμός X είναι 2 καθώς το T-RECs εντόπισε δύο ακολουθίες επερώτησης που ενδεχομένως να αποτελούν δέκτη κάποιου ανασυνδυασμού. Ακόμη, παρέχεται η δυνατότητα αναζήτησης κάποιας ακολουθίας μέσα στη λίστα αυτή γράφοντας τον accession number της και πατώντας το κουμπί «Search». Τέλος, υπάρχει η δυνατότητα φιλτραρίσματος αποτελεσμάτων με βάση το όνομα της ομάδας στην οποία ανήκουν που θα έχει σαν αποτέλεσμα την εμφάνιση ακολουθιών συγκεκριμένης ομάδας στη λίστα. Η επιλογή αυτή δίνεται μέσω του

«Filtering by group names» το οποίο αν ενεργοποιηθεί από το χρήστη θα εμφανιστεί η λίστα φιλτραρίσματος όπως φαίνεται στην εικόνα 11.

Εικόνα 11: Λίστα φιλτραρίσματος με βάση το όνομα των διαθέσιμων ομάδων.

Επιλέγοντας μία από τις διαθέσιμες επιλογές της λίστας φιλτραρίσματος και πατώντας το κουμπί «Select group name», η λίστα που περιέχει τους accession numbers των ακολουθιών επερώτησης που έχουν ανιχνευθεί ως πιθανά γεγονότα ανασυνδυασμού, ενημερώνεται με τέτοιο τρόπο ώστε να περιέχει τους accession numbers μόνο από τις ακολουθίες που ανήκουν στην ομάδα που επιλέχθηκε από τον χρήστη στη λίστα φιλτραρίσματος. Για την επαναφορά της λίστας γεγονότων στην αρχική της κατάσταση αρκεί να απενεργοποιηθεί η επιλογή «Filtering by group names».

Περιοχή απεικόνισης: Βρίσκεται στο πάνω μέρος της φόρμας αποτελεσμάτων. Στην περιοχή αυτή εμφανίζονται τρία βασικά στοιχεία: Η ακολουθία επερώτησης που είναι επιλεγμένη στη λίστα γεγονότων της περιοχής επιλογής γεγονότος και εμφανίζεται με ένα μοτίβο γκρι χρώματος στο πάνω μέρος της περιοχής, η ακολουθία του πρωτεύοντα δότη (major donor) με έντονο ροζ χρώμα επίσης στο πάνω μέρος της περιοχής και τέλος, ο/οι δευτερεύων/οντες δότης/ες (minor donors) με διαφορετικά χρώματα ο καθένας. Σημειώνεται πως τα χρώματα των minor donors μπορούν να αλλαχθούν από τον χρήστη κάνοντας δεξί κλικ πάνω σε οποιοδήποτε κομμάτι του επιθυμητού minor donor. Η ακολουθία του δέκτη εμφανίζεται πάντα σε όλο το μήκος της περιοχής απεικόνισης ενώ οι ακολουθίες των major και minor donors μπορεί να εμφανιστούν σε κομμάτια κατά μήκος της ακολουθίας του δέκτη. Αυτό οφείλεται στο φιλτράρισμα που λαμβάνει χώρα μετά την ολοκλήρωση του Blast, καθώς κρατούνται και απεικονίζονται μόνο τα τμήματα που ικανοποιούν συγκεκριμένες προϋποθέσεις όπως έχει ήδη αναφερθεί.

Περιοχή επεξεργασίας: Βρίσκεται στο κάτω δεξιό μέρος της φόρμας αποτελεσμάτων. Η περιοχή αυτή περιλαμβάνει τρεις καρτέλες (General, Similarity Plot, Annotation editor) μέσω των οποίων μπορεί να γίνει περαιτέρω επεξεργασία των αποτελεσμάτων από τον χρήστη. Οι τρεις καρτέλες που περιλαμβάνονται στην περιοχή επεξεργασίας αναλύονται στις ακόλουθες ενότητες.

3.7 Καρτέλα γενικών επιλογών (General Tab)

Information bar

Η καρτέλα αυτή περιλαμβάνει στο πάνω μέρος της μία γραμμή πληροφοριών. Τοποθετώντας τον δείκτη του ποντικιού επάνω σε οποιαδήποτε ακολουθία της περιοχής απεικόνισης, η γραμμή πληροφοριών εμφανίζει τα στοιχεία που αφορούν την εκάστοτε ακολουθία. Στο παράδειγμα της Εικόνας 10 ο δείκτης του ποντικιού βρίσκεται πάνω στην ακολουθία που αναπαρίσταται με πορτοκαλί χρώμα. Η γραμμή πληροφοριών εμφανίζει τα εξής στοιχεία για την συγκεκριμένη ακολουθία: «Possible Donor: AB447444 - GII-4 Coverage: 2350 - 3100 Total sequence size: 7511». Αρχικά δίνεται ο accession number της επιλεγμένης ακολουθίας (AB447444) και στη συνέχεια η εξελικτική ομάδα στην οποία ανήκει (GII-4). Έπειτα ακολουθεί η περιοχή που καλύπτει το επιλεγμένο κομμάτι στην ακολουθία επερώτησης (από 2350 έως 3100) και τέλος δίνεται η πληροφορία σχετικά με το μέγεθος ολόκληρης της ακολουθίας (7411) που αντιπροσωπεύεται από το επιλεγμένο κομμάτι.

Query position / Aln Position

Κάτω από τη γραμμή πληροφοριών βρίσκεται η γραμμή θέσης σύμφωνα με την ακολουθία επερώτησης (Query Position) η οποία κατά τη μετακίνηση του δείκτη του ποντικιού στην περιοχή απεικόνισης ενημερώνει τον χρήστη για το σημείο στο οποίο βρίσκεται πάνω στην ακολουθία επερώτησης. Ακριβώς κάτω από τη γραμμή αυτή βρίσκεται μία ακόμη γραμμή που δείχνει τη θέση της επιλεγμένης ακολουθίας με βάση τη στοίχιση. Η γραμμή ονομάζεται «Aln Position» (Alignment Position) και για να δείξει κάποια τιμή θα πρέπει πρώτα να γίνει στοίχιση των ακολουθιών όπως περιγράφεται σε επόμενη ενότητα.

Smart filtering by largest size

Στη συνέχεια παρατηρείται η επιλογή «Smart filtering by largest size». Πρόκειται για μια επιλογή η οποία φιλτράρει τα δεδομένα της περιοχής απεικόνισης βάσει του μεγέθους τους. Έτσι, όταν η επιλογή είναι ενεργοποιημένη, το πρόγραμμα θα αναζητήσει τα κομμάτια των ακολουθιών τα οποία έχουν επικάλυψη και θα κρατήσει εκείνο που έχει το μεγαλύτερο μέγεθος. Απενεργοποιώντας την επιλογή αυτή, τα περιεχόμενα της περιοχής απεικόνισης θα επανέλθουν στην αρχική τους μορφή.

Export image

Το T-RECs δίνει στο χρήστη, μέσω της καρτέλας γενικών επιλογών, τη δυνατότητα να εξάγει τα περιεχόμενα της περιοχής απεικόνισης σε μορφή εικόνας. Για την επίτευξη αυτού το σκοπού αρκεί να επιλέξει το κουμπί με όνομα «Export image».

Save blast results

Η επιλογή «Save blast results» δίνει τη δυνατότητα να αποθηκευτούν τα αποτελέσματα του Blast ύστερα από την εκτέλεση μίας ανάλυσης ώστε να μπορεί ο χρήστης να τα ξαναφορτώσει στο T-RECs χωρίς να χρειαστεί να εκτελέσει και πάλι την ανάλυση. Στην περίπτωση της «Pipeline Analysis», όπως περιγράφεται παραπάνω, τα αποτελέσματα του Blast αποθηκεύονται αυτόματα τόσο για την ανάλυση πριν την ομαδοποίηση των ακολουθιών όσο και μετά από αυτήν· όμως, όπως περιγράφεται σε επόμενη ενότητα, κατά την εκτέλεση των επιμέρους αναλύσεων χωρίς την «Pipeline Analysis» τα αποτελέσματα δεν αποθηκεύονται αυτόματα επομένως η χρήση του «Save blast results» είναι απαραίτητη.

Export Acceptors and Donors in fasta

Η τελευταία επιλογή που παρέχεται στη συγκεκριμένη καρτέλα ονομάζεται «Export Acceptors and Donors in fasta». Η επιλογή αυτή δίνει τη δυνατότητα στον χρήστη να εξάγει ένα αρχείο κειμένου για κάθε γεγονός που περιέχεται στη λίστα πιθανών γεγονότων ανασυνδυασμού το οποίο θα περιέχει όλες τις ακολουθίες που περιλαμβάνει το εκάστοτε γεγονός και πρόκειται για τις ακολουθίες των: Query, Major Donor, Minor donors αλλά και αυτές που ανήκουν στην ίδια εξελικτική ομάδα με αυτή της ακολουθίας επερώτησης και έδωσαν το καλύτερο αποτέλεσμα κατά το Blast στα σημεία όπου το T-RECs εντόπισε κάποιον πιθανό ανασυνδυασμό.

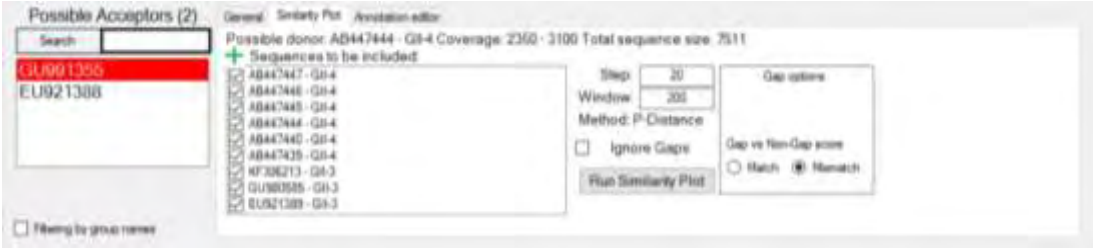
3.8 Διάγραμμα ομοιότητας (Similarity Plot)

3.8.1 Καρτέλα διαγράμματος ομοιότητας (Similarity Plot Tab)

Μέσω της καρτέλας δημιουργίας διαγραμμάτων ομοιότητας ο χρήστης έχει τη δυνατότητα να επιλέξει τις ακολουθίες που θα συμπεριληφθούν στην κατασκευή του κάθε διαγράμματος, αλλά και να επιλέξει τις παραμέτρους με τις οποίες θα γίνει η δημιουργία τους.

Information bar

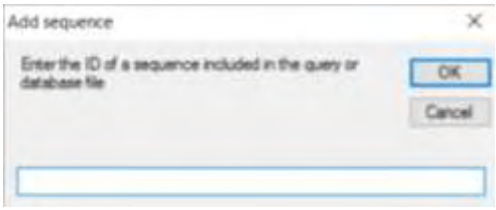
Όπως φαίνεται και στην εικόνα 12, η γραμμή πληροφοριών εξακολουθεί να υπάρχει και στην καρτέλα διαγράμματος ομοιότητας.



Εικόνα 12: Η καρτέλα επιλογών διαγράμματος ομοιότητας

Λίστα ακολουθιών (Sequences to be included list)

Βασικό στοιχείο της καρτέλας επιλογών διαγράμματος ομοιότητας αποτελεί η λίστα των ακολουθιών που θα συμπεριληφθούν στο διάγραμμα που θα δημιουργηθεί. Οι ακολουθίες που περιλαμβάνονται στη λίστα είναι αυτές των Major και Minor Donors καθώς επίσης και αυτές που ανήκουν στην ίδια εξελικτική ομάδα με αυτή της ακολουθίας επερώτησης και έδωσαν το καλύτερο αποτέλεσμα κατά το Blast στα σημεία όπου το T-RECs εντόπισε κάποιον πιθανό ανασυνδυασμό. Η ακολουθία επερώτησης περιλαμβάνεται αυτόματα και δεν περιέχεται στη λίστα. Ο χρήστης μπορεί να επιλέξει ποιες από τις ακολουθίες της λίστας θα συμπεριληφθούν κάνοντας κλικ σε αυτές. Η προκαθορισμένη επιλογή είναι να συμπεριληφθούν όλες οι ακολουθίες. Σημειώνεται ότι για να δημιουργηθεί ένα διάγραμμα ομοιότητας θα πρέπει να είναι επιλεγμένη τουλάχιστον μία ακολουθία. Ακόμη, δίνεται η δυνατότητα προσθήκης κάποιας ακολουθίας η οποία περιέχεται είτε στο αρχείο της βάσης δεδομένων είτε σε αυτό με τις ακολουθίες επερώτησης. Για να γίνει αυτό, ο χρήστης θα πρέπει να επιλέξει το κουμπί με τον πράσινο σταυρό (+) ώστε να εμφανιστεί η φόρμα που φαίνεται στην εικόνα 13. Στη φόρμα αυτή αρκεί να εισάγει ο χρήστης τον accession number της επιθυμητής ακολουθίας και να πατήσει το κουμπί «OK».



Εικόνα 13: Προσθήκη ακολουθιών στη λίστα με τις ακολουθίες που θα χρησιμοποιηθούν στη δημιουργία διαγράμματος ομοιότητας.

Παράμετροι βήματος/παραθύρου (Step/Window)

Για τη δημιουργία του διαγράμματος ομοιότητας είναι απαραίτητος ο καθορισμός του βήματος και του μεγέθους του συρόμενου παραθύρου. Οι δύο αυτές παράμετροι μπορούν να ρυθμιστούν από τον χρήστη στις επιλογές «Step» και «Window» αντίστοιχα. Οι προκαθορισμένες τιμές τους είναι 200 για το παράθυρο και 20 για το βήμα. Οι τιμές αυτές είναι υποκειμενικές και ο χρήστης θα πρέπει να είναι εκείνος που θα επιλέξει την κατάλληλη τιμή των παραμέτρων αυτών.

Η μέθοδος που χρησιμοποιείται για την κατασκευή των διαγραμμάτων ομοιότητας είναι η μέθοδος της παρατηρούμενης απόστασης (P-Distance).

Διαχείριση κενών

Πριν από τη δημιουργία του διαγράμματος ομοιότητας οι επιλεγμένες ακολουθίες θα πρέπει να έχουν στοιχηθεί. Όπως περιγράφεται σε επόμενη ενότητα το T-RECs προσφέρει τη δυνατότητα αυτή. Κατά τη δημιουργία της στοίχισης προστίθενται κενά (gaps) στις επιλεγμένες ακολουθίες ώστε να επιτευχθεί η βέλτιστη στοίχιση μεταξύ τους. Όπως φαίνεται στην εικόνα 12 παρέχονται επιλογές σχετικά με τη διαχείριση των κενών κατά τη δημιουργία του διαγράμματος ομοιότητας.

Ignore Gaps: Όταν η επιλογή αυτή είναι ενεργοποιημένη, κατά τη δημιουργία του διαγράμματος ομοιότητας όλες οι θέσεις που περιέχουν κενά αγνοούνται και δε συμπεριλαμβάνονται στη σύγκριση για τη δημιουργία του διαγράμματος ομοιότητας. Παρακάτω φαίνεται ένα παράδειγμα για τον τρόπο λειτουργίας της επιλογής αυτής:

Παράδειγμα 1: Οι δύο ακολουθίες συγκρίνονται για τη δημιουργία του διαγράμματος ομοιότητας. Με πράσινο χρώμα φαίνονται οι περιπτώσεις match ενώ με κόκκινο οι mismatch. Το μέγεθος της πάνω ακολουθίας είναι 20 nt ενώ της κάτω 10 nt (Χωρίς τα κενά). Εφόσον η επιλογή «Ignore Gaps» είναι ενεργοποιημένη, τα κενά δεν καταμετρούνται και έτσι, εικονικά, η επάνω ακολουθία συγκρίνεται με την κάτω σαν να έχει και εκείνη μέγεθος 10 nt. Έτσι η ομοιότητα των δύο ακολουθιών θεωρείται πως είναι 8/10 δηλαδή 80%.

Gap Options: Όταν η επιλογή «Ignore Gaps» δεν είναι επιλεγμένη, τα κενά συγκρίνονται κανονικά μεταξύ των δύο ακολουθιών. Για το λόγο αυτό, στο πλαίσιο «Gap Options» όπως φαίνεται στην εικόνα 12 παρέχεται μια επιλογή σχετικά με την αξιολόγηση περιπτώσεων σύγκρισης μιας θέσης που περιέχει κενό με μία θέση που δεν περιέχει. Η επιλογή αυτή ονομάζεται «Gap vs Non-Gap score» και δίνει στο

χρήστη τη δυνατότητα να επιλέξει αν μια τέτοια περίπτωση θα έχει σαν αποτέλεσμα το match ή το mismatch.

Run Similarity Plot

Αφότου έχει γίνει η ρύθμιση της κάθε παραμέτρου, για την εκκίνηση της διαδικασίας κατασκευής του διαγράμματος ομοιότητας θα πρέπει να επιλεγεί το κουμπί «Run Similarity Plot». Σε περίπτωση που υπάρχουν ακολουθίες στο αρχείο βάσης δεδομένων που έχει εισάγει ο χρήστης οι οποίες ανήκουν στην ίδια εξελικτική ομάδα με την ακολουθία επερώτησης, το T-RECs θα ενημερώσει το χρήστη για τον αριθμό των ακολουθιών αυτών μέσω του μηνύματος που φαίνεται στην εικόνα 14. Σε αυτό το σημείο ο χρήστης έχει τη δυνατότητα να επιλέξει πόσες από αυτές τις ακολουθίες επιθυμεί να συμπεριλάβει στο διάγραμμα ομοιότητας που πρόκειται να δημιουργηθεί εισάγοντας τον επιθυμητό αριθμό στο διαθέσιμο πλαίσιο κειμένου. Επίσης του δίνεται η δυνατότητα να επιλέξει το χρώμα με το οποίο θα εμφανίζονται όλες αυτές οι ακολουθίες στο διάγραμμα ομοιότητας κάνοντας κλικ στο μαύρο πλαίσιο.

Εικόνα 14: Μήνυμα ενημέρωσης του χρήστη για την ύπαρξη επιπρόσθετων ακολουθιών που ανήκουν στην ίδια εξελικτική ομάδα με αυτήν της ακολουθίας επερώτησης.

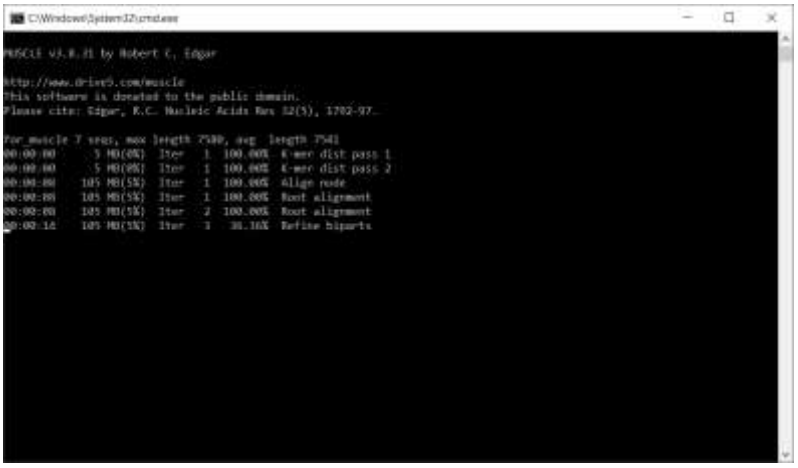
Ύστερα από την επιλογή του «OK» ένα νέο μήνυμα εμφανίζεται, όπως φαίνεται στην εικόνα 15, το οποίο ενημερώνει τον χρήστη πως οι ακολουθίες της λίστας δεν έχουν στοιχηθεί, ρωτώντας παράλληλα εάν ο χρήστης επιθυμεί τη στοίχισή τους. Όταν επιλεγεί η απάντηση «Ναι» τότε ξεκινά η στοίχιση των επιλεγμένων ακολουθιών με τη βοήθεια του προγράμματος muscle όπως φαίνεται στην εικόνα 16. Θα πρέπει να σημειωθεί πως ο χρόνος ολοκλήρωσης της διαδικασίας της στοίχισης εξαρτάται από τους εξής παράγοντες:

1) Το πλήθος των ακολουθιών που στοιχίζονται.

2) Το μέγεθος των ακολουθιών που στοιχίζονται.

3) Τα τεχνικά χαρακτηριστικά του υπολογιστή στον οποίο εκτελείται το πρόγραμμα.

Εικόνα 15: Μήνυμα ενημέρωσης του χρήστη για την ανάγκη στοίχισης των επιλεγμένων ακολουθιών.



Εικόνα 16: Το πρόγραμμα muscle στοιχίζει τις επιλεγμένες ακολουθίες για τη δημιουργία του διαγράμματος ομοιότητας.

Σημειώνεται πως μετά τη στοίχιση ενδέχεται να παρατηρηθούν αλλαγές στην περιοχή απεικόνισης της φόρμας αποτελεσμάτων (εν. 3.6) καθώς το T-RECs θα προσαρμόσει τα αποτελέσματα σύμφωνα με τη νέα στοίχιση.

3.8.2 Φόρμα διαγράμματος ομοιότητας (Similarity Plot form)

Μόλις ολοκληρωθεί η διαδικασία του προγράμματος muscle, το T-RECs θα εμφανίσει τη φόρμα που θα περιέχει το διάγραμμα ομοιότητας. Ένα παράδειγμα της φόρμας διαγράμματος ομοιότητας φαίνεται στην εικόνα 17. Η φόρμα αυτή χωρίζεται σε δύο βασικά μέρη:

1) Την περιοχή σχεδίασης (Plot Area)
2) Το υπόμνημα (Legend)

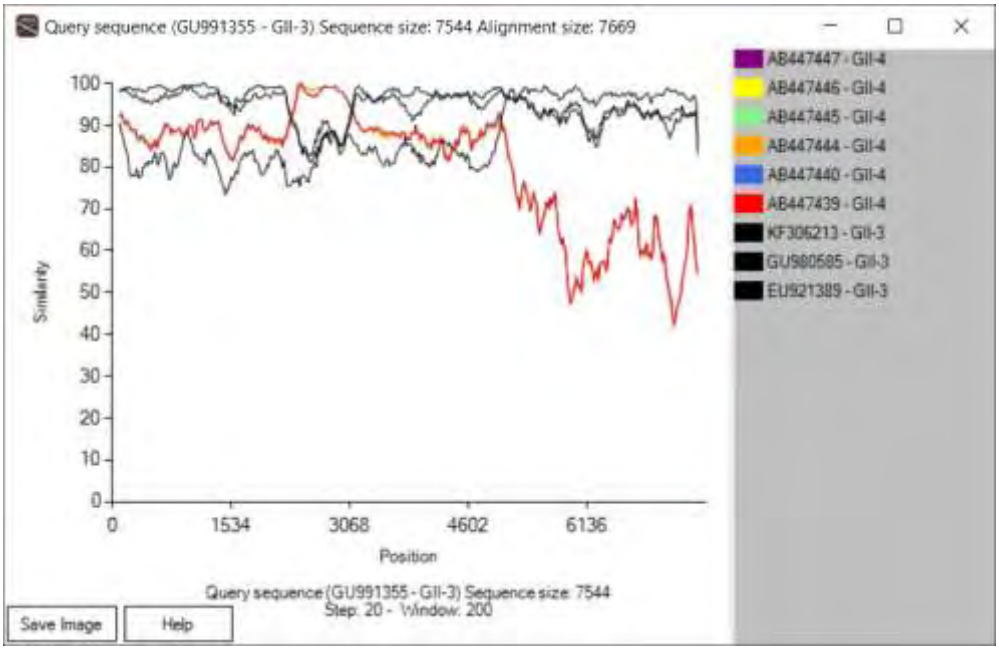
Περιοχή σχεδίασης

Πρόκειται για την περιοχή του γραφήματος όπου φαίνεται με τη μορφή γραμμών η ομοιότητα, συναρτήσει της θέσης, κατά μήκος κάθε ακολουθίας σε σχέση με την ακολουθία επερώτησης. Κάνοντας αριστερό κλικ σε οποιαδήποτε από αυτές τις γραμμές ο χρήστης μπορεί να προβάλει τον accession number αλλά και το όνομα

της εξελικτικής ομάδας στην οποία ανήκει η επιλεγμένη ακολουθία. Τόσο το υπόμνημα όσο και η περιοχή σχεδίασης αποτελούν περιοχές μεταβλητού μεγέθους. Για την αλλαγή του μεγέθους τους, ο χρήστης μπορεί να κάνει αριστερό κλικ και να μετακινήσει είτε αριστερά είτε δεξιά την κατακόρυφη γραμμή που χωρίζει τις δύο περιοχές.

Zoom in/out: Κρατώντας πατημένο το αριστερό κλικ σε οποιοδήποτε σημείο της περιοχής σχεδίασης και σύροντας προς οποιοδήποτε άλλο σημείο της περιοχής, εμφανίζεται ένα πλαίσιο εστίασης. Μόλις αφεθεί ελεύθερο το αριστερό κλικ, στο γράφημα θα εμφανίζεται μόνο η περιοχή που περικλείεται από το πλαίσιο εστίασης. Για την από-εστίαση του γραφήματος ένα βήμα πίσω ο χρήστης θα πρέπει να κάνει ένα μεσαίο κλικ με το ποντίκι.

Ακόμη, ο χρήστης έχει τη δυνατότητα να επιλέξει ένα σημείο της περιοχής σχεδίασης κάνοντας αριστερό κλικ και στη συνέχεια να προβάλει τις ακριβείς τιμές θέσης – ομοιότητας για το συγκεκριμένο σημείο κάνοντας δεξί κλικ. Σημειώνεται πως το δεξί κλικ δεν είναι απαραίτητο να γίνει ακριβώς στο ίδιο σημείο όπου έγινε το αριστερό.



Εικόνα 16: Η φόρμα διαγράμματος ομοιότητας.

Υπόμνημα

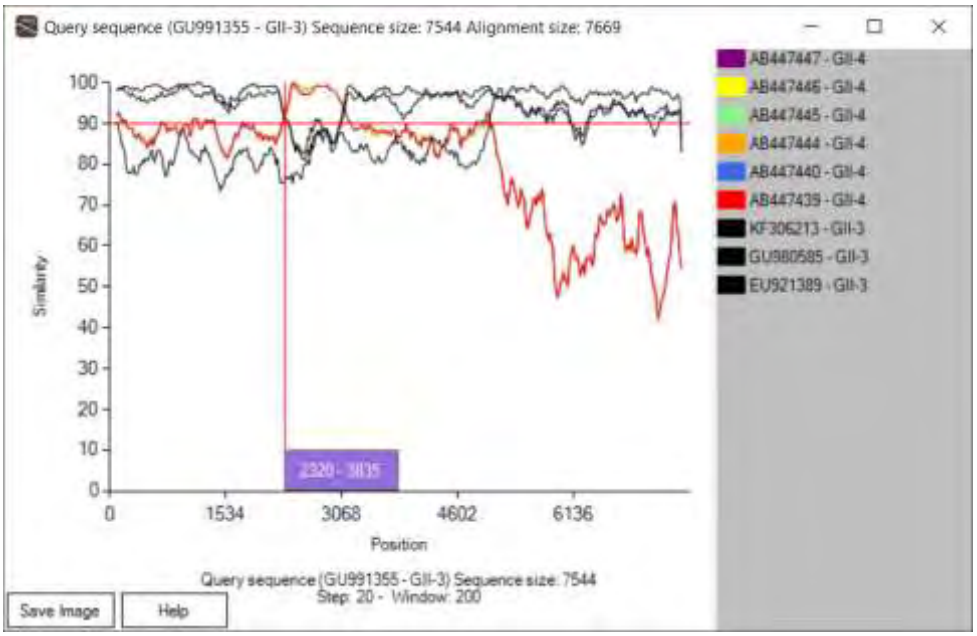
Πρόκειται για την περιοχή του γραφήματος όπου επεξηγούνται τα χρώματα των γραμμών της περιοχής σχεδίασης καθώς αντιστοιχίζονται στους accession numbers αλλά και στα ονόματα των εξελικτικών ομάδων στις οποίες ανήκουν οι ακολουθίες του γραφήματος.

Στην περιοχή του υπομνήματος ο χρήστης έχει επίσης ορισμένες επιλογές. Αρχικά, κάνοντας δεξί κλικ στο πλαίσιο χρώματος κάποιας ακολουθίας του δίνεται η δυνατότητα να αλλάξει το χρώμα της τόσο στο υπόμνημα όσο και στην περιοχή σχεδίασης αλλά και στην περιοχή απεικόνισης της φόρμας αποτελεσμάτων (εν. 3.6). Ακόμη, κάνοντας αριστερό κλικ στο πλαίσιο χρώματος κάποιας ακολουθίας, ο χρήστης είναι σε θέση να αποκρύψει μια συγκεκριμένη ακολουθία από την περιοχή σχεδίασης του γραφήματος. Η δυνατότητα αυτή μπορεί να φανεί αρκετά χρήσιμη σε περιπτώσεις όπου στο γράφημα υπάρχουν πολλές ακολουθίες.

Blast στους server του NCBI μέσω του similarity plot

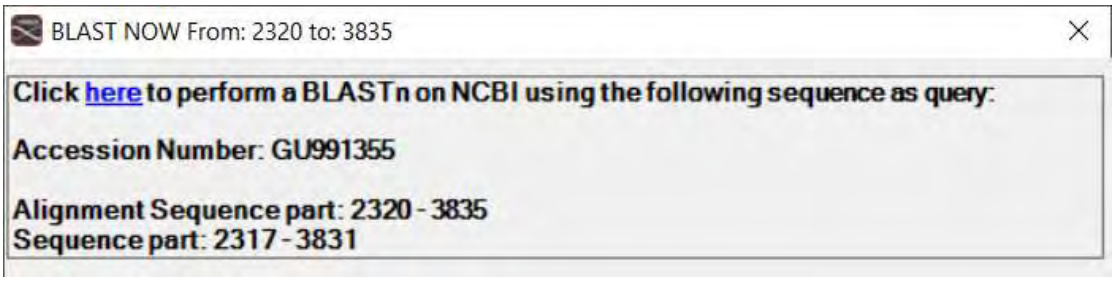
Μέσω του διαγράμματος ομοιότητας το T-RECs δίνει τη δυνατότητα στο χρήστη να επιλέξει ένα κομμάτι της ακολουθίας επερώτησης και να εκτελέσει το Blastn στους servers του NCBI χρησιμοποιώντας ως ακολουθία επερώτησης το κομμάτι αυτό. Για την επίτευξη του στόχου αυτού ο χρήστης θα πρέπει να ακολουθήσει τα παρακάτω βήματα:

- Κράτημα του πλήκτρου Alt και αριστερό κλικ στην περιοχή σχεδίασης ώστε να επιλεγεί το σημείο έναρξης του επιθυμητού τμήματος.
- Μετακίνηση του δείκτη του ποντικιού προς τα δεξιά ώστε να εμφανιστεί η μοβ μπάρα που αναπαριστά το επιλεγμένο τμήμα της ακολουθίας (εικόνα 17)
- Αριστερό κλικ στο επιθυμητό σημείο λήξης του τμήματος



Εικόνα 17: Δυνατότητα επιλογής τμήματος της ακολουθίας επερώτησης που θα χρησιμοποιηθεί ως νέα ακολουθία επερώτησης για Blast στους server του NCBI μέσω του διαγράμματος ομοιότητας του T-RECs.

Μόλις ολοκληρωθούν τα τρία παραπάνω βήματα θα εμφανιστεί μία φόρμα σύνοψης του επιλεγμένου τμήματος όπως φαίνεται στην εικόνα 18.



Εικόνα 18: Σύνοψη των πληροφοριών του επιλεγμένου τμήματος για Blast στους servers του NCBI.

Όπως φαίνεται στην εικόνα 18, οι πληροφορίες που δίνονται για το επιλεγμένο κομμάτι είναι ο accession number, το κομμάτι της ακολουθίας μετά από τη στοίχιση αλλά και το κομμάτι της ακολουθίας χωρίς τη στοίχιση.

Για την εκτέλεση του Blastn στους servers του NCBI χρησιμοποιώντας ως ακολουθία επερώτησης το επιλεγμένο κομμάτι αρκεί ο χρήστης να επιλέξει το σύνδεσμο «**here**» όπως απεικονίζεται στη φόρμα πληροφοριών της εικόνας 18. Έτσι, θα γίνει

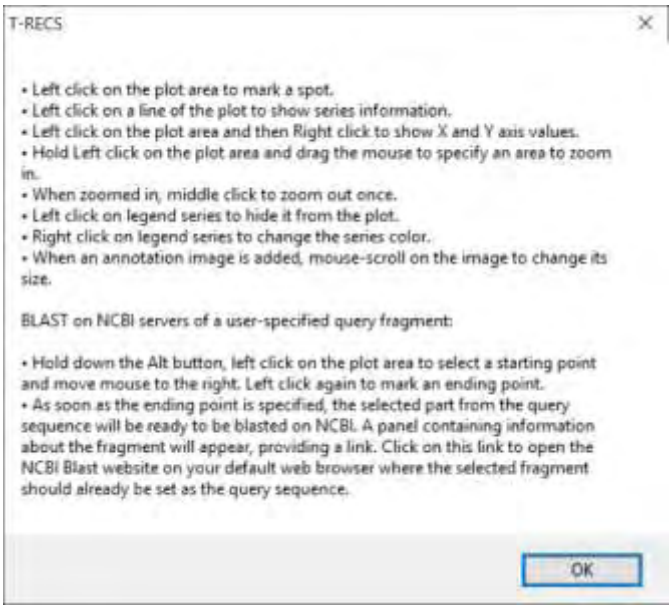
εκκίνηση του προεπιλεγμένου προγράμματος πλοήγησης στο διαδίκτυο το οποίο θα μεταβεί αυτόματα στους servers του NCBI.

Save Image

Ο χρήστης έχει τη δυνατότητα επιλέγοντας το κουμπί με τίτλο «Save Image» να αποθηκεύσει τα αποτελέσματα του διαγράμματος ομοιότητας σε μορφή εικόνας. Η αποθήκευση μπορεί να γίνει και με τη συντόμευση πληκτρολογίου Ctrl + S.

Help

Το κουμπί με τίτλο «Help» βρίσκεται εκεί ώστε να δίνει στο χρήστη τη δυνατότητα ανά πάσα στιγμή να προβάλει μια φόρμα βοήθειας όπως φαίνεται στην εικόνα 19 με κάθε δυνατή επιλογή που υπάρχει για το διάγραμμα ομοιότητας.

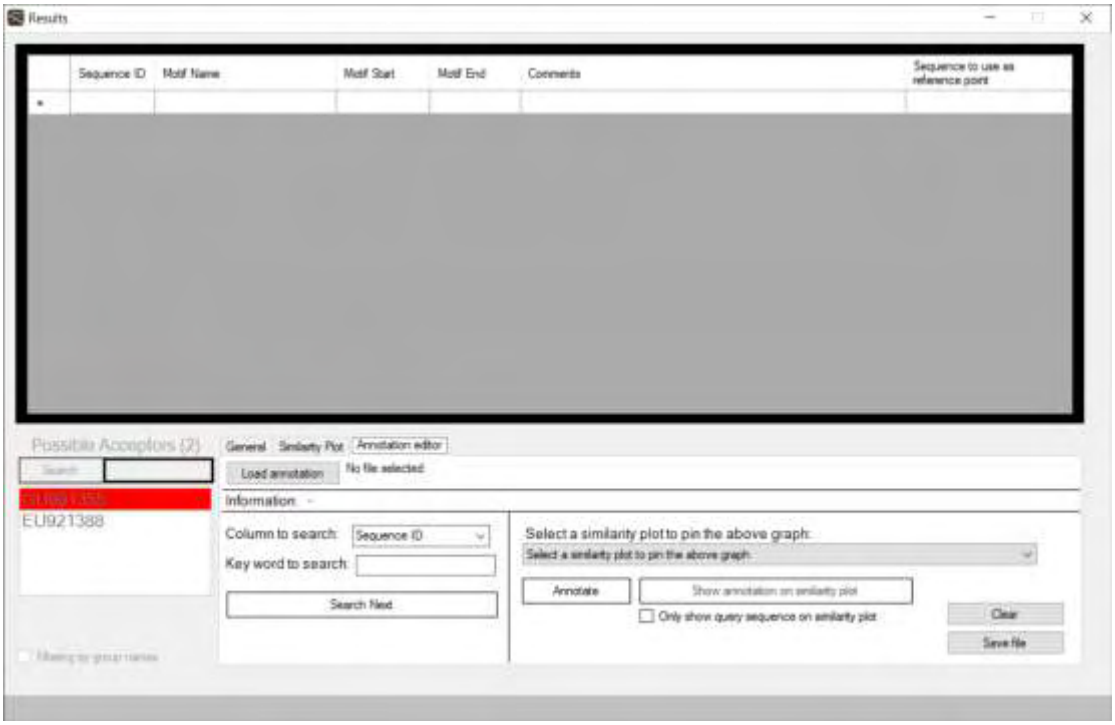


Εικόνα 19: Η φόρμα παροχής βοήθειας για τις επιλογές του χρήστη στο διάγραμμα ομοιότητας.

3.9 Καρτέλα επεξεργασίας σχολιασμού (Annotation editor tab)

Όπως φαίνεται στην εικόνα 20, κατά την επιλογή της καρτέλας επεξεργασίας σχολιασμού (Annotation editor tab), παρατηρείται αλλαγή στην εμφάνιση της περιοχής απεικόνισης της φόρμας των αποτελεσμάτων (εν. 3.6). Πλέον η περιοχή αυτή αποτελεί την περιοχή επεξεργασίας σχολιασμού. Με τον όρο «σχολιασμός» (annotation) περιγράφεται η δημιουργία ετικετών (labels) οι οποίες μπορούν να τοποθετηθούν τόσο στην περιοχή απεικόνισης όσο και στο διάγραμμα ομοιότητας. Οι ετικέτες αυτές βοηθούν στην καλύτερη οργάνωση της εικόνας των

αποτελεσμάτων καθώς μπορεί για παράδειγμα να αναπαριστούν τις περιοχές του γονιδιώματος της ακολουθίας επερώτησης. Ο χρήστης θα πρέπει να εισάγει τα χαρακτηριστικά της ετικέτας σχολιασμού (annotation label) στον πίνακα που εμφανίζεται στη θέση της περιοχής απεικόνισης της φόρμας των αποτελεσμάτων (εν. 3.6). Ο πίνακας περιλαμβάνει 6 στήλες καθεμιά από τις οποίες περιγράφεται αναλυτικά παρακάτω.



Εικόνα 20: Η καρτέλα επεξεργασίας σχολιασμού.

Sequence ID

Η στήλη αυτή θα πρέπει να περιλαμβάνει τον accession number της ακολουθίας πάνω από την οποία θα τοποθετηθεί η ετικέτα σχολιασμού. Για παράδειγμα σε περίπτωση που ο χρήστης εισάγει σε μία γραμμή της στήλης αυτής τον accession number της ακολουθίας επερώτησης τότε η ετικέτα σχολιασμού, όταν δημιουργηθεί, θα τοποθετηθεί ακριβώς επάνω από την ακολουθία επερώτησης.

Motif Name

Η στήλη «Motif Name» θα πρέπει να περιλαμβάνει το όνομα της εκάστοτε ετικέτας σχολιασμού. Σημειώνεται ότι πρόκειται για το όνομα της ετικέτας και όχι για το κείμενο που θα περιέχει. Το όνομα μιας ετικέτας σχολιασμού μπορεί να προβληθεί από τη γραμμή πληροφοριών της καρτέλας γενικών επιλογών (εν. 3.7) ή της καρτέλας δημιουργίας διαγράμματος ομοιότητας (εν. 3.8.1) μέσω μεταφοράς του δείκτη του ποντικιού πάνω στο πλαίσιο της ετικέτας.

Motif Start

Η στήλη «Motif Start» θα πρέπει να περιλαμβάνει τον αριθμό που υποδηλώνει το σημείο της ακολουθίας αναφοράς (βλ. Sequence to use as reference point) από το οποίο θα ξεκινά η ετικέτα.

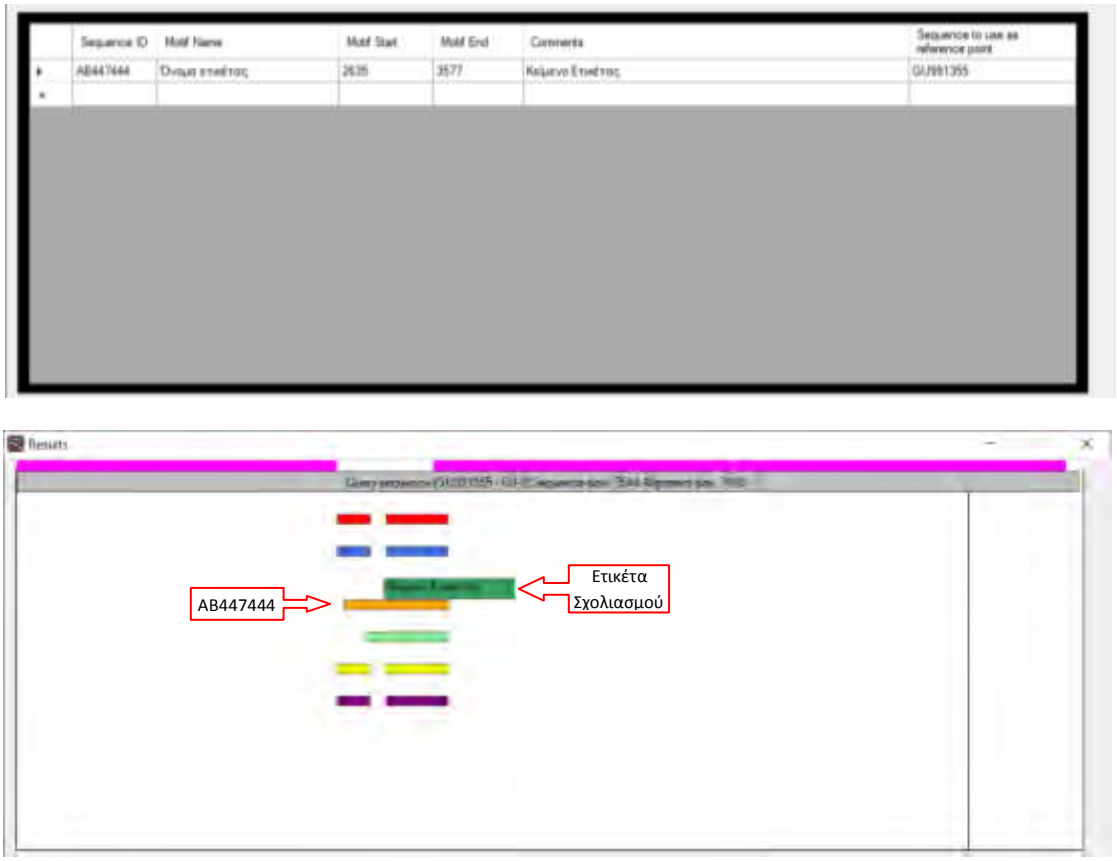
Motif end

Η στήλη «Motif end» θα πρέπει να περιλαμβάνει τον αριθμό που υποδηλώνει το σημείο της ακολουθίας αναφοράς (βλ. Sequence to use as reference point) στο οποίο θα καταλήγει η ετικέτα.

Sequence to use as reference point

Η στήλη αυτή θα πρέπει να περιέχει τον accession number της ακολουθίας που θα χρησιμοποιηθεί ως ακολουθία αναφοράς ώστε να υπολογιστεί η ακριβής θέση της ετικέτας που θα δημιουργηθεί.

Στο παράδειγμα 2 παρουσιάζεται σχηματικά η επεξήγηση της δημιουργίας μιας ετικέτας σχολιασμού.



Παράδειγμα 2: Δημιουργία ετικέτας σχολιασμού. Επάνω: εισαγωγή παραμέτρων ετικέτας. Κάτω: Η μορφή της ετικέτας στην περιοχή απεικόνισης.

Όπως φαίνεται στο παράδειγμα 2, στις παραμέτρους της ετικέτας έχουμε τις εξής τιμές:

Sequence ID: AB447444, που υποδηλώνει πως η ετικέτα θα σχεδιαστεί πάνω από την ακολουθία με accession number το AB447444. Πράγματι η ακολουθία που απεικονίζεται με πορτοκαλί χρώμα (πάνω από την οποία είναι σχεδιασμένη η ετικέτα) είναι η ακολουθία με accession number AB447444.

Motif Name: Όνομα ετικέτας, αυτό είναι το όνομα που χρησιμοποιείται από την ετικέτα.

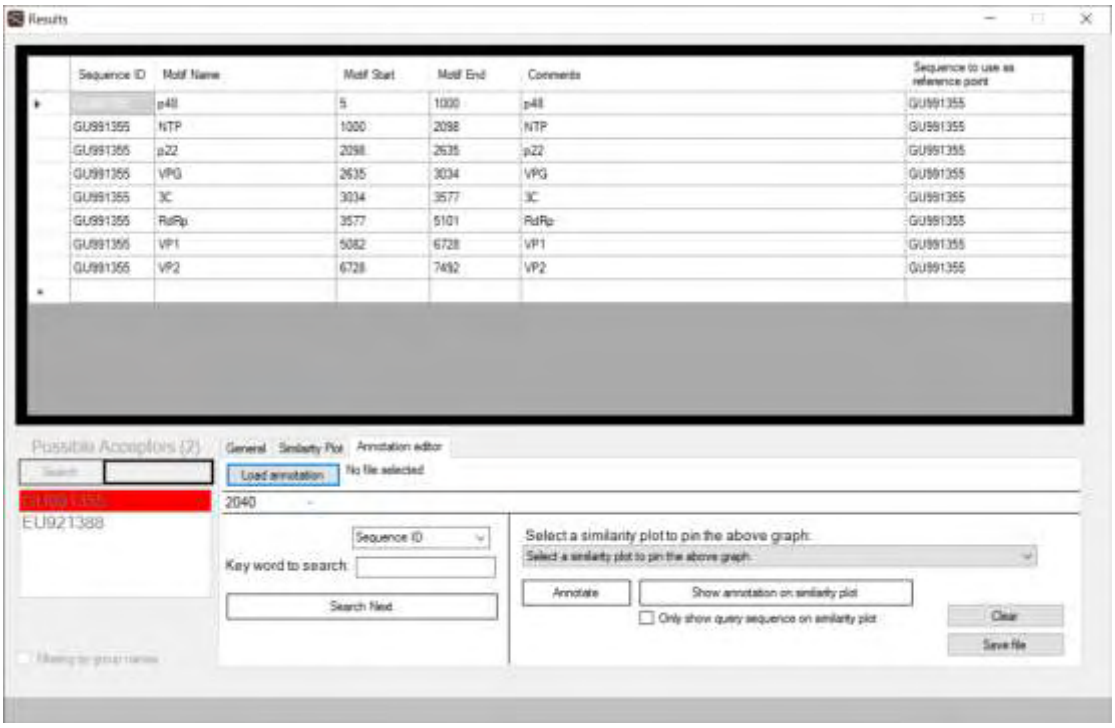
Motif Start: 2635, σημείο έναρξης.

Motif End: 3577, σημείο λήξης.

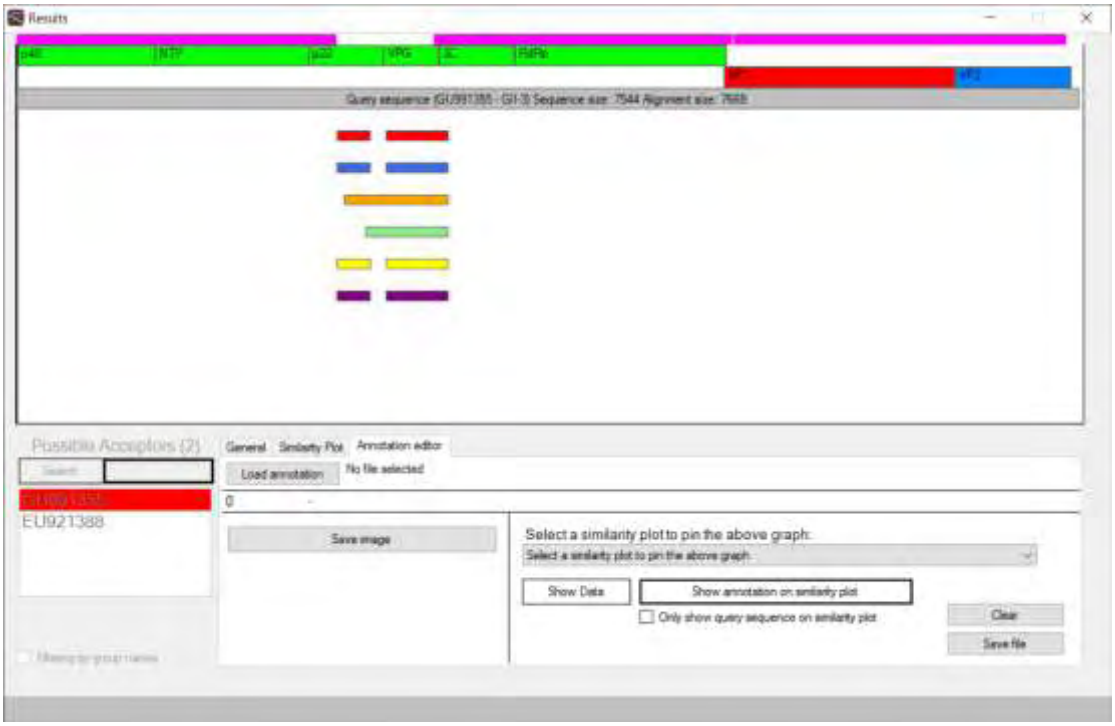
Comments: Κείμενο ετικέτας, πρόκειται για το κείμενο που εμφανίζεται πάνω στην ετικέτα που δημιουργήθηκε.

Sequence to use as reference point: GU991355, πρόκειται για τον accession number της ακολουθίας επερώτησης που υποδηλώνει ότι παρά το γεγονός πως η ετικέτα θα σχεδιαστεί πάνω από την ακολουθία με accession number AB447444, η θέση της (Motif Start / End) θα υπολογιστεί βάσει της ακολουθίας με accession number GU991355.

Στις εικόνες 21 και 22 φαίνεται μια ολοκληρωμένη δημιουργία ετικετών σχολιασμού που αναφέρεται στις περιοχές του γονιδιώματος της ακολουθίας ενός Norovirus GII-3 με accession number GU991355.



Εικόνα 21: Συμπλήρωση παραμέτρων των ετικετών σχολιασμού.



Εικόνα 22: Γραφική αναπαράσταση ετικετών σχολιασμού.

Annotate (κουμπί δημιουργίας ετικετών)

Από τη στιγμή που θα συμπληρωθεί ο πίνακας σχολιασμού από τις επιθυμητές τιμές, για να προκύψει η γραφική απεικόνιση των ετικετών, ο χρήστης θα πρέπει να επιλέξει το κουμπί με όνομα «Annotate». Τη θέση του πίνακα θα λάβει τότε και πάλι η περιοχή απεικόνισης εμφανίζοντας όλες τις ετικέτες με τον τρόπο που έχουν οριστεί στον πίνακα. Κατά το πάτημα του κουμπιού «Annotate», αυτό παίρνει το όνομα «Show Data» και αποτελεί πλέον την επιλογή επιστροφής στον πίνακα σχολιασμού.

Save file (κουμπί αποθήκευσης ετικετών)

Όταν ολοκληρωθεί η συμπλήρωση του πίνακα από τις επιθυμητές τιμές, ο χρήστης έχει τη δυνατότητα να αποθηκεύσει όλες αυτές τις τιμές ώστε να μη χρειαστεί να τις πληκτρολογήσει ξανά. Για να γίνει αυτό αρκεί να επιλέξει το κουμπί με τίτλο «Save file» που βρίσκεται στο κάτω δεξιό μέρος της καρτέλας επεξεργασίας σχολιασμού. Το αρχείο σώζεται σε μορφή απλού κειμένου σε σημείο που θα επιλέξει ο χρήστης.

Load annotation (κουμπί φορτωσης ετικετών)

Για τη φόρτωση ενός αρχείου που περιέχει ήδη τις τιμές για τη δημιουργία ετικετών, ο χρήστης θα πρέπει να επιλέξει το κουμπί με το όνομα «Load annotation» που βρίσκεται στο πάνω αριστερό μέρος της καρτέλας επεξεργασίας σχολιασμού και να υποδείξει στο πρόγραμμα τη θέση στην οποία βρίσκεται το αρχείο ώστε αυτό να εντοπιστεί και να φορτωθεί.

Clear

Το κουμπί αυτό που βρίσκεται στο κάτω δεξιό μέρος της καρτέλας επεξεργασίας σχολιασμού αναλαμβάνει να καθαρίσει τον πίνακα με τις παραμέτρους των ετικετών σχολιασμού, προετοιμάζοντάς τον για την εισαγωγή νέων τιμών.

Περιοχή αναζήτησης

Στην καρτέλα επεξεργασίας σχολιασμού και συγκεκριμένα στο αριστερό κομμάτι της παρατηρούνται τρία στοιχεία: Μια λίστα με το όνομα κάθε στήλης του πίνακα, ένα πλαίσιο κειμένου με τίτλο «Key word to search» και ένα κουμπί με τίτλο «Search Next». Πρόκειται για την πλατφόρμα αναζήτησης τιμών στα περιεχόμενα του πίνακα. Από τη λίστα θα πρέπει ο χρήστης να επιλέξει το όνομα της στήλης στην οποία επιθυμεί να πραγματοποιήσει την αναζήτηση, στη συνέχεια να εισάγει τη λέξη κλειδί που θα ήθελε να αναζητήσει και τέλος να επιλέξει το κουμπί «Search Next» ώστε να αναζητηθεί το αμέσως επόμενο κελί που περιέχει τη λέξη κλειδί.

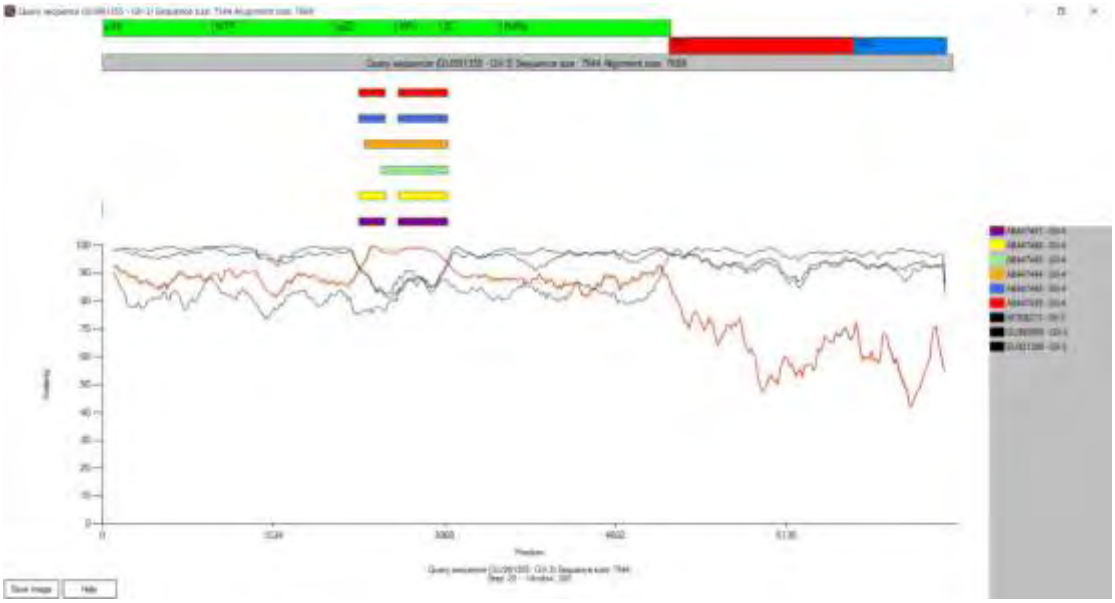
Ενσωμάτωση εικόνας σχολιασμού σε ένα διάγραμμα ομοιότητας

Μια από τις δυνατότητες που ξεχωρίζουν στο T-RECs είναι αυτή της ενσωμάτωσης μιας εικόνας που περιέχει τόσο τις ακολουθίες όσο και τις ετικέτες σχολιασμού σε ένα διάγραμμα ομοιότητας. Η διαδικασία αυτή μπορεί να γίνει μόνο μέσω της καρτέλας επεξεργασίας σχολιασμού ακολουθώντας τα παρακάτω βήματα:

1) Συμπλήρωση του πίνακα με τις παραμέτρους των ετικετών
2) Επιλογή του κουμπιού «Annotate» ώστε να δημιουργηθεί η εικόνα
3) Αλλαγή χρωμάτων των ετικετών ή/και οποιασδήποτε ακολουθίας με δεξί κλικ και επιλογή χρώματος (προαιρετικό)
4) Επιλογή από τη λίστα «Select a similarity plot to pin the above graph» του διαγράμματος ομοιότητας στο οποίο είναι επιθυμητή η ενσωμάτωση της εικόνας
5) Επιλογή του κουμπιού «Show annotation on similarity plot» ώστε να ολοκληρωθεί η ενσωμάτωση

Η επιλογή «Only show query sequence on similarity plot» είναι υπεύθυνη για την ενσωμάτωση μόνο της ακολουθίας επερώτησης (και των ετικετών σχολιασμού) στο διάγραμμα ομοιότητας. Η επιλογή αυτή είναι πολύ χρήσιμη ειδικά σε περιπτώσεις όπου οι ακολουθίες που περιέχονται στην εικόνα είναι πάρα πολλές. Για να λειτουργήσει σωστά θα πρέπει να ενεργοποιηθεί πριν από την επιλογή του κουμπιού «Show annotation on similarity plot».

Στην εικόνα 23 παρουσιάζεται ένα παράδειγμα ενσωμάτωσης των αποτελεσμάτων του T-RECs σε συνδυασμό με τις ετικέτες σχολιασμού (όπως φαίνονται στην εικόνα 22) στο αντίστοιχο διάγραμμα ομοιότητας.



Εικόνα 23: Ενσωμάτωση αποτελεσμάτων του T-RECs σε συνδυασμό με τις ετικέτες σχολιασμού. Συγκεκριμένα, οι ετικέτες σχολιασμού αποτελούν τις περιοχές του γονιδιώματος της ακολουθίας επερώτησης με accession number GU991355.

Μία επιπλέον δυνατότητα που προστίθεται στη φόρμα του διαγράμματος ομοιότητας ύστερα από την ενσωμάτωση της εικόνας είναι η κατακόρυφη αλλαγή μεγέθους της εικόνας. Για να επιτευχθεί αυτό, ο χρήστης θα πρέπει να μετακινήσει τον δείκτη του ποντικιού επάνω στην ενσωματωμένη εικόνα στο γράφημα και με τη βοήθεια της ροδέλας του ποντικιού να αυξήσει (Scroll down) ή να μειώσει (Scroll up) το μέγεθος της εικόνας.

Σημειώνεται πως και στην καρτέλα επεξεργασίας σχολιασμού περιλαμβάνεται η γραμμή πληροφοριών.

3.10 Individual Analyses

Η επιλογή «Pipeline Analysis» αποτελεί έναν αυτόματο τρόπο λειτουργίας του T-RECs ο οποίος συνδυάζει ορισμένες από τις μεθόδους αναλύσεων του προγράμματος με τις προεπιλεγμένες παραμέτρους. Στο σημείο αυτό θα εξηγηθούν οι τρεις επιμέρους αναλύσεις που μπορεί να πραγματοποιήσει ο χρήστης μέσα από το T-RECs δίνοντας μεγαλύτερη έμφαση στις παραμέτρους. Επίσης θα περιγραφούν τα εργαλεία επεξεργασίας των ακολουθιών που εισάγει ο χρήστης ώστε να βελτιώσει την αποτελεσματικότητα του προγράμματος.

Από το κεντρικό μενού του T-RECs ο χρήστης θα πρέπει να επιλέξει το κουμπί με τίτλο «Individual Analyses» ώστε να εμφανιστεί η αντίστοιχη φόρμα όπως φαίνεται στην εικόνα 24. Για την επιστροφή στο κεντρικό μενού του T-RECs αρκεί να επιλεγεί το κουμπί με τίτλο «< BACK».

Εικόνα 24: Κεντρικό μενού των επιμέρους αναλύσεων του T-RECs

Οι τρεις επιμέρους αναλύσεις που μπορεί να πραγματοποιήσει ο χρήστης μέσω του T-RECs είναι οι εξής:

1) Blast Based Genotyping Tool
2) Blast Based Intergroup Recombination Analysis
3) Similarity Plot Analysis

Επίσης το πρόγραμμα διαθέτει δύο επιπλέον εργαλεία για την επεξεργασία των ακολουθιών που εισάγει ο χρήστης:

1) Sequence editor
2) Sequence clustering

Για λόγους ευκολότερης κατανόησης του τρόπου λειτουργίας των επιμέρους αναλύσεων θα περιγραφεί πρώτα η μέθοδος «Blast Based Intergroup Recombination Analysis» έπειτα το «Blast Based Genotyping Tool» και τέλος η «Similarity Plot Analysis».

3.10.1 Blast Based Intergroup Recombination Analysis

Η ανάλυση με τίτλο «Blast Based Intergroup Recombination Analysis» είναι η βασική ανάλυση που διεξάγει το T-RECs για τον εντοπισμό πιθανών γεγονότων ανασυνδυασμού. Η λειτουργία της έχει ήδη περιγραφεί καθώς αποτελεί την μέθοδο που χρησιμοποιείται κατά τη λειτουργία της Pipeline Analysis (εν. 3.5). Όπως και

στην Pipeline Analysis έτσι και στην Blast Based Intergroup Recombination Analysis θα πρέπει ο χρήστης να εισάγει τα τρία απαραίτητα αρχεία:

***Query file**:* Ο χρήστης θα πρέπει, μετά το πάτημα του σχετικού κουμπιού, να επιλέξει το αρχείο με τις ακολουθίες επερώτησης.

***Database file**:* Ο χρήστης θα πρέπει, μετά το πάτημα του σχετικού κουμπιού, να επιλέξει το αρχείο με τις ακολουθίες της βάσης δεδομένων.

***Evolutionary Groups**:* Ο χρήστης θα πρέπει, μετά το πάτημα του σχετικού κουμπιού, να επιλέξει το αρχείο με τις αντιστοιχίσεις του κάθε accession number με το αντίστοιχο όνομα της ομάδας στην οποία ανήκει η εκάστοτε ακολουθία.

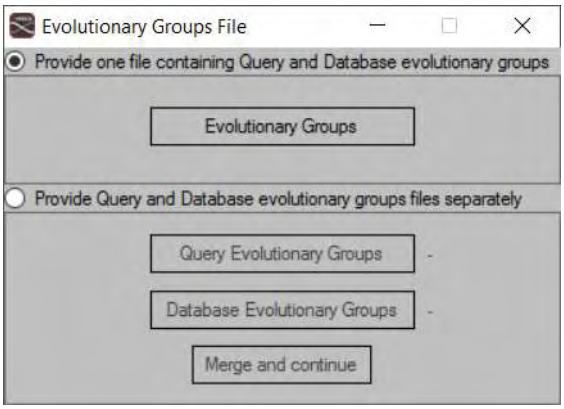
Η δομή που θα πρέπει να έχει το κάθε αρχείο περιγράφεται στο Παράρτημα 1.

Μόλις φορτώνεται κάθε ένα από τα παραπάνω αρχεία, εμφανίζεται δίπλα από το σχετικό κουμπί η πλήρης διαδρομή του αρχείου στο δίσκο. Έπειτα ακολουθεί ένα κουμπί με την ένδειξη «X» το οποίο εάν επιλεγεί σβήνεται το επιλεγμένο αρχείο από τη μνήμη του T-RECs και είναι απαραίτητη η φόρτωση ενός νέου αρχείου.

Μία ιδιαιτερότητα που έχει η Blast Based Intergroup Recombination Analysis αφορά τη φόρτωση του αρχείου που περιέχει την αντιστοίχιση των accession numbers με τις εξελικτικές ομάδες. Όπως φαίνεται και στην εικόνα 25, κατά τη φόρτωση του συγκεκριμένου αρχείου εμφανίζεται μία φόρμα που παρέχει 2 επιλογές για τη φόρτωσή του:

1) Φόρτωση ενός ενιαίου αρχείου που περιέχει τις αντιστοιχίσεις accession numbers – εξελικτική ομάδα τόσο για τις ακολουθίες που περιέχονται στο αρχείο με τις ακολουθίες επερώτησης όσο και για αυτές του αρχείου της βάσης δεδομένων. Πατώντας το κουμπί με όνομα «Evolutionary Groups» στη φόρμα αυτή ο χρήστης θα πρέπει να υποδείξει τη θέση του συγκεκριμένου αρχείου.

2) Φόρτωση δύο ξεχωριστών αρχείων καθένα από τα οποία παρέχει τις αντιστοιχίσεις accession numbers – εξελικτική ομάδα: ένα για τις ακολουθίες του αρχείου με τις ακολουθίες επερώτησης και ένα για αυτές του αρχείου της βάσης δεδομένων. Στην περίπτωση αυτή, αφού ο χρήστης επιλέξει το κουμπί με όνομα «Provide Query and Database evolutionary group files separately» θα πρέπει να φορτώσει τα δύο αρχεία ξεχωριστά επιλέγοντας τα σχετικά κουμπιά στη φόρμα (Query Evolutionary Groups και Database Evolutionary Groups αντίστοιχα) και στη συνέχεια να επιλέξει το κουμπί με όνομα «Merge and continue». Έτσι, το T-RECs θα δημιουργήσει ένα νέο, ενιαίο αρχείο το οποίο θα χρησιμοποιηθεί ως Evolutionary Groups file και ο χρήστης θα ερωτηθεί για την τοποθεσία όπου θέλει να αποθηκεύσει το νέο αυτό αρχείο για πιθανή μελλοντική χρήση.



Εικόνα 25: Φόρμα φόρτωσης αρχείου που περιέχει την αντιστοίχιση των accession numbers με τις εξελικτικές ομάδες.

Αφού φορτωθεί το αρχείο με τις εξελικτικές ομάδες, ένα μήνυμα θα ενημερώσει το χρήστη για τον συνολικό αριθμό των αντιστοιχίσεων που περιέχονται στο αρχείο αυτό. Πατώντας το κουμπί «OK» στο κάτω δεξιό μέρος του κεντρικού μενού των Individual Analyses, ο χρήστης μεταφέρεται στη φόρμα παραμέτρων του Blast. Όπως φαίνεται στην εικόνα 26, η φόρμα αυτή περιέχει τρεις καρτέλες: 1) Blast, 2) Sequence tool, 3) Results.

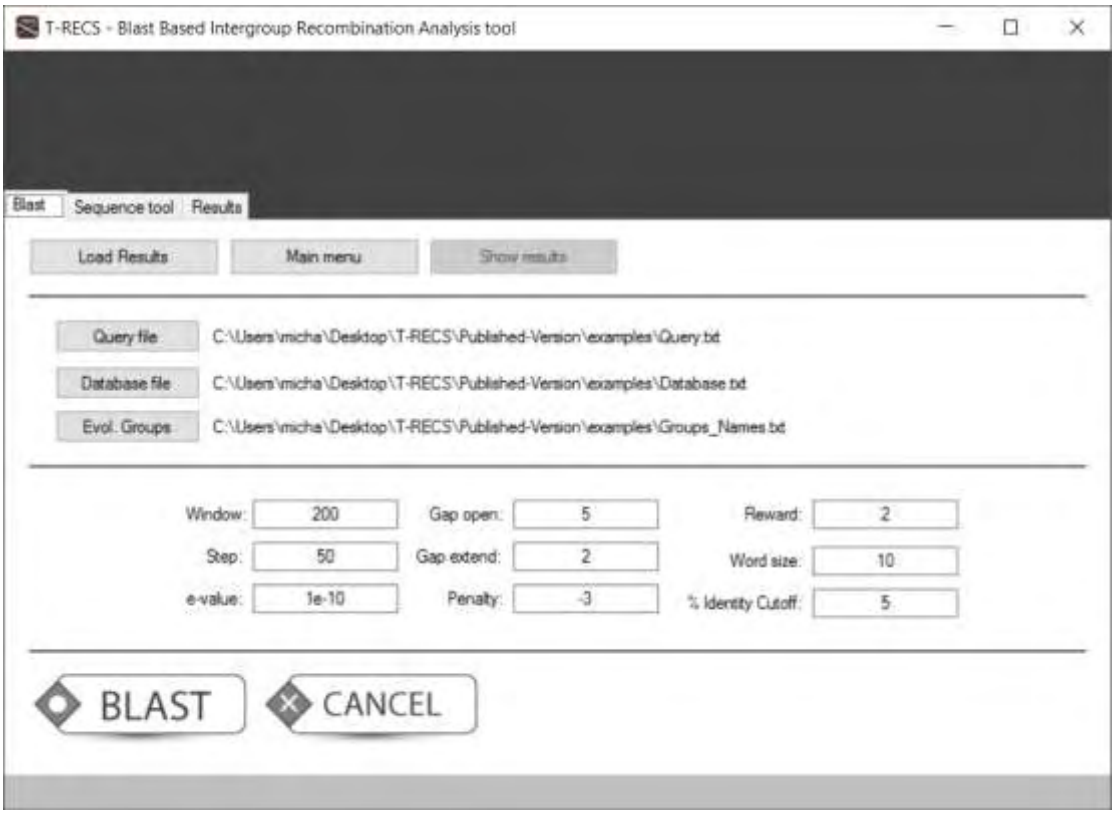
Καρτέλα παραμέτρων του Blast

Στην καρτέλα αυτή ο χρήστης έχει ακόμη την επιλογή να αλλάξει τα αρχεία εισόδου εάν επιθυμεί επιλέγοντας τα σχετικά κουμπιά (Query file, Database file, Evol. Groups). Ακριβώς κάτω από τα αρχεία εισόδου εντοπίζονται οι παράμετροι με τις οποίες θα γίνει το Blast. Εκτενέστερη περιγραφή των παραμέτρων γίνεται στο Παράρτημα 2. Οι τρεις παράμετροι που θα σχολιαστούν στην ενότητα αυτή είναι οι: 1) Window, 2) Step, 3) % Identity Cutoff.

Window: Πρόκειται για την παράμετρο που υποδεικνύει στο πρόγραμμα το μέγεθος του συρόμενου παραθύρου.

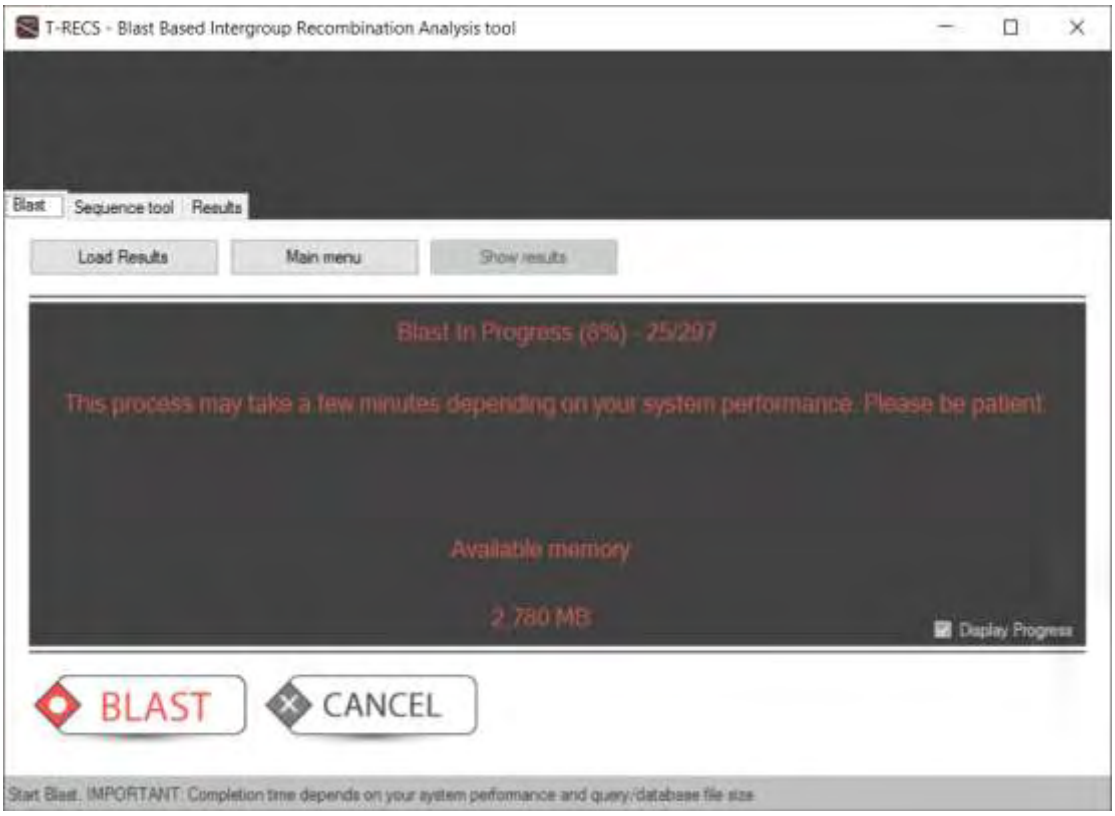
Step: Πρόκειται για την παράμετρο που υποδεικνύει στο πρόγραμμα το βήμα του συρόμενου παραθύρου.

% Identity Cutoff: Πρόκειται για την παράμετρο που υποδεικνύει στο πρόγραμμα το κατώφλι διαφοράς ταύτισης βάσει του οποίου θα γίνονται δεκτά τα αποτελέσματα (εν. 3.4).



Εικόνα 26: Φόρμα παραμέτρων του Blast.

Όταν οριστούν οι επιθυμητές τιμές για τις παραμέτρους με τις οποίες θα πραγματοποιηθεί το Blast, το T-RECs θα είναι σε θέση να εκτελέσει την ανάλυση. Για να γίνει αυτό ο χρήστης θα πρέπει να επιλέξει το κουμπί με όνομα «BLAST». Μόλις ξεκινήσει η ανάλυση, οι παράμετροι του Blast αλλά και οι επιλογές για τα αρχεία εισόδου θα καλυφθούν από τον πίνακα ενημέρωσης προόδου όπως φαίνεται στην εικόνα 27. Ο πίνακας αυτός περιλαμβάνει μια γραμμή ενημέρωσης για το ποσοστό της ανάλυσης που έχει ολοκληρωθεί. Σημειώνεται πως κατά την εκκίνηση του Blast, το T-RECs σπάει σε μικρότερα κομμάτια κάθε ακολουθία που βρίσκεται στο αρχείο ακολουθιών επερώτησης και χρησιμοποιεί το κάθε κομμάτι σαν μια ξεχωριστή ακολουθία επερώτησης. Δίπλα από το ποσοστό ολοκλήρωσης της διαδικασίας εμφανίζεται και η πρόοδος για τα κομμάτια των ακολουθιών για τα οποία έχει ολοκληρωθεί η διαδικασία του Blast. Παρά το γεγονός ότι εμφάνιση της προόδου είναι μια πολύ σημαντική ένδειξη, η καθυστέρηση που μπορεί να προκαλέσει στη συνολική διαδικασία είναι κάποιες φορές εξαιρετικά μεγάλη. Για το λόγο αυτό παρέχεται η επιλογή «Display Progress» η οποία όταν είναι ενεργοποιημένη επιτρέπει την εμφάνιση της προόδου ενώ όταν δεν είναι τότε η πρόοδος δεν εμφανίζεται και εξοικονομείται αρκετός χρόνος για τη διεκπεραίωση της ανάλυσης.



Εικόνα 27: Η διαδικασία του Blast βρίσκεται σε εξέλιξη.

Η διαδικασία μπορεί να ακυρωθεί οποιαδήποτε στιγμή πατώντας το κουμπί με το όνομα «Cancel».

Μετά την ολοκλήρωση του Blast, το T-RECs ενημερώνει το χρήστη για το χρόνο που χρειάστηκε η διαδικασία για να ολοκληρωθεί και στη συνέχεια εμφανίζει τη φόρμα αποτελεσμάτων (εν. 3.6). Από εκεί και έπειτα ο χρήστης είναι σε θέση να προβάλει και να επεξεργαστεί τα αποτελέσματα όπως έχει ήδη περιγραφεί στις ενότητες 3.6 έως και 3.9.

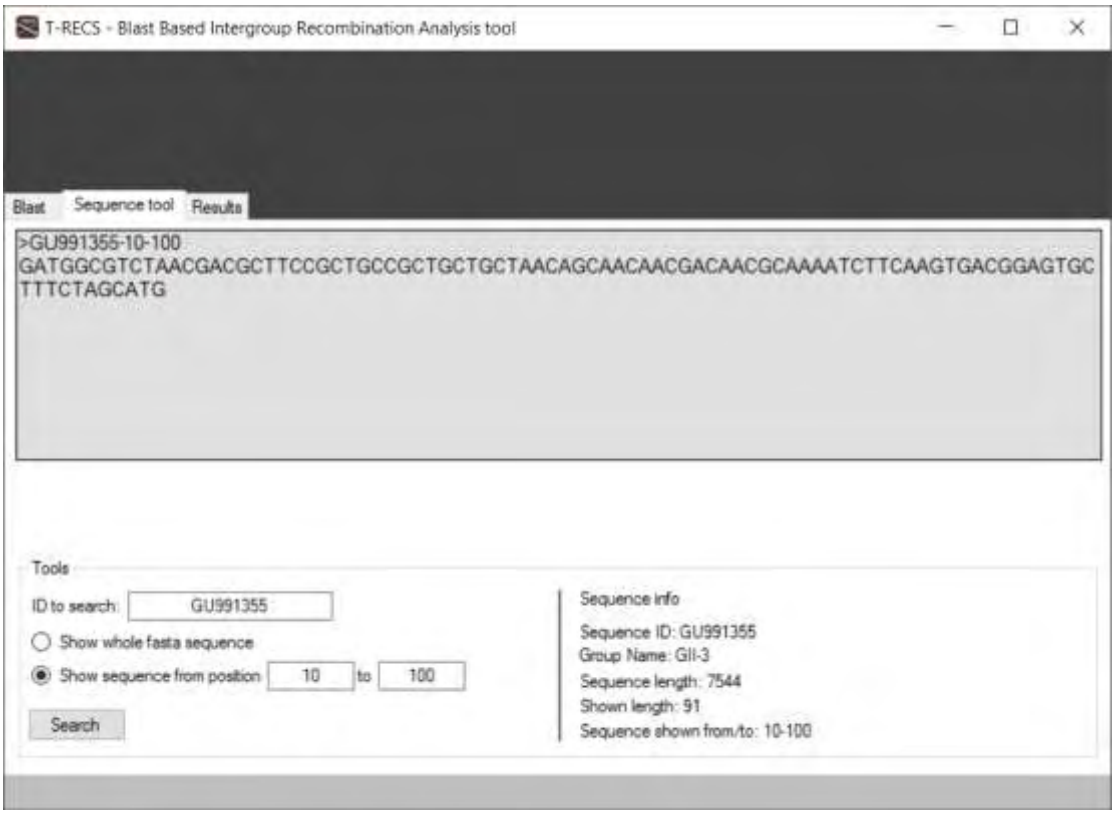
Load Results: Με την επιλογή του κουμπιού αυτού, ο χρήστης μπορεί να φορτώσει αποτελέσματα από μία παλαιότερη ανάλυση και έτσι να μη χρειαστεί να εκτελέσει ξανά τη διαδικασία του Blast. Πατώντας το κουμπί θα πρέπει να υποδείξει στο πρόγραμμα την τοποθεσία του επιθυμητού αρχείου. Απαραίτητη προϋπόθεση για τη σωστή λειτουργία αυτής της επιλογής είναι η παράλληλη φόρτωση των ίδιων αρχείων εισόδου (Query/Database file και Evolutionary Groups file) που είχαν χρησιμοποιηθεί στη συγκεκριμένη ανάλυση.

Main menu: Με την επιλογή του κουμπιού «Main menu» ο χρήστης επιστρέφει στο κεντρικό μενού των Individual Analyses.

Show results: Η επιλογή αυτή είναι απενεργοποιημένη μέχρι ο χρήστης να εκτελέσει μια ανάλυση και να εμφανιστεί η φόρμα των αποτελεσμάτων. Αν η φόρμα αυτή κλείσει, το κουμπί αυτό ενεργοποιείται και ο χρήστης μπορεί να την επαναφέρει με το πάτημά του.

Καρτέλα Sequence tool

Πρόκειται για ένα μικρό εργαλείο το οποίο μπορεί να προβάλει πληροφορίες σχετικές με μια επιθυμητή ακολουθία. Στην εικόνα 28 παρουσιάζεται ένα παράδειγμα για την ακολουθία με accession number GU991355. Αρχικά, ο χρήστης εισάγει στο πλαίσιο κειμένου «ID to search» τον κωδικό της επιθυμητής ακολουθίας. Έπειτα έχει να επιλέξει ανάμεσα στην εμφάνιση ολόκληρης της ακολουθίας ή ενός τμήματος αυτής. Στην εικόνα 28 είναι επιλεγμένο το «Show sequence from position X to Y» όπου X η θέση έναρξης του τμήματος και Y η θέση λήξης (στην εικόνα είναι 10 και 100 αντίστοιχα). Ακολούθως, με το πάτημα του κουμπιού search εμφανίζεται η επιθυμητή ακολουθία στο πλαίσιο που βρίσκεται στο πάνω μέρος της καρτέλας, καθώς επίσης και διάφορες πληροφορίες σχετικά με την επιλεγμένη ακολουθία στο κάτω δεξιό μέρος της καρτέλας.



Εικόνα 28: Η καρτέλα του Sequence tool

Καρτέλα Results

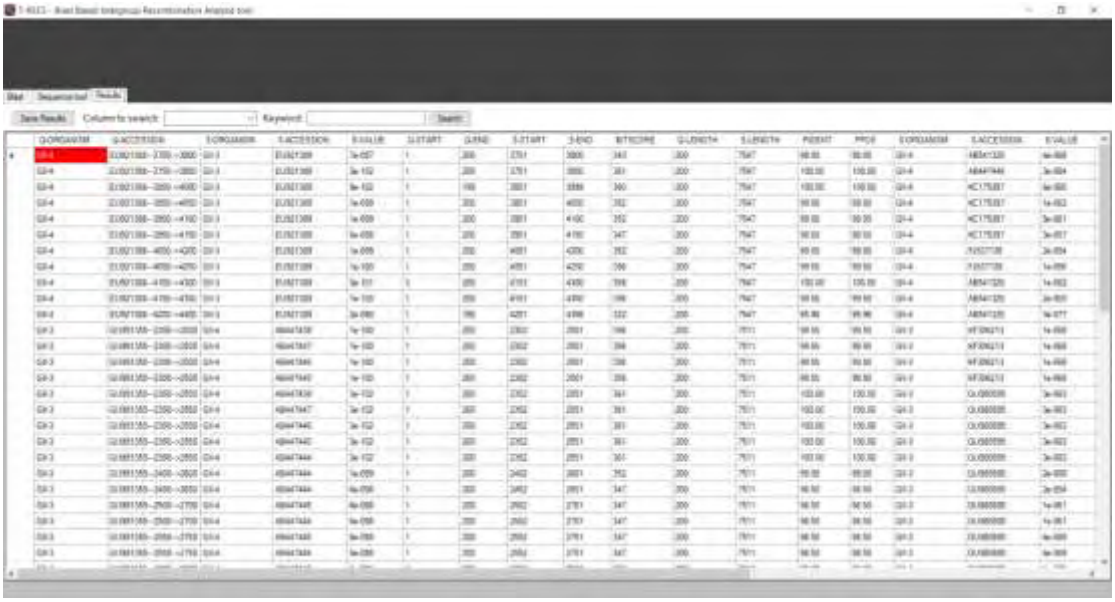
Η καρτέλα αποτελεσμάτων είναι κενή ώσπου να πραγματοποιηθεί μια ανάλυση. Έπειτα, ο πίνακας που περιέχεται στην καρτέλα αυτή γεμίζει με τα αποτελέσματα του Blast. Η καρτέλα αποτελεσμάτων παρουσιάζεται στην εικόνα 29. Οι στήλες που περιέχονται στον πίνακα αυτής της καρτέλας αναλύονται στο Παράρτημα 1 (αρχεία εξόδου)

Save Results

Στην καρτέλα αυτή, όπως και στη φόρμα αποτελεσμάτων, παρέχεται η επιλογή αποθήκευσης των αποτελεσμάτων του Blast για μελλοντική χρήση χωρίς να χρειαστεί να επαναληφθεί η διαδικασία του Blast. Η επιλογή αυτή δίνεται από το κουμπί με τίτλο «Save Results».

Επιλογές αναζήτησης

Η καρτέλα αποτελεσμάτων παρέχει ακόμη και τη δυνατότητα αναζήτησης μιας λέξης κλειδί σε μία επιλεγμένη στήλη. Αρχικά μέσω της λίστας «Column to search» θα πρέπει να επιλεγεί η επιθυμητή στήλη και έπειτα στο πλαίσιο κειμένου «Keyword» να πληκτρολογηθεί η λέξη κλειδί. Για να ξεκινήσει η αναζήτηση, ο χρήστης θα πρέπει να επιλέξει το κουμπί με όνομα «Search».



Εικόνα 29: Η καρτέλα των αποτελεσμάτων (Results tab)

Σημειώνεται πως η φόρμα παραμέτρων του Blast παρέχει μια γραμμή βοήθειας στο κάτω μέρος της. Έτσι, όταν ο δείκτης του ποντικιού μετακινείται σε ορισμένες

επιλογές, εμφανίζεται αυτόματα ένα κείμενο βοήθειας σχετικά με τη λειτουργία της εκάστοτε επιλογής.

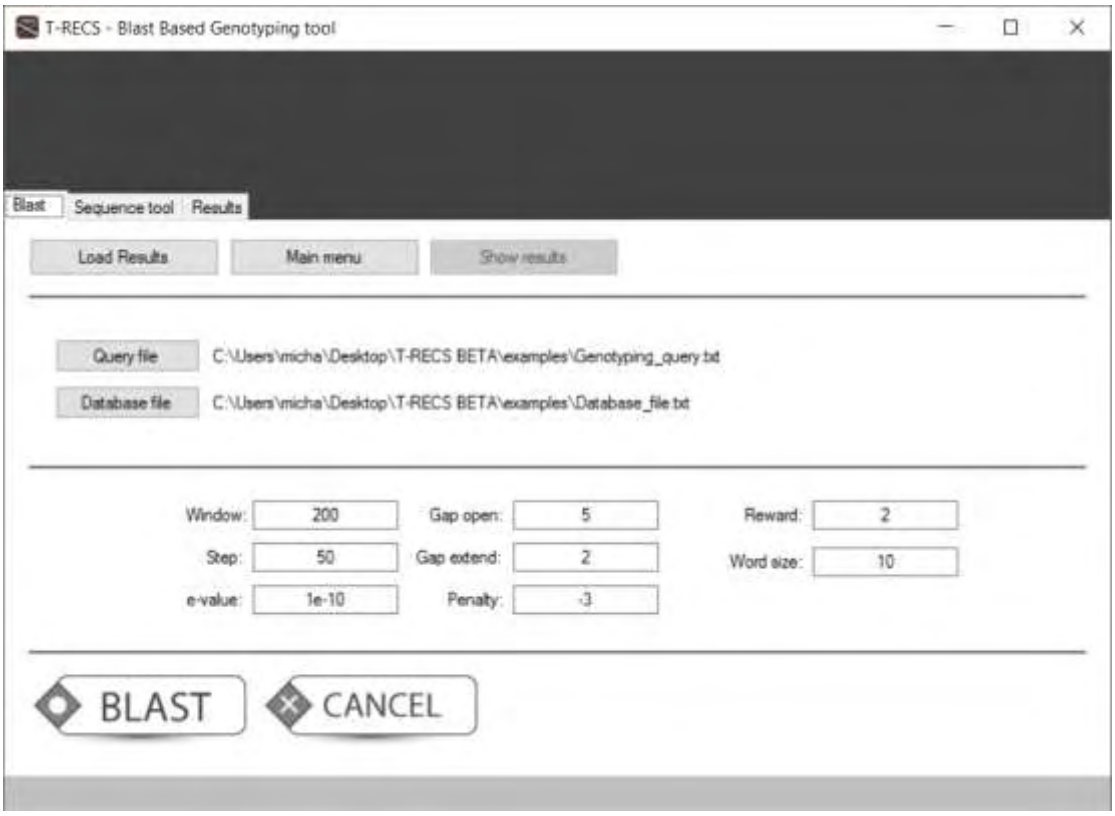
3.10.2 Blast Based Genotyping Tool

Κάποιες φορές μπορεί να μην είναι γνωστή η εξελικτική ομάδα στην οποία ανήκει μια ακολουθία. Για το λόγο αυτό το T-RECs περιλαμβάνει ένα τύπο ανάλυσης που βασίζεται στο Blast και πραγματοποιείται από το εργαλείο με όνομα «Blast Based Genotyping Tool». Πρόκειται για ένα εργαλείο που λειτουργεί με τη μέθοδο του συρόμενου παραθύρου και είναι σε θέση να αναζητά και να εντοπίζει την πιθανή εξελικτική ομάδα μιας ακολουθίας. Η διαδικασία εντοπισμού της πιθανότερης εξελικτικής ομάδας έχει ως εξής: Αρχικά η υπό εξέταση ακολουθία κόβεται σε μικρότερα τμήματα βάσει του μεγέθους του συρόμενου παραθύρου και του βήματός του. Στη συνέχεια μέσω του Blast εντοπίζεται το καλύτερο αποτέλεσμα για το κάθε κομμάτι της ακολουθίας. Τέλος, η εξελικτική ομάδα με την οποία φαίνεται να ταυτίζονται τα περισσότερα κομμάτια παρουσιάζεται ως η πιθανότερη εξελικτική ομάδα στην οποία ανήκει η υπό εξέταση ακολουθία.

Για τη χρήση του εργαλείου, ο χρήστης θα πρέπει να το επιλέξει από το κεντρικό μενού των Individual Analyses ενώ για τη σωστή λειτουργία του εργαλείου αυτού, θα πρέπει να εισάγει ένα αρχείο με μία ή περισσότερες ακολουθίες αγνώστου εξελικτικής ομάδας. Στη συνέχεια θα πρέπει να φορτωθεί από το χρήστη ένα αρχείο βάσης δεδομένων με ακολουθίες γνωστών εξελικτικών ομάδων. Σημειώνεται ότι όσο περισσότερες είναι οι ακολουθίες της βάσης δεδομένων και όσο πιο καλή εκπροσώπηση έχει η κάθε εξελικτική ομάδα από γνωστές ακολουθίες, τόσο πιο ικανοποιητικά θα είναι τα αποτελέσματα. Τέλος είναι απαραίτητο το αρχείο που θα αντιστοιχίζει την εξελικτική ομάδα κάθε μιας από τις ακολουθίες της βάσης δεδομένων με τον αντίστοιχο accession number. Αυτή τη φορά, σε αντίθεση με την Blast Based Intergroup Recombination Analysis, ο χρήστης καλείται να εισάγει ένα και μόνο αρχείο που θα περιέχει τις αντιστοιχίσεις για τις ακολουθίες της βάσης δεδομένων ενώ και στην περίπτωση αυτή, μόλις το αρχείο επιλεχθεί ο χρήστης θα ενημερωθεί για τον αριθμό των αντιστοιχίσεων που υπάρχουν στο αρχείο.

Η μορφή που θα πρέπει να έχουν τα παραπάνω αρχεία περιγράφεται στο Παράρτημα 1 (Input files).

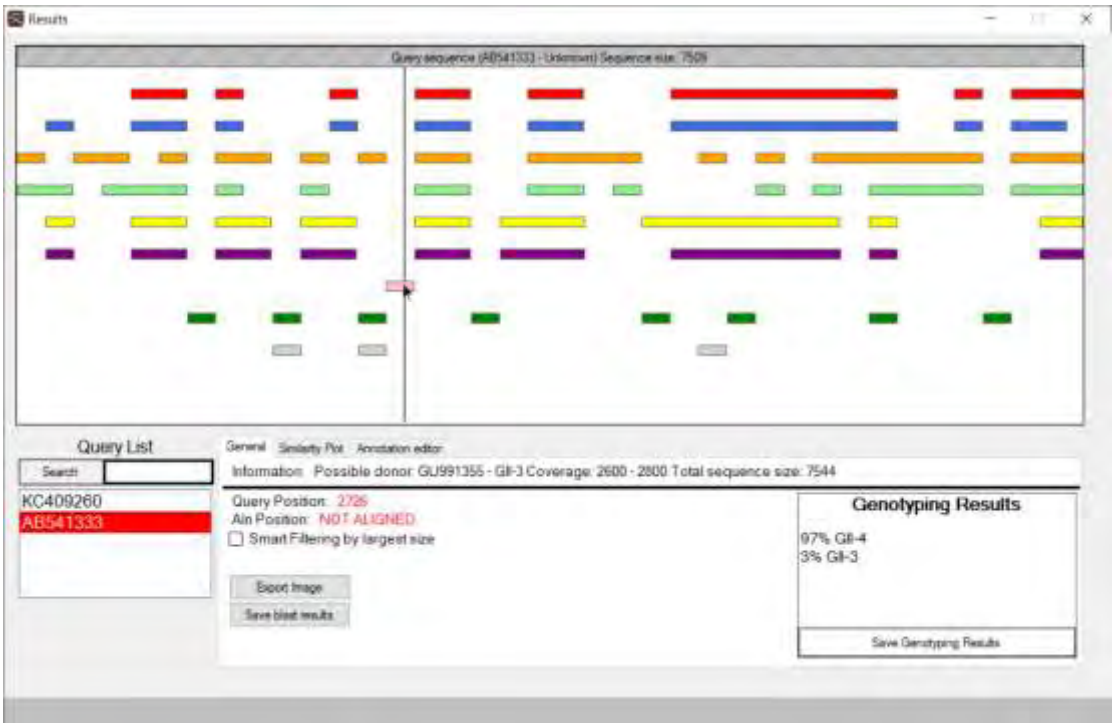
Αφού φορτωθούν τα αρχεία, ο χρήστης θα είναι σε θέση να ξεκινήσει το εργαλείο επιλέγοντας το κουμπί «OK» στο κάτω δεξιό μέρος του κεντρικού μενού των Individual Analyses.



Εικόνα 30: Φόρμα παραμέτρων του Blast στο Blast Based Genotyping tool.

Στην εικόνα 30 φαίνονται οι παράμετροι του Blast για το Blast Based Genotyping tool. Η διαφορά με την αντίστοιχη φόρμα για την Blast Based Intergroup Recombination Analysis είναι ότι τώρα δεν διατίθεται επιλογή για % Identity Cutoff καθώς κάτι τέτοιο δεν είναι απαραίτητο για τη λειτουργία του συγκεκριμένου εργαλείου. Από την άλλη πλευρά, όλες οι υπόλοιπες επιλογές είναι ακριβώς οι ίδιες όπως και στην Blast Based Intergroup Recombination Analysis.

Μία ακόμη διαφορά εντοπίζεται στη φόρμα αποτελεσμάτων η οποία για το Blast Based Genotyping tool παρέχει κάποιες επιπλέον δυνατότητες.



Εικόνα 31: Φόρμα αποτελεσμάτων μετά το Blast για το Blast Based Genotyping tool.

Αφού ολοκληρωθεί η διαδικασία του Blast εμφανίζεται η φόρμα αποτελεσμάτων όπως παρουσιάζεται στην εικόνα 31. Παρατηρείται πως σχεδόν σε όλο το μήκος της, η ακολουθία επερώτησης για την οποία η εξελικτική ομάδα είναι άγνωστη φαίνεται να έχει καλύτερη ταύτιση με ακολουθίες που ανήκουν στην εξελικτική ομάδα GII-4. Κάτω από το δείκτη του ποντικιού φαίνεται ένα μικρό κομμάτι όπου η ακολουθία παρουσιάζει καλύτερη ταύτιση με την εξελικτική ομάδα GII-3. Τα αποτελέσματα παρατίθενται σε μορφή ποσοστών στο κάτω δεξιό μέρος της καρτέλας γενικών επιλογών της φόρμας αποτελεσμάτων ως «Genotyping Results». Εκεί, ο χρήστης ενημερώνεται για το ποσοστό κάλυψης ανά εξελικτική ομάδα κατά μήκος ολόκληρης της ακολουθίας. Στο συγκεκριμένο παράδειγμα το 97% της ακολουθίας εντοπίζεται ως GII-4 ενώ το 3% ως GII-3. Από τη στιγμή που το επικρατέστερο ποσοστό αντιστοιχεί στην ομάδα GII-4, το T-RECs θεωρεί πως η ακολουθία με την άγνωστη εξελικτική ομάδα ανήκει τελικά σε αυτή των GII-4.

Μια επιπλέον λειτουργία που είναι διαθέσιμη στη φόρμα αποτελεσμάτων έπειτα από τη χρήση του εργαλείου Blast Based Genotyping tool είναι η επιλογή «Save Genotyping Results». Επιλέγοντας αυτή την επιλογή, ο χρήστης θα είναι σε θέση να αποθηκεύσει σε ένα νέο αρχείο τους accession numbers της κάθε ακολουθίας επερώτησης αντιστοιχισμένα με την εξελικτική ομάδα που έχει εντοπίσει το εργαλείο κατά την ανάλυση. Πρόκειται ουσιαστικά για τη δημιουργία ενός αρχείου «Evolutionary Groups» με τη μορφή που περιγράφεται στο Παράρτημα 1.

3.10.3 Similarity Plot Analysis

Η Τρίτη και τελευταία επιλογή ανάλυσης από το κεντρικό μενού των Individual Analyses είναι η «Similarity Plot Analysis». Αποτελεί μία απλή ανάλυση όπου ο χρήστης έχει τη δυνατότητα να δημιουργεί διαγράμματα ομοιότητας χρησιμοποιώντας τις ακολουθίες που ο ίδιος είναι σε θέση να εισάγει. Όταν η ανάλυση αυτή επιλεγεί από τον χρήστη, όπως φαίνεται στην εικόνα 32, απαιτεί ένα και μοναδικό αρχείο εισόδου. Πρόκειται για το «Query file» το οποίο θα περιέχει τις επιθυμητές από το χρήστη ακολουθίες που θα χρησιμοποιηθούν για την ανάλυση με διαγράμματα ομοιότητας. Η μορφή του αρχείου αυτού περιγράφεται στο Παράρτημα 1.

Εικόνα 32: Η Similarity Plot Analysis απαιτεί μόνο το αρχείο Query file με τις επιθυμητές ακολουθίες

Αφού ο χρήστης φορτώσει το αρχείο με τις επιθυμητές ακολουθίες μπορεί να επιλέξει το κουμπί «OK» στο κάτω δεξιό μέρος του κεντρικού μενού των Individual analyses για να ξεκινήσει την Similarity Plot Analysis.

Στην εικόνα 33 παρουσιάζεται η φόρμα επιλογών της Similarity Plot Analysis. Στο αριστερό μέρος είναι διαθέσιμη η λίστα με όλες τις ακολουθίες που περιλαμβάνει το αρχείο που έχει φορτώσει ο χρήστης στο πρόγραμμα. Από τη λίστα αυτή, ο χρήστης μπορεί να επιλέξει ποιες από τις διαθέσιμες ακολουθίες θα χρησιμοποιηθούν ώστε να δημιουργηθεί ένα διάγραμμα ομοιότητας. Στο δεξιό μέρος παρατίθενται όλες οι παράμετροι που έχουν να κάνουν με τη δημιουργία του

διαγράμματος ομοιότητας. Θα πρέπει να σημειωθεί πως οι επιλογές αυτές παραμένουν ανενεργές μέχρι τη στιγμή που ο χρήστης θα επιλέξει τις επιθυμητές ακολουθίες από τη λίστα και θα πατήσει το κουμπί «OK».

Περιοχή αναζήτησης

Στο επάνω αριστερό μέρος της φόρμας της Similarity Plot Analysis εντοπίζεται ένα πλαίσιο κειμένου με τίτλο «Search». Μέσω του πλαισίου αυτού, ο χρήστης έχει τη δυνατότητα να αναζητήσει κάποια ακολουθία πληκτρολογώντας τον κωδικό που τη χαρακτηρίζει. Στην περίπτωση αυτή δε χρειάζεται να επιλέξει κάποιο κουμπί για να γίνει η αναζήτηση καθώς αυτή λαμβάνει χώρα κάθε φορά που ο χρήστης πληκτρολογεί ένα νέο χαρακτήρα στο πλαίσιο κειμένου.

Select All

Με την επιλογή του κουμπιού «Select All» επιλέγονται αυτόματα όλες οι ακολουθίες που βρίσκονται μέσα στη λίστα ακολουθιών.

Deselect All

Με την επιλογή του κουμπιού «Deselect All» από-επιλέγονται αυτόματα όλες οι ακολουθίες που βρίσκονται μέσα στη λίστα ακολουθιών.

Κουμπί OK

Μόλις ο χρήστης επιλέξει όλες τις ακολουθίες που θέλει να συμπεριλάβει στη δημιουργία του διαγράμματος ομοιότητας από τη λίστα ακολουθιών, θα πρέπει να επιλέξει αυτό το κουμπί ώστε να ενεργοποιηθούν οι παράμετροι δημιουργίας του διαγράμματος ομοιότητας.

Λίστα Query

Στην περιοχή των παραμέτρων του διαγράμματος ομοιότητας εντοπίζεται μία λίστα με το όνομα «Query». Η λίστα αυτή περιλαμβάνει όλες τις ακολουθίες που έχει επιλέξει ο χρήστης να συμπεριλάβει στο διάγραμμα ομοιότητας που πρόκειται να δημιουργηθεί και ο ρόλος της είναι να δώσει στο χρήστη την επιλογή μιας εξ αυτών η οποία θα χρησιμοποιηθεί σαν ακολουθία επερώτησης. Έτσι, το διάγραμμα ομοιότητας που θα παραχθεί θα παρουσιάζει την ομοιότητα κάθε άλλης ακολουθίας που βρίσκεται σε αυτή τη λίστα σε σύγκριση με την επιλεγμένη ακολουθία επερώτησης.

Create Similarity Plot

Το κουμπί με όνομα «Create Similarity Plot» είναι υπεύθυνο για την έναρξη της κατασκευής του διαγράμματος ομοιότητας. Αφού επιλέξει μια ακολουθία που θα

χρησιμοποιηθεί ως ακολουθία επερώτησης, ο χρήστης μπορεί να προχωρήσει στη δημιουργία του διαγράμματος ομοιότητας πατώντας αυτό το κουμπί. Για την κατασκευή του διαγράμματος ομοιότητας είναι απαραίτητη η στοίχιση των ακολουθιών. Σε περίπτωση που αυτές δεν είναι στοιχισμένες, το πρόγραμμα ενημερώνει τον χρήστη ότι έχει την επιλογή να πραγματοποιήσει στοίχιση όπως έχει ήδη περιγραφεί στην ενότητα 3.8.1 με τη βοήθεια του προγράμματος muscle.

Go back to main menu

Για την επιστροφή στο κεντρικό μενού των Individual Analyses, ο χρήστης μπορεί να επιλέξει το κουμπί με τίτλο «Go back to main menu».

Η μέθοδος που χρησιμοποιείται και σε αυτή την περίπτωση για την κατασκευή του διαγράμματος ομοιότητας είναι η P-Distance ενώ ο χρήστης μέσω των πλαισίων κειμένου «Step» και «Window» είναι σε θέση να ορίσει τις τιμές για το βήμα και το μέγεθος του συρόμενου παραθύρου αντίστοιχα όπως έχει περιγραφεί στην ενότητα 3.8.1.

Ύστερα από την ολοκλήρωση της δημιουργίας του διαγράμματος ομοιότητας εμφανίζεται η φόρμα με το ολοκληρωμένο γράφημα. Η φόρμα αυτή είναι ίδια με τη φόρμα διαγράμματος ομοιότητας που περιγράφηκε στην ενότητα 3.8.2. Επίσης, ενεργοποιείται το κουμπί με όνομα «Annotation» στην κεντρική φόρμα επιλογών του διαγράμματος ομοιότητας το οποίο όταν επιλεγεί εμφανίζει μία φόρμα δημιουργίας ετικετών σχολιασμού όπως φαίνεται στην εικόνα 33. Η λειτουργία της είναι πανομοιότυπη με αυτή που έχει ήδη περιγραφεί, με μόνη ουσιαστική διαφορά το γεγονός ότι δε χρειάζεται ο χρήστης να επιλέξει το διάγραμμα ομοιότητας στο οποίο θα ενσωματωθεί η εικόνα καθώς στην περίπτωση αυτή πρόκειται για ένα και μόνο διάγραμμα.

Εικόνα 33: Η φόρμα επεξεργασίας ετικετών σχολιασμού της Similarity Plot Analysis

3.11 Sequence clustering

Σε αρκετές περιπτώσεις κάποια γεγονότα ανασυνδυασμού μπορεί να διαφύγουν τον εντοπισμό από το T-RECs. Το πρόβλημα αυτό μπορεί να προκύψει όταν τα γεγονότα ανασυνδυασμού δεν είναι πρόσφατα. Κατά το πέρασμα του χρόνου, μια ανασυνδυασμένη ακολουθία συσσωρεύει μεταλλάξεις και μπορεί να διαφέρει αρκετά τόσο από τον πρωτεύοντα (major) όσο και από τον δευτερεύοντα (minor) δότη. Σαν αποτέλεσμα, το T-RECs για μια τέτοια ακολουθία, κατά τη σύγκριση των αποτελεσμάτων του blast μπορεί να εντοπίσει ως αμέσως καλύτερη ακολουθία μία παρόμοια με αυτή της επερώτησης όπου και οι δύο ανήκουν στην ίδια ομάδα. Έτσι, το γεγονός ανασυνδυασμού καλύπτεται καθώς βάσει των κριτηρίων του (εν. 3.4) και το T-RECs δεν θεωρεί πως υπάρχει ανασυνδυασμός.

Για την αντιμετώπιση του προβλήματος αυτού, το T-RECs ενσωματώνει ένα εργαλείο το οποίο βοηθά στη δημιουργία ομάδων ακολουθιών βάσει της ομοιότητάς τους. Ο χρήστης ορίζει ένα κατώφλι ομοιότητας και το πρόγραμμα ομαδοποιεί τις ακολουθίες οι οποίες έχουν ομοιότητα μεγαλύτερη από αυτό. Το εργαλείο ονομάζεται «Sequence clustering» και για τη δημιουργία των ομάδων χρησιμοποιεί το πρόγραμμα UCLUST. Κατά την ομαδοποίηση, το UCLUST δημιουργεί μια λίστα που περιέχει τις ακολουθίες που αντιπροσωπεύουν την κάθε ομάδα (μία για κάθε ομάδα). Έτσι, το T-RECs είναι σε θέση να χρησιμοποιήσει μόνο αυτές τις ακολουθίες φιλτράροντας τα παλιά γεγονότα και κρατώντας μόνο έναν αντιπρόσωπο από αυτά. Η πρόσβαση στο εργαλείο αυτό είναι εφικτή από το κουμπί «Sequence clustering» που βρίσκεται στο κάτω αριστερό μέρος του κεντρικού

μενού των επιμέρους αναλύσεων (εν. 3.10, εικόνα 24). Προτού ο χρήστης επιλέξει το κουμπί αυτό θα πρέπει να φορτώσει στο πρόγραμμα το αρχείο με τις ακολουθίες που επιθυμεί να ομαδοποιηθούν βάσει της ομοιότητάς τους. Η φόρτωση του αρχείου γίνεται από το κουμπί «Query file». Μόλις εισαχθεί αυτό το αρχείο και επιλεγεί το κουμπί «Sequence clustering» το κεντρικό μενού των επιμέρους αναλύσεων θα πάρει τη μορφή που φαίνεται στην εικόνα 34. Προαιρετικά, ο χρήστης μπορεί να φορτώσει και το αρχείο που αντιστοιχίζει τους accession numbers με τις εξελικτικές ομάδες για κάθε ακολουθία του αρχείου που φορτώθηκε προηγουμένως.

Εικόνα 34: Παράμετροι του εργαλείου Sequence clustering

Στο αριστερό μέρος της εικόνας 34 τοποθετούνται οι επιλογές που έχει ο χρήστης όσον αφορά τη λειτουργία του εργαλείου Sequence clustering. Συγκεκριμένα, ο χρήστης μπορεί να εισάγει στο πλαίσιο κειμένου «Identity Threshold (%)» μία τιμή που θα αντιστοιχεί στο χαμηλότερη τιμή ομοιότητας που επιτρέπεται να έχουν οι ακολουθίες που θα εισαχθούν στην κάθε ομάδα. Αφού εισαχθεί αυτή η τιμή, ο χρήστης μπορεί να εκκινήσει τη διαδικασία πατώντας το κουμπί «Start». Αμέσως, θα ξεκινήσει να εκτελείται το UCLUST ενώ στο δεξιό μέρος της φόρμας θα εμφανίζεται η πρόοδός του. Σημειώνεται πως σε περίπτωση που ο χρήστης θελήσει να κλείσει τη φόρμα των παραμέτρων του Sequence clustering αρκεί να επιλέξει το κουμπί «X» που βρίσκεται στο πάνω αριστερό μέρος της.

3.11.1 Similarity Plot Analysis

Μόλις ολοκληρωθεί η διεργασία του UCLUST θα εμφανιστεί η φόρμα αποτελεσμάτων της ομαδοποίησης όπως φαίνεται στην εικόνα 35.

Εικόνα 35: Φόρμα αποτελεσμάτων ομαδοποίησης

Για την ευκολότερη κατανόηση της φόρμας αποτελεσμάτων ομαδοποίησης θα χωριστεί σε τρία μέρη: Καρτέλα Centroids, Καρτέλα Cluster members και Καρτέλα Cluster Information.

Καρτέλα Centroids

Η καρτέλα αυτή περιέχει μία λίστα με τους αντιπροσώπους (Centroids) της κάθε ομάδας. Στο αριστερό μέρος αναγράφονται οι accession numbers ενώ στο δεξιό τα ονόματα των εξελικτικών ομάδων εφόσον είναι διαθέσιμα.

Καρτέλα Cluster members

Κάθε φορά που επιλέγεται από τον χρήστη ένα στοιχείο της καρτέλας Centroids, η καρτέλα Cluster members γεμίζει με τους accession numbers και τα ονόματα των εξελικτικών ομάδων κάθε ακολουθίας που περιέχεται στην ομάδα με αντιπρόσωπο το επιλεγμένο στοιχείο από την καρτέλα Centroids.

Καρτέλα Cluster Information

Η καρτέλα αυτή περιέχει πληροφορίες τόσο για την επιλεγμένη ομάδα όσο και για την επιλεγμένη ακολουθία της καρτέλας Cluster members. Οι πληροφορίες αυτές είναι: 1) Cluster number, που υποδηλώνει τον αριθμό της ομάδας. 2) Cluster size, που υποδηλώνει το συνολικό αριθμό των μελών της ομάδας. 3) Centroid size, που

υποδηλώνει το μέγεθος της ακολουθίας που αντιπροσωπεύει την ομάδα αυτή. 4) % Similarity with Centroid, που υποδηλώνει την επί τοις εκατό ομοιότητα της ακολουθίας του επιλεγμένου στοιχείου από την καρτέλα Cluster members με την ακολουθία του αντιπροσώπου της ομάδας. Η ομοιότητα της κάθε ακολουθίας που βρίσκεται στην επιλεγμένη ομάδα παρουσιάζεται και γραφικά με ένα διάγραμμα όπου στον άξονα Χ τοποθετούνται οι τιμές ομοιότητας ενώ στον άξονα Υ οι accession numbers με τα ονόματα των εξελικτικών ομάδων κάθε ακολουθίας.

Export Centroids Fasta

Από τη φόρμα αποτελεσμάτων ομαδοποίησης ο χρήστης έχει τη δυνατότητα να δημιουργήσει ένα αρχείο που θα περιέχει μόνο τις ακολουθίες των αντιπροσώπων κάθε ομάδας σε μορφή fasta. Για να γίνει αυτό θα πρέπει να επιλέξει το κουμπί «Export Centroids Fasta» και να υποδείξει στη συνέχεια το σημείο όπου θα αποθηκευτεί το αρχείο αυτό.

Export Annotation file

Το κουμπί αυτό δίνει τη δυνατότητα δημιουργίας ενός αρχείου αντιστοίχισης ονομάτων εξελικτικών ομάδων με accession numbers αλλά ως ονόματα χρησιμοποιεί τον αριθμό της ομάδας στην οποία βρίσκεται η κάθε ακολουθία. Έτσι, ο χρήστης μπορεί να επιλέξει αυτό το κουμπί και στη συνέχεια να υποδείξει το σημείο όπου θα γίνει η αποθήκευση αυτού του αρχείου.

Use Centroids as Query file

Όταν ενεργοποιείται η επιλογή «Use Centroids as Query file», οι ακολουθίες των αντιπροσώπων κάθε ομάδας φορτώνονται αυτόματα στο κεντρικό μενού επιμέρους αναλύσεων του T-RECs ως ακολουθίες επερώτησης.

Use Centroids as Database file

Όταν ενεργοποιείται η επιλογή «Use Centroids as Database file», οι ακολουθίες των αντιπροσώπων κάθε ομάδας φορτώνονται αυτόματα στο κεντρικό μενού επιμέρους αναλύσεων του T-RECs ως ακολουθίες βάσης δεδομένων.

OK

Το κουμπί «OK» κλείνει τη φόρμα των αποτελεσμάτων της ομαδοποίησης

3.12 Sequence editor

Όπως έχει ήδη αναφερθεί, το T-RECs χρησιμοποιεί το πρόγραμμα BLASTN. Τα αρχεία εισαγωγής αυτού του προγράμματος τα οποία περιέχουν τις ακολουθίες επερώτησης και αυτές της βάσης δεδομένων αντίστοιχα θα πρέπει να είναι σε μορφή fasta. Επίσης, θα πρέπει η κάθε ακολουθία να βρίσκεται σε μία γραμμή αλλά

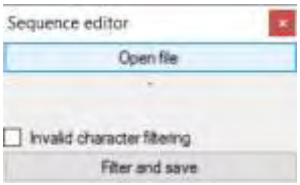
και να μην περιέχει μόνο έγκυρους χαρακτήρες όπως φαίνεται στην εικόνα 36. Ως έγκυροι χαρακτήρες θεωρούνται οι: A, T, G, C, a, t, g και c. Οποιοσδήποτε άλλος χαρακτήρας αναγνωρίζεται ως μη έγκυρος (invalid character) από το ίδιο το blastn και κατά συνέπεια και από το T-RECs. Στο σημείο αυτό θα πρέπει να σημειωθεί ότι ορισμένες ακολουθίες περιέχουν μη έγκυρους χαρακτήρες οι οποίοι παρουσιάζονται στον πίνακα 2.

Το T-RECs, με σκοπό τη διευκόλυνση του χρήστη, περιλαμβάνει ένα εργαλείο που τροποποιεί κατάλληλα τα αρχεία εισόδου έτσι ώστε να διορθωθούν πιθανές μη έγκυρες μορφές και δημιουργεί νέα αρχεία εισόδου που πληρούν τις παραπάνω προϋποθέσεις. Το εργαλείο ονομάζεται «Sequence editor» και ο χρήστης έχει πρόσβαση σε αυτό πατώντας το ομώνυμο κουμπί που βρίσκεται στο κάτω αριστερό μέρος του κεντρικού μενού των επιμέρους αναλύσεων (εν. 3.10, εικόνα 24). Μόλις ο χρήστης επιλέξει το συγκεκριμένο κουμπί θα εμφανιστεί η φόρμα φιλτραρίσματος όπως φαίνεται στην εικόνα 37.

Εικόνα 36: 1) Η μορφή που θα πρέπει να έχουν οι ακολουθίες που περιέχονται στο αρχείο με τις ακολουθίες επερώτησης και σε αυτό με τις ακολουθίες της βάσης δεδομένων. 2) Η ακολουθία έχει μη έγκυρη μορφή καθώς περιέχει μη έγκυρους χαρακτήρες (N, W και K) οι οποίοι απεικονίζονται με πιο έντονη γραφή. 3) Η ακολουθία έχει μη έγκυρη μορφή καθώς δεν είναι συνεχής αλλά χωρίζεται σε τρεις γραμμές.

Μη έγκυροι χαρακτήρες ακολουθιών	
Χαρακτήρας	Επεξήγηση
R	A ή G
Y	C ή T
S	G ή C
W	A ή T
K	G ή T
M	A ή C
B	C ή G ή T
D	A ή G ή T
H	A ή C ή T
V	A ή C ή G
N	A ή T ή G ή C
. ή -	κενό

Πίνακας 2: Χαρακτήρες που μπορεί περιέχονται σε ακολουθίες αλλά θεωρούνται ως μη έγκυροι τόσο από το blastn όσο και από το T-RECs



Εικόνα 37: Φόρμα επεξεργασίας ακολουθιών οι οποίες έχουν μη έγκυρη μορφή.

Open file

Ο χρήστης θα πρέπει να φορτώσει το αρχείο που περιέχει τις ακολουθίες οι οποίες έχουν μη έγκυρη μορφή. Για το σκοπό αυτό θα πρέπει να επιλεγεί το κουμπί με όνομα «Open file» και στη συνέχεια να υποδείξει στο πρόγραμμα τη θέση στην οποία βρίσκεται το επιθυμητό αρχείο.

Invalid character filtering

Μέσω της φόρμας επεξεργασίας ακολουθιών, δίνεται στο χρήστη η δυνατότητα να επιλέξει την αφαίρεση των μη έγκυρων χαρακτήρων ενεργοποιώντας την επιλογή «Invalid character filtering».

Filter and save

Επιλέγοντας το κουμπί με τίτλο «Filter and save» ο χρήστης θα είναι σε θέση να δημιουργήσει ένα νέο αρχείο το οποίο θα αποτελεί μια τροποποιημένη έκδοση του

αρχείου που εισήχθη κατά την επιλογή του κουμπιού «Open file». Η νέα έκδοση θα περιέχει τις ακολουθίες σε μία γραμμή ενώ σε περίπτωση που ο χρήστης έχει ενεργοποιήσει και την επιλογή «Invalid character filtering» οι ακολουθίες θα είναι απαλλαγμένες και από τους μη έγκυρους χαρακτήρες.

Το νέο αρχείο που θα δημιουργηθεί θα είναι πλέον κατάλληλο ώστε να φορτωθεί σε οποιαδήποτε από τις αναλύσεις που παρέχει το T-RECs ως αρχείο ακολουθιών επερώτησης ή/και βάσης δεδομένων.

4 Συζήτηση

Περισσότερα από 60 εργαλεία βιοπληροφορικής έχουν αναπτυχθεί τα τελευταία 20 χρόνια, τα οποία έχουν τη δυνατότητα να εντοπίζουν γεγονότα ανασυνδυασμού [27, 29, 41]. Αυτά τα εργαλεία βασίζονται σε τέσσερεις μεγάλες κατηγορίες μεθόδων όπως η σύγκριση ακολουθιών κατά ζεύγη, η πληθυσμιακή γενετική, η φυλογενετική και τα μοτίβα περιοχών. Ορισμένες μέθοδοι βασίζονται σε τοπική στοίχιση κατά ζεύγη με χρήση συρόμενου παραθύρου και έχουν τη δυνατότητα να αναλύουν με έναν απλό και γρήγορο τρόπο πληθώρα γενομικών δεδομένων. Για το λόγο αυτό η χρήση τους είναι ιδανική σαν ένα πρώτο βήμα σε μια στρατηγική εντοπισμού ανασυνδυασμών με πολλές μεθόδους. Ένα διαθέσιμο υπολογιστικό εργαλείο που βρίσκεται στην κατηγορία αυτή είναι το εργαλείο γενοτύπησης του NCBI (NCBI genotyping tool) που βασίζεται στο BLAST [42]. Ο χρήστης μπορεί να ορίσει μια δική του βάση δεδομένων ακολουθιών ενώ τα αποτελέσματα μπορούν να αναλυθούν/οπτικοποιηθούν περαιτέρω με διαγράμματα ομοιότητας. Προς το παρόν, αυτό το εργαλείο δε μπορεί να αναλύσει πάνω από μία ακολουθία επερώτησης κάθε φορά. Ένα άλλο υπολογιστικό εργαλείο της ίδιας κατηγορίας είναι το SWE-Blast, ένα εργαλείο που λειτουργεί μέσω της γραμμής εντολών. Το εργαλείο αυτό ξεκινάει σπάζοντας την ακολουθία επερώτησης σε κομμάτια με ένα συρόμενο παράθυρο, έπειτα καταθέτει κάθε κομμάτι στο NCBI blast και στη συνέχεια αποθηκεύει τα αποτελέσματα σε αρχεία κειμένου [43]. Η χρήση φυλογενετικών δέντρων και διαγραμμάτων ομοιότητας [44] είναι οι πιο διαδεδομένες μέθοδοι. Ένα κοινό πρόβλημα κατά τον εντοπισμό γεγονότων ανασυνδυασμού με φυλογενετικές αναλύσεις είναι ότι θα πρέπει να καθοριστούν τα όρια των περιοχών του γονιδιώματος που θα αναλυθούν. Αν το σημείο του ανασυνδυασμού βρίσκεται μεταξύ των ορίων δύο περιοχών, τότε η ακολουθία αυτή ενδέχεται να ταξινομηθεί λανθασμένα με άλλα τμήματα ακολουθιών στο δέντρο και συνήθως με χαμηλές τιμές bootstrap. Ένας τρόπος για να αντιμετωπισθεί το πρόβλημα αυτό είναι να χρησιμοποιηθεί η μέθοδος Bootscan [28] αλλά αν ο αριθμός και το μήκος των ακολουθιών είναι πολύ μεγάλος, τότε η ανάλυση γίνεται προβληματική όσον αφορά το χρόνο και την πολυπλοκότητά της. Από την άλλη πλευρά, τα διαγράμματα ομοιότητας επιτρέπουν τον εντοπισμό του σημείου ανασυνδυασμού αλλά το πρόβλημά τους έγκειται στον περιορισμένο αριθμό ακολουθιών που μπορούν να χρησιμοποιηθούν κάθε φορά. Το RDP είναι ένα πολύ δημοφιλές πρόγραμμα το οποίο συνδυάζει πολλές μεθόδους εντοπισμού ανασυνδυασμών αλλά προϋποθέτει ότι οι ακολουθίες που εισάγονται σε αυτό θα πρέπει να είναι στοιχισμένες [45, 46]. Έτσι, οι τρέχουσες μέθοδοι δεν είναι βελτιστοποιημένες για ανάλυση εκατοντάδων ή ακόμη και χιλιάδων ολόκληρων ιικών γονιδιωμάτων με ένα απλό βήμα και/ή σε γραφικό περιβάλλον φιλικό προς τον χρήστη. Το T-RECs είναι ένα εργαλείο με γραφικό περιβάλλον που δημιουργήθηκε με τη γλώσσα προγραμματισμού Visual Basic και η λειτουργία του

βασίζεται στο blast, το muscle και το Usearch/Uclust ενώ έχει τη δυνατότητα να αναζητά γεγονότα ανασυνδυασμού ανάμεσα σε εκατοντάδες ή ακόμη και χιλιάδες ιικά γονιδιώματα ή κομμάτια ακολουθιών. Βασίζεται στην στοίχιση ακολουθιών κατά ζεύγη του BLASTN με τη χρήση συρόμενων παραθύρων στις ακολουθίες επερώτησης έναντι μιας προκαθορισμένης από τον χρήστη βάσης δεδομένων ακολουθιών. Οι ακολουθίες επερώτησης μπορούν επίσης να γενοτυπηθούν ή να ομαδοποιηθούν με βάση την ομοιότητά τους με άλλες ακολουθίες. Τα εντοπισμένα πιθανά γεγονότα ανασυνδυασμού μπορούν επίσης να μελετηθούν γραφικά με τη χρήση διαγραμμάτων ομοιότητας μέσα από το T-RECs ή να καταχωρηθούν στην ιστοσελίδα του NCBI blast για περαιτέρω ανάλυση. Το T-RECs συνίσταται ως ένα υψηλής απόδοσης εργαλείο φιλτραρίσματος για γεγονότα ανασυνδυασμού που μπορούν να αναλυθούν περαιτέρω με άλλα πιο εξειδικευμένα προγράμματα όπως Simplot, RDP, HyPhy, GARD, Bootscan, Maxchi, Chimaera, 3SEQ, VISRD, LARD, Siscan και άλλα, [11, 28, 35, 36, 37, 38, 44, 45, 46, 47, 48, 49]. Η επαλήθευση της σωστής λειτουργίας του T-RECs έγινε με τη χρήση του RDP4 χρησιμοποιώντας δεδομένα από τη διπλωματική εργασία του κ. Γιώργου Μπαχούμη. Ουσιαστικά, τα περισσότερα πλεονεκτήματα του NCBI Genotyping tool, του SWE-Blast και του Simplot, ενσωματώνονται στο T-RECs και προσαρμόζονται με τέτοιο τρόπο ώστε να μπορούν να ανταπεξέλθουν με το βέλτιστο δυνατό τρόπο στην πρόκληση που εισάγει η πληθώρα δεδομένων που παράγεται από την αλληλούχιση νέας γενιάς (Next Generation Sequencing).

Παράρτημα 1 – Αρχεία Εισόδου/Εξόδου

Αρχεία εισόδου

Query file

Αρχείο ακολουθιών επερώτησης. Μπορεί να περιέχει μία ή περισσότερες ακολουθίες επερώτησης σε μορφή fasta. Οι ακολουθίες θα πρέπει να είναι νουκλεοτιδικές και να βρίσκονται σε μια γραμμή. Σε περίπτωση που περιέχονται παραπάνω από μία ακολουθίες θα πρέπει να είναι ταξινομημένες η μία κάτω από την άλλη χωρίς κενές γραμμές. Το όνομα της κάθε ακολουθίας δεν θα πρέπει να περιέχει κενά. Τέλος, οι μόνοι επιτρεπτοί χαρακτήρες που μπορούν να περιέχονται στις ακολουθίες είναι οι A, T, G και C. Παρακάτω φαίνεται ένα παράδειγμα με το περιεχόμενο ενός αρχείου ακολουθιών επερώτησης:

```
>Ακολουθία_1
ACATGCATCGTACGTACGTAGCTATCAGTCGATCGTAGCTAGCGACGACTCAGGTGCAGTC
>Ακολουθία_2
CAGTCAGTTGCGACGAGTCGATGCATGCTAGCTACGTATGCGATCGATCGATCGAGCATCA
>Ακολουθία_3
CGTAGTCAGTCGATCGATGCTAGCTAGCTAGTCGCTAGCTGTCGTAGCTAGTCAGTCGAGG
```

Database file

Αρχείο ακολουθιών της βάσης δεδομένων. Η μορφή του θα πρέπει να είναι όπως αυτή που περιγράφηκε παραπάνω για το Query file.

Evolutionary groups file

Αρχείο αντιστοίχισης accession number κάθε ακολουθίας με το όνομα της εξελικτικής ομάδας στην οποία αυτή ανήκει. Πρόκειται για ένα απλό αρχείο κειμένου όπου στην κάθε σειρά αναγράφεται ο accession number (όπως ακριβώς είναι γραμμένος για κάθε ακολουθία στο query και το database file) και το όνομα της εξελικτικής ομάδας. Οι δύο αυτές τιμές θα πρέπει να χωρίζονται με ένα TAB. Τα ονόματα των εξελικτικών ομάδων δε θα πρέπει να περιέχουν κενά. Παρακάτω φαίνεται ένα παράδειγμα περιεχομένων του αρχείου evolutionary groups:

Accession_number_1	Ομάδα_1
Accession_number_2	Ομάδα_2
Accession_number_3	Ομάδα_3
Accession_number_4	Ομάδα_3
Accession_number_5	Ομάδα_1
Accession_number_6	Ομάδα_2

Annotation file

Αρχείο κειμένου το οποίο περιέχει τα στοιχεία για τη δημιουργία ετικετών σχολιασμού για ένα συγκεκριμένο πιθανό γεγονός ανασυνδυασμού που έχει εντοπιστεί από το T-RECs. Περιέχει 6 στήλες ενώ η κάθε γραμμή του αντιστοιχεί σε μία ετικέτα σχολιασμού. Τα ονόματα των στηλών είναι: Sequence ID, Motif Name, Motif Start, Motif end και Sequence to use as reference point και περιγράφονται αναλυτικά στην ενότητα 3.9. Παρακάτω φαίνεται ένα παράδειγμα με τα περιεχόμενα ενός αρχείου σχολιασμού.

EU921388	p48	5	994	p48	EU921388
EU921388	NTP	994	2092	NTP	EU921388
EU921388	p22	2092	2629	p22	EU921388
EU921388	VPG	2629	3028	VPG	EU921388
EU921388	3C	3028	3571	3C	EU921388
EU921388	RdRp	3571	5104	RdRp	EU921388
EU921388	VP1	5085	6707	VP1	EU921388
EU921388	VP2	6707	7513	VP2	EU921388

Οι στήλες θα πρέπει να χωρίζονται μεταξύ τους με ένα TAB.

Αρχεία εξόδου

Blast results file

Πρόκειται για το αρχείο κειμένου που προκύπτει ύστερα από την εκτέλεση των Blast Based αναλύσεων που παρέχει το T-RECs. Αποτελείται από 4 περιοχές: Την περιοχή επικεφαλίδας, την περιοχή <WARNINGS>, την περιοχή <WINDOW_STEP> και την περιοχή <MAJOR_DONORS>.

• Περιοχή επικεφαλίδας
Η περιοχή αυτή περιέχει την επικεφαλίδα για κάθε στήλη των αποτελεσμάτων του blast. Οι επικεφαλίδες και η σημασία τους περιγράφονται παρακάτω:

Q-ORGANISM: Query evolutionary group
Q-ACCESSION: Query sequence ID
S-ORGANISM: Subject evolutionary group
S-ACCESSION: Subject sequence ID
E-VALUE: Expect value
Q-START: Start of alignment in query
Q-END: End of alignment in query
S-START: Start of alignment in subject
S-END: End of alignment in subject
BITSCORE: Bit score
Q-LENGTH: Query sequence length
S-LENGTH: Subject sequence length

PIDENT: Percentage of identical matches
PPOS: Percentage of positive-scoring matches

Υπάρχουν ακόμη 12 στήλες με τίτλο: S-ORGANISM, S-ACCESSION, E-VALUE, Q-START, Q-END, S-START, S-END, BITSCORE, Q-LENGTH, S-LENGTH, PIDENT και PPOS στο τέλος. Αυτές οι στήλες αναφέρονται στα αποτελέσματα του Blast για το καλύτερο blast hit όπου η ακολουθία Query και Subject ανήκουν στην ίδια εξελικτική ομάδα.

• Περιοχή <WARNINGS>
Αυτή η περιοχή είναι ορατή μόνο όταν υπάρχουν ενδείξεις για ψευδώς θετικά αποτελέσματα. Περιέχει μια λίστα με τις IDs των ακολουθιών που θεωρούνται ως ψευδώς θετικά αποτελέσματα.

• Περιοχή <WINDOW_STEP>
Η περιοχή αυτή αναφέρει το μέγεθος παραθύρου και βήματος που χρησιμοποιήθηκαν στην τρέχουσα ανάλυση.

• Περιοχή <MAJOR_DONORS>
Η τελευταία περιοχή περιλαμβάνει τους πρωτεύοντες δότες για κάθε πιθανό γεγονός ανασυνδυασμού. Κάθε γραμμή περιέχει την ID της ακολουθίας επερώτησης και την αντίστοιχη ID του πρωτεύοντα δότη. Επίσης περιέχει τις θέσεις όπου ο πρωτεύοντας δότης είναι το καλύτερο blast hit στα αποτελέσματα του blast.

Παράρτημα 2 – Παράμετροι του Blast

e-value

Αποτελεί το κατώφλι στατιστικής σημαντικότητας για την αναφορά των matches έναντι των ακολουθιών της βάσης δεδομένων: Η προκαθορισμένη τιμή είναι $1e^{-10}$ που σημαίνει ότι $1e^{-10}$ matches αναμένεται να εντοπισθούν απλώς κατά τύχη, σύμφωνα με το στοχαστικό μοντέλο των Karlin και Altschul (1990). Αν η στατιστική σημαντικότητα που δέχθηκε ένα match είναι μεγαλύτερη από το αναμενόμενο κατώφλι, τότε το match δεν θα αναφερθεί. Χαμηλές e-values είναι περισσότερο αυστηρές, οδηγώντας έτσι στην αναφορά λιγότερων matches.

Gap Open

Αποτελεί το κόστος για το άνοιγμα ενός κενού.

Gap Extend

Αποτελεί το κόστος για την επέκταση των κενών.

Penalty

Ποινή για ένα mismatch.

Reward

Τιμή που θα δίνεται για κάθε match.

Word Size

Αποτελεί το μέγεθος της αρχικής λέξης/ακολουθίας που θα πρέπει να ταιριάξει ανάμεσα στην ακολουθία επερώτησης και στη συγκρινόμενη ακολουθία της βάσης δεδομένων.

Παράρτημα 3 – Help page όπως εμφανίζεται στο T-RECs

Pipeline Analysis

By clicking this option you will be able To perform an automated recombination analysis. You need to specify a Query file, a Database file and an Evolutionary Groups file (See Input Files section). As soon as you load these files you will also need to specify a directory of your choice where the output files will be exported. Then, the START button will be enabled and the analysis will begin When you click it. On the right side of this form you will be able to monitor the progress of the analysis.

The stages of the analysis are the following:

• Searching for recombinations
On this first stage the software is performing a Blast Based Intergroup Recombination Analysis.

• Saving Results
When the first search for recombinations is finished the results will be saved on a plain text file to the directory that you have specified.

• Clustering Database Sequences
The sequences contained in the Database File are clustered by 99% identity

• Searching for recombinations in clustered sequences
Another Blast Based Intergroup Recombination Analysis takes place using the clustered sequences as the Database File.

• Saving Results
The new results are saved in another plain text file

When the analysis is done you will be able to view the results of the Blast based Intergroup Recombination Analysis before and after clustering by clicking the "Show results before clustering" and "Show results after clustering" buttons respectively.

You can then analyze the results as described on Blast based Intergroup Recombination Analysis section.

Manual Analysis

By clicking this option you will be able to perform manually any of the following types of analyses:

Blast Based Genotyping Tool
Blast Based Intergroup Recombination Analysis
Similarity Plot Analysis

Manual analysis does also include the "Sequence editor" and the "Sequence clustering" tools.

Blast Based Intergroup Recombination Analysis

This is the main recombination analysis that is performed by T-RECS. The software searches for possible recombination events through the supplied Query sequence(s) using blast against the supplied Database sequence(s).

Blast

This is the Blast window where you can change the parameters, the input files or load a results file. This window includes 3 Tabs:

• Blast

• Sequence Tool

• Results

Blast Tab

Here you can change the input files that you had previously loaded (in case you want to) and the parameters that will be used for the analysis.
The two most important parameters you should always take into consideration are the "Window" and the "Step" parameters:
The window parameter represents the size of the fragments that each query sequence will be cut into. These fragments will be used as query sequences and will be blasted against the Database sequence(s). Ensure that you do not enter a window size larger than the size of the query sequence(s).
The step parameter indicates the step that will be used in order to split the query sequence(s) into fragments.

When you are ready to proceed click the Blast button (lower left corner) and the blast process will start. T-RECS will inform you about the progress of the analysis by showing the number of fragments that have already been blasted. However, this may increase the total time of the analysis. You have the option to disable/enable the viewing of the progress by unchecking/checking the "Display Progress" check box.
You can cancel the process at any time by clicking the Cancel button.

Please note that the blast process may take a long time, according to your system performance and the size of the Query and Database files.

• Show Results Button
As soon as the blast process is completed, the Results Window will appear. If you accidentally close this window you can restore it by clicking the "Show Results" button.

• Load Results Button
If you have already performed a Blast Based Intergroup Recombination Analysis and you have saved your results, you can load the results file by clicking this button and specifying its location.

Sequence Tool Tab

On this tab you can view information about a sequence contained in a sequence file. You need to load a fasta sequence file by clicking the "Load sequences file" button and a file containing the IDs for each one of those sequences. For more information about the format of the files see the Input Files section.

After loading the required files you will be able to search for an ID by typing its name in the "ID to search" text box and clicking the "Search" button. The sequence will appear on the top textbox. You can either view the whole sequence or specify a part of the sequence that will be shown. To specify a part of the sequence check the "Show sequence from position:" button and specify the starting and ending points of the part of the sequence that you want to view.

Additional sequence information is shown on the right side of this tab.

Results Tab

On this tab you can view the Blast Based Intergroup Recombination Analysis results in text format (See the Output Files section for more information about the results file).

As soon as the blast process is finished, this tab will be filled with the analysis results. You can save the results by clicking the "Save" button.

You can also search through the results for a keyword by defining a keyword and the column through which the keyword will be searched. To start the search click the "Search" button.

Results Window

This window appears after blast process is done. It consists of 3 sections:

• Results Panel (Top panel)

• Possible Acceptors list (Lower left side)

• Options Panel (Bottom panel that contains 3 tabs: General, Similarity Plot, Annotation editor)

Here you can view, edit and save the T-RECS results.

Results Panel

The selected possible acceptor's sequence is shown on the top of this panel whereas the blast results are visualized below this sequence. Above the query sequence, the major donor sequence will appear with pink color. Note that only the major donor's fragments which had the best blast hit will appear. By moving the mouse in the white area you can view the current location of the query sequence on the options panel. Additional information is shown on the options panel if you move the mouse over any displayed sequence. By right clicking on any of the visualized T-RECS results you can change their display color.

Possible Acceptors list

This list contains the accession numbers of every query sequence that was detected by T-RECS as a possible recombination event. Click any of the contained accession numbers to view the results.

If there are sequences in the query sequence file that you have loaded which have a unique evolutionary group name, they may also appear in the Possible Acceptors list. In this case, when you select one of those a red warning label will appear to indicate that this event could probably be a false positive event.

"Filtering by group names" check box will filter the Possible Acceptors list by a specified evolutionary group. If you check this box a list with all the available evolutionary groups (that are represented in the Possible Acceptors list) will appear. You will then be able to select one of them and the Possible Acceptors list will change so that only the accessions that correspond to this evolutionary group are shown. If you uncheck this box the Possible Acceptors list will be restored.

You can also search for a specific accession number through this list by typing it in the search box and clicking the "Search" button.

Options Panel

This panel offers a variety of options for viewing, editing and analyzing the results.

General Tab

On the top of this panel you will be able to view information of the sequence that you have the mouse pointer on the Results Panel.
As you move the mouse pointer on the Results panel, the exact location of the query sequence will be displayed on the Query Position label. If the sequences are aligned and you have the mouse pointer over a possible donor sequence, the alignment position of the query will be displayed on the Alignment Position label.

• Smart Filtering by largest size

When this box is checked, the donor fragments (shown on the Results Panel) will be compared to each other. If there is a coverage among these fragments, only the one with the largest size will be shown on the results.

• Export Image button
By clicking on this button you will be able to save the Results Panel as an image file. Annotation of the sequences will also be included on the image.

• Save blast button
By clicking this button you will be able to save the T-RECS results in txt format.

• Export Acceptors and Donors in fasta
With this option you can export a file for each possible acceptor contained in the list which will include: The possible acceptor sequence, the major donor sequence, each possible donor sequence and the sequence(s) from the same evolutionary group with the query sequence that were found as best blast hits after each possible donor.

Similarity Plot Tab

Through this tab you are able to perform a similarity plot analysis of the displayed results. The sequences that will be used for the construction of the similarity plot are listed in the "Sequences to be included" list. When you click an event from the Possible Acceptors list, the "Sequences to be included" list is filled with the possible donors of this event. In addition, sequences that were found as the best blast hits of the same evolutionary group as the query for each possible recombination fragment, are also included. The query sequence is included for the construction of the similarity plot but it is not listed in that list.

You have the option to choose which of those sequences will be used for the construction of the similarity plot by simply checking or unchecking them. You can also add more sequences by clicking on the green cross button and entering the accession number of the desired sequence.

Parameters:

• Step
The value of the step that will be used for the construction of the similarity plot

• Window
The value of the sliding window that will be used for the construction of the similarity plot

• Gap options
You can specify whether the score of a Gap to a Non-Gap point comparison will be considered as a match or a mismatch.

• Ignore Gaps check box
When this box is checked, the Gap points will not be taken into account for the construction of the similarity plot.

• Run Similarity Plot
By clicking this button the similarity plot construction process will begin.

Note 1
As soon as you click the Run Similarity Plot button, if your database file contains sequences of the same evolutionary group as the query that were not included in the "Sequences to be included" list, an option box will appear. On this box you will be informed about the number of those sequences and you will have the option to include as many of them as you like. You will also have the option to choose the similarity plot line color of the sequences that have the same evolutionary group with the query.

Note 2
You may be prompted to align the included sequences with muscle. This step is required for the proper construction of the similarity plot.

On the similarity plot that will appear click the "Help" button for more help on how to handle the similarity plot window.

Annotation Editor Tab

On this tab you can create, load Or edit annotation which you can Then attach On a similarity plot (See Input Files/Annotation File For more information about the format Of the annotation).

• Load button
Click this button And specify an annotation file that you want To load.

• Clear button
Click this button To clear the current annotation panel.

• Save button
Click this button To save the current annotation panel In txt format.

• Annotate/Show Data button
Click the Annotate button To visualize the annotation. Click the Show Data button To Return To the annotation panel. (This button Is visible only When you have clicked the Annotate Button)

• Save image button

Click this button To save an image Of the displayed annotation. (This button Is visible only When you have clicked the Annotate Button)

• Search Next button
Specify a keyword And a column through which this keyword will be searched And click this button To find the Next occurrence Of the keyword.

• Show annotation On similarity plot button
Before you click this button you should Select a similarity plot On which you want To pin the annotation image that you have created. You can specify the desired similarity plot from the drop list above this button.

• Only show query sequence On similarity plot check box
Check this box If you want To exclude the possible donor sequence(s) from the annotation image that will be attached On the selected similarity plot.

Blast Based Genotyping Tool

This tool provides the option to identify the possible evolutionary group of the query sequence(s) that you provide. The software uses a sliding window to break the provided Query sequence(s) into fragments and searches for the closest sequence for each fragment using blast against the provided Database sequence(s). When the process is finished T-RECS will visualize the results and will give you the option to save them as an evolutionary groups file.

The user interface of this tool is similar to the Blast Based Intergroup Recombination Analysis user interface (For more help on the user interface see Blast Based Intergroup Recombination Analysis section).

The main user interface difference is that on the Results Window --> Options Panel --> General Tab, a Genotyping Results box is now visible. This box contains the genotyping results for the selected Query from the Query list. It shows the total coverage percentage for each genotyping group that is represented at least by one sequence in the blast results. You can save the genotyping results for all of the queries by clicking on the "Save Genotyping Results" button. The Genotyping Results will be saved as an evolutionary groups file.

Similarity Plot Analysis

This type of analysis gives you the option to construct similarity plots using sequences of a Query Sequences file that you provide. The user interface of this type of analysis provides a list where you can check/uncheck the accession numbers of the sequences that you want to include/exclude for the construction of the similarity plot.

• Search box
In this box you can type an accession number which will be automatically searched through the list and will be highlighted if found.

• Select/Deselect All buttons
Select or deselect all of the listed sequences by clicking on the corresponding button.

• OK button
When you have selected the accession numbers of the sequences that you want to use for the construction of the similarity plot, click this button to lock your preference and enable the similarity plot parameters.

• Query list
On this list you need to specify one of the included sequences to be used as the query sequence for the similarity plot.

• Step text box
Enter a step value for the construction of the similarity plot.

• Window text box
Enter a sliding window value for the construction of the similarity plot.

• Annotation button
This button will be enabled after the construction of the similarity plot. Its function is to provide an annotation form for the similarity plot (For more information about the user interface of this annotation form see Blast Based Intergroup Recombination Analysis --> Results Window --> Options Panel --> Annotation Editor Tab section).

• Ignore Gaps check box
When this box is checked, the Gap points will not be taken into account for the construction of the similarity plot.

• Create Similarity Plot button
By clicking this button the similarity plot construction process will begin.

Note 2
You may be prompted to align the included sequences with muscle. This step is required for the proper construction of the similarity plot.

On the similarity plot that will appear click the "Help" button for more help on how to handle the similarity plot window.

Sequence Editor

This tool will help you to edit the fasta sequence(s) input file(s) so that they are compatible to run with T-RECS. The function of this tool is to trim any new line

characters of the provided sequences so that they appear in one line. In addition you will have the option to remove invalid characters from the sequences.

First, you need to click on the "Sequence Editor" button. On the box that will appear you need to load the fasta file that you want to edit by clicking the "Open file" button. Once you do that, you have the option to enable/disable the invalid character filtering by checking/unchecking the corresponding check box. If this check box is checked, any sequence character other than A, T, G or C will be removed from the sequence. Note that in order for T-RECS to work properly, invalid sequence characters should be removed.

Example: Sequence before editing

```
>ACCESSION_NUMBER
ACTAG----TACGWNWCAT
CATCGTGTNN-N
CATCGATCAGTGTCGATCGTAGGT
```

Example: Sequence after editing

```
>ACCESSION_NUMBER
ACTAGTACGCATCATCGTGTCATCGATCAGTGTCGATCGTAGGT
```

Sequence Clustering

Through this tool you can cluster sequences contained into a fasta file based on a identity cutoff. First you have to load the fasta file. You can do this by loading the file through the "Query File" button on the main menu. In order to get extended information for the clusters that will be created you may also load an Evolutionary Groups file through the corresponding button on the main menu. Once you load the files, click on the "Sequence Clustering" button and on the panel that will appear define an Identity Threshold for the creation of the clusters. Then you can start the clustering process by pressing the start button.

Clustering Results Window

When the clustering process is completed, a new window will appear containing the clustering results. It consists of 3 sections: Centroids (x/y), Cluster members and Cluster Information.

• Centroids (x/y)
This section contains the centroid of each cluster. Inside the parenthesis, x stands for the number of centroids and y stands for the total number of the sequences that were loaded.

• Cluster members

When you select a centroid from the Centroids section, the Cluster members section will list all of the members of the cluster that is represented by the selected centroid.

• Cluster Information
This section includes various information for the selected cluster: Cluster number, Cluster size, Centroid size and a graph where the similarity of each cluster member with the centroid is illustrated.

On the bottom of this window more options are included:

• Export Centroids Fasta button
By clicking this button you will be able to save the sequences of each cluster in single fasta format files. You only need to specify the output directory.

• Export Annotation file button
By clicking this button you will be able to create a new Evolutionary Groups file using as an evolutionary group the number of the cluster that each sequence is contained.

• Use Centroids as Query file check box
If you check this box the Centroid sequences will be automatically loaded as a query file on the main menu.

• Use Centroids as Database file check box
If you check this box the Centroid sequences will be automatically loaded as a Database file on the main menu.

• OK button
Click this button to close the Clustering Results window.

Input Files

On this section the format of the input files will be described.

Query File
The Query File should contain the sequences that you want to scan for recombination. You may include more than one sequence in this file.

Each sequence should not contain any other characters than A, T, G, C. In addition, each sequence should be in one line. You can apply these changes through the Sequence Editor tool which is provided by T-RECS.

Query Sequence File example:

```
>1234
CATCGAAACCGCATCCAAGT
>5678
ACGATCGATCGCGTACACGT
>9012345_Group1
ACAAGCTCGGGATCGACTTA
```

Database File

Each of the sequences contained in this file will be compared to each of the sequence contained in the Query File during the recombination searching process. You may include more than one sequence in this file.

Each sequence should not contain any other characters than A, T, G, C. In addition, each sequence should be in one line. You can apply these changes through the Sequence Editor tool which is provided by T-RECS.

Database Sequence File example:

```
>1234
CATCGAAACCGCATCCAAGT
>5678
ACGATCGATCGCGTACACGT
>9012345_Group1
ACAAGCTCGGGATCGACTTA
```

Evolutionary Groups File

This file should be a plain text file containing the sequence ID and the evolutionary group of each sequence that takes place in the analysis:

For the Blast Based Intergroup Recombination Analysis you need to add the sequence IDs and the evolutionary groups for the Query AND the Database sequences.

For the Blast Based Genotyping you need to add the sequence IDs and the evolutionary groups only for the Database sequences.

Format:

Each line should contain the sequence ID, then a TAB and then the corresponding evolutionary group. The sequence ID should be exactly as it is in the Query or Database file (without the > character).

IMPORTANT: No space characters should be used.

Evolutionary Groups File example:

```
1234    Group_1
5678    Group_2
9012345_Group1      Group_1
```

Annotation File

This file can be loaded in the Annotation Editor tab on the results window.

Annotation File Format:

The file should be in plain text format containing 6 columns separated by a TAB character. The 6 columns are described below:

• Sequence ID
The ID of the sequence on which the annotation will be drawn. The sequence ID should be exactly as it is in the Query or Database file (without the > character).

• Motif Name
The name of the annotation.

• Motif Start
The starting point of the annotation

• Motif End
The ending point of the annotation

• Comments
Text that will be displayed on the annotation

• Sequence to use as reference point
The ID of the sequence that will be used as a reference point to calculate the starting and ending position of the annotation that will be created. The sequence ID should be exactly as it is in the Query or Database file (without the > character).

Annotation File example:

1234	A1	1	10	Region_1	1234
5678	A2	11	19	Region_2	5678

This file can be created either on a text editor or through the Annotation Editor tab on the results window.

Output Files

On this section the format of the output files will be described.

Results File

This is the T-RECS Blast Based Intergroup Recombination Analysis results file. It consists of 4 sections: The header, the <WARNINGS>, the <WINDOW_STEP> and the <MAJOR_DONORS>.

• Header section
This section contains the header for each column of the blast results. The headers and their meanings are:

Q-ORGANISM: Query evolutionary group
Q-ACCESSION: Query sequence ID
S-ORGANISM: Subject evolutionary group
S-ACCESSION: Subject sequence ID
E-VALUE: Expect value
Q-START: Start of alignment in query
Q-END: End of alignment in query
S-START: Start of alignment in subject
S-END: End of alignment in subject
BITSCORE: Bit score
Q-LENGTH: Query sequence length
S-LENGTH: Subject sequence length
PIDENT: Percentage of identical matches
PPOS: Percentage of positive-scoring matches

There are 12 more columns with the headers: S-ORGANISM, S-ACCESSION, E-VALUE, Q-START, Q-END, S-START, S-END, BITSCORE, Q-LENGTH, S-LENGTH, PIDENT and PPOS at the end of the Header section. These headers stand for the blast results for the best blast hit where the Query and the Subject have the same evolutionary group.

• <WARNINGS> section
This section is visible only when warnings for possible false positive results are present. It lists the sequence IDs for each false positive warning.

• <WINDOW_STEP> section
This section states the window and the step values that were used for this analysis.

• <MAJOR_DONORS> section
This section contains the major donor for each possible recombination event. Each line contains the Query sequence ID and the corresponding Major Donor sequence ID. It also contains the positions on which the Major Donor is the best blast hit on the blast results.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1) Kyriakopoulou, Z. *et al.* (2014) Recombination among human non-polio enteroviruses: implications for epidemiology and evolution. Virus Genes.

2) Simon-Loriere, E. & Holmes, E.C. (2011) Why do RNA viruses recombine? Nat. Rev. Microbiol., 9, 617–626.

3) Hurst, L. D. & Peck, J. R. (1996). Recent advances in understanding of the evolution and maintenance of sex. Trends in Ecology & Evolution 11, 46-52.

4) White, K. A. & Morris, T. J. (1994). Recombination between defective tombusvirus RNAs generates functional hybrid genomes. Proceedings of the National Academy of Sciences, USA 91, 3642-3646 .

5) Raju, R. *et al.* (1995). Genesis of Sindbis virus by in vivo recombination of nonreplicative RNA precursors. Journal of Virology 69, 7391-7401 .

6) Weiss, B. G. & Schlesinger, S. (1991). Recombination between Sindbis virus RNAs. Journal of Virology 65, 4017-4025 .

7) Borja, M. *et al.* (1999). Restoration of wild- type virus by double recombination of tombusvirus mutants with a host transgene. Molecular Plant–Microbe Interactions 12, 153-162.

8) Gal-On, A. *et al.* (1998). Recombination of engineered defective RNA species produces infective potyvirus in plants. Journal of Virology 72, 5268-5270 .

9) Greene, A. E. & Allison, R. F. (1994). Recombination between viral RNA and transgenic plant transcripts. Science 263, 1423-1425 .

10) Rubio, T. *et al.* (1999). Recombination with host transgenes and effects on virus evolution: an overview and opinion. Molecular Plant–Microbe Interactions 12, 87-92.

11) Holmes, E. C. *et al.* (1999). Phylogenetic evidence for recombination in dengue virus. Molecular Biology and Evolution 16, 405-409.

12) Suzuki, Y. *et al.* (1998). Intragenic recombinations in rotaviruses. FEBS Letters 427, 183-187.

13) Worobey, M. *et al.* (1999). Widespread intra- serotype recombination in natural populations of dengue virus. Proceedings of the National Academy of Sciences, USA 96, 7352-7357 .

14) Vennema, H. *et al.* (1998). Feline infectious peritonitis viruses arise by mutation from endemic feline enteric coronaviruses. Virology 243, 150-157.

15) Gao, F. *et al.* (1998). A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. Journal of Virology 72, 5680-5698 .

16) Hahn, C. S. *et al.* (1988). Western equine encephalitis virus is a recombinant virus. Proceedings of the National Academy of Sciences, USA 85, 5997-6001 .

17) Kusters, J. G. *et al.* (1990). Sequence evidence for RNA recombination in field isolates of avian coronavirus infectious bronchitis virus. Vaccine 8, 605-608.

18) Sibold, C. *et al.* (1999). Recombination in Tula hantavirus evolution: analysis of genetic lineages from Slovakia. Journal of Virology 73, 667-675.

19) Revers, F. *et al.* (1996). Frequent occurrence of recombinant potyvirus isolates. Journal of General Virology 77, 1953-1965 .

20) Banner, L. R. & Lai, M. M. C. (1991). Random nature of coronavirus RNA recombination in the absence of selection pressure. Virology 185, 441-445.

21) Kotier, S. A. *et al.* (1995). Experimental evidence of recombination in coronavirus infectious bronchitis virus. Virology 213, 569-580.

22) Mindich, L. (1996). Heterologous recombination in the segmented dsRNA genome of bacteriophage ?6. Seminars in Virology 7, 389-397.

23) Palasingam, K. & Shaklee, P. N. (1992). Reversion of QâRNA phage mutants by homologous RNA recombination. Journal of Virology 66, 2435-2442 .

24) Snijder, E. J. *et al.* (1991). Comparison of the genome organization of toroviruses and coronaviruses: evidence for 2 non homologous RNA recombination events during Berne virus evolution. Virology 180, 448-452.

25) Weaver, S. C. *et al.* (1997). Recombinational history and molecular evolution of western equine encephalomyelitis complex alphaviruses. Journal of Virology 71, 613-623.

26) Olsthoorn, R. C. L. & van Duin, J. (1996). Evolutionary reconstruction of a hairpin deleted from the genome of an RNA virus. Proceedings of the National Academy of Sciences, USA 93, 12256-12261 .

27) Martin, D.P. *et al.* (2011) Analyzing recombination in nucleotide sequences. Mol Ecol Resour, 11, 943–955.

28) Salminen, M.O. *et al.* (1995) Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res. Hum. Retroviruses, 11, 1423–1425.

29) Lemey, P. & Posada, D. (2009) Introduction to recombination detection. In, The phylogenetic handbook. Cambridge University press, pp. 493 – 517.

30) McGuire, G. & Wright, F. (1998). TOPAL: recombination detection in DNA and protein sequences. Bioinformatics 14, 219-220.

31) Worobey, M. & Holmes, EC. (1999) Evolutionary aspects of recombination in RNA viruses. J Gen Virol 80:2535–2543

32) Gao, F. *et al.* (1999). Origin of HIV-1 in the chimpanzee Pan troglodytes. Nature 397, 436-441.

33) Martin, DP. *et al.* (2005b). A modified BOOTSCAN algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses.21:98-102.

34) Padidam, M. *et al.* (1999). Possible emergence of new geminiviruses by frequent recombination. Virology.265:218-225.

35) Maynard Smith J. (1992). Analyzing the mosaic structure of genes. J Mol Evol.34:126-129.

36) Posada, D. & Crandall, KA. (2001). Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Natl Acad Sci USA. 98:13757-13762.

37) Gibbs, M. *et al.* (2000). Sister-Scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. Bioinformatics 16, 573-582.

38) Boni, M.F. *et al.* (2007). An exact nonparametric method for inferring mosaic structure in sequence triplets. Genetics 176, 1035-1047.

39) Lemey, P. *et al.* (2009). Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning.BMC Bioinformatics 10, 126.

40) Archer, J. *et al.* (2012) Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. BMC Bioinformatics, 13, 47.

41) Salminen, M. & Martin, D.P. (2009) Detecting and characterizing individual recombination events. In, The phylogenetic handbook. Cambridge University press, pp. 519 – 546.

42) Rozanov, M. *et al.* (2004) A web-based genotyping resource for viral sequences. Nucleic Acids Res., 32, W654–659.

43) Fourment, M. *et al.* (2008) SWeBLAST: a Sliding Window Web-based BLAST tool for recombinant analysis. J. Virol. Methods, 152, 98–101.

44) Lole, K.S. *et al.* (1999) Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. J. Virol., 73, 152–160.

45) Martin, D.P. *et al.* (2010) RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics, 26, 2462–2463.

46) Martin, D.P. *et al.* (2015) RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evolution, 1, vev003–vev003.

47) Kosakovsky Pond, S.L. *et al.* (2006) GARD: a genetic algorithm for recombination detection. Bioinformatics, 22, 3096–3098.

48) Lemey, P. *et al.* (2009) Identifying recombinants in human and primate immunodeficiency virus sequence alignments using quartet scanning. BMC Bioinformatics,10, 126.

49) Pond, S.L.K. *et al.* (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics, 21, 676–679.