Κεφάλαιο 11

Υπολογιστική Γονιδιωματική

Σύνοψη
Στο κεφάλαιο αυτό εξετάζονται οι υπολογιστικές τεχνικές που χρησιμοποιούνται στη μελέτη ολόκληρων γονιδιωμάτων. Εκτός από το ξεκάθαρο ενδιαφέρον που έχουν αυτές οι τεχνικές στη φυλογενετική ανάλυση και στην εξέλιξη, υπάρχει και άλλος πιο πρακτικός λόγος για τη χρησιμότητά τους. Παρόλο που τεχνικές που αναπτύχθηκαν σε προηγούμενα κεφάλαια όπως η στοίχιση και οι προγνώσεις είναι εύκολο να χρησιμοποιηθούν σε ολόκληρα γονιδιώματα, η ταυτόχρονη αξιολόγηση της θέσης του κάθε γονιδίου στα γονδιώματα συγγενικών οργανισμών (συγκριτική γονιδιωματική) μπορεί να δώσει πολλές επιπλέον πληροφορίες για μια σειρά από λειτουργικές ιδιότητες οι οποίες δεν θα μπορούσαν να είχαν προβλεφθεί με άλλον τρόπο.

Προαπαιτούμενη γνώση
Το κεφάλαιο απαιτεί κατανόηση των μεθόδων των κεφαλαίων 3, 4, 6 και 7.

11. Εισαγωγή

Γονιδιωματική, ονομάζουμε τον επιστημονικό κλάδο ο οποίος χρησιμοποιεί διαφορετικές τεχνικές της γενετικής, της μοριακής βιολογίας και της βιοπληροφορικής με σκοπό να βρει την αλληλουχία, να κάνει την συναρμολόγηση και να αναλύσει τη δομή και τη λειτουργία των γονιδιωμάτων, δηλαδή, ολόκληρης της γενετικής πληροφορίας που περιέχεται σε ένα κύτταρο ενός οργανισμού. Υπάρχουν πολλές υποδιαιρέσεις της γονιδιωματικής, κυρίως όσον αφορά τις διαφορετικές τεχνικές που είναι δυνατό να χρησιμοποιηθούν κάθε φορά. Για παράδειγμα, η δομική γονιδιωματική ασχολείται με το μαζικό προσδιορισμό τρισδιάστατων δομών πρωτεϊνών από ολόκληρα γονιδιώματα, ενώ η λειτουργική γονιδιωματική ασχολείται κυρίως με τη μελέτη των λειτουργικών περιοχών στα γονιδιώματα (υποκινητές, μικρά RNA κλπ).

Στο παρόν κεφάλαιο, θα εστιάσουμε στις υπολογιστικές τεχνικές που χρησιμοποιούνται στην ανάλυση γονιδωμάτων. Σε πρώτο επίπεδο, και με βάση τον γενικότερο ορισμό, υπολογιστική γονιδιωματική είναι και κάθε προσπάθεια ανάλυσης του γονιδιώματος ενός και μόνο οργανισμού, δηλαδή οι τεχνικές αλληλούχισης και συναρμολόγησης του γονιδιώματος (Zerbino & Birney, 2008), η εύρεση γονιδίων (Picardi & Pesole, 2010), η εύρεση ρυθμιστικών περιοχών (Harbison et al., 2004), η εύρεση μικρών RNA (Rigoutsos, 2010; Vlachos & Hatzigeorgiou, 2013) ή η εύρεση περιοχών οριζόντιας γονιδιακής μεταφοράς (Soucy, Huang, & Gogarten, 2015) και η εύρεση του τρόπου γονιδιακής ρύθμισης. Σε ένα επόμενο επίπεδο, οι τεχνικές που χρησιμοποιούνται είναι απλά εφαρμογές σε ολόκληρα γονιδιώματα, γνωστών μεθόδων και αλγορίθμων που σχεδιάστηκαν για αλληλουχίες (π.χ. μέθοδοι πρόγνωσης), και στη συνέχεια, στατιστική ανάλυση των αποτελεσμάτων με σκοπό την εξαγωγή γενικότερων κανόνων και συμπερασμάτων. Θα παρουσιάσουμε κάποια τέτοια παραδείγματα με σκοπό να εξοικειωθεί ο αναγνώστης με τη μεθοδολογία. Στο επόμενο στάδιο όμως, θα παρουσιαστούν οι πιο ενδιαφέρουσες τεχνικές της συγκριτικής γονιδιωματικής, οι οποίες προσφέρουν κάτι επιπλέον: αξιοποιώντας την πληροφορία για την ύπαρξη, τη θέση και την εσωτερική δομή των γονιδίων στα γονιδιώματα διαφόρων υπό σύγκριση οργανισμών, μπορούν να μας δώσουν επιπλέον πληροφορίες, πληροφορίες που από μια απλή ανάλυση ενός οργανισμού (και του γονιδιώματός του) δεν θα μπορούσαν να εξαχθούν.

Στο τέλος, θα παρουσιαστούν κάποια γνωστά παραδείγματα εφαρμογής των μεθόδων αυτών, αλλά και τα βασικά εργαλεία λογισμικού που χρησιμοποιούνται σε τέτοιου είδους αναλύσεις.

11.1. Υπολογιστική ανάλυση γονιδιωμάτων

Όπως είδαμε, ο όρος «υπολογιστική γονιδιωματική» είναι αρκετά γενικός, και πολλών ειδών υπολογιστικές αναλύσεις μπορούν να θεωρηθούν ως τέτοιες. Για παράδειγμα, κάποιος επιστήμονας μπορεί να μελετήσει ένα γονιδίωμα για να βρει πόσα γονίδια αυτό περιέχει ή ποιες είναι οι αποστάσεις μεταξύ τους ή ποια είναι η κατανομή κάποιου άλλου ειδικού χαρακτηριστικού (π.χ. ποια γονίδια ελέγχονται από κάποιο συγκεκριμένο

μεταγραφικό παράγοντα, ποια γονίδια κωδικοποιούν μεμβρανικές πρωτεΐνες κ.ο.κ.). Μπορεί επίσης να ενδιαφέρει η εύρεση μικρών RNA ή η εύρεση περιοχών οριζόντιας γονιδιακής μεταφοράς και η εύρεση του τρόπου γονιδιακής ρύθμισης. Επιπλέον, πολλές από τις αναλύσεις τις γενετικής όπως η εύρεση πολυμορφισμών ή η εύρεση επαναληπτικών αλληλουχιών μπορεί να εμπίπτει στον ορισμό της γονιδιωματικής.

Μία πιο μεγάλης κλίμακας ανάλυση θα λάβει χώρα όταν αναλυθούν παράλληλα πολλά γονιδιώματα για κάποια από τα παραπάνω χαρακτηριστικά (π.χ. για τη σύσταση GC ή για τον αριθμό των γονιδίων που κωδικοποιούν μεμβρανικές πρωτεΐνες κ.ο.κ.). Σε αυτή την περίπτωση, οδηγούμαστε τελικά σε μια ανάλυση στην οποία κάθε «γραμμή» στο αρχείο μας (δηλαδή, κάθε παρατήρηση όπως λέμε στη στατιστική) αντιστοιχεί σε ένα γονιδίωμα, ενώ κάθε στήλη (μεταβλητή) αποτελεί το χαρακτηριστικό που μελετάμε. Συνήθως τέτοιες αναλύσεις συνδυάζονται, έμμεσα ή άμεσα, με φυλογενετικά δεδομένα με σκοπό να δείξουν την κατανομή του υπό μελέτη χαρακτηριστικού στις διάφορες ταξινομικές βαθμίδες των οργανισμών υπό μελέτη. Ο σκοπός τέτοιων αναλύσεων, είναι η εξαγωγή συνολικών συμπερασμάτων και κανόνων από την ταυτόχρονη μελέτη πολλών διαφορετικών γονιδιωμάτων.

Μία πολύ απλή τέτοια γονιδιωματική ανάλυση, αλλά με τεράστια σημασία, αφορά τις αναλύσεις που έδειξαν ότι σε όλους τους οργανισμούς, οι α-ελικοειδείς διαμεβρανικές πρωτεΐνες αντιστοιχούν σε περίπου 20-30% των πρωτεϊνών που κωδικοποιούνται από τα γονιδιώματα αυτά, ενώ τα διαμεβρανικά β-βαρέλια αντιστοιχούν σε περίπου 1-2% των βακτηριακών γονιδιωμάτων. Άλλες τέτοιες αναλύσεις αφορούν τους GPCRs (οι οποίοι αποτελούν τη μεγαλύτερη οικογένεια διαμεμβρανικών υποδοχέων στα θηλαστικά, με περίπου 2% του γονιδιώματος) ή τις εκκρινόμενες πρωτεΐνες που αντιστοιχούν σε περίπου 15% των πρωτεϊνών που κωδικοποιούνται από τα γονιδιώματα όλων των οργανισμών. Επίσης, σημαντικές γονιδιωματικές αναλύσεις, για την ιδιαίτερα σημαντική αυτή ομάδα των πρωτεϊνών (διαμεμβρανικές πρωτεΐνες), έχουν γίνει για να απαντήσουν το ερώτημα του κατά πόσο στην εξέλιξή τους έχει γίνει εκτεταμένη χρήση του φαινομένου του εσωτερικού γονιδιακού διπλασιασμού. Στην αρχική ανάλυση, βρέθηκε από συσχετίσεις του μήκους των πρωτεϊνών με τον αριθμό των διαμεμβρανικών τμημάτων ότι κάτι τέτοιο είναι πιθανό (Arai, Ikeda, & Shimizu, 2003). Κατόπιν, με στοιχίσεις του πρώτου «μισού» των πρωτεϊνών αυτών με το δεύτερο μισό (δηλαδή, της μισής ακολουθίας προς το αμινοτελικό άκρο με την ακολουθία προς το καρβοξυτελικό άκρο), βρέθηκε ότι ανάμεσα σε 38,174 διαμεμβρανικές πρωτεΐνες από 87 γονιδιώματα, 377 ήταν δυνατό να έχουν παραχθεί από ένα μηχανισμό εσωτερικού διπλασιασμού και αφορούσαν κυρίως περιπτώσεις με 8, 10 και 12 διαμεμβρανικά τμήματα (Shimizu, Mitsuke, Noto, & Arai, 2004).

Φυσικά, σε κάθε ανάλυση γονιδιωμάτων είναι απαραίτητη και μια ανάλυση των πρωτεϊνικών οικογενειών με βάση τα δεδομένα κάποιων από τις βάσεις πρωτεϊνικών δεδομένων που είδαμε στο κεφάλαιο 2. Οι βάσεις αυτές μπορεί να είναι οι βάσεις γενική χρήσης όπως PFAM ή πιο εξειδικευμένες όπως η TCDB, η CAZy κ.ο.κ.

	X_1	X_2	X_3	X_4	…	X_k
Γονιδίωμα Α						
Γονιδίωμα Β						
Γονιδίωμα Γ						
Γονιδίωμα Δ						
…						

Εικόνα 11.1: *Παράδειγμα κωδικοποίησης πληροφορίας από γονιδιώματα.*

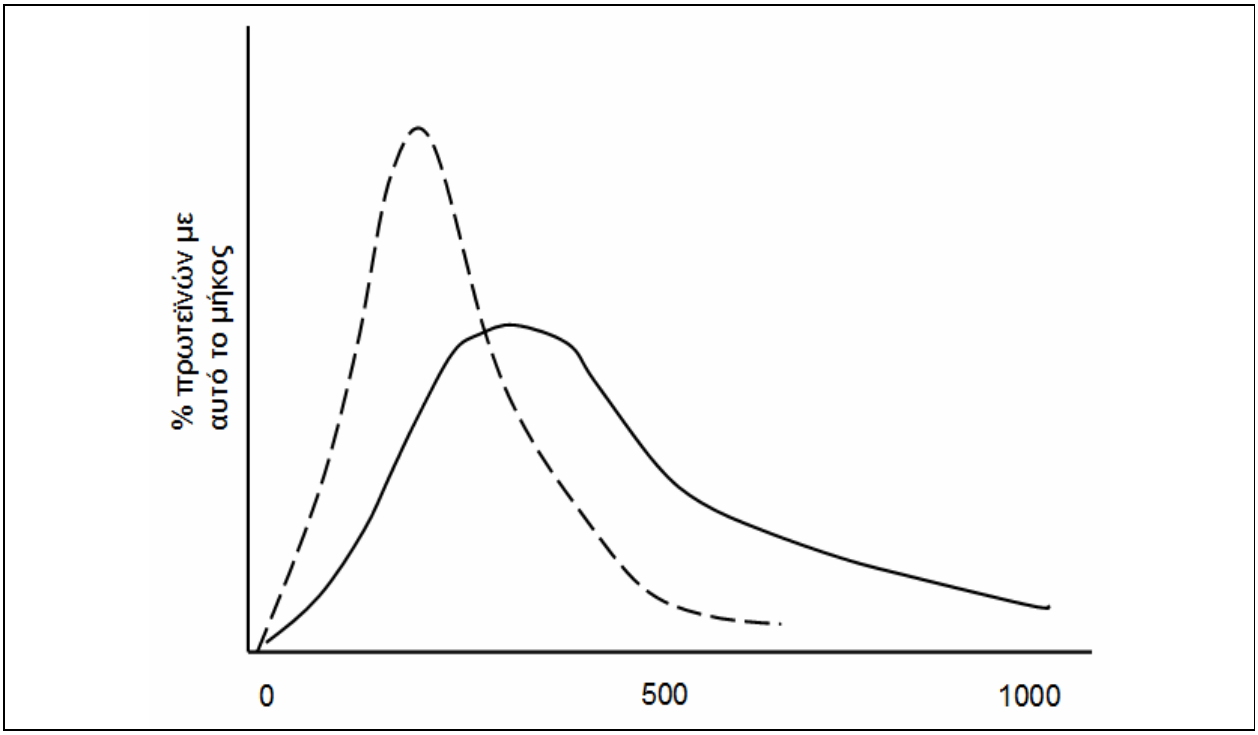
Σε άλλες περιπτώσεις, χρειάζεται να συμπτύξουμε την πληροφορία των γονιδιωμάτων και να την περιορίσουμε με χρήση μερικών μόνο παραμέτρων. Το ποσοστό των βάσεων GC είναι μία από τις πιο ευρέως χρησιμοποιούμενες παραμέτρους όταν επιθυμούμε να συγκρίνουμε διαφορετικούς οργανισμούς με βάση το γονιδίωμα τους (Li, 2011). Στα γονιδιώματα διαφόρων οργανισμών παρατηρούνται διαφορετικά ποσοστά εμφάνισης GC (το λεγόμενο GC% content). Οι συνέπειές του είναι πολλές, κυρίως γιατί τα διαφορετικά κωδικόνια για τα ίδια αμινοξέα εμφανίζονται με διαφορετικές συχνότητες και έτσι, τελικά γονιδιώματα με διαφορετικό ποσοστό GC καταλήγουν να κωδικοποιούν πρωτεΐνες με διαφορετική περιεκτικότητα σε αμινοξέα.

Το φαινόμενο αυτό, ονομάζεται «codon bias» (πολωμένη σύσταση των κωδικονίων) και παρατηρείται σε όλους τους οργανισμούς, τόσο στο γονιδιωματικό επίπεδο όσο και μεταξύ λειτουργικώς συνδεδεμένων γονιδίων (π.χ. οπερόνια), αλλά και σε μεμονωμένα γονίδια. Άλλες παραλλαγές του φαινομένου περιλαμβάνουν τα πολωμένα ζευγάρια κωδικονίων και την πολωμένη συν-εμφάνιση κωδικονίων. Παρόλο που είναι γενικά αποδεκτό ότι η έναρξη της μετάφρασης είναι το βασικό σημείο στην πρωτεϊνοσύνθεση, είναι επίσης αναγνωρισμένο ότι το codon bias παίζει ρόλο συνεισφέροντας στην αποδοτικότητα της μετάφρασης ρυθμίζοντας τη φάση της επιμήκυνσης. Επιπλέον, παίζει σημαντικό ρόλο στον έλεγχο πολλών άλλων κυτταρικών διεργασιών οι οποίες ποικίλουν, από τη διαφορική σύνθεση πρωτεϊνών, μέχρι το πρωτεϊνικό δίπλωμα (Quax, Claassens, Soll, & van der Oost, 2015).

Σε μια από τις πρώτες, αρκετά απλές αλλά ιδιαίτερα πληροφοριακές τέτοιες μελέτες, ο Ouzounis και ο Kreil, ανέλυσαν την αμινοξική σύσταση των πρωτεϊνών που κωδικοποιούν τα γονιδιώματα 6 θερμόφιλων αρχαιοβακτηρίων (αρχαίων), 2 θερμόφιλων βακτηρίων, 17 μεσόφιλων βακτηρίων και 2 ευκαρυωτικών οργανισμών. Στην ανάλυση χρησιμοποίησαν την αμινοξική σύσταση και το ποσοστό GC και πραγματοποίησαν ιεραρχική ομαδοποίηση και ανάλυση κύριων συνιστωσών (principal components analysis). Παρόλο που το ποσοστό GC είχε μια ξεκάθαρη επιρροή, τα θερμόφιλα είδη μπορούν να αναγνωριστούν με μόνη χρήση της ολικής αμινοξικής σύστασης (Kreil & Ouzounis, 2001). Αναλύοντας τα αποτελέσματα, φάνηκε ότι τα θερμόφιλα είδη έχουν λιγότερη Γλουταμίνη (Gln) και περισσότερο Γλουταμικό (Glu) σε σχέση με τα μεσόφιλα. Τα θερμόφιλα, έχουν επίσης περισσότερη Βαλίνη (Val) και λιγότερη Θρεονίνη (Thr) σε σχέση με τα μεσόφιλα. Για τα αμινοξέα Ιστιδίνη (His), Σερίνη (Ser) και Ασπαραγίνη (Asn) υπήρχαν επίσης ενδείξεις αλλά με μικρότερο στατιστικό βάρος.

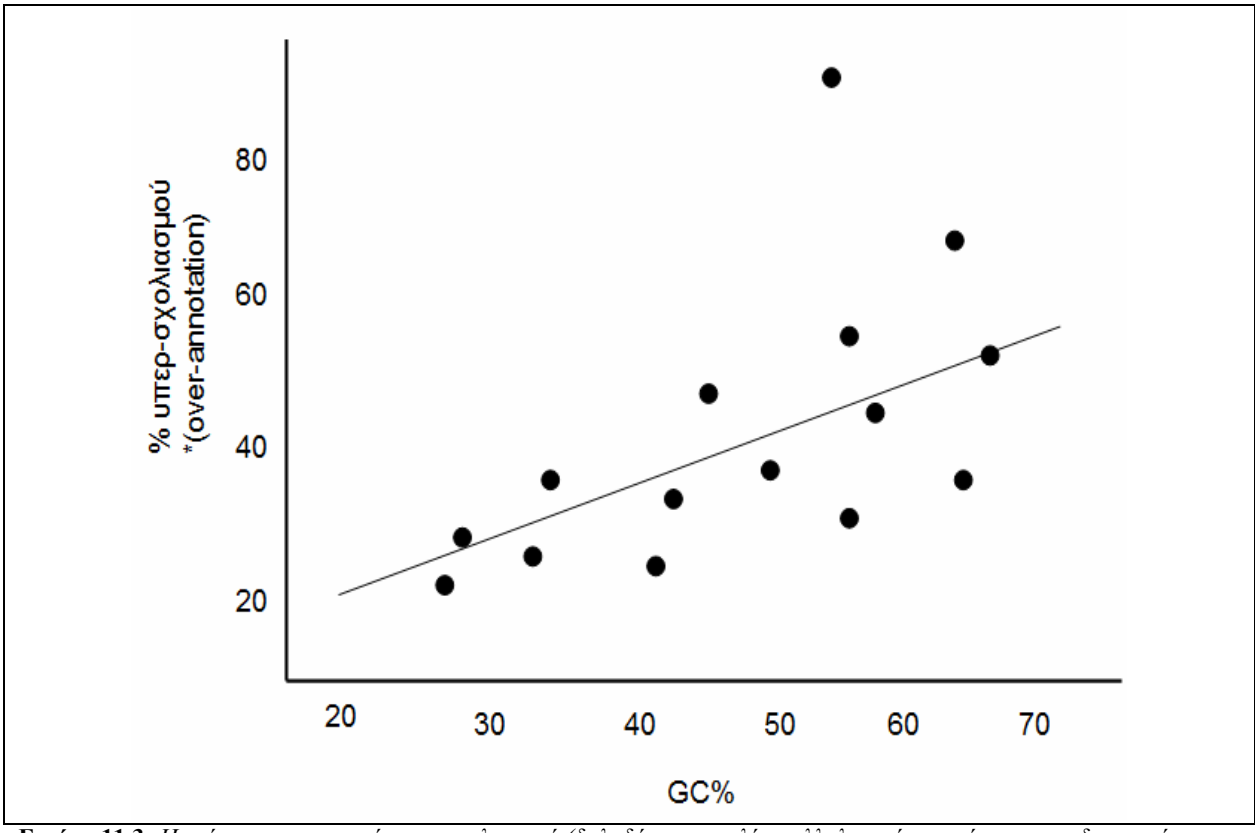
Μια άλλη ιδιαίτερα ενδιαφέρουσα εργασία, χρησιμοποίησε το ποσοστό GC με σκοπό να προσδιορίσει το μήκος των πραγματικών πρωτεϊνών στα γνωστά γονιδιώματα. Όπως είναι γνωστό, οι περισσότερες πρωτεΐνες είναι γνωστές όχι με πειραματικό τρόπο, αλλά έμμεσα κάνοντας χρήση της γονιδιωματικής πληροφορίας (conceptual translation). Έτσι, δημιουργείται το αναπόφευκτο ερώτημα αν όντως όλες αυτές οι «υποθετικές πρωτεΐνες» είναι πραγματικές ή αν αποτελούν τεχνητά προϊόντα, λάθη δηλαδή που προέκυψαν από τα προγράμματα εύρεσης γονιδίων. Η βασική σκέψη πίσω από την ανάλυση αυτή, είναι η εξής: Οι τριπλέτες είναι 64 (61 για αμινοξέα και 3 για λήξη). Αν τα νουκλεοτίδια θεωρηθούν τυχαία κατανεμημένα και ισοπίθανα (όπως απλουστευτικά είχαμε δει στο Κεφάλαιο 2), τότε θα έχουμε μια τριπλέτα λήξης περίπου μετά από 21 αμινοξέα. Όπως είδαμε, αυτό δίνει μια κατανομή για το μήκος των «ανοιχτών πλαισίων ανάγνωσης», η οποία αντιστοιχεί στη γεωμετρική κατανομή. Παρατηρούμε όμως ότι οι τριπλέτες λήξης είναι πλούσιες σε AT (TAA, TGA, TAG). Άρα, σε γονιδιώματα με μεγάλο λόγο AT οι τριπλέτες αυτές θα είναι πιο συχνές, ενώ στα γονιδιώματα με υψηλό λόγο GC, αυτές θα είναι πιο σπάνιες. Κατά συνέπεια το μήκος των «τυχαίων» ανοιχτών πλαισίων ανάγνωσης θα αυξάνει στα πλούσια σε GC γωνιδιώματα.

Οι ερευνητές λοιπόν, προχώρησαν βρίσκοντας όλα τα ORF από τα γνωστά βακτηριακά γονιδιώματα (34 εκείνη την εποχή). Κατόπιν, αφαίρεσαν τις πολύ ομόλογες πρωτεΐνες (Redundancy Reduction) και στη συνέχεια πραγματοποίησαν μια απλή σύγκριση με αναζήτηση ομοιότητας έναντι των πραγματικών (non-hypothetical) πρωτεϊνών της SwissProt (E-value$<10^{-6}$). Τα αποτελέσματα αναλύθηκαν με γραφικές παραστάσεις και στατιστικές μεθοδολογίες, ειδικά για να μπορέσει να εκτιμηθεί το ποσοστό του «υπερ-σχολιασμού» (over-annotation), δηλαδή των επιπλέον πρωτεϊνών που είχαν προβλεφθεί για το κάθε γονιδίωμα (Skovgaard, Jensen, Brunak, Ussery, & Krogh, 2001).



Εικόνα 11.2: *Ιστόγραμμα της κατανομής των μηκών των υποθετικών πρωτεϊνών που είχαν ομοιότητα με πραγματική πρωτεΐνη της SwissProt (συνεχής γραμμή), και αυτών που δεν είχαν ομοιότητα. (διακεκομμένη γραμμή).*

Τα αποτελέσματα της ανάλυσης ήταν εντυπωσιακά. Οι πρωτεΐνες των γονιδιωμάτων που είχαν ξεκάθαρη ομοιότητα με κάποια «σίγουρη» πρωτεΐνη της Swissprot, είχαν διαφορετική κατανομή του μήκους τους από αυτές οι οποίες δεν εμφάνισαν τέτοια ομοιότητα. Για την ακρίβεια, η δεύτερη ομάδα, αυτές που πιθανότατα ήταν αποτελέσματα ψευδών προβλέψεων των προγραμμάτων εύρεσης γονιδίων, ήταν μικρότερες κατά μέσο όρο και με μια κατανομή που προσέγγιζε τη γεωμετρική (Εικόνα 11.2). Το πρόβλημα αυτό μας θυμίζει αρκετά το πρόβλημα της «μίξης κατανομών» (mixture of distributions) στη στατιστική και στην ομαδοποίηση. Παρόλο που η διαφορά ήταν εμφανής οπτικά, δεν είναι και τόσο εύκολο να προβλεφθεί η ταυτότητα μιας συγκεκριμένης πρωτεΐνης, γιατί οι κατανομές δεν διαχωρίζονται επαρκώς. Για παράδειγμα, για μια πολύ μικρή (π.χ. 100 αμινοξέα) ή για μια αρκετά μεγάλη πρωτεΐνη (π.χ. 500 αμινοξέα), είναι αρκετά εύκολο να κάνουμε μια πρόβλεψη, αλλά για τις περισσότερες πρωτεΐνες που έχουν μήκος στην περιοχή 200 με 300 αμινοξέα, αυτό δεν είναι εύκολο. Παρ' όλα αυτά, είναι εύκολο αλλά και σημαντικό να προβλεφθεί με μεγάλη ακρίβεια ο συνολικός αριθμός των πρωτεϊνών που θα ανήκουν σε κάθε ομάδα. Με βάση αυτή την εκτίμηση, οι ερευνητές προχώρησαν σε μια απλή γραφική παράσταση του ποσοστού GC με το ποσοστό υπερ-σχολιασμού (δηλαδή, το ποσοστό των «ψεύτικων» πρωτεϊνών που αναμένουμε να υπάρχουν στο γονιδίωμα).



Εικόνα 11.3: *Η σχέση του ποσοστού υπερ-σχολιασμού (δηλαδή των επιπλέον αλληλουχιών που έχουν προσδιοριστεί λανθασμένα ως πραγματικές πρωτεΐνες) με το ποσοστό GC των γονιδιωμάτων.*

Τα αποτελέσματα (Εικόνα 11.3), επιβεβαίωσαν πλήρως το θεωρητικό μοντέλο, καθώς τα γονιδιώματα με υψηλό GC ήταν και αυτά με το μεγαλύτερο ποσοστό «ψεύτικων» πρωτεϊνών. Το ένα γονιδίωμα που ξεφεύγει από το διάγραμμα, καθώς εμφανίζει ένα ιδιαίτερα υψηλό ποσοστό «ψεύτικων» πρωτεϊνών, περίπου μία στις δύο πρωτεΐνες (τέτοιες παρατηρήσεις ονομάζονται outlier στη στατιστική), βρέθηκε μετά από αναζήτηση στη βιβλιογραφία ότι ανήκε στο βακτήριο *A. pernix*, στον προσδιορισμό του γονιδιώματος του οποίου, οι ερευνητές δεν χρησιμοποίησαν καν κάποιο πρόγραμμα εύρεσης γονιδίων, αλλά απλά ονόμασαν «πρωτεΐνη» κάθε ανοιχτό πλαίσιο ανάγνωσης. Φυσικά, δεν πρέπει να ξεχνάμε ότι η απλή αυτή γραμμική σχέση δεν εξηγεί 100% την μεταβλητότητα του δείγματος (με άλλα λόγια, τα σημεία είναι διασκορπισμένα αρκετά εκατέρωθεν της ευθείας γραμμής). Ο λόγος είναι προφανής: μιλάμε για διαφορετικά γονιδιώματα, τα οποία προσδιορίστηκαν και αναλύθηκαν από διαφορετικές ερευνητικές ομάδες, σε διαφορετικές εποχές, και με χρήση διαφορετικών εργαλείων. Αν μια παρόμοια ανάλυση γινόταν, για παράδειγμα, στα γονιδιώματα που προσδιορίστηκαν πρόσφατα, θα περιμέναμε ότι το ποσοστό των «ψεύτικων» πρωτεϊνών θα ήταν μειωμένο, γιατί τα σύγχρονα εργαλεία εύρεσης γονιδίων λειτουργούν καλύτερα.

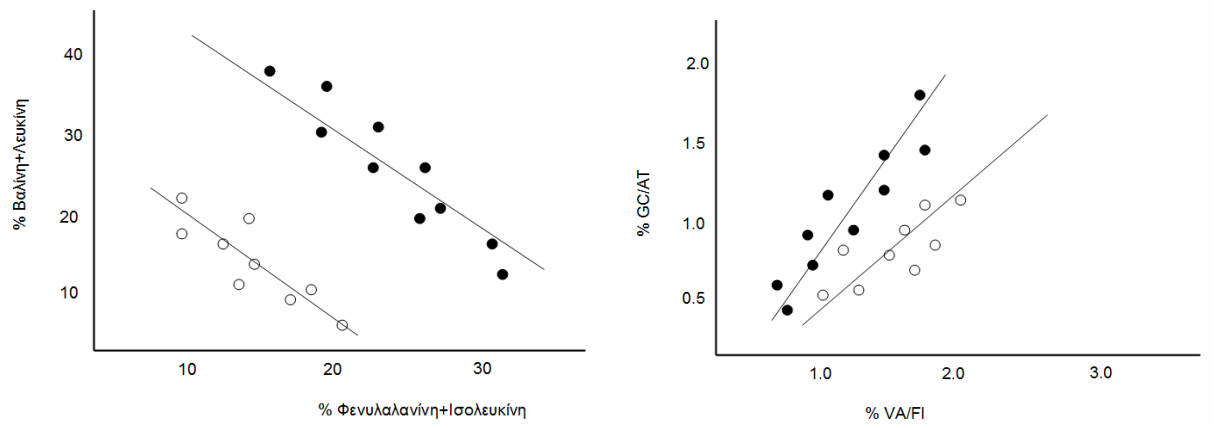
Ένα άλλο παράδειγμα γονιδιωματικής ανάλυσης που δίνει πολύ ενδιαφέροντα συμπεράσματα που σχετίζονται με το ποσοστό GC, αφορά τις διαμεμβρανικές πρωτεΐνες. Η εύρεση των α-ελικοειδών διαμεμβρανικών πρωτεϊνών στηρίζεται στη, με διάφορους τρόπους, αναζήτηση περιοχών πλούσιων σε υδρόφοβα κατάλοιπα. Τα γονιδιώματα όμως διαφέρουν στο ποσοστό GC όπως είδαμε πριν. Επιπλέον, τα κωδικόνια των υδρόφοβων αμινοξέων περιέχουν GC σε διαφορετικό βαθμό. Για την ακρίβεια, όπως φαίνεται στην Εικόνα 11.4, η Αλανίνη και η Γλυκίνη έχουν περισσότερα GC στα κωδικόνια τους, η Βαλίνη και η Λευκίνη έχουν ίδιο αριθμό GC και AT, ενώ η Ισολευκίνη και η Φενυλαλανίνη έχουν περισσότερο AT. Επιπλέον δε, η Αλανίνη και η Γλυκίνη, αν και υδρόφοβα αμινοξέα είναι «λιγότερο» υδρόφοβα σύμφωνα με τις περισσότερες κλίμακες υδροφοβικότητας. Κατά συνέπεια, ένας «γενικής χρήσης» αλγόριθμος πρόγνωσης μπορεί να υπερ- ή υπό-εκτιμά την πρόγνωση διαμεμβρανικών τμημάτων όταν εφαρμόζεται σε πρωτεΐνες από οργανισμούς με διαφορετικό GC. Θεωρητικά, αναμένουμε ότι σε γονιδιώματα πλούσια σε GC, η Αλανίνη και

η Γλυκίνη θα βρίσκονται σε μεγαλύτερη συχνότητα και κατά συνέπεια οι προγνώσεις για διαμεμβρανικά τμήματα θα είναι πιο «σπάνιες» (ή, πιο δύσκολες).

Αμινοξύ	**Κωδικόνιο**	**min(A+T)**	**min(G+C)**
Αλανίνη (Ala)	G-C-X	0	2
Γλυκίνη (Gly)	G-G-X	0	2
Βαλίνη (Val)	G-T-X	1	1
Λευκίνη (Leu)	C-T-X, T-T-[AG]	1	1
Ισολευκίνη (Ile)	A-T-[ACT]	2	0
Φενυλαλανίνη (Phe)	T-T-[CT]	2	0

Εικόνα 11.4: *Κατανομή των κωδικωνίων για τα υδρόφοβα αμινοξέα.*

Στην ανάλυση των τότε γνωστών γονιδιωμάτων, οι ερευνητές χρησιμοποίησαν ένα σχετικά απλό τρόπο εύρεσης των διαμεμβρανικών τμημάτων (έναν αλγόριθμο κυλιόμενου παραθύρου με χρήση κλίμακας υδροφοβικότητας), και συσχέτισαν τα ποσοστά Βαλίνης και Λευκίνης (VL) με αυτά της Φενυλαλανίνης και της Ισολευκίνης (FI). Το ίδιο έγινε όχι μόνο για τις ακολουθίες των διαμεμβρανικών τμημάτων, αλλά και για το σύνολο του πρωτεόματος. Τα αποτελέσματα έδειξαν μια πολύ ισχυρή συσχέτιση, όπως αναμενόταν. Επιπλέον δε, ο λόγος αυτός (VL/FI) εμφανίζει μια ξεκάθαρη συσχέτιση με το λόγο GC/AT. Με άλλα λόγια, επαληθεύεται η αρχική υπόθεση ότι τα γονιδιώματα που είναι πλούσια σε GC, έχουν συγκριτικά περισσότερες Βαλίνες και Αλανίνες σε σύγκριση με Ισολευκίνες και Φενυλαλανίνες. Αυτό έχει σαν συνέπεια τα διαμεμβρανικά τμήματα των διαμεμβρανικών πρωτεϊνών που κωδικοποιούνται σε αυτά τα γονιδιώματα, να είναι λιγότερο υδρόφοβα σε σχέση με αυτά των πρωτεϊνών που προέρχονται από οργανισμούς φτωχούς σε GC. Όλα τα παραπάνω, σημαίνουν ότι σε ακραίες περιπτώσεις, αυτές οι διαφορές θα πρέπει να λαμβάνονται υπόψη και (αν είναι δυνατόν) η πληροφορία αυτή να ενσωματωθεί ακόμα και στους αλγόριθμους πρόγνωσης διαμεμβρανικών τμημάτων.

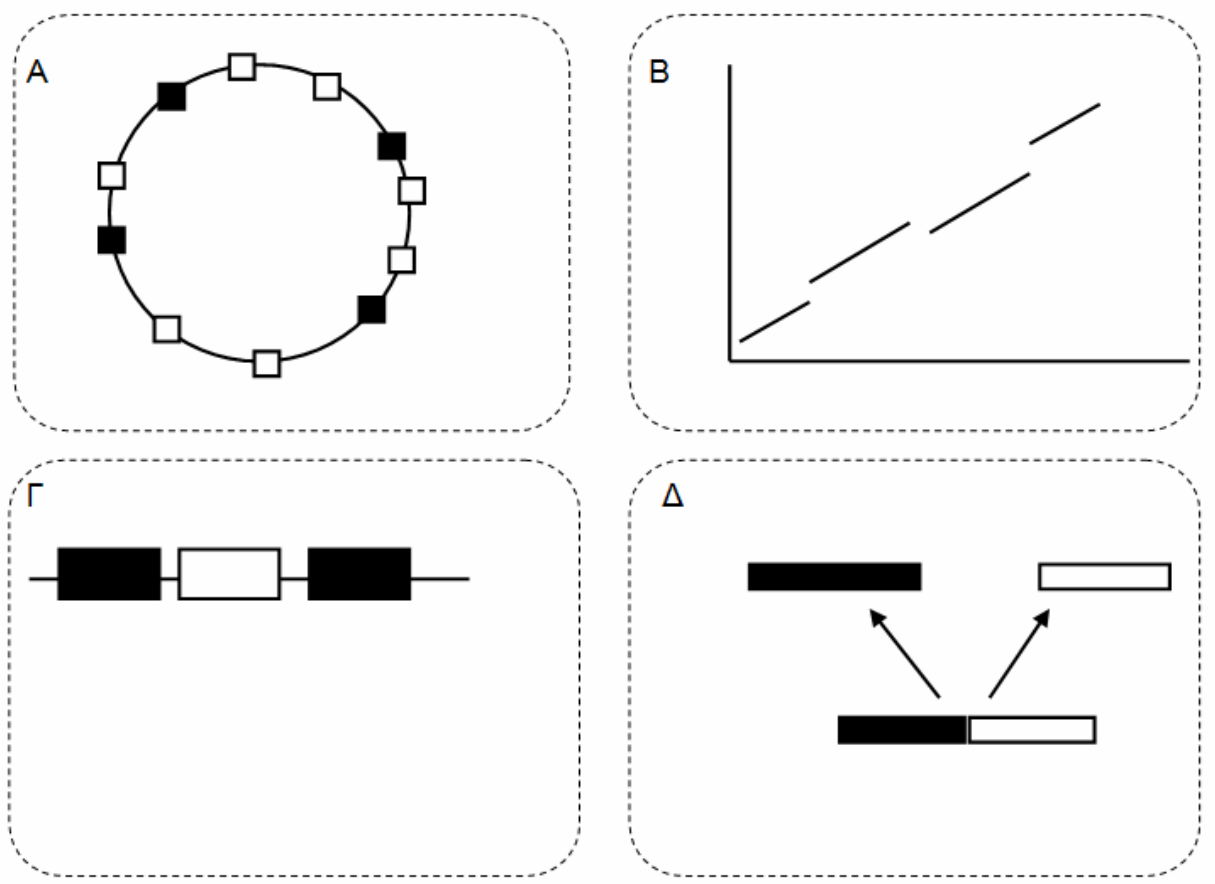


Εικόνα 11.5: *Αριστερά, η συσχέτιση του ποσοστού Βαλίνης καιΛευκίνης με το ποσοστό Φενυλαλανίνης και Ισολευκίνης. Δεξιά, η συσχέτιση του λόγου αυτών των δύο ποσοστών με τον λόγο GC/AT. Τα μαύρα σημεία αντιστοιχούν στις διαμεμβρανικές πρωτεΐνες, ενώ τα λευκά στο σύνολο του γονιδιώματος.*

11.2. Συγκριτική Γονιδιωματική

Η συγκριτική γονιδιωματική, πάει ένα βήμα παραπέρα την υπολογιστική ανάλυση των γονιδιωμάτων. Αντί να εστιάζει μόνο στα συνολικά στατιστικά μέτρα από κάθε γονιδίωμα, όπως π.χ. το ποσοστό GC ή κάποιο άλλο μέτρο, επιχειρεί να χρησιμοποιήσει τη βασική αρχή της φυλογενετικής ανάλυσης, ότι δηλαδή τα γονιδιώματα όλων των οργανισμών προέρχονται από προγονικές μορφές και έχουν διαμορφωθεί έτσι όπως είναι σήμερα μετά από αλλεπάλληλες αλλαγές που έγιναν μέσα σε εκατομμύρια χρόνια. Οι αλλαγές αυτές, αφορούν τόσο τα αντίστοιχα ορθόλογα γονίδια και τις αλληλουχίες τους, όσο και το ίδιο το γονιδίωμα, τη δομή του, και τη διάταξή των γονιδίων πάνω σε αυτό.

Βασικά, η συγκριτική γονιδιωματική κάνει χρήση των κλασικών αλγορίθμων στοίχισης και εύρεσης ομοιότητας μεταξύ γονιδίων ή/και πρωτεϊνών, αλλά συνδυάζοντας αυτή την πληροφορία με τη δομή του γονιδιώματος και τη διάταξη των γονιδίων πάνω σε αυτό, καταφέρνει να εξάγει πολύ σημαντικά συμπεράσματα, που δεν θα μπορούσαν να έχουν εξαχθεί με άλλον τρόπο (ούτε καν με πρόγνωση). Οι βασικές τεχνικές που χρησιμοποιούνται στη συγκριτική γονιδιωματική είναι τέσσερις (Tsoka & Ouzounis, 2000)



Εικόνα 11.6: *Οι τέσσερις κλασικές μέθοδοι συγκριτικής γονιδιωματικής. Η μέθοδος «αφαίρεσης» γονιδίων (Α), η μέθοδος στοίχισης ολόκληρων γονιδιωμάτων(Β), η μέθοδος σύγκρισης της σειράς των γονιδίων (Γ) και η μέθοδος εντοπισμού προϊόντων γονιδιακής σύντηξης (Δ).*

Η μέθοδος «αφαίρεσης» γονιδίων, στην οποία συγκρίνονται σε μια σειρά οργανισμούς τα κοινά γονίδια και εντοπίζονται τα μοναδικά γονίδια.

Η μέθοδος σύγκρισης της σειράς των γονιδίων, σύμφωνα με την οποία εντοπίζονται γονίδια που έχουν την τάση να βρίσκονται κοντά σε όλα τα υπό μελέτη γονιδιώματα,

Η μέθοδος στοίχισης ολόκληρων γονιδιωμάτων, σύμφωνα με την οποία στοιχίζονται ολόκληρα γονιδιώματα και εντοπίζονται οι περιοχές στις οποίες έχουν μεγάλη ομοιότητα, και τέλος

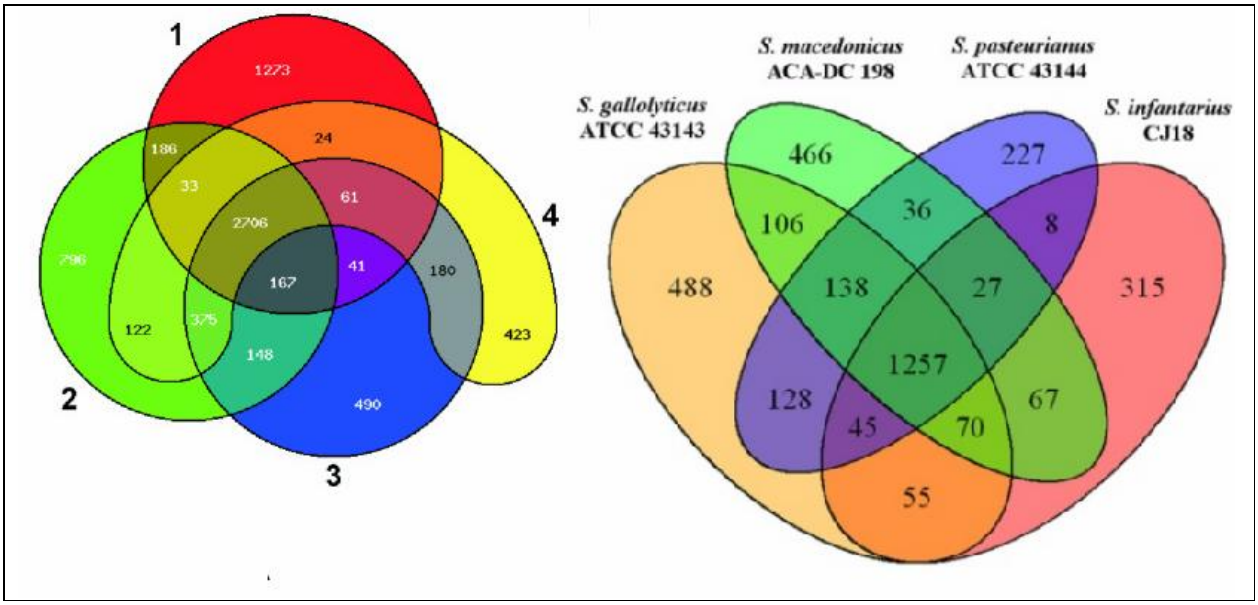
Η μέθοδος εντοπισμού προϊόντων γονιδιακής σύντηξης, στην οποία εντοπίζονται με υπολογιστικό τρόπο γονίδια τα οποία σε κάποιον άλλον οργανισμό βρίσκονται ενωμένα (σύντηξη), λειτουργούν δηλαδή σαν ανεξάρτητες πρωτεϊνικές περιοχές (domains).

Όλες οι παραπάνω μεθοδολογίες λειτουργούν με χρήση της ομοιότητας των γονιδίων και των πρωτεϊνικών προϊόντων τους και κάνουν χρήση της πληροφορίας από τη σχετική θέση των γονιδίων (ή και την ίδια την ύπαρξή τους) σε διαφορετικούς οργανισμούς. Παρόλα αυτά, οι μεθοδολογίες αυτές εντοπίζουν διαφορετικού είδους λειτουργικές συσχετίσεις μεταξύ των γονιδίων. Προσφέρουν δηλαδή διαφορετικά αποτελέσματα, γι' αυτό και στη μεγάλη τους πλειοψηφία δρουν συμπληρωματικά, όπως θα δούμε παρακάτω.

11.2.1 Η μέθοδος «αφαίρεσης» γονιδίων

Η μέθοδος αυτή, βασίζεται στην εύρεση κοινών, ομόλογων δηλαδή, γονιδίων σε μια σειρά υπό σύγκριση οργανισμών. Η βασική αρχή, είναι η γνωστή από παλιά αρχή στη φυλογενετική, ότι οι πιο συγγενικοί οργανισμοί θα έχουν και περισσότερα κοινά χαρακτηριστικά (δηλαδή, γονίδια στην περίπτωση μας). Με την ενσωμάτωση της γνώσης για τη μοριακή λειτουργία αυτών των γονιδίων, μπορούμε να εντοπίσουμε ποια γονίδια είναι χαρακτηριστικά για μια ομάδα οργανισμών και να εξάγουμε χρήσιμα συμπεράσματα για τη φυλογένεση (λειτουργούν δηλαδή ως απομορφικοί χαρακτήρες). Εξετάζοντας τα γονίδια που είναι μοναδικά σε κάποιον οργανισμό (ή σε κάποιους οργανισμούς) μπορούμε επίσης να εντοπίσουμε ειδικές λειτουργίες που επιτελεί αυτός ο οργανισμός για να επιβιώσει (π.χ. τα μεθανότροφα βακτήρια έχουν ειδικά μεταβολικά μονοπάτια για να αποικοδομούν το μεθάνιο που βρίσκεται σε περίσσεια στο περιβάλλον τους).

Οι μέθοδοι αναπαράστασης τέτοιων αναλύσεων, ξεκινούν από απλά διαγράμματα Venn και φτάνουν μέχρι περίπλοκες αναπαραστάσεις μεταβολικών δρόμων στις οποίες ο κάθε οργανισμός απεικονίζεται με κάποιο χρώμα, έτσι ώστε να φανεί ποιοι οργανισμοί έχουν κάποια συγκεκριμένα ένζυμα ή άλλα μοριακά συστήματα.



Εικόνα 11.7: *Παραδείγματα διαγραμμάτων Venn. Αριστερά: Σύγκριση στελεχών του Xanthomonas Oryzae με το EDGAR (Blom et al., 2009) Δεξιά: Σύγκριση διαφορετικών ειδών Streptococcous με το R (Papadimitriou et al., 2014)*.

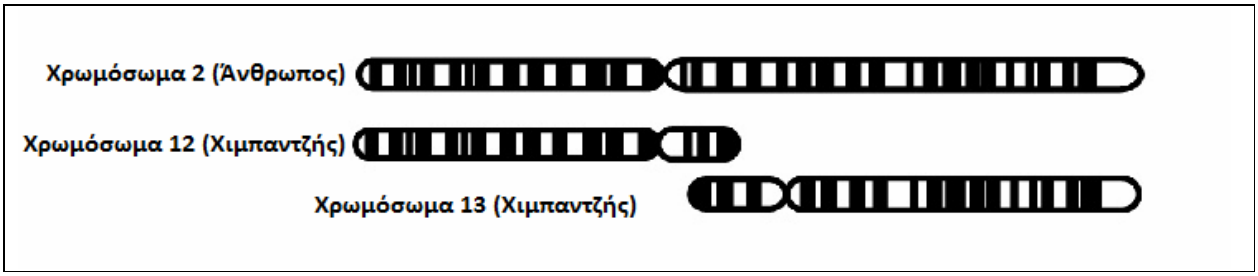
Τέτοιου είδους αναλύσεις, έχουν χρησιμοποιηθεί για να διαλευκανθεί το ερώτημα που αφορά τον τελευταίο κοινό πρόγονο όλων των σύγχρονων οργανισμών (Last Universal Common Ancestor-LUCA). Οι αναλύσεις ξεκίνησαν με τη μελέτη του οργανισμού με το μικρότερο γονιδίωμα, του βακτηρίου *Mycoplasma genitalium* το οποίο είναι υποχρεωτικό ενδοκυτταρικό παράσιτο και κωδικοποιεί μόλις 468 γονίδια που παράγουν πρωτεΐνες. Ακόμα και σε σύγκριση με κάποιο άλλο βακτήριο, π.χ. με το *Haemophilus influenzae* (1703 γονίδια) γίνεται εμφανές ότι μόνο 240 γονίδια του *M. genitalium* έχουν ορθόλογα γονίδια στον *H. influenzae*. Το ερώτημα λοιπόν ήταν αν ο LUCA ήταν ένας οργανισμός με λίγα γονίδια (όπως π.χ. το *Mycoplasma*) ή αν, αντίθετα, ήταν οργανισμός με περισσότερα γονίδια (όπως τα περισσότερα βακτήρια) και τελικά η εξέλιξη οδήγησε κάποιους οργανισμούς να χάσουν τα γονίδια αυτά και άλλους να αποκτήσουν κάποια νέα. Συγκριτικές αναλύσεις, με κάποιες παραδοχές (όπως π.χ. ότι δεν αναμένουμε σε όλους τους οργανισμούς να είναι συντηρημένα όλα τα γονίδια), έδειξε ότι μάλλον η δεύτερη εκδοχή είναι η σωστή. Για παράδειγμα, όταν στην ανάλυση συμπεριλήφθηκαν μόνο προκαρυώτες, βρέθηκε ότι ο κοινός πρόγονος όλων των οργανισμών πρέπει να είχε γονίδια μεταξύ 1006 και 1189, ενώ όταν συμπεριλήφθηκαν και οι ευκαρυώτες, ο αριθμός ανέβηκε στο 1344 με 1529, -αριθμοί που είναι πιο κοντά στο μέσο όρο των σημερινών βακτηρίων παρά στο ελάχιστο (δηλαδή στο *Mycoplasma*) (Ouzounis, Kunin, Darzentas, & Goldovsky, 2006).

Παρόμοιες αναλύσεις, έχουν μεγάλο ενδιαφέρον και στη λεγόμενη «εξωβιολογία», τον κλάδο δηλαδή που μελετάει θεωρητικά το πώς αναμένουμε να είναι οι οργανισμοί που ενδεχομένως βρεθούν σε άλλους πλανήτες, αλλά και στη συνθετική βιολογία και τη γενετική μηχανική. Για παράδειγμα, τέτοιου είδους

αναλύσεις, έκαναν δυνατό τον υπολογισμό των απαραίτητων γονιδίων που απαιτούνται για να συντηρήσουν τη ζωή σε ένα βακτήριο και εφαρμόστηκαν πρόσφατα όταν επιστήμονες συνέθεσαν εξ' ολοκλήρου ένα βακτηριακό γονιδίωμα 1Mbp και το ενσωμάτωσαν σε ένα βακτηριακό κύτταρο από το οποίο είχαν αφαιρέσει το γονιδίωμα. Το «νέο» βακτήριο, το οποίο χρησιμοποιεί αποκλειστικά το συνθετικό DNA (*Mycoplasma mycoides* JCVI-syn1.0), είχε τις αναμενόμενες φαινοτυπικές λειτουργίες και ήταν ικανό να αναπαράγεται (Gibson et al., 2010).

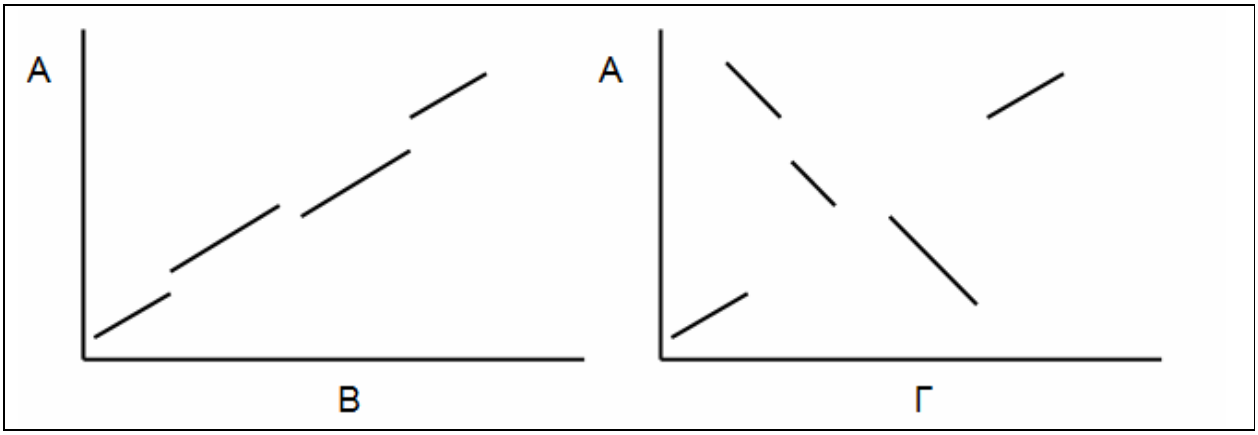
11.2.2 Η μέθοδος στοίχισης ολόκληρων γονιδιωμάτων

Η μέθοδος αυτή βασίζεται στην ίδια αρχή με τις στοιχίσεις αλληλουχιών (οι συγγενικοί οργανισμοί είναι πιο πιθανό να έχουν μεγάλες ομοιότητες στο γονιδίωμα τους). Με τη μέθοδο αυτή στοιχίζονται ολόκληρα γονιδιώματα και εντοπίζονται οι περιοχές στις οποίες έχουν μεγάλη ομοιότητα. Τέτοιες τεχνικές σε πιο πρώιμη μορφή ήταν γνωστές από παλιά, π.χ. από παρατηρήσεις ότι το ανθρώπινο DNA υβριδοποιείται με το αντίστοιχο του χιμπατζή, είχε γίνει γνωστό ότι τα γονιδιώματα του ανθρώπου και των άλλων μεγάλων πιθήκων έχουν μεγάλη ομοιότητα. Παρόμοιες ανακαλύψεις είχαν γίνει και με τη χρήση καρυότυπου, όταν για παράδειγμα έγινε γνωστό ότι το χρωμόσωμα 2 του ανθρώπου εμφανίζει μερική ομοιότητα με το χρωμόσωμα 12 και 13 του χιμπατζή, και έγινε κατανοητό ότι στο απώτατο παρελθόν είχε προκύψει από σύντηξη τελομερών.



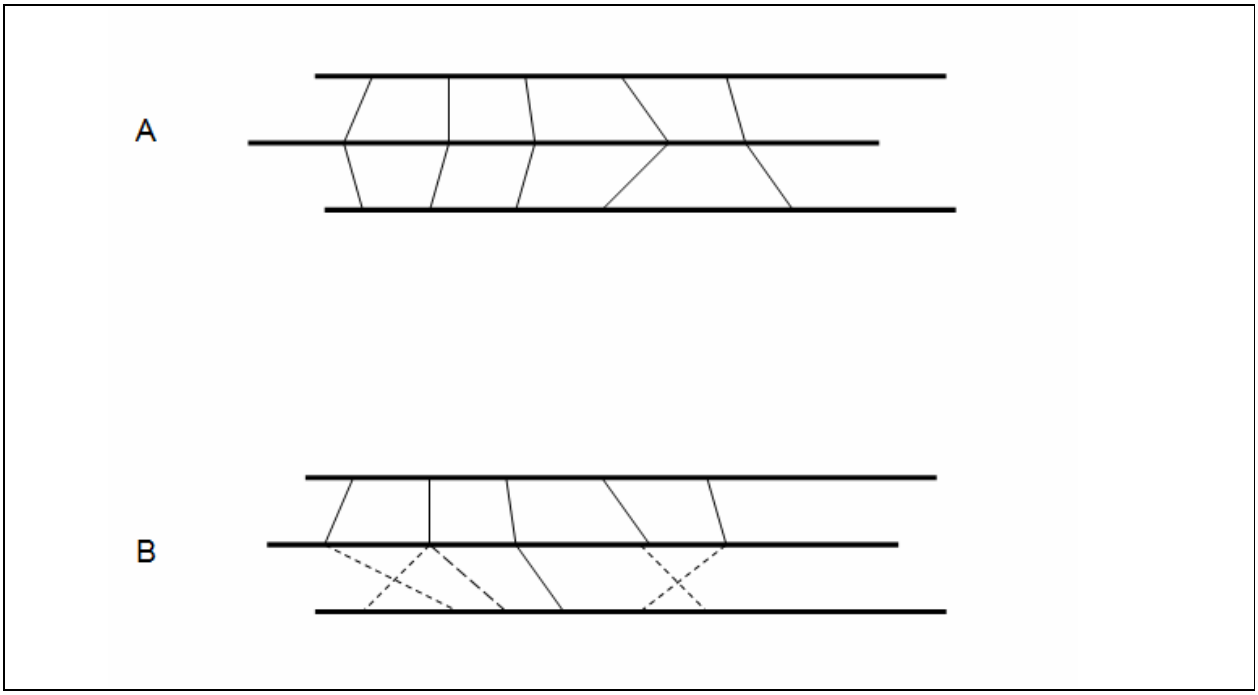
Εικόνα 11.8: *Σύγκριση των χρωμοσωμάτων του ανθρώπου και του χιμπατζή.*

Στοίχιση ολόκληρων γονιδιωμάτων, μπορεί να γίνει με διαφορετικούς τρόπους. Ο πιο απλός είναι με μια παραλλαγή του γνωστού διαγράμματος σημείων (dot-plot) η οποία επεκτείνεται σε όλο το γονιδίωμα ή με κάποια επέκταση κάποιου γνωστού αλγόριθμου στοίχισης (όπως το BLAST) η οποία να επιτρέπει χρήση μεγάλων ακολουθιών. Οι πιο σύγχρονες μεθοδολογίες, συνδυάζουν τους αλγορίθμους τοπικής ή ολικής στοίχισης (για κάθε ζευγάρι ομόλογων γονιδίων) με τη θέση των γονιδίων αυτών στο γονιδίωμα, δείχνοντας π.χ. με διαφορετικό χρωματισμό τα ζευγάρια, ενώ κάποιες από τις τεχνικές αυτές επιτρέπουν και πολλαπλή στοίχιση. Όπως γίνεται εύκολα αντιληπτό, οι τεχνικές αυτές είναι πολύ πιο εύκολο να εφαρμοστούν σε βακτηριακά ή ϊικά γονδιώματα, τόσο γιατί είναι πιο μικρά όσο και γιατί είναι ενιαία, καθώς τα πολλαπλά χρωμοσώματα των ευκαρυωτικών οργανισμών απαιτούν σύγκριση ένα με ένα.



Εικόνα 11.9: *Παραδείγματα στοίχισης γονιδιωμάτων. Στοίχιση που δείχνει συνταινικότητα (Α-Β), και στοίχιση που δείχνει αναστροφή (Α-Γ).*

Οι μεθοδολογίες ολικής στοίχισης γονιδιωμάτων είναι δυνατό να δώσουν πολλές πληροφορίες για τις αλλαγές που έχουν συμβεί στα γονιδιώματα στο πέρασμα του εξελικτικού χρόνου. Για παράδειγμα, μια στοίχιση και ένα διάγραμμα σημείων περίπου στο ύψος της διαγωνίου δείχνει την κοινή προέλευση και τη στενή σχέση των δύο οργανισμών. Επιπλέον, αλλαγές μεγάλης κλίμακας όπως αναστροφές και διπλασιασμοί είναι ιδιαίτερα εύκολο να εντοπιστούν. Τέλος, περιοχές μη ομοιότητας ανάμεσα σε 2 κατά κανόνα «όμοια» γονιδιώματα είναι δυνατό να δείξουν πρόσφατη απόκτηση γενετικού υλικού (είτε με οριζόντια μεταφορά είτε με κάποιον άλλο τρόπο ενσωμάτωσης DNA).

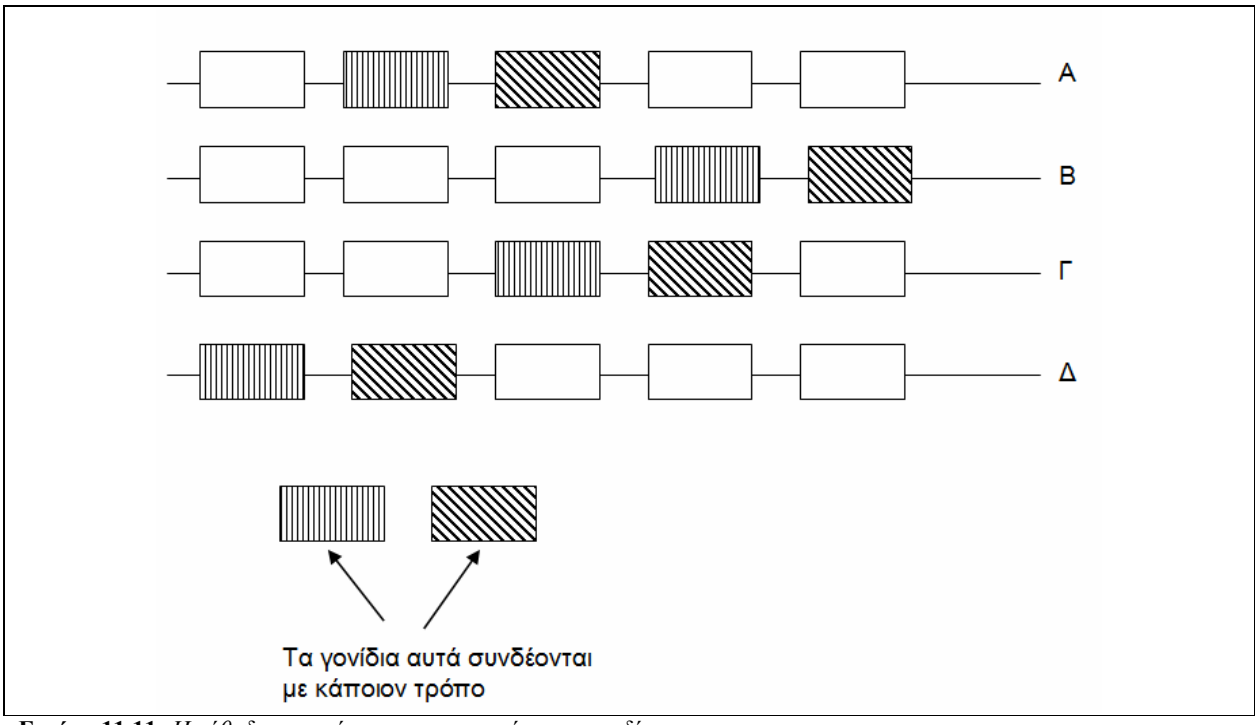


Εικόνα 11.10: *Τρόπος αναπαράστασης της πολλαπλής στοίχισης γονιδιωμάτων. Οι συνεχείς γραμμές δείχνουν ζεύγη γονιδίων που είναι σε συνταινία, ενώ οι διακεκομμένες γραμμές δείχνουν ζεύγη που έχουν υποστεί αναστροφή.*

11.2.3 Η μέθοδος σύγκρισης της σειράς των γονιδίων

Σύμφωνα με μέθοδο αυτή εντοπίζονται γονίδια που έχουν την τάση να βρίσκονται κοντά σε όλα ή στα περισσότερα τα υπό μελέτη γονιδιώματα. Η βασική αρχή της μεθόδου μοιάζει διαισθητικά με την αρχή της σύνδεσης στη γενετική, μόνο που εδώ χρησιμοποιείται σε μεγαλύτερη κλίμακα χρόνου. Η ιδέα είναι ότι γονίδια που βρίσκονται σε πολλούς οργανισμούς δίπλα-δίπλα, το κάνουν για κάποιο λόγο (π.χ. εκφράζονται μαζί ή συμμετέχουν σε κάποιο κοινό μεταβολικό μονοπάτι). Ειδικά στα βακτήρια, είναι γνωστό ότι ομάδες γονιδίων που συμμετέχουν στο ίδιο μονοπάτι, βρίσκονται οργανωμένα σε ομάδες που ονομάζονται οπερόνια, ομάδες οι οποίες εκφράζονται και ελέγχονται ταυτόχρονα.

Με τη μέθοδο αυτή είναι δυνατό να εντοπιστούν συσχετίσεις μεταξύ γονιδίων που κωδικοποιούν τελείως διαφορετικές πρωτεΐνες. Για παράδειγμα, αν υποθέσουμε ότι στο οπερόνιο της λακτόζης, ξέραμε τη λειτουργία της γαλακτοσιδάσης (lacZ) αλλά όχι αυτή της περμεάσης (lacY), με την παρατήρηση ότι σε μια σειρά από οργανισμούς τα δύο γονίδια βρίσκονται πάντα μαζί, θα μπορούσαμε να συμπεράνουμε ότι αποτελούν και τα δύο τμήμα κάποιου οπερονίου. Δεν θα ξέραμε φυσικά ακριβώς τη λειτουργία του νέου γονιδίου, αλλά συνδυάζοντας κάποιες απλές μεθόδους πρόγνωσης, όπως για παράδειγμα την πρόγνωση διαμεμβρανικών τμημάτων, θα βλέπαμε ότι πρόκειται για διαμεμβρανική πρωτεΐνη με 12 πιθανά διαμεμβρανικά τμήματα και αμέσως θα υποθέταμε ότι πρόκειται για κάποιον διαμεμβρανικό υποδοχέα που πιθανότατα εμπλέκεται στο μεταβολισμό της λακτόζης. Μια τόσο λεπτομερής πρόβλεψη για τη λειτουργία μιας πρωτεΐνης δεν θα μπορούσε με κανέναν τρόπο να γίνει δυνατή με χρήση μόνο της ακολουθίας της, αλλά βλέπουμε ότι αυτό συμβαίνει όταν χρησιμοποιήσουμε την πληροφορία από τη σειρά των γονιδίων και τη συντήρησή της στα γονδιώματα.

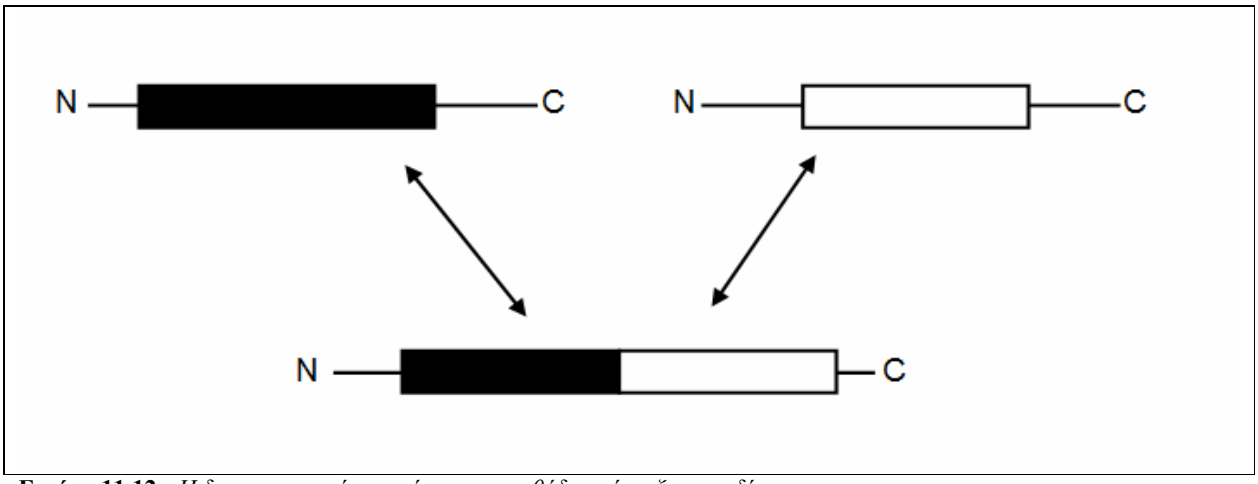


Εικόνα 11.11: *Η μέθοδος της σύγκρισης της σειράς των γονιδίων.*

11.2.4 Η μέθοδος εντοπισμού προϊόντων γονιδιακής σύντηξης

Η βασική αρχή αυτής της μεθόδου βασίζεται στη σπονδυλωτή φύση των πρωτεϊνών, δηλαδή, στην ύπαρξη ανεξάρτητων δομικών και λειτουργικών περιοχών (domains). Έτσι, με τη μέθοδο αυτή εντοπίζονται με υπολογιστικό τρόπο γονίδια ενός οργανισμού Α τα οποία σε κάποιον άλλον οργανισμό Β βρίσκονται ενωμένα, λειτουργούν δηλαδή σαν ανεξάρτητες περιοχές της ίδιας πρωτεΐνης. Η εξήγηση είναι ότι σε κάποια προγονική μορφή, είτε τα γονίδια βρίσκονταν ανεξάρτητα και συνενώθηκαν (σύντηξη γονιδίων) με το πέρασμα του χρόνου στον οργανισμό Β, είτε ότι σε κάποια προγονική μορφή τα γονίδια βρίσκονταν ενωμένα, ήταν δηλαδή πρωτεϊνικές περιοχές και κατόπιν στην πορεία της εξέλιξης αυτή η σχέση διακόπηκε στον οργανισμό Α (Enright, Iliopoulos, Kyrpides, & Ouzounis, 1999). Με τη μέθοδο αυτή, δεν μπορούμε να διακρίνουμε ποια από τις δύο εναλλακτικές όντως συνέβη, αλλά αυτό δεν αποτελεί πρόβλημα σε αυτές τις αναλύσεις, γιατί μπορούμε να εξάγουμε ούτως ή άλλως σημαντικά συμπεράσματα για πρωτεΐνες που ούτε ομοιότητα έχουν, αλλά και ούτε βρίσκονται κοντά στο γονιδίωμα.

Συνήθως τέτοιες περιπτώσεις γονιδίων αφορούν ένζυμα τα οποία εμπλέκονται στον ίδιο μεταβολικό δρόμο, πιθανότατα το προϊόν του ενός να είναι αντιδρόν στο άλλο και με αυτόν τον τρόπο διευκολύνονται οι μεταβολικές οδοί. Ένα κλασικό παράδειγμα, είναι η διυδροφολική αναγωγάση (DHFR) η οποία στους ευκαρυωτικούς οργανισμούς αποτελεί μια πρωτεΐνη με μια μοναδική πρωτεϊνική περιοχή, αλλά στα βακτήρια στο ίδιο μόριο συνυπάρχει και η λειτουργική περιοχή της θυμιδιλικής συνθετάσης (TS) η οποία συμμετέχει στο ίδιο μονοπάτι (σύνθεση νουκλεοτιδίων) και η οποία στους ευκαρυωτικούς οργανισμούς βρίσκεται σε διαφορετικό γονίδιο.

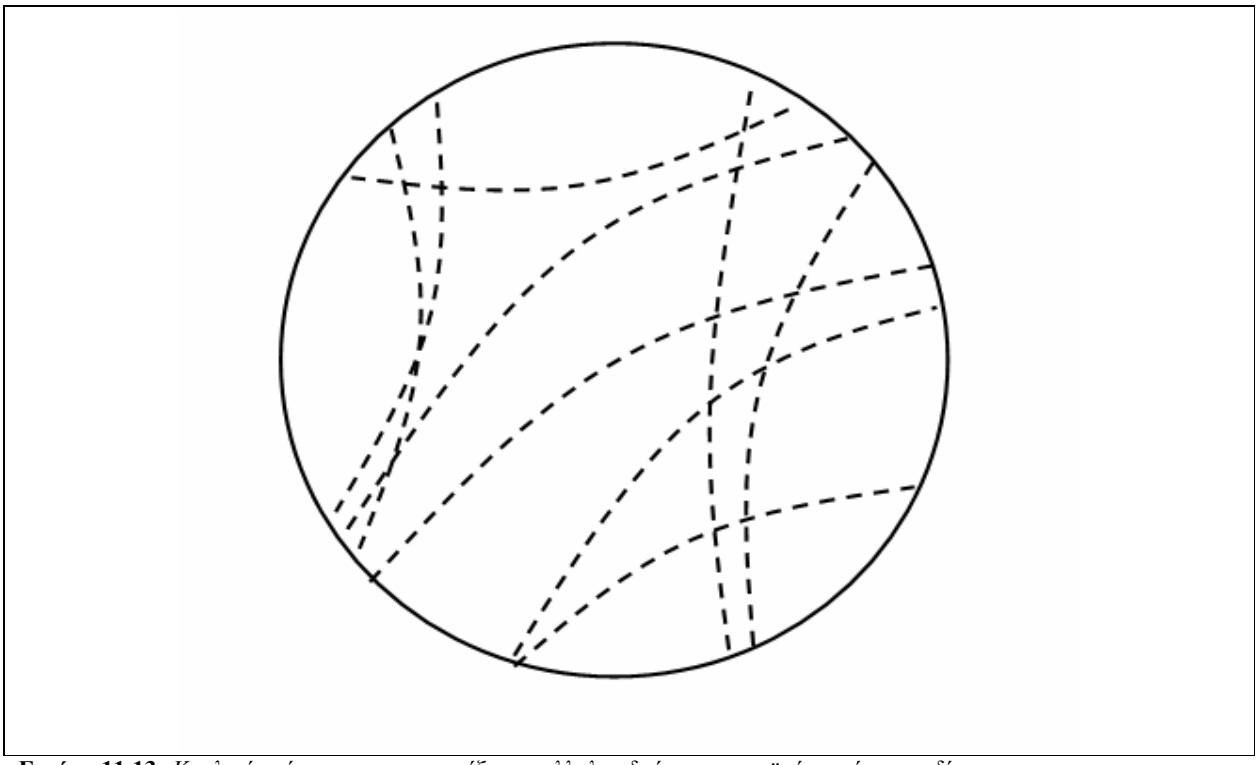


Εικόνα 11.12: *Η διαγραμματική απεικόνιση της μεθόδου σύντηξης γονιδίων.*

Η μέθοδος αυτή, είναι υπολογιστικά απαιτητική καθώς απαιτεί μία προς μία στοιχίσεις όλων των πρωτεϊνών του ενός οργανισμού, με όλες τις πρωτεΐνες του άλλου οργανισμού, ενώ απαιτείται και επιπλέον επεξεργασία για να διασφαλιστεί ότι οι δύο υποψήφιες πρωτεΐνες μοιάζουν μεν με μια άλλη πρωτεΐνη του άλλου οργανισμού αλλά σε διαφορετική περιοχή (δηλαδή, ότι δεν μοιάζουν μεταξύ τους). Από την άλλη μεριά, ένα σημαντικό πλεονέκτημα της μεθόδου σε σχέση με τις υπόλοιπες μεθόδους που αναλύθηκαν παραπάνω, είναι ότι καθώς δεν χρησιμοποιεί τη σειρά των γονιδίων, μπορεί να εφαρμοσθεί με ακριβώς τον ίδιο τρόπο σε κάθε είδους ζευγάρια ή ομάδες οργανισμών, ανεξάρτητα τόσο της εξελικτικής τους απόστασης όσο και του αριθμού χρωμοσωμάτων τους. Μπορεί με άλλα λόγια, να χρησιμοποιηθεί για τη σύγκριση του ανθρώπου με ένα βακτήριο και να δώσει χρήσιμα συμπεράσματα σε αντίθεση με τις προηγούμενες μεθόδους οι οποίες αποδίδουν καλύτερα και πρέπει να χρησιμοποιούνται κυρίως σε συγγενικούς οργανισμούς (και κατά βάση, σε βακτήρια).

Προφανώς το ποιος αλγόριθμος στοίχισης θα χρησιμοποιηθεί είναι ένα ανοιχτό ζήτημα (στην αρχική εργασία χρησιμοποιήθηκε το BLAST και έγινε εκ των υστέρων επεξεργασία με τον αλγόριθμο Smith-Waterman), όπως επίσης και το ποιες θα είναι οι παράμετροι (ποιο E-value θα χρησιμοποιηθεί σαν όριο για την εύρεση της ομοιότητας κ.ο.κ.). Κάτι ακόμα που πρέπει να γίνει σαφές, είναι ότι όσο περισσότεροι οργανισμοί χρησιμοποιούνται στην ανάλυση, τόσο περισσότερες πρωτεϊνικές αλληλεπιδράσεις θα εντοπιστούν στο δεδομένο γονιδίωμα επερώτησης. Αυτό γίνεται, γιατί έστω και σε έναν από τους οργανισμούς αυτούς να βρεθεί μια πρωτεΐνη με τις δύο περιοχές ενωμένες, τότε σε όλους τους υπόλοιπους θα αναγνωριστεί αυτή η «αλληλεπίδραση».

Προσοχή βέβαια χρειάζεται στη χρήση της έννοιας αυτής της «αλληλεπίδρασης» (interaction το ονομάζουν οι συγγραφείς), καθώς μπορεί να γίνει σύγχυση με τη φυσική αλληλεπίδραση των δύο πρωτεϊνών (τη φυσική επαφή). Παρόλο που κάτι τέτοιο είναι φυσικά πολύ πιθανό να συμβαίνει, η μέθοδος δεν προβλέπει απευθείας αυτό, αλλά μόνο μια λειτουργική αλληλεπίδραση, όμοια με αυτές που προβλέπει η προηγούμενη μέθοδος της «σειράς των γονιδίων». Βλέπουμε επομένως ότι οι μέθοδοι αυτές, λειτουργούν περισσότερο συμπληρωματικά παρά ανταγωνιστικά και αυτό είναι κάτι που πρέπει να το έχουμε πάντα στο μυαλό μας. Για παράδειγμα, η μέθοδος της σειράς των γονιδίων εντοπίζει λειτουργικά συνδεδεμένες πρωτεΐνες των οποίων τα γονίδια βρίσκονται πάντα μαζί, ενώ η μέθοδος σύντηξης γονιδίων εντοπίζει παρόμοιες συνδέσεις μεταξύ γονιδίων που βρίσκονται σε διαφορετικά μέρη στο γονιδίωμα. Επιπλέον δε, μπορεί με τη δημιουργία ενός κυκλικού χάρτη των αλληλεπιδράσεων να βρεθούν θερμές περιοχές (hot-spots), περιοχές δηλαδή με μεγάλη πυκνότητα τέτοιων αλληλεπιδράσεων και αυτές οι περιοχές να συσχετιστούν με πιθανά σημεία γονιδιωματικών ανακατατάξεων τα οποία θα εντοπιστούν με τη μέθοδο στοίχισης γονιδιωμάτων.



Εικόνα 11.13: *Κυκλικός χάρτης που απεικονίζει τις αλληλεπιδράσεις πρωτεϊνών σε ένα γονιδίωμα.*

11.3. Λογισμικό

Με τη ραγδαία ανάπτυξη της αλληλούχισης και την πρόοδο της γονιδιωματικής συνολικά, έχουν υλοποιηθεί τα τελευταία χρόνια εκατοντάδες εργαλεία συγκριτικής γονιδιωματικής που υλοποιούν κάποιους από τους αλγόριθμους και τις μεθόδους που αναλύσαμε παραπάνω, με έναν τρόπο εύκολο και βολικό για τον τελικό χρήστη (Edwards & Holt, 2013). Τα πιο πετυχημένα από αυτά τα εργαλεία συνδυάζουν την απλότητα με την ευελιξία καθώς προσφέρουν ένα ολοκληρωμένο περιβάλλον για να διευκολύνει πολλαπλές αναλύσεις, ενώ συνήθως προσφέρονται σαν διαδικτυακές εφαρμογές. Παρακάτω, αναλύουμε τα πιο γνωστά από αυτά τα εργαλεία.

Το **ACT** (https://www.sanger.ac.uk/resources/software/act/) είναι ένα εργαλείο βασισμένο στη Java το οποίο επιτρέπει την οπτικοποίηση γονιδιωμάτων και τη σύγκρισή τους. Για τη στοίχιση των αλληλουχιών χρησιμοποιεί το BLAST. Κατόπιν τα δύο γονιδιώματα και το αποτέλεσμα από την αναζήτηση του BLAST εισάγονται στο ACT για οπτικοποίηση της σύγκρισης. Επιπλέον, το εργαλείο μπορεί να οπτικοποιήσει ταυτόχρονα περισσότερες από μία συγκρίσεις γονδιωμάτων. Οι ομόλογες περιοχές οι οποίες βρίσκονται στην ίδια κατεύθυνση στο γονιδίωμα χρωματίζονται με κόκκινο ενώ αυτές που βρίσκονται σε αντίθετες κατευθύνσεις, με μπλε. Η ένταση του χρωματισμού αντικατοπτρίζει το επίπεδο ομοιότητας. Τα πλεονεκτήματα του ACT περιλαμβάνουν τη δυνατότητα να απεικονίζει τη στοίχιση σε διαφορετικές μεγεθύνσεις (zoom in – zoom out) έτσι ώστε να μπορεί να απεικονίσει είτε τη στοίχιση ολόκληρου του γονιδιώματος, είτε να εστιάσει σε συγκεκριμένα γονίδια ενδιαφέροντος, αλλά και τη δυνατότητα που προσφέρει στο χρήστη να προσθέσει δικό του σχολιασμό για τα γονιδιώματα που αναλύονται (Carver et al., 2005).

Το **MAUVE** (http://darlinglab.org/mauve/mauve.html) είναι επίσης ένα εργαλείο βασισμένο στη Java κατάλληλο για συγκρίσεις γονιδιωμάτων. Διαθέτει ενσωματωμένο σύστημα απεικόνισης αλλά και τη δυνατότητα να εξάγει την πληροφορία από τη σύγκριση των γονιδιωμάτων σε διάφορες μορφές. Το MAUVE μπορεί να εργαστεί με δεδομένα αλληλούχισης νέας γενιάς, και έτσι παρέχει τη δυνατότητα να τοποθετήσει και να διατάξει μια σειρά από contigs απέναντι σε ένα ολόκληρο γονιδίωμα. Το εργαλείο δέχεται σαν είσοδο τις τελικές μορφές των γονιδιωμάτων και δημιουργεί μια στοίχιση αυτών. Αναγνωρίζει περιοχές με μεγάλη ομολογία και αναθέτει ένα ξεχωριστό χρώμα σε κάθε μία. Κατόπιν, κάθε γονιδίωμα απεικονίζεται σαν μια

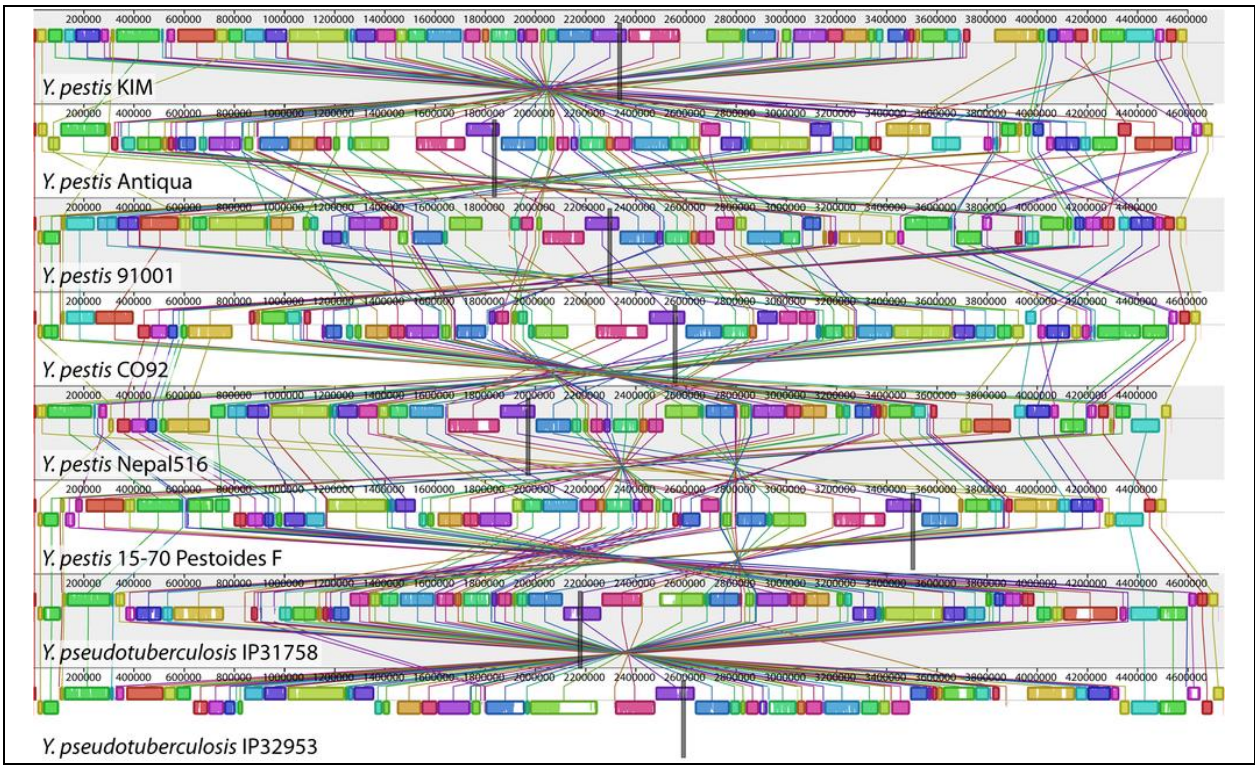
ακολουθία τέτοιων χρωματιστών περιοχών. Με τον τρόπο αυτό, γίνεται εύκολος ο εντοπισμός περιοχών με μοναδικά γονίδια. Επίσης, το MAUVE μπορεί χρησιμοποιηθεί (καθώς δουλεύει όπως αναφέραμε και με δεδομένα αλληλούχισης νέας γενιάς) και για τον εντοπισμό νουκλεοτιδικών πολυμορφισμών (SNPs) οι οποίοι μπορούν να χρησιμοποιηθούν παρακάτω για φυλογενετικές, εξελικτικές ή ιατρικές αναλύσεις (Darling, Mau, & Perna, 2010).

Το **EDGAR** (http://edgar.cebitec.uni-bielefeld.de) είναι ένα ακόμα σύγχρονο διαδικτυακό εργαλείο συγκριτικής γονιδωματικής το οποίο μπορεί να δεχτεί και δεδομένα αλληλούχισης. Το EDGAR είναι σχεδιασμένο έτσι ώστε να διευκολύνει το χρήστη και να απλοποιεί τις διαδικασίες. Ενσωματώνει τις βάσεις δεδομένων του NCBI και έχει στη βάση δεδομένων του όλα τα αποτελέσματα γνωστών γονιδιωμάτων προ-υπολογισμένα, ενώ έχει και τη δυνατότητα να απεικονίσει εξελικτικές και φυλογενετικές σχέσεις οι οποίες πολλές φορές διαλευκάνουν υποθέσεις σύγκρισης γονιδιωμάτων. Επίσης, υποστηρίζει μια σειρά από τρόπους απεικόνισης των αποτελεσμάτων όπως τα διαγράμματα στοίχισης γονιδιωμάτων (synteny plots) και διαγράμματα Venn για τα κοινά γονίδια (Blom, et al., 2009).

Το **CGAT** (http://mbgd.genome.ad.jp/CGAT/) είναι ένα ακόμα παρόμοιο εργαλείο που δημιουργήθηκε για να διευκολύνει τις συγκρίσεις συγγενικών βακτηριακών γονιδιωμάτων. Το CGAT λειτουργεί με αρχιτεκτονική client-server, στην οποία ο client AlignmentViewer (μια εφαρμογή Java) συνεργάζεται με τον DataServer (προγράμματα Perl). Το εργαλείο οπτικοποιεί στοιχίσεις γονιδιωμάτων τόσο στη μορφή των διαγραμμάτων σημείων όσο και στη μορφή των στοιχίσεων. Ο χρήστης μπορεί να προσθέσει πληροφορία στη σύγκριση, όπως για παράδειγμα την ύπαρξη επαναληπτικών αλληλουχιών και αλλαγές στη συχνότητα κωδικονίων έτσι ώστε να διευκολυνθεί στην εξαγωγή συμπερασμάτων. Εκτός από την οπτικοποίηση, ένα πλεονέκτημα του CGAT είναι η ευελιξία του καθώς επιτρέπει τη χρήση πολλών διαφορετικών αλγόριθμων στοίχισης γονιδιωμάτων (Uchiyama, Higuchi, & Kobayashi, 2006).

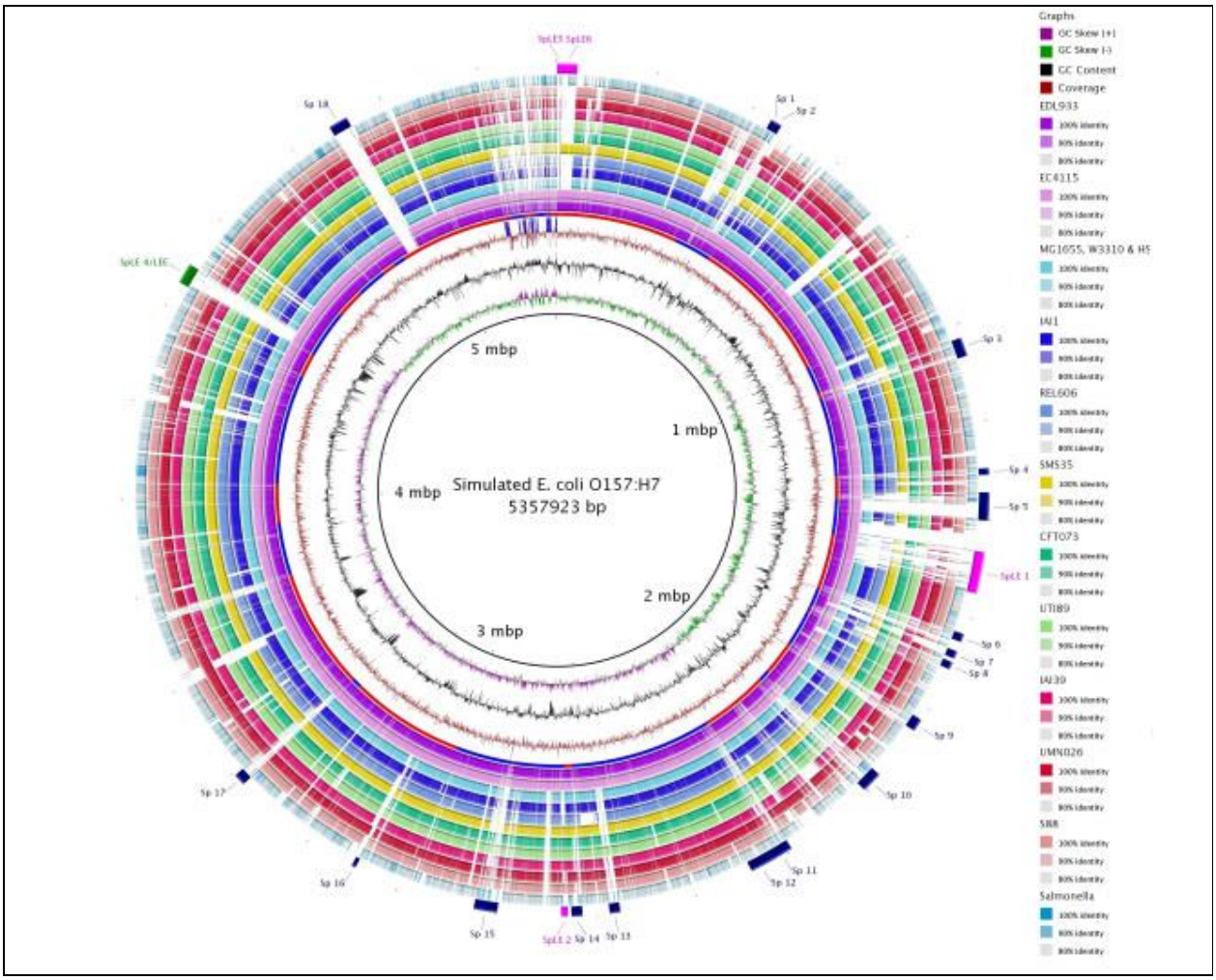
Το **BRIG** (BLAST Ring Image Generator, http://sourceforge.net/projects/brig/) είναι ένα άλλο εργαλείο βασισμένο στη Java, το οποίο οπτικοποιεί τη σύγκριση ενός γονιδιώματος αναφοράς με μία ή περισσότερες άλλες αλληλουχίες. Χρησιμοποιεί έναν ιδιαίτερο τρόπο οπτικοποίησης, σύμφωνα με τον οποίο τα γονιδιώματα αναπαρίστανται ως σειρές από επάλληλους κύκλους (δαχτυλίδια), με ειδικό χρωματισμό, για να δηλώνει την παρουσία μιας περιοχής ή ενός γονιδίου στο γονιδίωμα αναφοράς. Το BRIG είναι αρκετά ευέλικτο και μπορεί να χρησιμοποιηθεί για να απαντήσει πλήθος ερωτημάτων, ανάλογα με την επιλογή των γονιδιωμάτων υπό σύγκριση. Αυτό που πρέπει να τονιστεί είναι το γεγονός ότι η αναπαράσταση είναι εξαρτώμενη από το γονιδίωμα αναφοράς. Με άλλα λόγια, ενώ το εργαλείο απεικονίζει ποιες περιοχές είναι παρούσες ή απούσες από τα γονιδιώματα σύγκρισης, δεν μπορεί να δείξει περιοχές των γονιδιωμάτων αυτών που λείπουν από το γονιδίωμα αναφοράς. Γι' αυτό το λόγο η επιλογή του γονιδιώματος αναφοράς είναι ιδιαίτερα σημαντική (Alikhan, Petty, Ben Zakour, & Beatson, 2011).

Το **VISTA** (http://genome.lbl.gov/vista/index.shtml) ήταν ένα από τα πρώτα εργαλεία οπτικοποίησης στοιχίσεων γονιδιωμάτων και είχε παρουσιαστεί το 2000. Σήμερα, έχει εξελιχθεί σε μια ολοκληρωμένη σουίτα προγραμμάτων τα οποία καλύπτουν κάθε ανάγκη συγκριτικής ανάλυσης γονιδιωμάτων. Διαθέτει ειδικά εργαλεία για διάφορες συγκριτικές αναλύσεις γονιδιωμάτων, διασύνδεση με τις βάσεις δεδομένων γονιδιωμάτων, ενώ διαθέτει και αποθηκευμένα προ-υπολογισμένα αποτελέσματα για τα γνωστά γονιδιώματα (ακόμα και των σπονδυλοτών). Διαθέτει ειδικό σύστημα οπτικοποίησης (VISTA Browser) το οποίο επιτρέπει στο χρήστη να υποβάλει και το δικό του γονιδίωμα για ανάλυση στους διάφορους εξυπηρετητές (VISTA servers, rVista, mVISTA, phyloVISTA, gVISTA κ.ο.κ.) στους οποίους ο χρήστης μπορεί να επιτελέσει στοιχίσεις με διαφορετικούς αλγόριθμους, οπτικοποίηση με διαφορετικούς τρόπους, αλλά και ενσωμάτωση διαφορετικών ειδών πληροφορίας όπως φυλογενετικές σχέσεις, ρυθμιστικές περιοχές κ.ο.κ. (Frazer, Pachter, Poliakov, Rubin, & Dubchak, 2004). Μια επιπλέον δυνατότητα του VISTA είναι ότι διαθέτει και μια ανεξάρτητη (standalone) εφαρμογή με σχεδόν τις ίδιες δυνατότητες, το GenomeVISTA, το οποίο μπορεί να εγκατασταθεί ελεύθερα στον υπολογιστή του χρήστη και να εκτελέσει εκεί τις ίδιες λειτουργίες με τη διαδικτυακή εκδοχή, προσφέροντας μεγαλύτερη ασφάλεια των δεδομένων και ίσως και ταχύτητα (Poliakov, Foong, Brudno, & Dubchak, 2014).

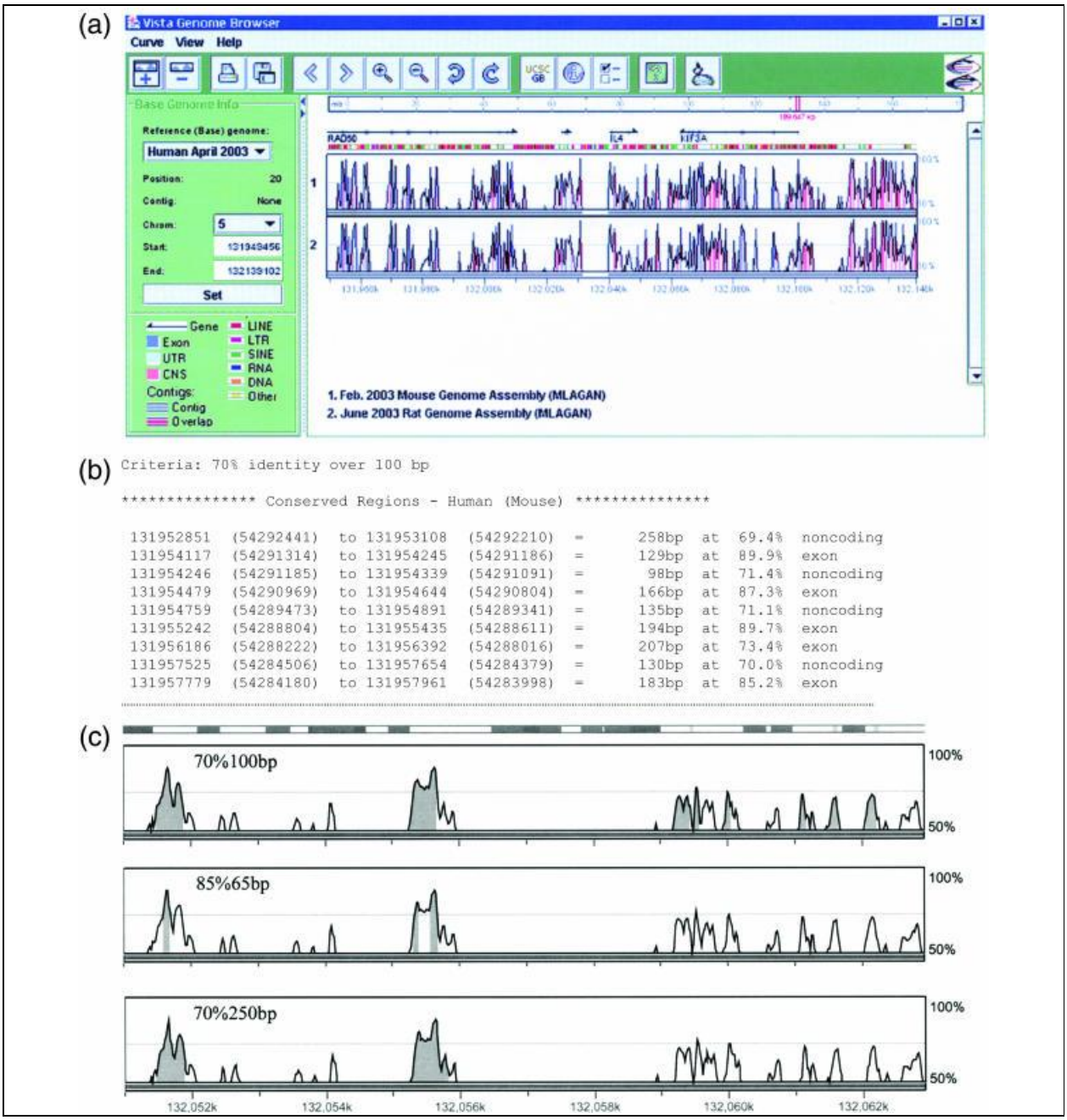


Εικόνα 11.14: *Ολική στοίχιση 8 γονιδιωμάτων της Yersinia με το MAUVE (Darling, Miklos, & Ragan, 2008).*

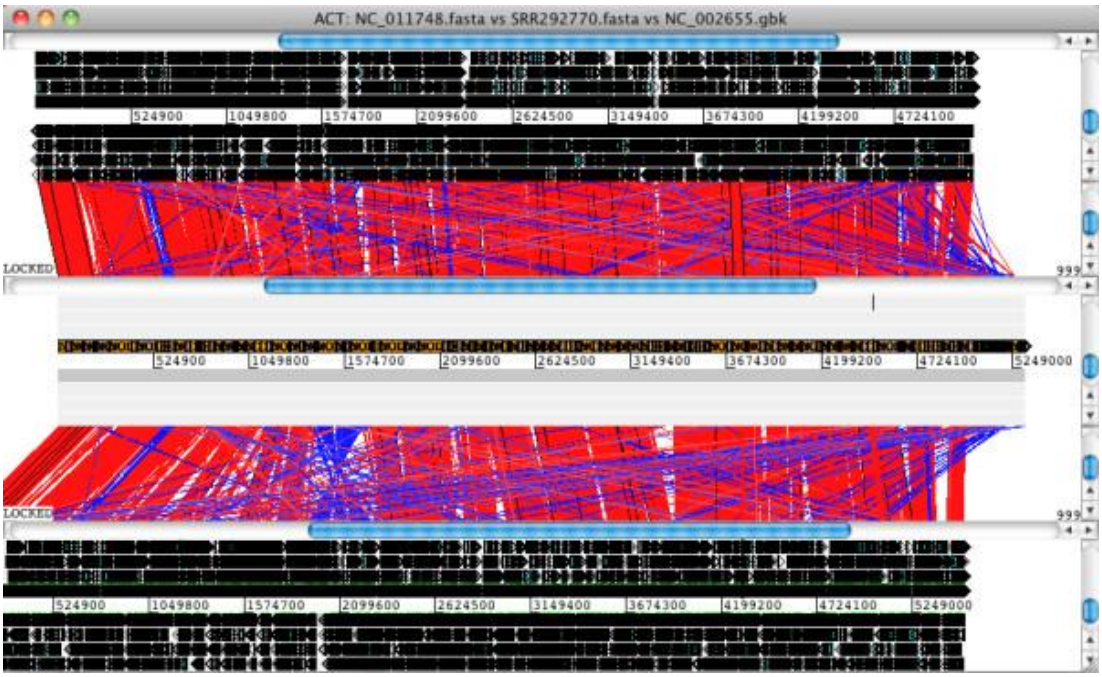
Όπως είδαμε, τα περισσότερα από τα προαναφερθέντα πακέτα λογισμικού παρέχουν τη δυνατότητα χρήσης διαφορετικών αλγορίθμων στοίχισης γονιδιωμάτων. Κάποια, διαθέτουν και δικούς τους αλγόριθμους στοίχισης αλλά τα περισσότερα δίνουν τα δυνατότητα ενσωμάτωσης και άλλων εξειδικευμένων αλγορίθμων. Οι πιο γνωστοί από αυτούς είναι το **MUMMER** (http://mummer.sourceforge.net/), το **MEGA-BLAST** (http://www.ncbi.nlm.nih.gov/BLAST/), το **LAGAN** (http://bioperl.org/wiki/LAGAN) και το **MGA** (http://bibiserv.techfak.uni-bielefeld.de/mga/). Όσον αφορά τους αλγόριθμους εύρεσης σύντηξης γονιδίων, η οποία σαν μέθοδος είναι και η πιο «απόμακρη» (ή ξεχωριστή) από τις υπόλοιπες, υπάρχουν επίσης διαθέσιμες μια σειρά από επιλογές, οι οποίες έχουν πολλαπλασιαστεί ιδιαίτερα τα τελευταία χρόνια με την έλευση της αλληλούχισης νέας γενιάς με τη χρήση τέτοιων τεχνικών σε διάφορες άλλες εφαρμογές, ακόμα και ιατρικές (Carrara et al., 2013). Ενδεικτικά, αναφέρουμε τον αρχικό αλγόριθμο των Ouzounis και συνεργατών, το **GeneRAGE** (Enright & Ouzounis, 2000), αλλά και μερικές νεότερες εφαρμογές όπως το **FusionMap** (http://www.omicsoft.com/fusionmap) (Ge et al., 2011) και το **MosaicFinder** (http://sourceforge.net/projects/mosaicfinder) (Jachiet, Pogorelcnik, Berry, Lopez, & Bapteste, 2013).



Εικόνα 11.15: *Κυκλική αναπαράσταση της στοίχισης του γονιδιώματος της* E. coli *O157:H7 str. Sakai και η σύγκριση με 27 άλλα προκαρυωτικά γονιδιώματα με του BRIG.*



Εικόνα 11.16: ***(a)*** *Διάγραμμα μιας χρωμοσωμικής περιοχής του ανθρώπινου γονιδιώματος που περιέχει το γονίδιο KIF3A (chr5:131949456–132139102) με το VISTA Η σύγκριση δείχνει συντηρημένες περιοχές μεταξύ ανθρώπου και ποντικού και μεταξύ ανθρώπου και αρουραίου.* ***(b)*** *Το VISTA παράγει μια λίστα με τα συντηρημένα στοιχεία μεταξύ ανθρώπου και ποντικού στην περιοχή του KIF3A* ***(c)*** *Η γονιδιωματική περιοχή πριν από το γονίδιο KIF3A, στην οποία εμφανίζονται συντηρημένες μη κωδικές περιοχές (Frazer, et al., 2004).*



Εικόνα 11.17: *Στοίχιση γονιδιωμάτων με το ACT, Το γονιδίωμα της E. coli O104:H4 είναι στη μεσαία σειρά, αυτό της E. coli Ec55989 φαίνεται πάνω, ενώ το γονιδίωμα της E. coli EDL933 είναι κάτω (Edwards & Holt, 2013).*

Βιβλιογραφία

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics, 12*, 402. doi: 10.1186/1471-2164-12-402

Arai, M., Ikeda, M., & Shimizu, T. (2003). Comprehensive analysis of transmembrane topologies in prokaryotic genomes. [Research Support, Non-U.S. Gov't]. *Gene, 304*, 77-86.

Blom, J., Albaum, S. P., Doppmeier, D., Pühler, A., Vorhölter, F.-J., Zakrzewski, M., & Goesmann, A. (2009). EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics, 10*(1), 154.

Carrara, M., Beccuti, M., Lazzarato, F., Cavallo, F., Cordero, F., Donatelli, S., & Calogero, R. A. (2013). State-of-the-art fusion-finder algorithms sensitivity and specificity. *Biomed Res Int, 2013*, 340620. doi: 10.1155/2013/340620

Carver, T. J., Rutherford, K. M., Berriman, M., Rajandream, M. A., Barrell, B. G., & Parkhill, J. (2005). ACT: the Artemis Comparison Tool. *Bioinformatics, 21*(16), 3422-3423. doi: 10.1093/bioinformatics/bti553

Darling, A. E., Mau, B., & Perna, N. T. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One, 5*(6), e11147. doi: 10.1371/journal.pone.0011147

Darling, A. E., Miklos, I., & Ragan, M. A. (2008). Dynamics of genome rearrangement in bacterial populations. *PLoS Genet, 4*(7), e1000128. doi: 10.1371/journal.pgen.1000128

Edwards, D. J., & Holt, K. E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microb Inform Exp, 3*(1), 2. doi: 10.1186/2042-5783-3-2

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., & Ouzounis, C. A. (1999). Protein interaction maps for complete genomes based on gene fusion events. *Nature, 402*(6757), 86-90. doi: 10.1038/47056

Enright, A. J., & Ouzounis, C. A. (2000). GeneRAGE: a robust algorithm for sequence clustering and domain detection. *Bioinformatics, 16*(5), 451-457.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., & Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res, 32*(Web Server issue), W273-279. doi: 10.1093/nar/gkh458

Ge, H., Liu, K., Juan, T., Fang, F., Newman, M., & Hoeck, W. (2011). FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics, 27*(14), 1922-1928. doi: 10.1093/bioinformatics/btr310

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R. Y., Algire, M. A., . . . Venter, J. C. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. *Science, 329*(5987), 52-56. doi: 10.1126/science.1190719

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., . . . Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature, 431*(7004), 99-104. doi: 10.1038/nature02800

Jachiet, P. A., Pogorelcnik, R., Berry, A., Lopez, P., & Bapteste, E. (2013). MosaicFinder: identification of fused gene families in sequence similarity networks. *Bioinformatics, 29*(7), 837-844. doi: 10.1093/bioinformatics/btt049

Kreil, D. P., & Ouzounis, C. A. (2001). Identification of thermophilic species by the amino acid compositions deduced from their genomes. *Nucleic Acids Res, 29*(7), 1608-1615.

Li, W. (2011). On parameters of the human genome. [Review]. *J Theor Biol, 288*, 92-104. doi: 10.1016/j.jtbi.2011.07.021

Ouzounis, C. A., Kunin, V., Darzentas, N., & Goldovsky, L. (2006). A minimal estimate for the gene content of the last universal common ancestor--exobiology from a terrestrial perspective. *Res Microbiol, 157*(1), 57-68. doi: 10.1016/j.resmic.2005.06.015

Papadimitriou, K., Anastasiou, R., Mavrogonatou, E., Blom, J., Papandreou, N. C., Hamodrakas, S. J., . . . Pot, B. (2014). Comparative genomics of the dairy isolate Streptococcus macedonicus ACA-DC 198 against related members of the Streptococcus bovis/Streptococcus equinus complex. *BMC Genomics, 15*(1), 272.

Picardi, E., & Pesole, G. (2010). Computational methods for ab initio and comparative gene finding. *Methods Mol Biol, 609*, 269-284. doi: 10.1007/978-1-60327-241-4_16

Poliakov, A., Foong, J., Brudno, M., & Dubchak, I. (2014). GenomeVISTA--an integrated software package for whole-genome alignment and visualization. *Bioinformatics, 30*(18), 2654-2655. doi: 10.1093/bioinformatics/btu355

Quax, T. E., Claassens, N. J., Soll, D., & van der Oost, J. (2015). Codon Bias as a Means to Fine-Tune Gene Expression. *Mol Cell, 59*(2), 149-161. doi: 10.1016/j.molcel.2015.05.035

Rigoutsos, I. (2010). Short RNAs: how big is this iceberg? *Curr Biol, 20*(3), R110-113. doi: 10.1016/j.cub.2009.12.036

Shimizu, T., Mitsuke, H., Noto, K., & Arai, M. (2004). Internal gene duplication in the evolution of prokaryotic transmembrane proteins. [Comparative Study

Research Support, Non-U.S. Gov't]. *J Mol Biol, 339*(1), 1-15. doi: 10.1016/j.jmb.2004.03.048

Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D., & Krogh, A. (2001). On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet, 17*(8), 425-428. doi: S0168-9525(01)02372-1 [pii]

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nat Rev Genet, 16*(8), 472-482. doi: 10.1038/nrg3962

Tsoka, S., & Ouzounis, C. A. (2000). Recent developments and future directions in computational genomics. *FEBS Lett, 480*(1), 42-48.

Uchiyama, I., Higuchi, T., & Kobayashi, I. (2006). CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes. *BMC Bioinformatics, 7*, 472. doi: 10.1186/1471-2105-7-472

Vlachos, I. S., & Hatzigeorgiou, A. G. (2013). Online resources for miRNA analysis. *Clin Biochem, 46*(10-11), 879-900. doi: 10.1016/j.clinbiochem.2013.03.006

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res, 18*(5), 821-829. doi: 10.1101/gr.074492.107