Βιοπληροφορική
Δεύτερη Άσκηση 2014-2015

Συνεργάτες: Κατερίνα Ποντζόλκοβα, 5405
Αθανασία Ζαχαριά, 5295

Αλγόριθμοι και Εφαρμογές στις Συγκρίσεις μεγάλων γονιδιωματικών ακολουθιών

Εισαγωγή

Η σύγκριση γονιδιωματικών ακολουθιών μεταξύ συγγενικών ειδών είναι συνήθως μια πολύ αποδοτική για την βιολογία διαδικασία, εξαιτίας της ομοιότητας που τείνουν να παρουσιάζουν τα λειτουργικά στοιχεία, όπως τα εξόνια, ενώ οι περιοχές που δεν είναι λειτουργικές συνήθως έχουν λιγότερη ομοιότητα. Το πρώτο βήμα στην σύγκριση των γονιδιωματικών ακολουθιών είναι η ευθυγράμμισή τους. Υπάρχουν αρκετές κατηγορίες ευθυγράμμισης: η τοπική ευθυγράμμιση η οποία συντελεί στην εύρεση τοπικών ομοιοτήτων μεταξύ δυο ακολουθιών, καθολική ευθυγράμμιση, για την αντιστοίχηση όλων των γραμμάτων μεταξύ των ακολουθιών. Ευθυγραμμίσεις μπορούν αν είναι είτε σε ζεύγη, μεταξύ δυο ακολουθιών, είτε πολλαπλές, που συγκρίνουν αρκετές ακολουθίες. Η κύρια πρόκληση στην ανάπτυξη των αλγορίθμων για τις γονιδιωματικές ευθυγραμμίσεις είναι ότι πρέπει να είναι αρκετά γρήγοροι για να μπορούν να επεξεργαστούν δεδομένα από μεγαβάσεις και και γίγαβάσεις, αλλά επίσης να εντοπίζουν με ακρίβεια ξεχωριστά ζευγάρια βάσεων. Η παραγωγή των ευθυγραμμίσεων είναι μια απαιτητική υπολογιστικά διαδικασία, όπως επίσης η απεικόνιση της ευθυγράμμισης, καθώς οι χρήστες πρέπει να έχουν την δυνατότητα να αλληλεπιδράσουν

με τα δεδομένα και τα προγράμματα επεξεργασίας στο πλαίσιο των κολοσσιαίων βάσεων δεδομένων. Η απεικόνιση των δεδομένων και των συμπερασμάτων παρέχουν αρκετές πολύτιμες γνώσεις σχετικά με τις μεταλλάξεις που έχει υποστεί μια συγκεκριμένη περιοχή.

Τοπική Ευθυγράμμιση

Η τοπική ευθυγράμμιση είναι η εύρεση παρόμοιων κομματιών από δυο ακολουθίες, ανεξάρτητα από την διάταξη και την θέση αυτών των ομοιοτήτων. Συνεπώς η τοπική ευθυγράμμιση επιτρέπει τον εντοπισμό των αναδιατάξεων μεταξύ δύο ακολουθιών, είναι κατάλληλη για τον σχηματισμό ολοκληρωμένου γενετικού υλικού στα κύτταρα. Ο αυθεντικός αλγόριθμος τοπικής ευθυγράμμισης είναι μία δυναμική προγραμματιστική προσέγγιση με το όνομα Smith-Watermann, ο οποίο τρέχει σε χρόνο ανάλογο ως προς το γινόμενο των μηκών των ακολουθιών. Ενώ αυτό δεν είναι πρακτικό για σύγκριση δυο μεγάλων γονιδιωματικών διαστημάτων έχει γίνει εκτεταμένη έρευνα στην εξέλιξη γρήγορων προσεγγίσεων για την τοπική ευθυγράμμιση γονιδιωματικών αλυσίδων. Σχεδόν κάθε αλγόριθμος ξεκινά με ένα προ-επεξεργαστικό βήμα το οποίο εντοπίζει την θέση κάθε σύντομου, ακριβούς ή προσεγγιστικού ταιριάσματος μεταξύ δυο ακολουθιών. Κάτι τέτοιο μπορεί να γίνει πολύ γρήγορα με δεικτοδότηση μιας εκ των δυο ακολουθιών σε μια κατάλληλη δομή δεδομένων, όπως ένας πίνακας αναζήτησης ή μια παραλλαγή του δέντρου επιθεμάτων, μειώνοντας έτσι την περιοχή αναζήτησης για τον αλγόριθμο τοπικής ευθυγράμμισης στις περιοχές που είναι πιο πιθανό να βρεθούν ομοιότητες. Αφού δημιουργηθεί ο κατάλληλος πίνακας, οι γειτονικοί δείκτες (seeds) μπορούν να συγχωνευτούν, αφού η παρουσία πολλών κοντινών αποδεικνύει περισσότερη ομοιογένεια, από ότι αν υπήρχε ένας δείκτης σε μια περιοχή. Περιοχές που αντιστοιχούν σε έναν δείκτη, ή σε μια ομάδα δεικτών, επεκτείνονται προς εύρεση περιοχών που δεν έχουν ακριβές ταίριασμα, αλλά είναι παρόλα αυτά όμοιες.

Τα τρία βήματα αυτά αποτελούν την βάση των περισσότερων αλγορίθμων της τοπικής ευθυγράμμισης για ακολουθίες DNA.

Δημιουργία δεικτών(Seed generation)

Η πιο εύκολη ίσος τεχνική παραγωγής των seeds μεταξύ δυο ακολουθιών είναι ένας πίνακας αναζήτησης (k-mers): οι δείκτες για κάθε λέξη μήκους k της ακολουθίας (βάσης) αποθηκεύονται σε έναν πίνακα, και οι λέξεις μήκους k της άλλης ακολουθίας (ερώτημα) χρησιμοποιούνται για την ανάκτηση των θέσεων από τον πίνακα στις οποίες μια συγκεκριμένη λέξη του ερωτήματος είναι παρούσα στην ακολουθία της βάσης. Αυτή η προσέγγιση χρησιμοποιήθηκε στους πρώτους και ίσος πιο γνωστούς ευθυγραμμιστές για μεγάλες ακολουθίες, FASTA και BLAST. Μια διαφορετική προσέγγιση είναι η χρήση πινάκων επιθεμάτων ή μιας προσέγγισης του πίνακα επιθεμάτων, όπως το αυτόματο Aho-Corasick, για την εύρεση των seeds. Ωστόσο αυτή η προσέγγιση έχει έχει μεγάλες απαιτήσεις μνήμης και χρόνου, αλλά κάνει πιο εύκολη την ανίχνευση μεγαλύτερων σε μήκος ομοιοτήτων: υπάρχουν 4^k πιθανοί συνδυασμοί DNA μήκους k, αλλά μόνο ένα μικρό θραύσμα μπορεί να υπάρχει σε μια πρόταση. Το γεγονός αυτό κάνει την μέθοδο του δέντρο επιθεμάτων πιο κατάλληλη για μεγαλύτερα ταιριάσματα, ενώ ο πίνακας αναζήτησης είναι προτιμότερος για μικρότερα ταιριάσματα. Αντί όμως να ψάξουμε για ταιριάσματα k-μήκους που να ταιριάζουν απόλυτα και στις δυο ακολουθίες, θα μπορούσαμε να χρησιμοποιήσουμε εκφυλισμένες λέξεις, οι οποίες επιτρέπουν έναν συγκεκριμένο αριθμό γραμμάτων να μην ταιριάξουν, αλλά σαφώς υπολογιστικά είναι πολύ πιο περίπλοκες. Υπάρχουν διάφοροι τρόποι να εφαρμοστεί αυτή η μέθοδος, όλοι τους όμως αυξάνονται με εκθετικό ρυθμό ανάλογα μα τον αριθμό των επιτρεπόμενων λαθών(μη ταιριασμάτων). Μια δημοφιλής εναλλακτική λύση είναι οι spaced seeds, οι οποίοι είναι πανόμοιοι με τις εκφυλισμένες λέξεις, αλλά οι θέσεις στις οποίες επιτρέπονται τα μη ταιριάσματα, είναι προκαθορισμένες. Οι spaced seeds (ή

PatternHunter) είναι πιο κατάλληλοι από το ακριβές ταίριασμα όταν πρόκειται για μεγάλου μήκους ομοιότητες και η αποδοτικότητα τους πέφτει όταν συγκρίνουμε ακολουθίες, οι οποίες έχουν μικρού μήκους ομοιότητες. Μια άλλη μέθοδος εύρεσης των seeds είναι η εύρεση των ομοιοτήτων μέχρι να βρεθεί το πρώτο μη ταίριασμα, η οποία έχει ονομαστεί max-mers, δηλαδή ταίριασμα που έχουν παραταθεί στο μέγιστο και έχουν ένα συγκεκριμένο ελάχιστο μήκος. Τα max-mers μπορούν να βρεθούν σε γραμμικό χρόνο σε σχέση με το μήκος της ακολουθίας χρησιμοποιώντας τα δέντρα επιθεμάτων. Ενώ τα max-mers δεν προσφέρουν μεγαλύτερη ευαισθησία από τα k-mers(τα ταιριάσματα k-μήκους), έχουν το πλεονεκτήματα ότι επιστρέφουν μόνο έναν δείκτη για κάθε κομμάτι της ακολουθίας όπου υπάρχει ταίριασμα. Έτσι όταν συγκρίνουμε 2 ακολουθίες που είναι πολύ κοντά, η μέθοδος αυτή θα επιστρέψει πολύ λιγότερες ομοιότητες που πρέπει να αναλυθούν. Παρόλα αυτά όταν συγκρίνουμε πιο μακρινές ακολουθίες, με αποτέλεσμα οι ομοιότητες να είναι σχετικά μικρού μήκους η μέθοδος των max-mers δεν προσφέρει πιο γρήγορα ή πιο αποδοτικά αποτελέσματα από τα k-mers, και δεδομένης της προεργασίας που πρέπει να γίνει για την εύρεση των k-mers η δεύτερη μέθοδος είναι προτιμότερη σε αυτές τις περιπτώσεις.

Συγχώνευση γειτονικών seeds

Ενώ ένας seed είναι καλή ένδειξη για την ομολογία μεταξύ δυο ακολουθιών, πολλοί μικρότεροι seeds υποδεικνύουν ακόμα μεγαλύτερη πιθανότητα ομοιογένειας. Το γεγονός αυτό έχει χρησιμοποιηθεί ευρέως από διάφορα προγράμματα, όπως το FASTA, το οποίο απαιτεί έναν συγκεκριμένο αριθμό seeds σε μια περιοχή για να ξεκινήσει να κάνει τοπική ευθυγράμμιση. Είναι μια συνηθισμένη πρακτική για την BLAST και μερικά άλλα προγράμματα να βρίσκουν δυο γειτονικά seeds πριν ξεκινήσουν μια επέκταση. Οι τεχνικές αυτές εφαρμόζονται συνήθως σε πίνακες αναζήτησης ή πίνακες κατακερματισμού για όλες τις διαγώνιες του μητρώου δυναμικού

προγραμματισμού. Όταν βρεθεί ένα seed, το πρόγραμμα ψάχνει την διαγώνιο στην οποία ανήκει, και αν υπάρχει ήδη ένα seed και είναι αρκετά κοντά σε αυτό που βρέθηκε, τα δυο seeds συγχωνεύονται. Το πρόγραμμά CHAOS χρησιμοποιεί μια διαφορετική προσέγγιση που προβλήματος. Έχει πολλές ομοιότητες με το FASTA, αλλά επιτρέπει να υπάρχουν ομάδες από seeds σε γειτονικές διαγώνιες. Κάθε seed, όταν βρεθεί, αποθηκεύεται σε μια λίστα μαζί με τις θέσεις του στις δυο ακολουθίες, που αντιπροσωπεύονται από τον αριθμό της διαγωνίου. Για κάθε καινούριο seed γίνει έλεγχος περιοχής στην λίστα, για να βρεθούν seeds που έχουν αποθηκευτεί προηγουμένως έχοντας αριθμό διαγωνίου που απέχει κάποια συγκεκριμένη απόσταση από την διαγώνιο στην οποία βρίσκεται το seed που βρήκαμε. Ο έλεγχος αυτός έχει ως αποτέλεσμα την απομάκρυνση κάθε seed που βρίσκεται πολύ μακρυά από την θέση στην οποία τα καινούρια seeds δημιουργούνται. Έτσι δημιουργούνται αλυσίδες, και όταν είναι να προστεθεί κάποιο καινούριο seed σε μια αλυσίδα, προτιμάται εκείνη με το μεγαλύτερο σκορ.

Προέκταση με ή χωρίς κενά

Αφού βρεθεί ένα seed ή μια ομάδα από seed, συνήθως γίνεται μια επέκταση αυτής της περιοχής, έτσι ώστε να βρεθούν τα όρια αυτής της ομολογίας. Η πιο συνηθισμένη μέθοδος είναι να δημιουργούνται οι χωρίς κενά προεκτάσεις, οι οποίες οποίες παρουσιάστηκαν για πρώτη φορά στο αρχικό BLAST πρόγραμμα. Σε μια ήδη υπάρχουσα ευθυγράμμιση προστίθεται ένα γράμμα από την κάθε μια ακολουθία, έχοντας στο μυαλό το συνολικό σκορ. Το σκορ αυξάνεται αν τα δυο καινούρια γράμματα ταιριάζουν, ή μειώνεται σε αντίθετη περίπτωση. Όταν το σκορ πέσει αισθητά, η προέκταση σταματά και η ευθυγράμμιση με το μέγιστο σκορ επιστρέφεται. Η διαδικασία αυτή είναι ευρέως γνωστή σαν BLAST ή χωρίς κενά προέκταση. Η εναλλακτική είναι η προέκταση του Smith-Waterman, όπου επιτρέπονται κενά γύρω από το seed. Η διαδικασία αυτή είναι πολύ πιο απαιτητική ως προς τον χρόνο

και πιο χρήσιμη όταν γίνεται σύγκρισή μεταξύ 2 μακρινών γονιδιωματικών ακολουθιών.

BLAST

Συνήθως οι αλγόριθμοι που αναζητούν τοπική ομοιότητα ψάχνουν μόνο τοπικά συντηρημένες ανακολουθίες, και ακόμα και με την πρώτη σύγκριση μπορούν να επιστρέψουν αρκετές ξεχωριστές ευθυγραμμίσεις. Είναι γνωστό ότι οι μη συντηρημένες περιοχές δεν συμβάλλουν στην μέτρηση της ομοιότητας. Η BLAST μετρά την ομοιότητα με τον εξής τρόπο: ξεκινά με έναν πίνακα που περιέχει τις βαθμολογίες για κάθε πιθανό ζευγάρι ταιριασμάτων. Ορίζεται το μεγίστου μήκους ζευγάρι ακολουθιών (MSP- maximal segment pair) να είναι το υψηλότερα βαθμολογημένο ζευγάρι ίδιου μήκους από διαφορετικές ακολουθίες. Καθώς από την πλευρά της βιολογίας μας ενδιαφέρουν όλες οι συντηρημένες περιοχές μεταξύ δυο πρωτεϊνών και όχι μόνο το υψηλότερα βαθμολογημένο ζευγάρι, ορίστηκε ότι το σκορ του ζευγαριού των ακολουθιών είναι μόνο τοπικά μέγιστο, αν δεν μπορεί να βελτιωθεί είτε με την επέκταση είτε με την σμίκρυνση.
Το σκορ για τα MSP μπορεί να υπολογιστεί σε χρόνο ανάλογο με το μήκος τους χρησιμοποιώντας δυναμικό προγραμματισμό. Αναζητώντας μέσα στην βάση για ταίριασμα, μέσα σε χιλιάδες ακολουθίες, ελάχιστες ή καμία θα ταιριάζουν με αποτέλεσμα να αναζητάμε μόνο εκείνες τις ακολουθίες που έχουν σκορ πάνω από ένα συγκεκριμένο όριο. Μέσα στις ακολουθίες αυτές συμπεριλαμβάνονται οι ακολουθίες που έχουν πολύ σημαντική ομοιότητα, αλλά και εκείνες με λιγότερη που βρίσκονται ίσα-ίσα πάνω από το όριο, οι οποίες μπορεί να είναι τυχαία ακριβή ή όχι και τόσο ταιριάσματα. Η κύρια στρατηγική της BLAST είναι να ψάχνει ζευγάρια ακολουθιών με σκόρ το λιγότερο T. Η σάρωση όλης της ακολουθίας μας δίνει γρήγορα την δυνατότητα να δούμε αν η ακολουθία περιέχει μια λέξη μήκους w που μπορεί να ταιριάξει με την ακολουθία αναζήτησης, ώστε το ταίριασμά τους να έχει σκορ ίσο ή μεγαλύτερο του ορίου T.

Όσο μικρότερο είναι αυτό το όριο, τόσο περισσότερα αποτελέσματα θα προκύψουν, κάτι που συντελεί στην αύξηση του χρόνου της εκτέλεσης του προγράμματος.
Ο αλγόριθμος BLAST εκτελείται σε βασικά τρία βήματα: δημιουργία λίστας με τα υψηλότερα σκορ, σάρωση της βάσης για ταιριάσματα, επέκταση των ταιριασμάτων. Ωστόσο ο αλγόριθμος διαφοροποιείται σε κάποια σημεία, αν η βάση περιέχει πρωτεΐνες ή ακολουθίες DNA.
Για πρωτεΐνες η σάρωση γίνεται με δυο προσεγγίσεις. Η πρώτη φτιάχνει έναν πίνακα με κάθε πιθανό συνδυασμό λέξεων συγκεκριμένου μήκους, και ψάχνει τον αριθμό των εμφανίσεων της κάθε λέξης. Η δεύτερη προσέγγιση χρησιμοποιεί ντετερμινιστικό αυτόματο, και είναι προτιμότερη αφού είναι πιο οικονομική από άποψη χρόνου και χώρου.
Η διαδικασία επέκταση προς μια μεριά του ταιριάσματος για την εύρεση του μέγιστου τοπικού ζευγαριού έχει οριστεί να σταματά για εξοικονόμηση χρόνου όταν η επέκταση γίνει πάνω από ένα συγκεκριμένο όριο, το οποίο επιβεβαιώθηκε και πειραματικά και θεωρητικά.
Για ακολουθίες DNA, χρησιμοποιείται πιο απλή λίστα με όλα τα πιθανά k-mers, τα οποία συνήθως ανέρχονται σε μερικές χιλιάδες λέξεις. Η σχεδίαση των εργαλείων αναζήτησης DNA βάσεων βασίστηκε κατά ένα μεγάλο μέρος στο γεγονός ότι οι ακολουθίες DNA είναι εξαιρετικά μη τυχαίες, με επαναλαμβανόμενες ακολουθίες-μοτίβα και τοπικά πιο πλούσιες σε συγκεκριμένα στοιχεία. Έτσι στην περίπτωση αναζήτησης κάποιας λέξης που ανήκει σε αυτές τις δυο κατηγορίες τα αποτελέσματα των ταιριασμάτων θα είναι λιγότερο ισχυρά. Το πρόγραμμα αποθηκεύει τις εμφανίσεις της λέξης, και αν είναι πιο πολλές από την πιθανότητα εμφάνισης, η οποία μπορεί να ρυθμιστεί από τον χρήστη, παράγει ανάλογα αποτελέσματα.
Οι στρατηγικές που χρησιμοποιεί η BLAST έχουν δεχτεί πολλές παραλλαγές. Έχει αναπτυχθεί μια έκδοση της BLAST που χρησιμοποιεί δυναμικό προγραμματισμό, για να επεκτείνει τα ταιριάσματα, ώστε να επιτρέψει κενά στην ευθυγράμμιση που θα προκύψει. Αυτό όμως αυξάνει σημαντικά τον χρόνο

εκτέλεσης του αλγορίθμου επέκτασης, αλλά και μειώνει σημαντικά την επιλεξιμότητα. Δοθέντων αυτών των μειονεκτημάτων, είναι αμφίβολο αν αυτή η εκδοχή BLAST μπορεί να θεωρηθεί ως βελτίωση. Επίσης έχει αναπτυχθεί και η εναλλακτική του πίνακα με όλα τα k-mers, δηλαδή μια βάση, στην οποία γίνεται σάρωση για ταιριάσματα και επεξεργασία των αποτελεσμάτων. Οι απαιτήσεις μνήμης είναι σημαντικές, αλλά πιο καταστροφικό ήταν ότι για αναζήτηση λέξης τυπικού μήκους, υπήρξε η ανάγκη για τυχαία πρόσβαση στην βάση κάτι που έκανε την διαδικασία πιο αργή στα υπολογιστικά συστήματα που χρησιμοποιήθηκαν από την σειριακή σάρωση.

Καθολική Ευθυγράμμιση

Η καθολική ευθυγράμμιση βρίσκει την αντιστοιχία μεταξύ των ακολουθιών, δημιουργώντας έναν μονοτονικά αυξανόμενο πίνακα μεταξύ των γραμμάτων της κάθε ακολουθίας. Έτσι παρέχεται μια πιο ακριβή ευθυγράμμιση μεταξύ των δυο ακολουθιών. Ο αρχικός αλγόριθμος καθολικής ευθυγράμμισης είναι ο Needleman-Wunsh, ο οποίος χρειάζεται χόνο ανάλογο με το γινόμενο των μηκών των 2 ακολουθιών. Ωστόσο, αυτός αλγόριθμος δεν είναι καθόλου αποδοτικός όταν πρόκειται να χειριστεί τεράστιες ακολουθίες, που ανήκουν σε μεγαβάσεις. Πρόσφατα έχουν αναπτυχθεί πιο γρήγοροι και πιο ακριβείς μέθοδοι όπως: DI-ALIGN, MUMmer, GLASS, WABA, AVID, LAGAN. Όλες αυτές οι μέθοδοι βασίζονται στην προσέγγιση της αγκύρωσης. Οι άγκυρες στην καθολική ευθυγράμμιση είναι το αντίστοιχο των seeds στην τοπική ευθυγράμμιση, και τα δυο μειώνουν τον χώρο αναζήτησης για ταιριάσματα. Η όλη διαδικασία μπορεί να συνοψιστεί σε 1) παραγωγή αποσπασμάτων των 2 ακολουθιών (τοπικών κομματιών που έχουν μεγάλη ομοιότητα), 2) επιλογή του κατάλληλου αποσπάσματος με βάση την ομοιότητα, χρησιμοποιώντας δυναμικό προγραμματισμό αραιής προσέγγισης ή κάποια παραλλαγή, 3) εκτέλεση ενός διεξοδικού αλγορίθμου καθολικής

ευθυγράμμισης είτε μεταξύ των αγκυρών, είτε σε κάποια περιοχή γύρω από αυτές.

Η Εύρεση Πιθανών Αγκυρών

Ίσως η πιο ευθεία μέθοδος εύρεσης πιθανών κομμάτων ταιριάσματος είναι η χρήση k-mers και η εφαρμογή αυτής της μεθόδου υπάρχει στο πρόγραμμα ευθυγράμμισης GLASS. Εξαιτίας της αναξιοπιστίας των μεμονωμένων k-mers, συμπληρώθηκαν με μια επέκταση χρησιμοποιώντας τον αλγόριθμο Needleman-Wunsh σε ένα 12 επί 12 παράθυρο γύρο από το κάθε k-mer. Παράλληλα οι δημιουργοί του MUMmer προγράμματος πρότειναν την χρήση των μέγιστων δυνατών ταιριασμάτων όπως τα κομμάτια που χρησιμοποιούνται για τον αλγόριθμο αλυσίδων. Τα MUM είναι συμβολοσειρές ταιριάσματος μέγιστης ακρίβειας μεταξύ δυο ακολουθιών και εμφανίζονται ακριβώς μια φορά, σε κάθε μια από τις δυο ακολουθίες. Το γεγονός ότι το MUM είναι μια μοναδική λέξη σε κάθε ακολουθία, μειώνει την πιθανότητα λανθασμένου ταιριάσματος, με μοναδικό μειονέκτημα την ανικανότητα να βρεθούν άγκυρες μεταξύ ανόμοιων γονιδιωμάτων, όπου τα μέγιστα πιθανά ταιριάσματα είναι υπερβολικά μικρά για να είναι μοναδικά. Αυτές οι ιδέες συνδυάστηκαν σε ένα πρόγραμμα που ονομάζεται AVID και ασχολείται με το ψάξιμο ταιριασμάτων μέγιστης ακρίβειας χρησιμοποιώντας παράλληλα τα παράθυρα του Needleman-Wunsh για να πιστοποιήσει την ποιότητα των αποτελεσμάτων. Η πιο πρόσφατη προσέγγιση παραγωγής κομματιών των ακολουθιών ήταν χρήση των πλήρως τοπικών ευθυγραμμίσεων με εφαρμογές στα προγράμματα DI-ALIGN και LAGAN χρησιμοποιώντας ευθυγραμμίσεις τύπου CHAOS, ενώ αντίστοιχα το ORCA χρησιμοποιούσε BLAST ευθυγραμμίσεις. Προσεγγίσεις οι οποίες αξιοποίησαν γνώσεις από την βιολογία στη διεργασία της επιλογής των αγκυρών ήταν οι WABA και CONREAL, οι οποίες αναζητούν άγκυρες που είναι πιθανότατα περιοχές πρωτεϊνικής κωδικοποίησης. Ενώ και οι δυο μέθοδοι έχουν

δείξει ότι είναι αποτελεσματικές στο κύριο σκοπό τους, που είναι η ευθυγράμμιση περιοχών πρωτεϊνικής κωδικοποίησης, δεν θεωρούνται εργαλεία γενικού σκοπού. Τα κύρια μειονεκτήματα αυτών των προγραμμάτων είναι πως το WABA δεν λειτουργεί σωστά για τους προωθητές και το CONREAL δεν ευθυγραμμίζει σωστά περιοχές κωδικοποίησης γονιδίων.

Εύρεση μιας έγκυρης ομάδας αγκυρών

Ο πιο εύκολος τρόπος για να βρεθεί μια σειρά από ταιριάσματα τοπικής ευθυγράμμισης, έτσι ώστε να μπορούν να χρησιμοποιηθούν σαν άγκυρες είναι να χρησιμοποιήσουμε μια άπληστη προσέγγιση, όπου το καλύτερο (το πιο δυνατό) ταίριασμα είναι και το πρώτο που γίνεται δεκτό, αλλά και κάθε λιγότερο δυνατό ταίριασμα επίσης περιλαμβάνεται, αρκεί να μην αντικρούεται με κάποιο προηγούμενο ταίριασμα. Αυτό ωστόσο, οδηγεί στο να αγνοούνται κάποια λιγότερο δυνατά ταιριάσματα από αλλά πιο ισχυρά εξαιτίας του μη σωστού σκορ ταιριάσματος. Έχει δειχθεί ότι είναι δυνατόν αν βρεθεί το σετ κομματιών με το μέγιστο σκορ σε χρόνο O(nlogn) χρησιμοποιώντας δυναμικό προγραμματισμό αραιής διαδικασίας αλυσιδοποίησης. Είναι συνηθισμένο να εκτελούνται αρκετοί γύροι αγκύρωσης: στον πρώτο βήμα επιστρέφονται μόνο τα πιο ισχυρά ταιριάσματα, σε κάθε επόμενο βήμα επιστρέφονται όλο και λιγότερο ισχυρά. Η προσέγγιση αυτή είναι γνωστή ως ιεραρχική αγκύρωση και είναι πολύ χρήσιμη, αφού σε κάθε γύρο βρίσκονται λιγότερες άγκυρες, κάτι που επιτρέπει την αναλυτικότερη ανάλυση του κάθε πιθανού ταιριάσματος.

Συμπλήρωση των κενών

Όταν πλέον έχει βρεθεί η έγκυρη ομάδα αγκυρών, είναι δυνατόν να μειώσουμε το διάστημα αναζήτησης για τον τελικό και αργό αλγόριθμο καθολικής ευθυγράμμισης στις περιοχές που αντιστοιχούν στις άγκυρες. Και τα k-mer και τα max-mers

μπορούν να χρησιμοποιηθούν ως ισχυρότατα ταιριάσματα, αφού δεν περιέχουν κενά, η βέλτιστη ευθυγράμμιση μπορεί να περάσει ακριβώς από πάνω τους. Αντίθετα τοπικές ευθυγραμμίσεις δεν είναι αξιόπιστες άγκυρες, αφού επιτρέπουν μικρά λάθη ή μη-ταιριάσματα. Για αυτό το λόγο το LAGAN έχει εφαρμόσει μια πιο ευέλικτη προσέγγιση, όπου μια άγκυρα δεν έχει αμετάβλητη θέση, απλά έχει σημειωθεί, και η καθολική ευθυγράμμιση είναι υποχρεωμένη να περάσει δίπλα από την άγκυρα, αλλά όχι απαραίτητα ακριβώς από πάνω της.

Πολλαπλή Καθολική Ευθυγράμμιση

Οι ομοιότητες ανάμεσα σε εξελικτικά μακρινά είδη μπορεί να αποκαλύψει αντισημιτικά βιολογικά χαρακτηριστικά. Οι πιο πρόσφατες έρευνες όμως έχουν δείξει ότι η σύγκριση μεταξύ μακρινών ειδών ίσως είναι περιττή, αφού αρκεί να συγκρίνουμε πολλά κοντινά είδη, και μέσα από αυτή την διαδικασία να ξεχωρίσουμε τις συντηρημένες ακολουθίες από τις ουδέτερες. Αυτή η συγκριτική ανάλυση βασίζεται σε πολλαπλή καθολική ευθυγράμμιση. Πολλαπλές ευθυγραμμίσεις έχουν αποδειχθεί να είναι πιο ισχυρές από αυτές που γίνονται σε ζεύγη, αφού δείχνουν τις συντηρημένες περιοχές με μεγαλύτερη ακρίβεια, καθώς επίσης έχουν σε γενικές γραμμές πιο ακριβή ευθυγράμμιση, αλλά και δίνουν στοιχεία για να μπορούν να εκτιμηθούν τοπικοί δείκτες εξέλιξης. Οι πολλαπλές ευθυγραμμίσεις είναι σημαντικά πιο δύσκολες να υπολογιστούν από τις ευθυγραμμίσεις σε ζεύγη, αφού ο χρόνος εκτέλεσης είναι ανάλογος με το γινόμενο των μηκών όλων των ακολουθιών.

Βαθμολογώντας μια Πολλαπλή Ευθυγράμμιση

Ίσως το πιο βασικό, αλλά ταυτόχρονα και πιο σημαντικό θέμα στο πρόβλημα της καθολικής ευθυγράμμισης είναι το πως να την βαθμολογήσεις. Η πιο κοινή μέθοδος είναι η βαθμολόγηση

με βάση το άθροισμα από το σκορ των ζευγών. Εναλλακτικά, μπορεί να χρησιμοποιηθεί η μέθοδος της πλειοψηφίας: για κάθε στήλη βρίσκουμε τον πιο πιθανό χαρακτήρα και βαθμολογούμε αρνητικά κάθε απόκλιση από αυτόν τον χαρακτήρα. Μια άλλη προσέγγιση είναι να μετρήσουμε την εντροπία της στήλης, όπως επίσης μπορεί να εφαρμοστεί και ένας συνδυασμός αυτών των μεθόδων, αλλά ακόμα το πρόβλημα της βαθμολόγησης της πολλαπλής ευθυγράμμισης παραμένει ως ένα ανοικτό πρόβλημα.

Αλγόριθμοι ευθυγράμμισης της MGA και της DIALIGN

Για να επιτύχουμε μια καθολική ευθυγράμμισή είναι απαραίτητο να παράγουμε σημεία αγκυρών μεταξύ αρκετών ακολουθιών. Έτσι έχει προταθεί μόνο μια πραγματικά πολλαπλή μέθοδο: multi-MEMs, που είναι πολλαπλά ακριβή ταιριάσματα μεταξύ όλων των ακολουθιών που ευθυγραμμίζονται. Είναι δυνατή η εύρεση μιας συνεχής αλυσίδας σε έναν αυθαίρετο αριθμό συμβολοσειρών σε τετραγωνικό χρόνο, κάτι που κάνει τα multi-MEMs να μειώνουν σημαντικά τον χώρο αναζήτησης για πολλαπλή ευθυγράμμιση. Είναι αναγκαίο να τρέξουμε μια ευαίσθητη μέθοδο για πολλαπλή ευθυγράμμιση ακολουθίας στο ενδιάμεσο διάστημα μεταξύ δυο αγκυρών για να ευθυγραμμίσουμε μεμονωμένα ζευγάρια βάσεων. Τέτοιου είδους προσέγγιση εφαρμόστηκε αρχικά στο πρόγραμμα MGA, με μοναδικό μειονέκτημα την απαίτηση κάθε άγκυρα να είναι παρούσα σε όλες τις ακολουθίες, το οποίο είναι πολύ δύσκολο όταν κάποιος συγκρίνει ανόμοιες ακολουθίες.
Όσον αφορά στην ευθυγράμμιση με το DIALIGN πρόγραμμα, για την δημιουργία πολλαπλής ευθυγράμμισης οι διαγώνιες που προκύπτουν από ζευγάρωμα κομματιών των ακολουθιών χωρίς κενά, ταξινομούνται ανάλογα με τα βάρη τους με άπληστο τρόπο, ξεκινώντας από κείνη με το μέγιστο βάρος, δεδομένου

ότι δεν έρχεται σε σύγκρουση με κάποια από αυτές που έχουν ήδη ενσωματωθεί.

Προοδευτική ευθυγράμμιση

Ο πιο κοινός τρόπος προσέγγισης για πολλαπλή ευθυγράμμιση ακολουθιών είναι κάποιου είδους προοδευτική στρατηγική η οποία χρησιμοποιεί συνεχείς εφαρμογές κάποιου αλγορίθμου ευθυγράμμισης ανά ζεύγη. Το πιο γνωστό σύστημα που βασίζεται σε αυτήν την λογική είναι το CLUSTALW, ενώ υπάρχουν και άλλα συστήματα όπως το MULTALIGN, MULTAL και το PRRP.
Η βασική ιδέα πίσω από όλα αυτά είναι πως τεχνικές ευθυγράμμισης ανά ζεύγη μπορούν να γενικευτούν από ευθυγράμμιση δυο ακολουθιών σε ευθυγράμμιση δυο προφίλ, όπου προφίλ ονομάζονται ακολουθίες στις οποίες κάθε θέση αποτελείται από ένα κομμάτι αδενίνης, κυτοσίνης, θυμίνης, γουανίνης. Επειδή κάθε ευθυγράμμιση μπορεί να θεωρηθεί σαν προφίλ και το κενό μπορεί να θεωρηθεί σαν πέμπτος χαρακτήρας είναι πολύ πιθανό να παράγουμε πολλαπλή ευθυγράμμιση μέσω περάσματος από κάτω προς τα πάνω ενός φυλογενετικού δέντρου. Ενώ μπορούμε μέσω προοδευτικής προσέγγισης να χτίσουμε ευθυγράμμιση αυτού του είδους σε χρόνο ανάλογο ως προς το γινόμενο των μηκών τους, υπάρχει τεράστιο πρόβλημα καθυστέρησης για μακριές ακολουθίες και έτσι χρησιμοποιούνται τεχνικές με άγκυρες για να μειωθεί ο συνολικός χρόνος εκτέλεσης, με τρόπο παρόμοιο όπως και στο πρόβλημα του ζευγαρώματος με άγκυρες.

Ευθυγράμμιση ακολουθιών με αναμεταθέσεις

Συνήθης τρόπος εξέλιξης του γονιδιώματος είναι μέσω αναμεταθέσεων διαφόρων κομματιών DNA. Οι πιο συνηθισμένες ευθυγραμμίσεις είναι: η αλλαγή κατεύθυνσης

ενός κομματιού DNA, χωρίς όμως να αλλάζει η θέση του, η επανατοποθέτηση και οι διπλασιασμοί.
Βασική μέθοδος ευθυγράμμισης είναι η καθολική, η οποία και δείχνει πως μπορούμε να μεταλλάξουμε μια ακολουθία σε μια άλλη χρησιμοποιώντας συνδυασμό απλών τροποποιήσεων, ενώ υπάρχει και η τοπική ευθυγράμμιση η οποία αναδεικνύει τοπικές ομοιότητες μεταξύ περιοχών, αλλά καμιά τους δεν μπορεί να χειριστεί ικανοποιητικά γεγονότα επανατοποθετήσεων. Στην περίπτωση που δυο ακολουθίες έχουν n αντίγραφα συγκεκριμένου γονιδίου, οι τοπικοί ευθυγραμμιστές επιστρέφουν n^2 τοπικές ευθυγραμμίσεις ανάμεσα σε όλα τα ζεύγη, ενώ μια καθολική ευθυγράμμιση είναι ξεκάθαρα πιο εξελιγμένη διαδικασία. Τα πρώτα προγράμματα ευθυγράμμισης μεγάλων γονιδιωματικές ακολουθιών με επανατοποθετήσεις ήταν το Shuffle-LAGAN, για ακολουθίες σε ζεύγη και το Mauve για πολλαπλές ακολουθίες.

Ευθυγράμμιση σε ζεύγη με το εργαλείο Shuffle-LAGAN

Ο αλγόριθμος Shuffle-LAGAN χτίστηκε πάνω στο πλαίσιο της καθολικής ευθυγράμμισης LAGAN, επιτρέποντας ωστόσο μεταθέσεις χρησιμοποιώντας μια πρωτοποριακή τεχνική αλυσίδας. Αποτελείται από τρία στάδια: στο πρώτο στάδιο με το εργαλείο CHAOS βρίσκονται οι τοπικές ευθυγραμμίσεις, στο δεύτερο στάδιο επιλέγεται ένα υποσύνολο από τις ευθυγραμμίσεις από τον χάρτη 1-μονοτονικής συντήρησης με βάση το σκορ τους έτσι ώστε να επιτρέπονται κάποια συγκεκριμένα κενά. Συγκεκριμένα, είναι η δομή αυτού του χάρτη που κάνει τον αλγόριθμο Shuffle-LAGAN να ξεχωρίζει από τους υπόλοιπους αλγορίθμους καθολικής ευθυγράμμισης με χρήση των αγκυρών. Τέλος, οι τοπικές ευθυγραμμίσεις από τον πίνακα συντήρησης που μπορούν να είναι μέρος της συνολικής καθολικής ευθυγράμμισης ενώνονται σε μεγίστου μήκους

συνεχόμενα υποτμήματα, τα οποία ευθυγραμμίζονται με την βοήθεια του LAGAN καθολικού ευθυγραμμιστή.

Πολλαπλές ευθυγραμμίσεις με το Mauve

Το πρόβλημα της ευθυγράμμισης των πολλαπλών γονιδιωμάτων, που έχουν υποστεί μεταθέσεις, αντιστροφές και χάσιμο στην αντιγραφή παραμένει ένα ανοικτό πρόβλημα. Μια αρχική πολλά υποσχόμενη μέθοδος έχει εφαρμοστεί στο πακέτο του Mauve για γονιδιωματικές ευθυγραμμίσεις.
Όπως άλλες μέθοδοι που έχουν περιγραφεί ως τώρα, το Mauve για να βρεί πιθανές άγκυρες χρησιμοποιεί seeds, ύστερα συνδέει τις άγκυρες και τέλος παράγει μια προοδευτική πολλαπλή ευθυγράμμιση. Η διαφορά του με τις υπόλοιπες μεθόδους είναι ότι δεν δημιουργεί ένα μοναδικό συνεχόμενο σετ από άγκυρες, αλλά χτίζει συνεχόμενο σετ από άγκυρες για κάθε συγγραμμικά αποσπάσματα των γονιδιωματικών ακολουθιών. Κάθε συνεχόμενο σετ από άγκυρες ονομάζεται Locally Collinear Block (LCB) - Τοπικά Συγγραμμικό Μπλοκ.
Ο αρχικός Mauve αλγόριθμος χρησιμοποιούσε μια τεχνική δεικτοδότησης και επέκτασης για να παράγει πολλαπλά-MUMs (multi-MUMs) τα οποία χρησιμοποιούνταν ως πιθανές άγκυρες. Τα multi-MUMs πρέπει να είναι ακριβή, μοναδικά ταιριάσματα τα οποία συμβαίνουν σε κάθε γονιδίωμα του οποίου ο βαθμός ευαισθησίας είχε περιοριστεί από την αρχική μέθοδο αγκύρωσης. Οι σύγχρονες εκδόσεις του προγράμματος χρησιμοποιούν seed μοτίβο με κενά για ταίριασμα πολλαπλών ακολουθιών ταυτόχρονα. Η μέθοδος του ταιριάσματος που επιτρέπει λάθη αυξάνει σημαντικά τον βαθμό ευαισθησίας των αγκυρών. Δεδομένου του πλήθους των αγκυρών που ταιριάζουν σε όλα τα προς ευθυγράμμιση γονιδιώματα ο Mauve χρησιμοποιεί μια άπληστη μέθοδο για αποκλεισμό των ορίων για να φιλτράρει τα ταιριάσματα εξαιτίας της τυχαίας ομοιότητας και ολοκληρώνει τις διεργασίες σε τρία βασικά στάδια επαναλήψεων. Το πρώτο στάδιο αναλύει τα όρια για αν προσδιορίσει τα όρια στη σειρά των αγκυρών παράγοντας LCB.

Το δεύτερο υπολογίζει το βάρος του κάθε LCB σαν σύνολο των μηκών των αγκυρών του. Το τρίτο στάδιο προσδιορίζει το LCB με το μικρότερο βάρος και αν αυτό είναι μικρότερο από το ελάχιστο όριο, διαγράφει τις άγκυρες και επιστρέφει στο πρώτο βήμα αλλιώς τερματίζει την διεργασία.
Στην αρχική δημοσίευση που αφορούσε τον Mauve αλγόριθμο, ο αλγόριθμος γονιδιωματικών ευθυγραμμίσεων εφαρμόστηκε σε μια ομάδα εννιά συγγενικών εντεροβακτηριδίων, προσδιορίζοντας πλήθος γονιδιωματικών επανατοποθετήσεων και περιοχές διαφορικού γονιδιωματικού περιεχομένου. Καθώς συνεχίζεται η έρευνα υπολογίζεται πως θα αυτοματοποιηθούν περισσότερο οι μέθοδοι και η ευθυγράμμιση με επανατοποθετήσεις θα οδηγήσει στην κατανόηση των εξελικτικών δυνάμεων όσον αφορά τα γονίδια και γονιδιωματικές λειτουργίες.

Πλήρης Γονιδιωματική Ευθυγράμμιση

Η διάθεση ολόκληρων γονιδιωματικών συλλογών από γενετικό υλικό ανθρώπων, ποντικιών και αρουραίων παρουσίασε την πρόκληση για χτίσιμο πολλαπλών ευθυγραμμίσεων από διάφορα μεγάλα γονιδιώματα. Το πρόβλημα που εμφανίζεται είναι πιο δύσκολο από εκείνο της απλής ευθυγράμμισης και λόγο του μεγέθους αλλά και εξαιτίας της ανάγκης να βρεθούν κομμάτια μπλοκ, περιοχές γονιδιωμάτων που ταιριάζουν μεταξύ τους, για να εφαρμοστούν πάνω τους αλγόριθμοι ευθυγράμμισης. Η εύρεση αυτών των περιοχών ανάμεσα σε 2 είδη δεν είναι συνηθισμένη υπολογιστικά διαδικασία. Εργαλεία τοπικών ευθυγραμμίσεων βρίσκουν κομμάτια με αρκετά υψηλές βαθμολογίες ταιριάσματος και επιπλέον αναγνωρίζουν διάφορες μη λογικές σχέσεις ή ακόμα και λάθος ευθυγραμμίσεις που οδηγούν σε επαναλήψεις μιας απλής ακολουθίας και άλλα προβλήματα. Οι πρώτες προσεγγίσεις για σύγκριση ολόκληρου του γονιδιώματος ανθρώπου και ποντικών βασίστηκαν είτε σε τοπική ευθυγράμμιση είτε σε τοπική-καθολική ευθυγράμμιση όπου κομμάτια του ενός γονιδιώματος χαρτογραφούνται πάνω

στο δεύτερο από έναν τοπικό ευθυγραμμιστή, επιβεβαιώνοντας το ταίριασμα μέσω ενός καθολικού ευθυγραμμιστή.

Τοπική ευθυγράμμιση σε κλίμακα ολόκληρου γονιδιώματος

Ίσως η πιο απλή προσέγγιση για να ευθυγραμμίσουμε δυο ολόκληρα γονιδιώματα είναι να κάνουμε μια τοπική ευθυγράμμιση. Αυτό είναι μια απαιτητική διαδικασία εξαιτίας του όγκου των πράξεων που πρέπει να γίνουν, αλλά επίσης και του προβλήματος να θέσουμε το σωστό κατώφλι για τις μεμονωμένες τοπικές ευθυγραμμίσεις: αν το κατώφλι είναι πολύ χαμηλό υπάρχει πιθανότητα να προκύψουν πολλές λανθασμένες τοπικές ευθυγραμμίσεις. Αντιθέτως αν είναι πολύ ψηλό, μπορεί ο αλγόριθμος να απορρίψει κάποιες σωστές ευθυγραμμίσεις. Επίσης, οι κλασσικές μέθοδοι για την τοπική ευθυγράμμιση δεν παίρνουν υπόψιν ότι μια συγκεκριμένη ευθυγράμμιση είναι μέρος μιας ευρύτερης συνταινιακής περιοχής, το οποίο οδηγεί σε δυσκολίες στις τοπικές ευθυγραμμίσεις που είναι τα αποτελέσματα των επαναλήψεων που δεν καλύφτηκαν, είτε παράλογα αντίγραφα. Παρ'όλ'αυτά οι τοπικές ευθυγραμμίσεις ήταν η μέθοδος που χρησιμοποιήθηκε αρχικά για τις συγκρίσεις ολόκληρου γονιδιώματος, καθώς έδειχναν καλύτερα τις αναδιατάξεις μεταξύ δυο μεγάλων γονιδιωμάτων από θηλαστικά, όπως ανθρώπου και ποντικιού.

Σύγκριση Γονιδιακής Δομής Ανθρώπου και Ποντικιού

Το βασικό πρόβλημα στην ανάλυση γονιδιωμάτων είναι η εύρεση γονιδίων. Είναι σχετικά εύκολο να τα βρούμε σε οργανισμούς με σχετικά μικρά γονιδιώματα, όπως βακτήρια ή σκουλήκια, αφού τα εξόνια τείνουν να είναι μεγαλύτερα και τα εσώνια είναι είτε ανύπαρκτα είτε πολύ μικρά. Μεγαλύτερη

πρόκληση αποτελούν τα μεγάλα γονιδιώματα, όπως των θηλαστικών, μιας και τα εξόνια είναι σκορπισμένα σε μια θάλασσα από πληροφορία και θόρυβο. Έτσι οι ακολουθίες εξονίων που μας ενδιαφέρουν αποτελούν το 75% στους μύκητες, και μόνο το 3% στο ανθρώπινο γονιδίωμα.
Η τεχνική που μπορεί να χρησιμοποιηθεί στην γονιδιακή αναγνώρηση είναι η σύγκριση γονιδιωμάτων ανάμεσα σε διαφορετικά είδη, δηλαδή η ταυτόχρονη ανάλυση των όμοιων loci σε δύο συγγενικά είδη, συγκεκριμένα ανθρώπου και ποντικιού. Όπως είναι γνωστό η σύγκριση γονιδιωμάτων ανάμεσα σε συγγενικά είδη μπορεί να τονίσει σημαντικά λειτουργικά στοιχεία όπως τα εξόνια, αφού αυτά τα στοιχεία τείνουν να συντηρούνται από την εξέλιξη καλύτερα ανάμεσα στα είδη από άλλες τυχαίες γονιδιακές ακολουθίες.
Αρχικά, έγινε η συστηματική σύγκριση 117 γονιδιωματικών ζευγαριών, ώστε να υπάρχει μια κατανόηση του επιπέδου της συντήρησης καθώς και του πλήθους και μήκους των εξονίων και εσωνίων. Με βάση αυτά τα αποτελέσματα αναπτύχθηκε ένα νέο πρόγραμμα για καθολική ευθυγράμμιση μεγάλων γονιδιωματικών περιοχών με την χρήση ιεραρχικής ευθυγράμμισης, GLASS και το ROSSETTA, πρόγραμμα αναγνώρισης κωδικοποιημένων εξονίων και στα δυο είδη με βάση την πιθανότητα της γονιδιωματικής δομής.
Η σύγκριση αποκάλυψε ενα φαινομενικό επίπεδο εξελικτικής συντήρησης. Ο αριθμός των εξονίων και στα δυο είδη που μελετήθηκαν ήταν όμοιος σε βαθμό 95%. Επίσης τα μήκη των ανάλογων εξονίων ήταν ισχυρά συντηρημένα και όμοια σε βαθμό 73%. Σε αντίθεση με τα εξόνια, τα εσώνια διέφεραν αρκετά στο μήκος τους. Τα ανθρώπινα εσώνια τείναν να είναι μεγαλύτερα από ποντικιού στο 68% των περιπτώσεων. Μέσα από την σύγκριση προέκυψε ότι τα εξόνια είχαν πολλές ισχυρές ομοιότητες, γύρω στο 85%, σε αντίθεση με τα εσώνια που είχαν πολύ αδύναμη ομοιότητα στο 35%, που δεν είναι πολύ μεγαλύτερη από τον θόρυβο του παρασκηνίου που προκύπτει από την ευθυγράμμιση τυχαίων ακολουθιών με κενά. Το επίπεδο της συντήρησης διέφερε σημαντικά ανάμεσα στα γονίδια.

Οπτικοποίηση

Ύστερα από την ευθυγράμμιση δυο ή περισσοτέρων γονιδιωματικών ακολουθιών το επόμενο βήμα είναι η ανάλυση των επιπέδων της συνολικής ομοιότητας, η κατανομή των άκρως συντηρημένων περιοχών και άλλων συγκριτικών χαρακτηριστικών. Το στάδιο της οπτικοποίησης των αποτελεσμάτων είναι ύψιστης σημασίας στην διαδικασία της σύγκρισης, αφού ο χειροκίνητος έλεγχος σε κλίμακα μεγαβάσης δεν είναι δυνατός.

Οπτικοποίηση ευθυγράμμισης δυο ακολουθιών

Υπάρχουν αρκετά εργαλεία οπτικοποίησης διαθέσιμα στο κοινό για μεγάλες ευθυγραμμίσεις σε ζεύγη ακολουθιών DNA. Το εργαλείο PIPMarker αναπαριστά το επίπεδο της συντήρησης σε περιοχές χωρίς κενά της BLASTZ τοπικής ευθυγράμμισης με 2 διαφορετικούς τρόπους: γραφικές παραστάσεις με ποσοστά ταυτότητας(pips) και διάγραμμα σημείων. Τα pips παρουσιάζουν μια συμπαγή και κατανοητή απεικόνιση τοπικών ευθυγραμμίσεων. Το VISTA απεικονίζει τα δεδομένα της σύγκρισης σε κυματομορφή, όπου η συντήρηση υπολογίζεται σε ένα συρόμενο παράθυρο καθολικής ευθυγράμμισης με κενά και είναι ένα εργαλείο που βασίζεται σε καθολικές ευθυγραμμίσεις. Στο διάργαμμά του ο άξονας Χ αναπαριστά την ακολουθία βάση και ο άξονας Υ αναπαριστά το ποσοστό ομοιότητας.
Προς το παρόν, η κατηγοριοποίηση των εργαλείων που υπάρχουν γίνεται με βάση το είδος της ευθυγράμμισης που χρησιμοποιούν για την ανάλυση, αλλά όσο περνά ο καιρός οι οι αλγόριθμοι ευθυγράμμισης γίνονται όλο και πιο εξεζητημένοι,

με αποτέλεσμα η διάκριση των εργαλείων και προγραμμάτων τοπικής και καθολικής ευθυγράμμισης να είναι πιο δύσκολη.

Οπτικοποίηση πολλαπλών ευθυγραμμίσεων

Και το Vista και το PIPMarker έχουν υιοθετήσει μια προσέγγιση που κάνει δυνατή την οπτικοποίηση πολλαπλών ευθυγραμμίσεων με την προβολή της ευθυγράμμισης σε μια συγκεκριμένη ακολουθία βάση με αποτέλεσμα την οπτικοποίηση ευθυγραμμίσεων σε ζεύγη μεταξύ της βάσης ακολουθίας και όμοιων κομματιών. Η προσέγγιση αυτή, όμως, δείχνει μόνο ένα μέρος της καθολικής ευθυγράμμισης, αφού μεταξύ των κομματιών των ακολουθιών που δεν υπάρχουν στην ακολουθία βάση δεν γίνεται ταίριασμα. Για παράδειγμα, αν ευθυγραμμίσουμε γονιδιώματα από άνθρωπο, ποντίκι και αρουραίο, χρησιμοποιώντας τον άνθρωπο, ως βάση, οι περιοχές συντήρησης μεταξύ ποντικιού και αρουραίου, που δεν υπάρχουν στον άνθρωπο, δεν θα απεικονιστούν. Η πλήρης οπτικοποίηση πολλαπλών ευθυγραμμίσεων είναι δύσκολη και σε μεγάλο βαθμό άλυτο πρόβλημα και προς τα παρόν είναι αντικείμενο ερευνητικού ενδιαφέροντος.

Το πρώτο εργαλείο το οποίο απεικόνιζε πολλαπλές ευθυγραμμίσεις ήταν το SynPlot. Το γραφικό αποτέλεσμα περιλαμβάνει ευθυγραμμίσεις μεγάλου μήκους μαζί με διαγραμματική αναπαράσταση και των δυο loci. Σε αντίθεση με το PIPMarker και το VISTA, το SynPlot χρησιμοποιεί την ευθυγράμμιση ως βάση συντεταγμένη, έτσι ώστε οι θέσεις όλων των χαρακτηριστικών των μεμονωμένων ακολουθιών να είναι χαρτογραγημένες στην ευθυγράμμιση. Τα αρχεία που παράγονται περιέχουν τις θέσεις των εξονίων και τον αριθμό των επαναλήψεων και μπορούν να εξαχθούν άμεσα ώς στην οθόνη. Έτσι το SynPlot εκφράζει σε μια γραφική γραμμική παράσταση την συγκριτική γονιδιωματική δομή, το μοτίβο των επαναλήψεων και την σχετική ομολογία των ακολουθιών. Το κύριο μειονέκτημά του είναι ότι δεν επιτρέπει στον χρήστη να διακρίνει την πηγή της ομοιότητας μέσα σε πολλαπλές ευθυγραμμίσεις, αφού μια ισχυρά συντηρημένη περιοχή σε 3

από τις 5 ακολουθίες θα μοιάζει ίδια με μια όχι τόσο καλά συντηρημένη περιοχή ομοιότητας που υπάρχει και στις 5 ακολουθίες. Το Phylo-VISTA, ένα πρόγραμμα που αναπτύχθηκε πρόσφατα από την Vista, χρησιμοποιεί την φυλογενετική σχέση σαν οδηγό για να απεικονίσει και να αναλύσει τον βαθμό της συντήρησης στους εσωτερικούς κόμβους του δέντρου. Η χρήση όλης της πολλαπλής ευθυγράμμισης, όχι απλά μιας ακολουθίας αναφοράς, σαν βάση στον άξονα Χ δίνει επιπλέον δυνατότητες στην απεικόνιση, όπως παρουσίαση των συγκριτικών δεδομένων μαζί με διαθέσιμα σχόλια για όλες τις ακολουθίες και υπολογισμό του μέτρου της ομοιότητας για οποιοδήποτε κόμβο του δέντρου. Η φυλογενετική σχέση μεταξύ των ειδών είναι σημαντική για την πολλαπλή ευθυγράμμιση και την ανάλυσή της, και συνεπώς την απεικόνιση των δεδομένων της ευθυγράμμισης.

Οπτικοποίηση ευθυγράμμισης πλήρους γονιδιώματος

Η πρόκληση της ευθυγράμμισης πλήρους γονιδιώματος, εμπεριέχει και την πρόκληση της οπτικοποίησης, δηλαδή πως να απεικονίζουν την πληροφορία μιας τεράστιας βάσης και πως να δώσουμε την δυνατότητα στον χρήστη να να αλληλεπιδράσει με τα δεδομένα και το πρόγραμμα επεξεργασίας. Μέθοδος του να επιλέγουμε ένα ολόκληρο γονιδίωμα σαν βάση ακολουθία χρησιμοποιήθηκε σε κλίμακα γονιδιώματος στον UCSC γονιδιωματικό περιηγητή και στον VISTA περιηγητή. Αυτά τα εργαλεία παρέχουν επιπρόσθετη πληροφορία για αρκετά γονιδιώματα, όπως ανθρώπου, ποντικιού, αρουραίου. Ο περιηγητής UCSC αναπαριστά τα σχόλια σαν οριζόντια ίχνη πάνω στην γονιδιωματική ακολουθία. Κάθε ίχνος αντιπροσωπεύει συγκεκριμένου είδους σχόλιο και υπάρχουν δύο είδη: συγκριτικά δεδομένα και διάφορες στατιστικές μετρήσεις των ευθυγραμμίσεων.
Ο περιηγητής Vista είναι μια εφαρμογή γραμμένη σε Java και ο σκοπός του είναι διαδραστική αναπαράσταση των

αποτελεσμάτων της ευθυγράμμισης ολόκληρων γονιδιωμάτων σε κλίμακα πλήρους χρωμοσώματος μαζί με σχόλια. Ο χρήστης μπορεί να επιλέξει οποιοδήποτε γονιδίωμα σαν σημείο αναφοράς ή βάση, και να δει το επίπεδο συντήρησης μεταξύ της βάσης και ακολουθίας που ανήκει σε άλλο είδος σε συγκεκριμένο διάστημα.

Εφαρμογές της Ευθυγράμμισης

Η αναγνώριση των ρυθμιστικών στοιχείων συχνά εμφανίζει μεγάλη πρόκληση στον σχολιασμό των γονιδιωμάτων των μεγαλύτερων σπονδυλωτών, εξαιτίας του μήκους αυτών των στοιχείων που είναι πολύ μικρό, αλλά και της λίγης πληροφορίας που περιέχουν. Η πρόσθεση των μεθόδων της συγκριτικής ανάλυσης ακολουθιών επέτρεψε την βελτίωση και αναβάθμιση της αναζήτησης για σήματα. Αυτές οι μέθοδοι βοηθούν στο φιλτράρισμα των υπολογιστικών προβλέψεων αφού μειώνουν τον θόρυβο των λανθασμένων προβλέψεων με κόστος την μείωση της ευαισθησίας.
Ενδείξεις για την αναγνώριση ακολουθιών που εμπεριέχονται στα ρυθμιστικά δίκτυα των ευκαριωτικών γονιδιωμάτων παρέχονται από την παρουσία συγκεκριμένων μοτίβων(TFBS), συμπλέγματα από τέτοια μοτίβα, και η συντήρηση αυτών των περιοχών που υπάρχει ανάμεσα στα είδη. Το εργαλείο rVista εκμεταλλεύεται όλες αυτές τις καθιερωμένες στρατηγικές για να βελτιώσει την ανίχνευση των ρυθμιστικών ακολουθιών που ελέγχουν την έκφραση του γονιδιώματος με την χρήση της ικανότητας του να αναγνωρίζει εξελικτικά συντηρημένα TFBSs. Το πλεονεκτήματα του αλγορίθμου rVista είναι η ικανότητα του να αναλύει αποδοτικά μεγάλες ακολουθίες γονιδιωμάτων και πιθανόν και ολόκληρα γονιδιώματα. Η οπτικοποίηση η οποία καθορίζεται από τον χρήστη το κάνει να είναι το πλέον σωστό εργαλείο για έρευνα των TFBS. Εκμεταλλευόμενο τον συνδυασμό της αναγνώρισης των μοτίβων με την πολλαπλή ευθυγράμμιση των ορθολογικών περιοχών, rVista κάνει την ανάλυσή του σε τέσσερα βήματα: την αναγνώριση των

ταιριασμάτων των TFBSs στις ξεχωριστές ακολουθίες, την αναγνώριση των καθολικά ευθυγραμμισμένων TFBS, τον υπολογισμό της τοπικής συντήρησης, την οπτικοποίηση των ξεχωριστών ή συμπλεγμένων TFBSs. Ένα άλλο εργαλείο, το Consite, χρησιμοποιεί την ίδια αρχή του συνδυασμού της πρόβλεψης των TFBSs και πληροφορίας για την συντήρηση της ακολουθίας και παρέχει μια αποδοτική διαδικτυακή εφαρμογή με γραφικό περιβάλλον για την οπτικοποίηση των αποτελεσμάτων της ανάλυσης.

Εν κατακλείδι

Σε αυτήν την εργασία έγινε εκτεταμένη αναφορά σε καθιερωμένα εργαλεία και μεθοδολογίες που έχουν χρησιμοποιηθεί για την δημιουργία των ευθυγραμμίσεων των γονιδιωματικών ακολουθιών, αλλά και των αποτελεσμάτων που αποφέρουν αυτά τα εργαλεία. Πρέπει να τονιστεί ότι η Συγκριτική Γονιδιωματική είναι ένα έντονο και εξελισσόμενο πεδίο, το οποίο έχει να προσφέρει πολλά και να ρίξει φως στα μυστήρια της εξέλιξης.

Πηγές

Srinivas Aluru, Handbook of Computational Molecular Biology: Comparison of Long Genomic Sequenced: Algorithms and Applications.

S. Batzoglou, L.Pachter, J.P. Mesirov. Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. Genome Res, 10(7):950-958, Jul 2000

S.F. Altschul, W. Gish, W. Miller, Basic Local Alignment Search Tool. Journal of Molecural Biology, 215:403-410, 1990