ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΑΣ
ΤΜΗΜΑ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ
ΙΕΡΑ ΟΔΟΣ 75, 118 55 ΑΘΗΝΑ

Μεταπτυχιακή Διατριβή για το Μεταπτυχιακό Δίπλωμα Ειδίκευσης
(Μ.Δ.Ε) "Εφαρμογές της Βιοτεχνολογίας στη Γεωπονία"

Μέθοδοι Ανάλυσης Συγκριτικής Γονιδιωματικής

ΚΩΝΣΤΑΝΤΙΝΟΣ ΜΠΟΥΓΙΟΥΚΟΣ

Αθήνα, 2014

Table of Contents

List of Figures

List of Tables

Appendices

List of Abbreviations and Symbols

JGI	Joint Genome Institute

1 ABSTRACT

In the context of this study we deal with 3 concepts in studying analysis methods of comparative genomics. Firstly, we review a set of online resources that provide data and valuable precomputed results of comparative genomics studies. We critically review and evaluate a wide range of analysis tools and methods for comparative genomics and elaborate with regards to the usage of these methods in the routine work of researchers of any level of expertise. The set of methods that we present in this study create a sufficient corpus of knowledge and can provide a researcher with the necessary familiarity they need in order to perform everyday advanced comparative genomic analysis. We focus more intensively into methods that involve any kind of phylogenetics analysis in comparative genomics such as phylogenomics and ortholog/paralog detection, as we reckon that evolutionary biology has been gravely transformed with the advent of genomics and vice versa. Finally we present the outcome of two experimental methods that we have developed independently, are heavily relied on comparative genomics analysis and have been used in the identification of microRNA targets in mammalian genomes and on the novel molecular markers detection in plants that lack a complete reference genomic sequence.

ΠΕΡΙΛΗΨΗ

Στο πλαίσιο αυτής της μελέτης θα ασχοληθούμε με 3 έννοιες της μελέτης των μεθόδων ανάλυσης της συγκριτικής γονιδιωματικής. Πρώτον, εξετάζουμε ένα σύνολο online πηγών που παρέχουν στοιχεία και πολύτιμα προϋπολογισμένα αποτελέσματα από μελέτες της συγκριτικής γονιδιωματικής. Δευτερον, έχουμε εξετάσει με κριτικό πνεύμα και αξιολογήσει ένα ευρύ φάσμα εργαλείων ανάλυσης και μεθόδων για την συγκριτική γονιδιωματική και τα έχουμε επεξεργαστεί σε σχέση με τη χρήση των μεθόδων αυτών στην καθημερινή εργασία ερευνητών με οποιοδήποτε επίπεδο εξειδίκευσης. Το σύνολο των μεθόδων που παρουσιάζονται σε αυτή τη μελέτη μπορεί να δημιουργήσει ένα επαρκές σώμα γνώσεων και μπορεί να παρέχει στους ερευνητής την αναγκαία γνώση που χρειάζονται για να πραγματοποιούν καθημερινές προηγμένες αναλυσεις συγκριτικής γονιδιωματικής. Έχουμε επικεντρωθεί περισσότερο εντατικά σε μεθόδους που περιλαμβάνουν κάθε είδους φυλογενετική ανάλυση στη συγκριτική γονιδιωματική όπως phylogenomics και ανίχνευση ορθολόγων/παραλόγψν ακολουθειών, καθώς εκτιμούμε ότι η εξελικτική βιολογία έχει μετασχηματιστεί σε πολύ μεγάλο βαθμό με την έλευση της γονιδιωματικής και αντίστροφα. Τέλος, παρουσιάζουμε τα αποτελέσματα των δύο πειραματικών μεθόδων που έχουμε αναπτύξει ανεξάρτητα, και οι οποίες στηρίχθηκαν σε μεγάλο βαθμό στη χρήση μεθόδων συγκριτικής γονιδιωματικής ανάλυσης και έχουν χρησιμοποιηθεί για τον εντοπισμό των στόχων των microRNA σε γονιδιώματα θηλαστικών και στον εντοπισμό νέων μοριακών δεικτών σε φυτά που δεν έχουν πλήρη γονιδιωματική αλληλουχία αναφοράς.

2 Εισαγωγή

Μία από τις πιο ιστορικές και διαδεδομένες μεθόδους στην βιολογία είναι η συγκριτική μέθοδος [Harvey & Pagel 1991]. Η συγκριτική μέθοδος αποτελεί παράδοση για τη βιολογία από τις εποχές ακόμα και πριν τον Δαρβίνου. Ελλείψει άλλων δεδομένων, η συγκριτική ανατομία αποτελούσε ένα από τα πιο ισχυρά εργαλεία στα χέρια των βιολόγων των περασμένων αιώνων και αποτέλεσε τη βάση για την μελέτη των απολιθωμάτων και την ανάπτυξη της εξελικτικής βιολογίας. Υπήρξε πάντα προσφιλής και συνηθισμένη διαδικασία για τους βιολόγους η σύγκριση χαρακτήρων διαφορετικών οργανισμών. Χρησιμοποιώντας διαφόρων ειδών ομόλογους χαρακτήρες, χαρακτήρες που έχουν αποκλίνει από κάποια κοινή προγονική δομή λόγω διαφόρων εξελικτικών αλλαγών (ειδογένεση, συσσώρευση μεταλλάξεων, αποκλίνουσα ή συγκλίνουσα εξέλιξη), η βιολογική έρευνα ήταν σε θέση να ερμηνεύσει μια σειρά φαινομένων αναφορικά με την φυσιολογία και την εξέλιξη των οργανισμών.

Με την αλλαγή παραδείγματος [Gilbert 1991] που επέφερε στη βιολογία η έναρξη και ολοκλήρωση των προγραμμάτων αλληλούχησης εκατοντάδων γονιδιωμάτων πολλών ευκαριωτικών και πολλών περισσοτέρων προκαρυωτικών οργανισμών αλλά και την αναμονή ολοκλήρωσης των εργασιών αλληλούχησης πολλών περισσοτέρων, οι

βιολογικές επιστήμες βρέθηκαν μπροστά σε μια δεξαμενή δεδομένων άνευ προηγουμένου μεγέθους. Η εφαρμογή λοιπόν της πιο ιστορικής, διαδεδομένης και πιο συνηθισμένης για τους βιοεπιστήμονες μεθόδου, της συγκριτικής μεθόδου, ήταν αναμενόμενη και προέκυψε χωρίς καθυστέρηση.

Η εφαρμογή όλων των αρχών αλλά και του θεωρητικού υπόβαθρου της συγκριτικής μεθόδου στη σύγχρονη μεταγονιδιωματική εποχή (post-genomics era), η ανάλυση, η σύγκριση και ο σχολιασμός δηλαδή μεγάλων τμημάτων των τεράστιων ακολουθιών που έχουν κάνει διαθέσιμες τα προγράμματα των γονιδιωμάτων, έδωσε γέννηση σε ένα εξαιρετικά ενδιαφέρον αντικείμενο της σύγχρονης βιολογίας τη συγκριτική γονιδιωματική (comparative genomics). Ως συγκριτική γονιδιωματική μπορούμε να ορίσουμε πολύ απλά τη συγκριτική μελέτη ολοκληρωμένων ή τμημάτων (διαφόρων μεγεθών και δομής) γονιδιωμάτων που προέρχονται από δύο ή περισσοτέρων είδη.

Χρησιμοποιώντας τα λόγια ενός από τους πρωτοπόρους αυτού του πεδίου του Russ Altman από το πανεπιστήμιο του Stanford μπορούμε να δώσουμε έναν ελεύθερο ορισμό στον όρο συγκριτική γονιδιωματική.

"Η συγκριτική γονιδιωματική μπορεί να οριστεί ως η σύγκριση γονιδιωμάτων σε μεγάλη κλίμακα, με σκοπό την κατανόηση της βιολογίας των γονιδιωμάτων ξεχωριστά όσο και την εξαγωγή γενικών συμπερασμάτων που αφορούν ομάδες γονιδιωμάτων"

Η βασικές αρχές που διέπουν τη συγκριτική γονιδιωματική είναι εύλογα κατανοητές. Τα κοινά χαρακτηριστικά δύο (ή περισσοτέρων) οργανισμών συχνά κωδικοποιούνται από τις ακολουθίες DNA που είναι συντηρημένες μεταξύ των ειδών. Πιο συγκεκριμένα, τα τμήματα ακολουθιών DNA, που κωδικοποιούν για πρωτεΐνες και RNA που είναι υπεύθυνα για διάφορες λειτουργίες και είναι συντηρημένες από τον τελευταίο κοινό πρόγονο πρέπει να έχουν συντηρηθεί εξελικτικά στις σύγχρονες ακολουθίες γονιδιωμάτων. Παρόμοια, τα τμήματα εκείνα του γονιδιώματος που ρυθμίζουν την έκφραση των γονιδίων, τα οποία εκφράζονται παρόμοια μεταξύ δύο (ή περισσοτέρων) συγγενικών ειδών αναμένεται να έχουν και αυτές ένα βαθμό ομοιότητας. Αντιθέτως, ακολουθίες που κωδικοποιούν (ή ρυθμίζουν την έκφραση) πρωτεϊνών και RNA υπεύθυνων για διαφορετικές λειτουργίες ανάμεσα σε είδη με τη σειρά τους είναι δυνατόν να χαρακτηριστούν από τη συγκριτική ανάλυση, ως τμήματα που δεν μοιράζονται κάποιες ελληλουχικές ομοιότητες.

Η συγκριτική γονιδιωματική έχει χαρακτηριστεί σαν το "Ιερό Δισκοπότηρο" της σύγχρονης έρευνας στις βιοεπιστήμες. Παρουσιάστηκε ως το επόμενο μεγάλο βήμα

μετά την ολοκλήρωση των προγραμμάτων αλληλούχησης και σαν τον τομέα εκείνο που θα μπορέσει να αξιοποιήσει και να δώσει όσο το δυνατόν περισσότερες πληροφορίες μέσα από τον κυκεώνα των δεδομένων (sequencing data) που έχουν κατακλύσει τη σύγχρονη βιολογική έρευνα. Δεν είναι τυχαίο ότι συμπεριλήφθηκε στον κατάλογο των 10 πιο καυτών ερευνητικών πεδίων του κορυφαίου επιστημονικού περιοδικού Science για το 2003 [Magazine 2002].

Μέχρι σήμερα έχουν εφαρμοστεί αρκετές μέθοδοι συγκριτικής γονιδιωματικής αλλά χωρίς να ονομάζονται έτσι. Οι σύγχρονοι ερευνητές των βιοεπιστημών χρησιμοποιούν σε καθημερινή βάση και σε επίπεδο ρουτίνας πλέον μεθόδους για τοπική ή και συνολική στοίχιση ακολουθιών (local and global alignment tools) και πολλαπλής στοίχισης ακολουθιών (multiple sequence alignment) τεχνικές και αλγόριθμοι που έχουν αναπτυχθεί εδώ και αρκετά χρόνια αλλά δεν κατατάσσονταν συστηματικά στο πεδίο της συγκριτικής γονιδιωματικής. Στο γενικό πλαίσιο της ομολογίας μεταξύ ακολουθιών και της εξελικτικής τους συγγένειας αυτές οι μέθοδοι αποτελούν τον πρόδρομο της συγκριτικής γονιδιωματικής. Επίσης όσον αφορά την πρόβλεψη της τριτοταγούς δομής πρωτεϊνών οι μέθοδοι που στηρίζονται στους αλγόριθμους του PSI-BLAST (Position Specific Iteration BLAST) αποτελούν και αυτές προάγγελους αλλά και την υποδομή των πιο εξελιγμένων σημερινών μεθόδων. Τέλος μέθοδοι συγκριτικής γονιδιωματικής χωρίς να ονομάζονται έτσι αποτελούν και οι μέθοδοι που ασχολούνται με την πρόβλεψη και την αναζήτηση μοτίβων πρόσδεσης μεταφραφικών παραγόντων (transcription factor binding sites ή TFBS) χρησιμοποιώντας πολλαπλές ακολουθίες και χρησιμοποιούνται για την εύρεση cis-ρυθμιστικών περιοχών ανωφορικά (upstream) των προαγωγέων (promoters) γονιδίων.

Συνοπτικά οι σύγχρονες μέθοδοι της συγκριτικής γονιδιωματικής στοχεύουν στα ακόλουθα:

- Στην ανακατασκευή των προτύπων εξέλιξης οικογενειών γονιδίων (gene families). Μέσα από αυτό είναι δυνατή η ταξινόμηση των γονιδίων σε ομάδες (gene clusters) αλλά και η αποσαφήνιση του λειτουργικού τους ρόλου. Μέσα από τη συστηματική μελέτη της εξέλιξης και της διασποράς των γονιδίων σε μεγάλες οικογένειες είναι δυνατή ο λειτουργικός τους σχολιασμός (functional annotation) και ο προσδιορισμός της λειτουργίας τους (functional assignment).

- Στην αποτίμηση του ρόλου εξελικτικών διαδικασιών στο επίπεδο της δομής του γονιδιώματος (διπλασιασμοί, εισαγωγές, ελλείψεις, μεταθέσεις γονιδίων και φαινόμενα οριζόντιας μεταφοράς) και στην διαμόρφωση των οργανισμών.
- Στον προσδιορισμό cis-ρυθμιστικών στοιχείων μέσω φυλογενετικών αποτυπωμάτων. Η μέθοδος συγκριτικής αλληλούχησης (comparative sequencing), όπου γονιδιωματικές ακολουθίες συγκρίνονται μεταξύ διαφορετικών ακολουθιών για συντηρημένες μη κωδικές περιοχές, έχει αποδειχθεί σαν ένα δυνατό εργαλείο για την αναγνώριση cis-ρυθμιστικών στοιχείων.
- Στην αναγνώριση ομάδων γονιδίων που μας δίνουν ενδείξεις για:
  1. Θετική επιλογή (positive selection).
  2. Μετατροπή γονιδίων (gene conversion).
  3. Οριζόντια μεταφορά (horizontal transfer).
  4. Συντονισμένη εξέλιξη (concerted evolution).
- Στην ανακατασκευή φυλογενετικών σχέσεων χρησιμοποιώντας πολλαπλές αλληλουχίες γονιδίων (ευκαρυωτικούς) ή/και το σύνολο του γονιδιώματος των οργανισμών (προκαρυωτικούς). Οι προσπάθειες αυτές εντάσσονται στον κλάδο της φυλογονιδιωματικής (phylogenomics) [Eisen 1998].
- Τέλος σημαντικά τεχνικής φύσης προβλήματα και πεδία έρευνας πάνω στη συγκριτική γονιδιωματική αποτελούν:
  1. Οι αλγόριθμοι για τη στοίχιση πολύ μεγάλων γονιδιωματικών ακολουθιών.
  2. Προσεγγίσεις μέσω της συγκριτικής γονιδιωματικής για τον εντοπισμό γονιδίων (gene finding). Και την δημιουργία γενετικών δεικτών (genetic markers).
  3. Προσεγγίσεις μέσω της συγκριτικής γονιδιωματικής στον καθορισμό λειτουργιών (funtional genomics)
  4. Συγκριτικές προσεγγίσεις στην αναγνώριση ρυθμιστικών περιοχών στο γονιδίωμα.
  5. Συγκριτικές προσεγγίσεις για την δράση της φυσικής επιλογής στα γονιδιώματα.

Στα πλαίσια της παρούσας εργασίας έχει γίνει μια προσπάθεια να παρουσιαστούν οι πιο διαδεδομένες εφαρμογές, εργαλεία και μέθοδοι της συγκριτικής γονιδιωματικής καθώς και να ταξινομηθούν και να αξιολογηθούν οι εκάστοτε δυνατότητες τους. Το

πρώτο μέρος της μελέτης αποτελεί μια ενδελεχή ανασκόπηση όλων των σύγχρονων (και όχι τόσο σύγχρονων, μερικές από τις βασικές μεθόδους έχουν εμφανιστεί εδώ και πάνω από 15 χρόνια) μεθόδων και εφαρμογών συγκριτικής γονιδιωματικής. Στο δεύτερο μέρος γίνεται μια ιδιαίτερη αναφορά, παρουσίαση και αξιολόγηση όλων εκείνων των μεθόδων της συγκριτικής γονιδιωματικής που περιλαμβάνουν κάποιου είδους φυλογενετική/εξελικτική ανάλυση (φυλογονιδιωματική phylogenomics, φυλογενετικά αποτυπώματα phylogenetic footprinting κ.α.) η ιδιαίτερη σχέση ανάμεσα στη συγκριτική γονιδιωματική και την εξελικτική βιολογία και η αλληλοϋποστήριξη των δύο πεδίων μεταξύ τους είναι από τα πιο ενδιαφέροντα και ελπιδοφόρα επιτεύγματα της σύγχρονης βιολογικής έρευνας [Koonin,Aravind & Kondrashov 2000]. Στο τρίτο μέρος παρουσιάζονται: α) ένα μοντέλο για τη διενέργεια αρνητικών πειραμάτων ελέγχου (negative control experiments) σε εφαρμογές συγκριτικής γονιδιωματικής και β) μια εαφρμογή μια σειράς μεθόδων της συγκριτικής γονιδιωματικής στην χαρτογράφηση γονιδίων και την δημιουργία γενετικών χαρτών και δεικτών με Next Generation Sequence (NGS) ανάλυση σε οργανισμούς χωρίς δημοσιευμένα γονιδιώματα.

3 Πρώτο Μέρος

3.1 Εισαγωγή

Στο πρώτο μέρος της εργασίας θα γίνει μια παρουσίαση των πιο σημαντικών ζητημάτων που άπτονται της συγκριτικής γονιδιωματικής. Θα παρουσιαστούν οι σύγχρονοι τρόποι πρόσβασης στις βάσεις δεδομένων των πιο σημαντικών γονιδιωμάτων και οι τρόποι να εκμεταλλευτούμε τα πρωτογενή εργαλεία για συγκριτική ανάλυση που μας παρέχουν πολλές από αυτές. Η ανάπτυξη των συγχρόνων βάσεων δεδομένων των γονιδιωμάτων δεν σταματά στην απλή αποθήκευση των γενομικών αλληλουχιών αλλά έχει καταφέρει να παρέχει και αρκετά εργαλεία τόσο όσον αφορά τον σχολιασμό των γονιδιωμάτων (genome annotation) όσο και τη σχεδίαση λεπτομερών γενετικών χαρτών (maping). Στη συνέχεια θα παρουσιαστούν οι μέθοδοι σύγκρισης ολόκληρων γονιδιωμάτων (whole genome comparisons) και τα επιτεύγματα τους στη λειτουργική γονιδιωματική (functional genomics) και την διερεύνηση της εξέλιξης των γονιδιωμάτων (genome evolution). Το κεφάλαιο θα κλείσει με μια ανασκόπηση, ταξινόμηση και αξιολόγηση των πιο σημαντικών και ελπιδοφόρων μεθόδων ανάλυσης και των εφαρμογών τους.

Σκοπός του κομματιού αυτού είναι τόσο να κάνει μια ενδελεχή παρουσίαση των πηγών και των μεθόδων που χρησιμοποιούνται για την πρόσκτηση και την ανάκτηση

δεδομένων και δομημένης γνώσης καθώς και να τις αξιολογήσει όσο και να επισημάνει κάποιες γενικές αρχές λειτουργίας τους που θα βοηθήσουν ερευνητές που δεν είναι εξοικειωμένοι με την βιοπληροφορική να είναι σε θέση να κατανοήσουν και να μπορούν να εντάξουν τις μεθόδους και τις πηγές αυτές στην καθημερινή ερευνητική τους πρακτική. Το σύνολο των δεδομένων που παρουσιάζονται στο πρώτο μέρος είναι χρήσιμο να διαβαστεί και να κατανοηθεί κάτω από το πρίσμα ενός απλού μαθηματικού μοντέλου που συνδέει τις φυλογενετικές αποστάσεις και το μέγεθος των συντηρημένων περιοχών που χρειάζεται να μελετηθούν με τον αριθμό των γονιδίων που απαιτείται να συγκριθούν [Eddy 2005]. Το μοντέλο προσδιορίζει κάτι που διαισθητικά είναι αναμενόμενο, ότι για μια δεδομένη φυλογενετική απόσταση ο αριθμός των γονιδιωμάτων που απαιτείται κλιμακώνεται αντίστροφα με το μέγεθος των στοιχείων σύγκρισις που μελετώνται, δηλαδή για την μελέτη μονηρών νουκλεοτιδικών πολυμορφισμών (SNPs) απαιτούνται πολύ περισσότερα γονιδιώματα από ότι για τη μελέτη ρυθμιστικών περιοχών ή τη μελέτη της δομής (εξόνια-εσόνια) των γονιδίων. Αυτή η αντίστροφα κλιμακούμενη συμπεριφορά είναι χρήσιμη για την κατανόηση της φύσης των πηγών δεδομένων καθώς και για το σχεδιασμό πειραμάτων συγκριτικής γονιδιωματικής οργανσιμών.

3.2 Βάσεις Δεδομένων Συγκριτικής Γονιδιωματικής

Η σύγχρονη συγκριτική γονιδιωματική εκμεταλλευόμενη τις εξελίξεις στην διαχείριση συστημάτων βάσεων δεδομένων καθώς και στις τεχνολογίες διαχείρισης και απεικόνισης δεδομένων έχει κάνει διαθέσιμες πολλές ενδιαφέρουσες υλοποιήσεις συγκέντρωσης δεδομένων σε βάσεις δεδομένων. Βάσεις οι όποιες πλέον εκτός από αποθήκευση απλά βιολογικών ακολουθιών στοχεύουν στο να παραγάγουν γνώση και πληροφορία και να την κάνουν διαθέσιμη αλλά και ορατή (μέσω σύγχρονων τεχνικών αναπαράστασης) στου ερευνητές των βιοεπιστημών. Τέτοιες πηγές πληροφορίας υπάρχουν ελεύθερες στο διαδίκτυο τόσο για προκαρυωτικούς οργανισμούς όσο και για ζώα και φυτά. Στην συνέχεια αυτού του μέρους θα παρουσιαστούν τα σημαντικότερα παραδείγματα που είναι διαθέσιμα, με μια έμφαση σε φυτά και μικροοργανισμούς, σε μια προσπάθεια αυτές οι πηγές να γίνουν εύκολα διαθέσιμες αλλά και να χρησιμοποιηθούν από ερευνητές των βιοεπιστημών που δεν είναι εξοικειωμένοι με τις μεθόδους και τις πηγές δεδομέων της βιοπληροφορικής.

3.2.1 Βάσεις δεδομένων γονιδιωμάτων φυτών

3.2.1.1 Η βάση γονιδιωμάτων φυτών PlantGDB

http://www.plantgdb.org/ [Duvick et al. 2008] είναι η νο1 κατεξοχήν σύγχρονη βάση που εξειδικεύεται πλήρως στην συγκριτική γονιδιωματική των φυτών.

3.2.1.2 Η βάση συγκριτικής γονιδιωματικής φυτών GreenPhyllDB

http://www.greenphyl.org [Rouard et al. 2011]. Η βάση αυτή εξειδικεύεται περισσότερο στη λειτουργική γονιδιωματική και στην ανάλυση των εξελικτικών σχέσεων των φυτών μέσω της συγκριτικής γονιδιωματικής. Είναι εξαιρετικά χρήσιμη για de-novo εύρεση γονιδίων σε μη χαρτογραφημένα είδη.

3.2.2 Βάσεις δεδομένων μικροοργανισμών

3.2.2.1 Η ολοκληρωμένη βάση γονιδιωμάτων μικροοργανισμών

http://mbgd.genome.ad.jp [Uchiyama,Mihara,Nishide & Chiba 2013]. Περιέχει ολοκληρωμένα γονιδιώματα μικροοργανισμών και επιτρέπει την απευθείας ανάλυση ορθόλογων – παράλογων ακολουθιών ελεύθερα από το web-interface της.

3.2.2.2 Η βάση συγκριτικής γονιδιωματικής Xbase

http://www.xbase.ac.uk/ [Chaudhuri & Pallen 2006]. Περιέχει τόσο ολοκληρωμένα όσο και μη ολοκληρωμένα και κομμάτια από γονιδιώματα. Αποτελεί μια συλλογή απο άλλες βάσεις δεδομένων, μια μέτα-βάση δεδομένων.

3.3 Συγκριτική Γωνιδιωματική Φυτών

Πολλά φυτά έχουν ορισμένες ιδιαιτερότητες στα γονιδιώματα τους που τα κάνουν να αποτελούν ενδιαφέροντα πεδία για την εφαρμογή σύγχρονων μεθόδων συγκριτικής γονιδιωματικής. Τα χαρακτηριστικά αυτά περιλαμβάνουν: το συγκριτικά πολύ μεγάλο μέγεθος των γονιδιωμάτων πολλών καλλιεργούμενων φυτών (π.χ. Όλα τα σιτηρά εκτός από το ρύζι) είναι γεμάτα με επαναλαμβανόμενες ακολουθίες DNA και πολλών ειδών μεταθετά στοιχεία. Επιπλέον τα γονιδιώματα όλων των συγχρονων καλλιεργούμενων φυτών έχουν προκύψει από πολλαπλούς αρχαίους συνολικούς διπλασιασμούς τους γονιδιώματος (whole genome duplications) [Cui et al. 2006] οι οποίοι έχουν συμβεί πέραν της μιας φοράς στο σύνολο της εξελικτικής τους ιστορίας και σε

κάποιου από τους παρελθόντες κοινούς προγόνους των σύγχρονων φυτών [Schmidt 2002].

3.3.1 Συγκριτική γονιδιωματική για τον προσδιορισμό της προέλευσης των γονιδιωμάτων καλλιεργούμενων φυτών.

Η προέλευση των γονιδιωμάτων των σημαντικότερων καλλιεργούμενων φυτών μας προσφέρει δύο πλεονεκτήματα. Πρώτον, η προέλευση και οι εξελικτικές σχέσεις των γονιδιωμάτων μεταξύ τους μπορεί να δώσουν πολύτιμες πληροφορίες σε πειράματα βελτίωση φυτών αλλά και σε εισαγωγή αγρονομικών χαρακτήρων με οικονομικό ενδιαφέρον. Δεύτερον, γνωρίζοντας την εξελικτική πορεία ενός σύγχρονου καλλιεργούμενου φυτού αλλά και των παθογόνων του μπορεί κανείς να αναπτύξει πολύ πιο αποτελεσματικά και ολοκληρωμένα προγράμματα τόσο καλλιέργειας όσο και καταπολέμησης των παθογόνων των φυτών.
Συνεξέλιξη φυτού / παθογοόνου... σημασίες της ανακατασκευής της εξελικτικής ιστορίας των φυτικών γονιδιωμάτων.

3.3.2 Συκριτική γονιδιωματική στην χαρτογράφηση γενετικών χαρακτήρων στα φυτά.

Ανάπτυξη COS (Conserved Ortholog Set) markers. Χαρτογράφηση με χρήση συντενικών χαρακτηριστικών σε είδη που δεν έχουν πλήρως αναγνωσμένα γονιδιώματα... κλπ. Κλπ.

3.3.3 Σύγχρονες πηγές συγκριτικής γονιδιωματικής στα φυτά.

3.3.3.1 Η βάση *Phytozome*

http://www.phytozome.net/ [Goodstein et al. 2012] αναπτύχθηκε από το Joint Genome Institute (JGI) και αποτελεί την πιο σύγχρονη πηγή δεδομένων και γνώσης όσον αφορά την συγκριτική γονιδιωματική φυτών. Έχει εξαιρετική ικανότητα για πολλά προ-υπολοσμένα καθημερινές δουλείες ρουτίνας που μπορεί να χρειαστεί ένας γονιδωματικός βιολόγος και πολύ καλή προγραμματική πρόσβαση. Προσφέρει την εξελικτική ιστορία κάθε αναγνωρισμένου γονιδίου στα φυτά που περιέχει τόσο στο επίπεδο της ακολουθίας όσο και στο επίπεδο της πρωτεϊνικής οικογένειας μέχρι την γονιδιωματική οικογένεια.

4 Δεύτερο Μέρος

4.1 Εισαγωγή

Στο μέρος αυτό θα επιχειρηθεί μια ευρεία ανασκόπηση, ερμηνεία και αξιολόγηση των αναλυτικών (μαθηματικών και υπολογιστικών) μεθόδων ανάλυσης στη συγκριτική γονιδιωματική. Ξεκινώντας από την υλοποίηση της εφαρμογής της συγκριτικής μεθόδου πάνω στις ελεύθερα προσβάσιμες πλήρης γονιδιωματικές ακολουθίες εκατοντάδων οργανισμών, και τις πρώτες προσπάθειές στοίχισης ακολουθιών η ερμηνία των μεθόδων θα φτάσει μέχρι τις πιο εξελιγμες σύγχρονες μεθόδους συγκριτικής φυλογονιδιωματικής.

Τα προγράμματα αποκωδικοποίησης των γονιδιωμάτων οργανισμών έχουν φτάσει πια στην μέτα- εποχή τους (post-genomics era) που σημαίνει ότι τα προβλήματα της ανάγνωσης και συναρμολόγησης ενός γονιδιώματος αποτελούν πλέον προβλήματα ρουτίνας και τα ανοιχτά ζητήματα και ερωτήματα αιχμής συγκεντρώνονται στο πεδίο της λειτουργικής, ρυθμιστικής και εξελικτικής ανάλυσης των γονιδιωμάτων. Και για τις τρεις αυτές κατηγορίες προβλημάτων (που αυθαίρετα τις ξεχωρίζουμε εδώ καθώς στην πράξη αποτελούν κομμάτια της έρευνας της ολοκληρωμένης βιολογίας -integrative biology) οι μέθοδοι συγκριτικής γονιδιωματικής προσφέρουν ανεκτίμητα εργαλεία ανάλυσης. Οι διάφορες αυτές ερωτήσεις μπορεί να αρχίσουν να διερευνόνται αποτελεσματικά συγκρίνοντας απλά -μέσω των διαδεδομένων αλγόριθμων στοίχισης- γονιδιώματα από οργανισμούς σε διαφορετικές φυλογενετικές ποστάσεις μεταξύ τους. Ξεχωρίζουμε 3 κατηγορίες οι οποίες αλληλοεπικαλύπτονται με τις 3 κατηγορίες ανάλυσρων που μόλις αναφέρθηκαν.

Μακρινές φυλογενετικές αποστάσεις (πάνω από 1 δισεκατομμύριο χρόνια)

Συγκρίνοντας γονιδιώματα οργανισμών που έχουν διαχωριστεί εξελικτικά για πάνω απο 1 δις χρόνια μπορούμε να αποκτήσουμε μια πλατιά εικόνα για τα είδη των γονιδίων που μοιράζονται μεταξύ τους οι οργανισμοί. Συνήθως τέτοια γονίδια έχουν λειτουργίες διαχείριστικες (housekeeping) και εκφράζονται συνεχώς στους οργανισμούς. Για παράδειγμα το μη πλεονάζων σετ γονιδίων από τη δροσόφιλα (*Drosophila melanogaster*) και τον σκουλήκι (*Caenorabditis elegans*) έχει σχεδόν το ίδιο μέγεθος και είναι το διπλάσιο (και για τους δύο οργανισμούς) από αυτό της ζύμης (*Saccharomyces serevisiae*). Παρατηρούμε λοιπόν ότι ακόμα και για 3 τόσο καλά μελετημένους οργανισμούς μπορεί κανείς να δει ότι η μεγαλύτερη αναπτυξιακή πολυπλοκότητα και τα πιο σύνθετα μονοπάτια μεταγωγής σήματος στα δύο είδη του ζωικού βασιλείου απαιτεί τα διπλά γονίδια από ότι ο μονοκύτταρος ευκαρυότης της ζύμης [Rubin et al. 2000]. Η συγκριτική γονιδιωματική μας δείχνει μια εικόνα της δροσόφιλας, που είναι ένα μετάζωο με πολύπλοκη ανάπτυξη και εξελιγμένο νευρικό σύστημα απαιτεί μόλις 2 φορές περισσότερα γονίδια από τον απλό ευκαρυώτη της ζύμης, αυτό βοήθα στο να κατανοήσουμε ότι η πολυπλοκότητα και η μεταβλητότητα μεταξύ των ευκαρυώτικών οργανισμών δεν οφείλετε αποκλειστικά στο γονιδίωμα τους αλλά στο πως αυτό εκφράζεται και ρυθμίζεται. Σημειώνουμε εδώ ότι σε τέτοιες μεγάλες φυλογενετικές αποστάσεις οι γονιδιακές ακολουθίες που κωδικοποιούν πρωτεΐνες εμφανίζουν ομοιότητες που μπορούν να ανιχνευθούν από τα προγράμματα στοίχισης αλλά τόσο η σειρά των γονιδίων στο γονιδίωμα (synteny) όσο και η συντήρηση ρυθμιστικών περιοχών δεν είναι συντηριμένες.

Ενδιάμεσες φυλογενετικές αποστάσεις (μεταξύ 80 και 200 εκατομμυρίων χρόνων)

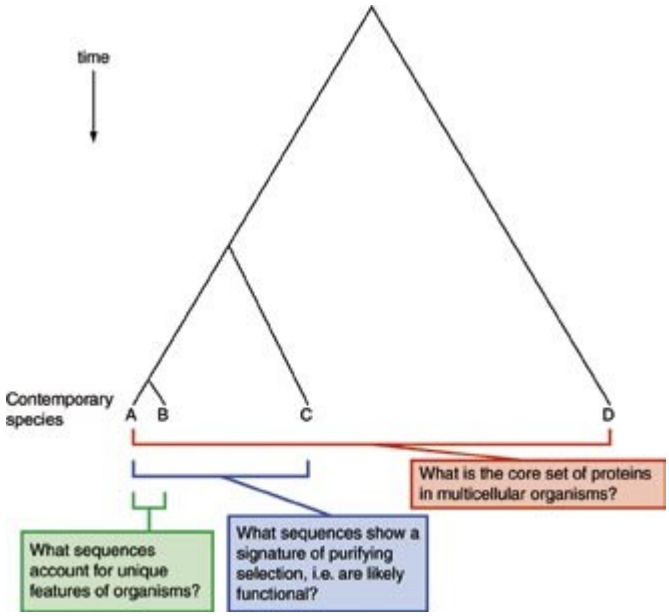
Σε αυτό το βαθμό διαχωρισμού των οργανισμών αναμένεται οι κωδικές ακολουθίες να εμφανίζουν ένα βαθμό απόκλισης και συσσώρευσης μεταλλαγών στατιστικά μικρότερο από ότι οι μη κωδικές περιοχές (σύμφωνα με τη θεωρία της ουδετερότητας των μεταλλαγών). Οι συγκριτική γονιδιωματική μπορεί να αποκαλύψει όχι μόνο τις λειτουργικές (και ως εκ τούτου συντηρημένες) περιοχές των γονιδιωμάτων αλλά συμβάλει στην αναγνώριση της δομής των κωδικών περιοχών (όπως εξόνια - εσόνια), του σχολιασμού της λειτουργίας τους καθώς και στον προσδιορισμό ορισμένων ρυθμιστικών τους περιοχών. Χαρακτηριστικά παραδείγματα αποτελούν οι συγκρίσεις που γο-

νιδιώματος του ποντικιού με το ανθρώπινο [Consortium et al. 2002] αλλά και την πρόβλεψη της λειτουργίας γονιδιωματικών περιοχών με τη σύγκριση των γονιδιωμάτων των 2 πιο διαδεδομένων οργανισμών στις ζύμες [Cliften et al. 2001].

Κοντινές φυλογενετικές αποστάσεις (μεταξύ 5 και 10 εκατομμυρίων χρόνων).

Οι συγκρίσεις γονιδιωμάτων σε αυτή την κλίμακα διαχωρισμού συμβάλει στο μεγαλύτερο βαθμό στο να μελετηθούν οι κομβικές διαφορές στις ακολουθίες που κάνει τους οργανισμούς διαφορετικούς. Για παράδειγμα μεταξύ του ανθρώπου και του χιμπαντζή που τους χωρίζει μόλις 1% διαφορές στο γονιδίωμα, οι συγκριτικ μελέτες μπορούν να αποκαλύψουν πέρα από τις νουκλεοτιδικές (και αμινοξικές διαφορές) στις κωδικές περιοχές και τις διαφορές στις ρυθμιστικές περιοχές αλλά και πως αυτές επηρεάζουν την διαφορική έκφραση των γονιδίων στην οποία οφείλετε το συντριπτικά μεγαλύτερο μέρος των διαφορών μεταξύ των δύο αυτών ανεπτυγμένων οργανισμών.

Συνοπτικά οι συγκρίσεις μεταξύ των γονιδιωμάτων σε διαφορετικές κλίμακες διαχωρισμού των ειδών και φυλογενετικές αποστάσεις απαντά σε 3 διαφορετικές όσο και σημαντικές ομάδες προβλημάτων και αποτελεί ένα από τα πιο πολύπλευρα και ισχυρά εργαλεία συγκριτικής ανάλυσης των γονιδιωμάτων όπως παρουσιάζονται στην εικόνα 4.



Εικόνα 4: Η σύγκριση γονιδιωμάτων σε διάφορες φυλογενετικές αποστάσεις επιτρέπει την αντιμετώπιση ανάλογων ερωτημάτων (τροποποιημένη από [Hardison 2003]).

Στη συνέχεια του δεύτερου μέρους θα παρουσιαστούν και αναλυθούν μέθοδοι ανάλυσης συγκριτικής γονιδιωματικής που βασίζονται στην αλληλουχικής ομοιότητας, αντιθέτως μέθοδοι που δεν στηρίζονται στην αλληλουχική ομοιότητα, μέθοδοι προσδιορισμού λειτουργίας και θα γίνει μια μικρή αναφορά και ανασκόπηση στα αλγοριθμικά προβλήματα που αντιμετωπίζει η συγκριτική γονιδιωματική. Σκοπός αυτών των μεθόδων είναι να χρησιμοποιήσουν ένα σύνολο γονιδιωμάτων προερχόμενο από διάφορες φυλογενετικές κλίμακες από κοινού με τέτοιο τρόπο ώστε να κατανοήσουμε καλύτερα τα μεμονωμένα γονιδιώματα [Haubold & Wiehe 2004].

4.2 Μέθοδοι ανάλυσης ομοιότητας αλληλουχιών

Ο βασικός κορμός των μεθόδων ανάλυσης στη συγκριτική γονιδιωματική αποτελείται απο μεθόδους που βασίζονται στην ανάλυση, ποσοτικοποίηση και στατιστική σημαντικότητα της σύγκρισης ομοιοτήτων και διαφορών μεταξύ ακολουθιών (τόσο γονιδίων και γονιδιωμάτων όσο και πρωτεϊνών). Ο τομέας της ανάλυσης ομοιότητας και ομολογίας μεταξύ ακολουθιών αποτελεί τη βάση ανάπτυξης και δομικό λίθο ανάπτυξης της συγκριτικής γονιδιωματικής. Οι βιολογικές ακολουθίες προέρχονται και έχουν δομηθεί από την συνεχιζόμενη βιολογική εξέλιξη των οργανισμών και ως εκ τούτου η ομοιότητα που εμφανίζουν στην ακολουθία τους είναι αποτέλεσμα κάποιας κοινής καταγωγής. Ομολογία λοιπόν για τη βιολογία σημαίνει κοινή καταγωγή και η ομοιότητα είναι απλά ένα αποτέλεσμα. Οι δύο έννοιες δεν πρέπει να συγχέονται καθώς η ομοιότητα αναφέρεται στην ποσοτικοποίηση της σύγκρισης ακολουθιών ενώ η ομολογία στην κοινή τους καταγωγή.

Οι βιολογικές ακολουθίες μπορεί να έχουν δύο κοινή καταγωγή με δύο τρόπους: είτε από κάποιο γεγονός ειδογένεσης οπότε οι ομόλογες ακολουθίες ονομάζονται ορθόλογες είτε από κάποιο γεγονός διπλασιασμού οπότε και ονομάζονται παράλογες. Ο διαχωρισμός ορθόλογων παράλογων ακολουθιών είναι σημαντικός για την αναγνώριση λειτουργιών των γονιδίων, απασχολεί ένα μεγάλο κομμάτι των μεθόδων της συγκριτικής γονιδιωματικής και θα αναφερθούμε εκτενέστερα σε επόμενο τμήμα αύτου του μέρους.

4.2.1 Μέθοδοι στοίχισης βιολογικών ακολουθιών.

Το πρώτο και σημαντικό βήμα στην συγκριτική ανάλυση ακολουθιών αποτελούν οι μέθοδοι στοίχισης (sequence alignment methods). Οι υπολογιστικές μέθοδοι στοίχισης προσβλέπουν στην λύση του προβλήματος της μέγιστης κοινής υποακολουθίας

μεταξύ δύο ακολουθιών (the largest common substring problem [Gusfield 1997]) ο τρόπος με τον οποίο το λύνουν τις ξεχωρίζει σε:

1. Τοπική η ολική στοίχιση: Με την τοπική στοίχιση (local alignment) οι αλγόριθμοι στοίχισης επιδιώκουν την μέγιστη δυνατή κάλυψη όσο το δυνατό περισσοτέρων νουκλοετιδίων/αμινοξέων. Οι αλγόριθμοι ολικής στοίχισης (global alignment) επιδιώκουν την στοίχιση όλων των αμινοξέων/νουκλεοτιδίων στις ακολουθίες.
2. Πολλαπλή ή δυαδική στοίχιση: Η δυαδική στοίχιση (pairwise alignment) υπολογίζει περιοχές ομοιότητας μεταξύ δύο μόνο ακολουθιών ενώ η πολλαπλή στοίχιση (multiple alignment) περιλαμβάνει την στοίχιση πολλών ακολουθιών μεταξύ τους και απαιτεί ως εκ τούτου την δυαδική στοίχιση όλων των ζευγαριών ακολουθιών μεταξύ τους.

Στη συνέχεια θα εστιάσουμε στην (τοπική αλλά και ολική μερικές φορές) πολλαπλή στοίχιση καθώς αποτλεί μια κομβική κατηγορία μεθόδων στη συγκριτική γονιδιωματική. Τόσο από μόνη της αποτελεί ένα σημαντικό εργαλείο για να προσδιοριστούν συντηρημένες περιοχές μεταξύ ομόλογων ακολουθιών αλλά και για να αποκαλυφθεί η εξελικτική ιστορία γονιδιωμάτων αλλά η πολλαπλή στοίχιση ακολουθιών αποτελεί το εναρκτήριο βήμα για αρκετές πιο ανεπτυγμένες μεθόδους συγκριτικής γονιδιωματικής, όπως το φυλογενετικο προφίλ και η δημιουργία πρωτεϊνικών/γονιδιακών οικογενειών (COGs [Tatusov et al. 2003]).

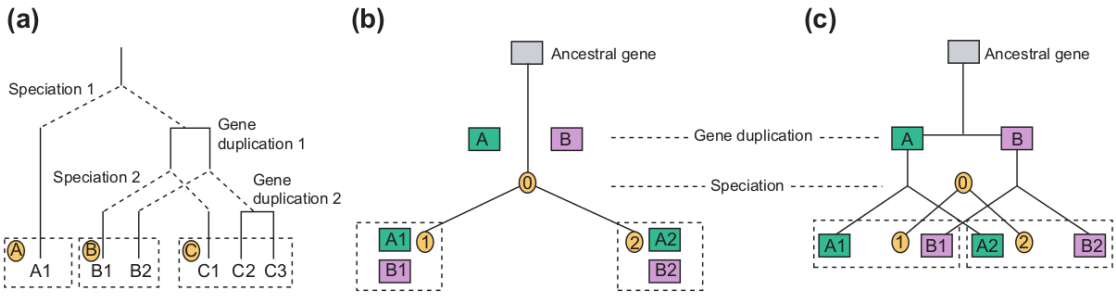
4.2.2 Τα COGs (clusters of orthologus groups) και η βάση δεδομένω τους.

Ένα από τα πιο επιτυχημένα εγχειρήματα της συγκριτικής γονιδιωματικής και ταυτόχρονα μια από τις πιο διαδεδομένες πηγές της είναι τα GOGs [Tatusov,Koonin & Lipman 1997] (clusters of orthologues groups) και η βάση δεδομένων του GOG database [Tatusov et al. 2003]. Ένα GOG αποτελείτε από ακολουθίς γονιδίων που έχουν χαρακτηριστεί ως ορθόλογες με υπολογιστικές μεθόδους (Best Reciprocal Blast Hit) και τα ορθόλογα αυτά γονίδια απαντώνται σε τουλάχιστον 3 διαφορετικά φύλα οργανισμών. Ξεκινώντας από αυτόν τον απλό ορισμό και αναπτύσσοντας τη ανάλυση και τη βάση δεδομένων τους τα GOGs αποτέλεσαν ένα από τα πρώτα δομικά στοιχεία της συγκριτικής γονιδιωματικής και ενώ ξεκίνησαν από ~720 οικογένειες GOG στην αρχική δημοσίευση αυτή τη στιγμή η GOG database περιέχει εκατομμύρια COGs από πολλές εκατοντάδες οργανισμούς καθώς και έναν βελτιωμένο αλγόριθμο εύρεσης COGs [Kristensen et al. 2010].

4.2.3 Διαχωρισμός ορθόλογων / παράλογων.

Ο διαχωρισμός ορθόλογων και παράλογων ακολουθιών στην συγκριτική γονιδιωματική αποτελεί ένα από τα πιο δύσκολα προβλήματα, που όμως η επίλυση του θα έχει να προσφέρει εξαιρετική βελτίωση στον τρόπο προσδιορισμού της λειτουργίας των γονιδιακών ακολουθιών. Πέρα από την αναγνώριση των COGs που περιγράφτηκε στην προηγούμενη ενότητα υπάρχουν πολύ πιο εξειδικευμένες μέθοδοι της συγκριτικής γονιδιωματικής και πηγές στο διαδίκτυο που προσφέρουν εξειδικευμένη αναγνώρηση, σχολιασμό και ανάλυση ορθόλογων ακολουθιών.

Ο ορισμός μιας ορθόλογης ακολουθίας δόθηκε πρώτη φορά από τον W. Fitch [] και μάλιστα χρειάστηκε να ξαναεπαναλάβει την αρχική του θέση καθώς με την εμφάνιση πολλών νέων πλήρως αναγνωσμένων γονιδιωμάτων άρχισε να επικρατεί σύγχυση στον κλάδο της εξελικτικής γονιδιωματικής. Η περιγραφή του Jensen στο [Jensen 2001] στηρίζεται στην ερμηνεία του Fitch και προσφέρει μια συμπαγή και ακέραια απάντηση στο πρόβλημα που προέκυψε μεταξυ εξελικτικών και γονιδιωματικών βιολόγων. Η εικόνα 5, παρουσιάζει την εξήγηση του Jensen που διευκρινίζει πλήρως τον διαχωρισμό ορθόλογων και παράλογων ακολουθιών.



Εικόνα 5: Ο διαχωρισμός σε ορθόλογα και παράλογα στηρίζεται στην εξελικτική πορείς των ακολουθιών και στο αν είναι προϊόντα ειδογένεσης ή διπλασιασμού. Στο παράδειγμα μας το γονίδιο A1 έχει 3 ορθόλογα γονίδια στο είδος C αλλά από αύτο μόνο το ένα είναι ορθόλογο με το B1. Επίσης το B2 έχει δύο ορθόλογα στο είδος C και ένα παράλογο το C1. Τέλος όλα τα γονίδια στο είδος C είναι παράλογα μεταξύ τους.

Μετά την απαραίτητη διευκρίνιση στην ορολογία ήταν ζήτημα χρόνου να αρχίσουν να αναπτύσσονται οι πρώτοι αλγόριθμοι εύρεσης ορθόλογων που να χρησιμοποιούν και να μπορούν να αντεπεξέλθουν στην υψηλή απόδοση των high-throuput μεθόδων. Ο πιο κλασικος αλγόριθμος χρησιμοποιεί μια τεχνική ομαδοποίησης που ονομάζεται Markov Clustering (MCL) για να προσδιορίσει ευριστικά ορθόλογες ακολουθίες σε μεγάλα σετ δεδομένων, ονομάζεται OrthoMCL [Li,Stoeckert & Roos 2003] και είναι ο πιο ευρέως διαδεδομένος αλγόριθμος για high-throuput χαρακτηρισμό ορθόλογων.

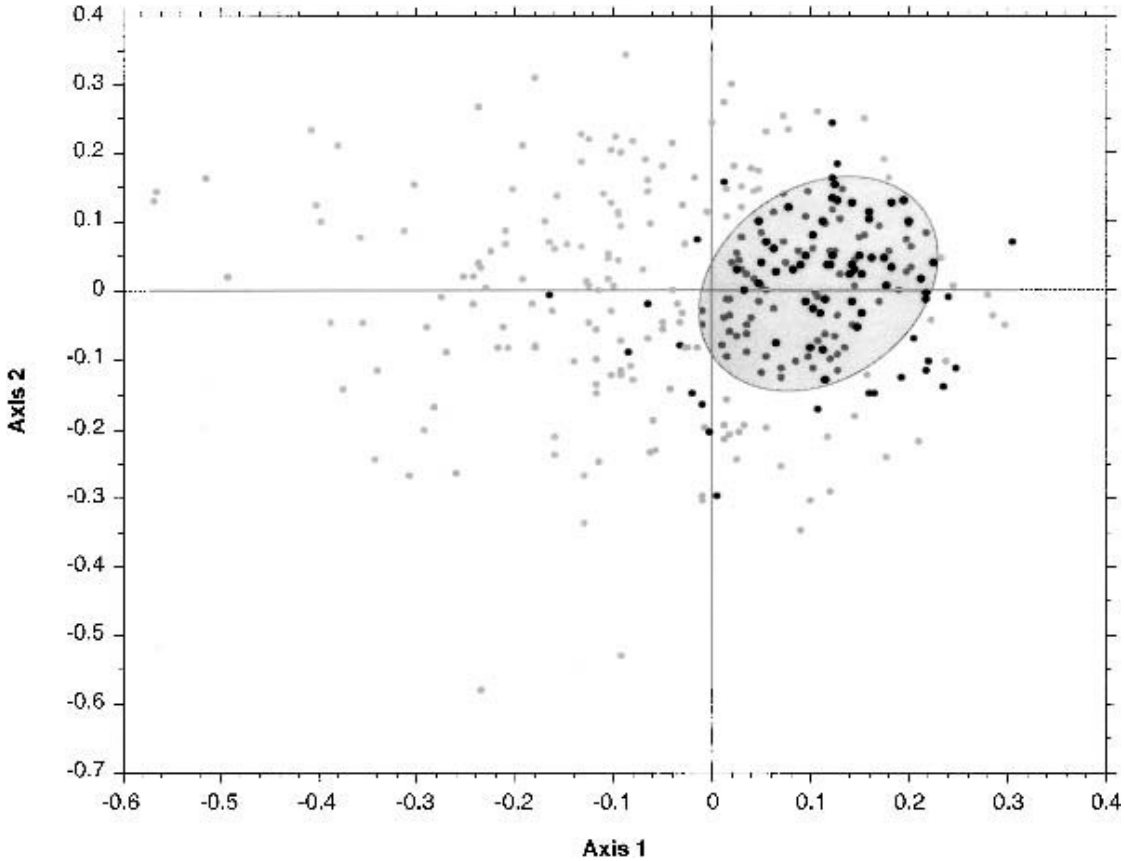
Παράλληλα έχουν αναπτυχθεί πλούσιες διαδικτυακές πηγές που περιέχουν ήδη υπολογισμένα μεγάλα σετ ορθόλογων ακολουθιών από πολλά μοντέλα οργανισμούς, η πιο διαδεδομένη βάση που περιέχει ήδη υπολογισμένες ορθόλογες ακολουθίες και από όπου μπορεί να πραγματοποιηθεί απευθείας αναζήτηση με μια οποιαδήποτε αλληλουχία (πρωτεΐνης ή DNA) που έχει ενδιαφέρον, είναι η OrthoDB[Kriventseva et al. 2014]. Η βάση ήδη περιέχει στην τελευταία της έκδοση 3027 (2627 βακτήρια) ολοκληρωμένα γονιδιώματα με τις ομάδες των ορθόλογων τους ιεραρχικά ομαδοποιημένες.

4.3 Μέθοδοι ανακατασκευής εξελικτικών σχέσεων που βασίζονται σε ολόκληρα γονιδιώματα.

Η ανάγνωση της πλήρους ακολουθίας πολλών γονιδιωμάτων από όλα τα εξελικτικά φύλα προσδίδει πλέον στη φυλογενετική μια σημαντική ευκαιρία να αναπτυχθεί από μια επιστήμη που μελετούσε την εξέλιξη συγκεκριμένων κομματιών DNA, (κυριώς γονίδια) σε μια επιστήμη που πλέον χρησιμοποιεί το σύνολο της γενετικής πληροφορίας που κληρονομεί ένας οργανισμός για να προσδιορίζει την εξελικτική του πορεία. Έτσι η φυλλογενετική ολοκληρωμένων γονιδιωμάτων (whole genome phylogenetics) είναι πλέον η κυρίαρχη προσέγγιση για την ανακατασκευή φυλογενετικών δέντρων και δικτύων. Στο πεδίο αυτό οι μέθοδοι της συγκριτικής γονιδιωματικής βρίσκουν σημαντικές εφαρμογές στον προσδιορισμό ενός "πυρήνα γονιδίων" (core gene set) που θα μπορεί να οδηγεί στη σύγκλιση των πολλών διαφορετικών φυλογενετικών δέντρων γονιδίων σε ένα συναινετικό δέντρο που θα αντιπροσωπεύει την εξελικτική σχέση των οργανισμών που μελετώνται.

Η μεθοδολογία προσδιορισμού του κοινού "πυρήνα γονιδίων" οποίος μπορεί να αντιρποσωπεύει την κοινή εξελικτική ιστορία ενός συνόλου οργανισμών βασίζεται σε μια σημαντική αναλυτική μέθοδο της συγκριτική γονιδιωματικής. Η μέθοδος ξεκινά απο την κατασκευή εξελικτικών δέντρων για κάθε γονίδιο ξεχωριστά από το σύνολο των κοινών γονιδίων των προς ανάλυση οργανισμών. Στη συνέχεια υπολογίζεται μια τοπολογική απόσταση για το σύνολο των δέντρων γονιδίων. Με βάση την μήτρα των τοπολογικών αποστάσεων μεταξύ κάθε ζεύγους δέντρων γονιδίων επιχειρείται μια αναδιάταξη (decomposition) της μήτρας αυτής με μια τεχνική αναδιάταξης μήτρων. Από την προβολή των δύο κυρίων συνιστωσών της αναδιάταξης γίνεται δυνατός ο εντοπισμός γονιδιών που δεν έχουν επηρεαστεί σημαντικά από οριζόντιες μεταφορές (καθώς αυτός είναι ο κύριος λόγος της ασυμφωνίας των δέντρων γονιδίων με τα

δέντρα οργανισμών). Με τον εντοπισμό αυτού του σετ γονιδίων μπορεί να ανασκευαστεί η εξελικτική ιστορία των προς μελέτη οργανισμών χρησιμοποιώντας οποιαδήποτε από τις σύγχρονες μεθόδους φυλογενετικής, μέγιστη πιθανοφάνεια (maximum likelhood ML) ή τη φειδωλή μέθοδο (maximum parsimony MP). Στο παράδειγμα που παραθέτουμε στην εικόνα 3 οι [Daubin,Gouy & Perrière 2002] χρησιμοποίησαν την principal coordinate analysis (PCO) για τον εντοπισμό του "πυρήνα των γονιδίων" και μεθόδους που βασίζονται σε γενετικές αποστάσεις αλλά και την ML (εικόνα 3). Στην ερευνητική μας ομάδα χρησιμοποιούμε μια παρόμοια μεθοδολογία που όμως στηρίζεται στην Canonical Correlation Analysis (CCA) για τον προσδιορισμό του "πυρήνα των γονιδίων" και στην MP αλλά και σύγχρονες ML μεθόδους για την ανακατασκευή του φυλογενετικού δέντρου. Χρησιμοποιώντας την CCA κανείς αποφεύγει την ανάγκη ορισμού κάποιου "αυθαίρετου" πλαφόν προκειμένου να αναγνωριστεί ο "πυρήνας των γονιδών" απο την προβολή των συνιστωσών.



Εικόνα 3: Προβολή της ανάλυσης κυρίων συντεταγμένων (PCO) του συνόλου των γονιδίων μεταξύ 45 οργανισμών. Ο πυρήνας των "πληροφοριακών" γονιδίων μπορεί να προσδιοριστεί. (μετασχηματισμένο απο [Daubin et al. 2002])

Οι σύγχρονες προσεγγίσεις της φυλογονιδιωματικής (αναλυτικά στο [Baurain & Philippe 2010]) αναδεικνύουν κάποιες ελλείψης στην ανάλυση της δημιουργίας του "πυρήνα των γονιδίων" ανεξάρτητα από το ποια μέθοδος θα χρησιμοποιηθεί. Οι πυ-

ρήνες των "φυλογενετικά πληροφοριακών" γονιδίων έχουν ένα μείγμα τόσο φυλογενετικού όσο και μη-φυλογενετικού σήματος (που μάλιστα είναι ενισχυμένο στο επίπεδο του γονιδιώματος). Ο βασικότερος λόγος ασυμβατότητας των δέντρων στη φυλογονιδιωματική είναι το artifact της προσέλκυσης των μακρών κλάδω των δέντρων (long brunch attraction) το οποίο φανερώνει το μείγμα φυλογενετικών και μη-φυλογενετικών στοιχειών (π.χ. ομοπλασία) τα οποία ενισχύονται όσο περισσότερα δεδομένα προσφέρουμε στις μεθόδους. Βασική η μείξη των δύο αυτών σημάτων μπορεί να επηρεάσει τόσο την ανακατασκευή των φυλογενειών με τέτοιο τρόπο ώστε πολλές φορές χρησιμοποιώντας ολόκληρα γονιδιώματα μπορεί να έχουμε ένα στατιστικά πολύ ισχυρά υποστηριζόμενο αλλά εξελικτικά εντελώς λανθασμένο δέντρο.

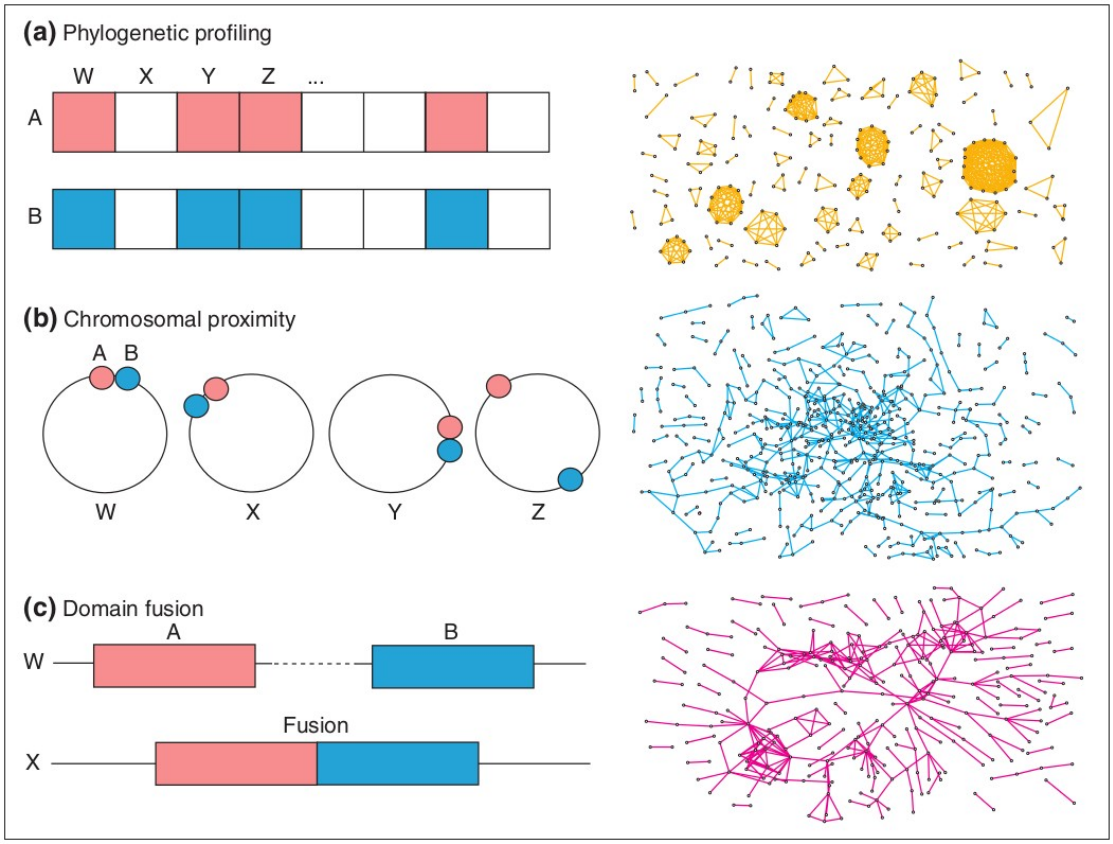
4.4 Μέθοδοι συγκριτικής γονιδιωματικής που δεν βασίζονται στην ομοιότητα μεταξύ ακολουθιών

4.4.1 Μέθοδοι λειτουργικού προσδιορισμού βιολογικών ακολουθιών.

Τρεις μέθοδοι συγκριτικής γονιδιωματικής έχουν αναπτυχθεί για την πρόβλεψη της λειτουργίας βιολογικών ακολουθιών και οι οποίες δεν στηρίζονται στην παραδοσιακή σύγκριση ομοιοτήτων μεταξύ ακολουθιών [Yanai & DeLisi 2002].

1. Η σύγκριση κοινών φυλογενετικών κατανομών
2. Η συντηρημένη μικρο-συνένια μεταξύ χρωμοσωμάτων και η σειρά οργάνωσης των γονιδίων
3. Τα γεγονότα σύντηξης γονιδίων σε πόλυ-τμηματικά (multi-domain) γονίδια.

Οι μέθοδοι αυτές είναι απλές στη σύλληψη τους αλλά καταφέρνουν σε συνδιασμό και οι τρεις να καλύψουν ένα πολύ μεγάλο ποσοστό (~70%) των προτείνων των πιο καλά μελετημένων βακτηρίων. Σχηματικά παρουσιάζονται στην εικόνα 2 και αναλυτικά ως εξής:



Είκονα 2: Μέθοδοι ανάλυσης συγκριτικής γονιδιωματικής που δεν βασίζονται σε ομοιότητα ακολουθιών. Α) Φυλογενετικό προφίλ Β) Χρωμοσωμική εγγύτητα C) Σύντηξη ενεργών κέντρων. Αναλυτική περιγραφή στο κείμενο. (τροποποιημένο από [Yanai & DeLisi 2002])

Φυλογενετικό προφίλ: Ο όρος προφίλ χρησιμοποιείται αναφορικά με την φυλογενετική κατανομή κατά μήκος των γονιδιωμάτων γονιδίων ή και μιας συλλογής (cluster) γονιδίων. Στην εικόνα 2Α για παράδειγμα δύο διαφορετικά γονίδια το ροζ και το μπλε απαντώνται διαδοχικά στους οργανισμούς W,Y και Z, ενώ λειπούν και τα δύο απο τον οργανισμό X. Με βάση αυτή τη φυλογενεντική κατανομή οι τέσσερις οργανισμοί στο παράδειγμα μας θα φτιάξουν ένα δίκτυο 4 κόμβων (αυθαίρετα προσδιορισμένων) όπου όλοι θα είναι συνδεδεμένοι μεταξύ τους (μιας και όλοι έχουν την ίδια κατανομή μεταξύ του μπλε και του ροζ γονιδίου), συνεπώς το αντίστοιχο δίκτυο που αναπαριστά τις σχέσεις μεταξύ γονιδίων με το ίδιο φυλογενετικό προφίλ θα έχει πολλά πλήρως συνδεδεμένα clusters (δεξί μέρος εικόνας 2Α).

Χρωμοσωμική εγγύτητα: Η μέθοδος αυτή στηρίζεται στην αρχή ότι η χρωσομομική γειτνίαση υπονοεί και λειτουργική γειτνίαση, με βάση αυτή την αρχή αναμένεται ότι γονίδια που έχουν παρόμοια λειτουργία τείνουν να βρίσκονται κοντά το ένα στο άλλο σε διάφορους οργανισμούς που συνδέονται με κοινή εξελικτική καταγωγή. Στο αριστερό μέρος της εικόνας 2 βλέπουμε ότι η εγγύτητα παρουσιάζεται ξεκάθαρα για τους οργανισμούς W,X και Y αλλά όχι για τον Z. Αντίστοιχα τα γενετικά δίκτυα γονιδιων που εμφανίζουν χρωμοσωμική εγγύτητα αναμένετε να αποτελούνται από αλυσιδωτές

δομές, πράγμα που αντικατοπτρίζει και την γονιδιωματική δομή τους (δεξί μέρος εικόνας 2B)

Σύντηξη ενεργών κέντρων: Η μέθοδοι αυτές βασίζονται στην παρατήρηση ότι διαφορετικά μή-ορθόλογα γονίδια τείνουν να σχετίζονται λειτουργικά και να αλληλεπιδρούν αν τα ορθόλογα τους παρουσιάζονται σαν ένα συντηγμένο γονίδιο σε έναν άλλο οργανισμό [Marcotte,Pellegrini,Ng,Rice,Yeates & Eisenberg 1999]. Το παράδειγμα της εικόνας 2C δείχνει ακριβώς αυτό το μηχανισμό σχηματικά και κατά συνέπεια τα αντίστοιχα γενετικά δίκτυα που προκύπτουν από την ανάλυση της σύντηξης γονιδίων παρουσιάζουν χαρακτηριστικά πολύπλοκων δικτύων με πολλούς κόμβους να έχουν λίγες συνδέσεις και λίγους να έχουν πολλές. Χαρακτηριστικό μοτίβο της κατανομής power-law (δεξί τμήμα εικονας 2C).

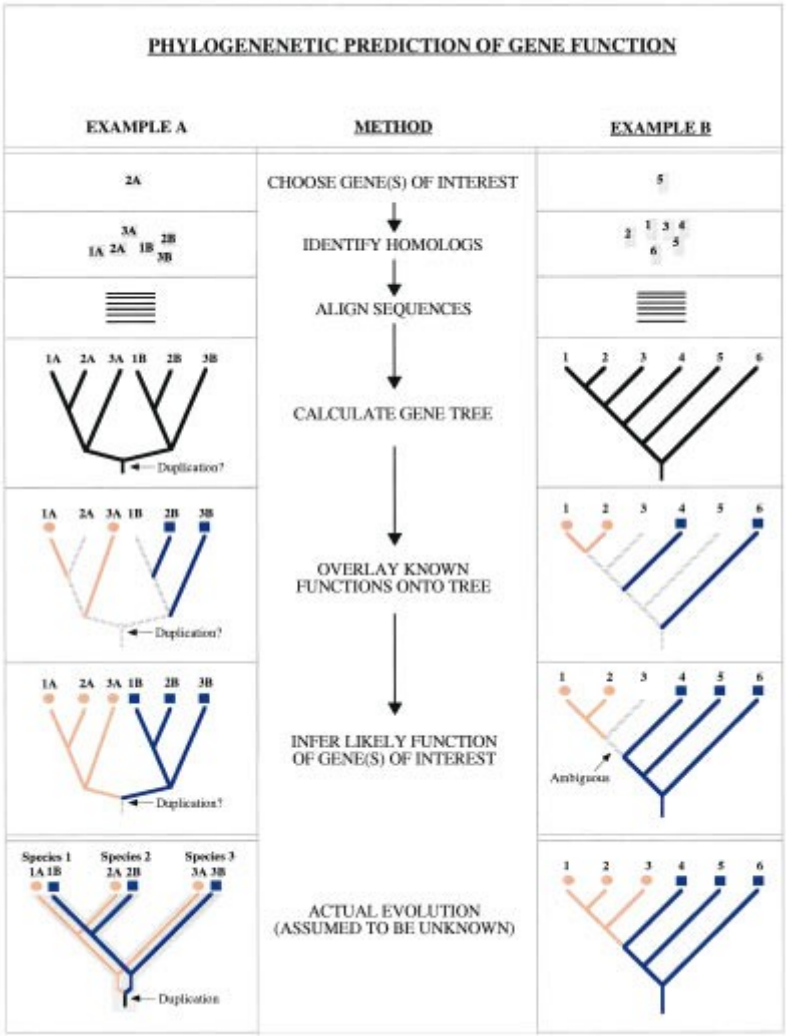
4.4.2 Μέθοδοι λειτουργικής φυλογονιδιωματικής

Μέθοδοι προσδιορισμού εξελικτικών σχέσεων και φυλογενετικών δέντρων με τη χρήση πολύ μεγάλων γονιδιωματικών περιοχών ή και ολόκληρων γονιδιωμάτων. Μολονότι είναι γενικά αποδεκτό ότι οι αλληλουχίες ολόκληρων γονιδιωμάτων είναι εξαιρετικά εργαλεία για εξελικτική ανάλυση, αποτελεί λιγότερο κοινό τόπο ότι και η εξελικτική ανάλυση αποτελεί εργαλείο για την καλύτερη μελέτη ολόκληρων των γονιδιωμάτων. Σαν φυλογονιδιωματική ορίζεται οι μέθοδοι και οι προσεγγίσεις εκείνες που χρησιμοποιούν εξελικτικά και φυλογενετικά εργαλεία προκειμένου να μελετήσουν ολόκληρα γονιδιώματα [Eisen & Fraser 2003]. Για παράδειγμα, δεν θα μπορούσε κάποιος να μελετήσει την πρόσληψη/χάσιμο γονιδίων σε γονιδιακές η πρωτεϊνικές οικογένειες ανάμεσα σε δύο είδη αν δεν είναι γνωστό το σύνολο του γονιδιώματος και για τα δύο αυτά είδη. Συνεπώς με την πληθώρα των ολοκληρωμένων γονιδίων οργανισμών που έχουμε σήμερα στη διάθεση μας οι φυλογονιδιωματικές μέθοδοι ανάλυσης μας εφοδιάζουν με τη δυνατότητα να: ολοκληρώνουμε μελέτες λειτουργικότητας των γονιδίων (εικόνα 1) χρησιμοποιώντας γονιδιωματικές αλληλουχίες, να βελτιώνουμε την ανάλυση των εξελικτικών σχέσεων μεταξύ των ειδών (και ειδικά ειδών που βρίσκονται κοντά σε σημαντικές εξελικτικές μεταβάσεις π.χ. ευ-προκαρυώτες) και να να ανακαλύπτουμε τις δομές που προσδιορίζουν τα γονιδιώματα.

Πιο συγκεκριμένα η φυλογονιδιωματική έχει χρησιμοποιηθεί εκτενώς για τον προσδιορισμό σε μεγάλη κλίμακα της λειτουργίας γονιδίων με μόνη πληροφορία την νουκλοτιδική (ή και πρωτεϊνική) τους ακολουθία. Το εγγονός αυτό είναι εξαιρετικά σημαντικό μιας και μας βοηθά στον λειτουργικό σχολιασμό γονιδιωμάτων οργανι-

σμών που είναι αδύνατο να τους καλλιεργήσουμε στο εργαστήριο και που περιλαμβάνουν σχεδόν το 95% των μικροοργανισμών που ζουν γύρω μας.

Μια ολοκληρωμένη μελέτη φυλογονιδιωματικής που μπορεί να προσδιορίσει λειτουργικές ιδιότητες γονιδίων παρουσιάζεται συνοπτικά στην εικόνα 1.



Είκονα 1: Περίγραμμα φυλογονιδιωματικών μεθοδολογίων (μετασχηματισμένο από [Eisen 1998]).

Σε αυτή την περίπτωση παρουσιάζονται σχηματικά δύο μονοπάτια που μπορούν να ακολουθηθούν για την επίτευξη της πρόβλεψης λειτουργίας άγνωστων γονιδίων από εξελικτικά δεδομένα. Στην εικόνα 1, παρατηρούμε την πορεία για την πρόβλεψη της λειτουργίας γονιδίων κάτω από δύο διαφορετικά εξελικτικά σενάρια: Α) Η γονιδιακή οικογένεια να έχει υποστεί κάποιο διπλασιασμό γονιδίων που ακολουθήθηκε από αποκλίνουσα λειτουργική πορεία των γονιδίων και Β) Η οικογένεια των μελετούμενων γονιδίων υπέστη κάποια μεταλλαγή σε μια και μόνη προγονική γραμμή (lineage).

Και στις δύο περιπτώσεις η μέθοδος της φυλογενετικής ανάλυσης μπορεί να καταφέρει να συμπεράνει τη λειτουργία του αγνώστου γονιδίου. Η φυλογονιδιωματική εξυπηρετεί και επιπλέον στόχους μιας ολοκληρωμένης συγκριτικής ανάλυσης γονιδιωμάτων, συμβάλει στο φυλογενετικό προσδιορισμό ορθόλογων/παράλογων ακολουθιών αλλά και η εξελικτική ανάλυση των προτύπων, των ρυθμών αλλά και του είδους της εξέλιξης των γονιδίων [Eisen & Wu 2002]. Οι αναλύσεις φυλογονιδιωματικής είναι απαιτητικές αναλύσεις, δεν είναι γρήγορες και απαιτούν εξειδικευμένους ερευνητές με μια ευρεία γκάμα γνώσεων, μολαταύτα υπάρχουν προσπάθειες κατά τις οποίες οι ερευνητές μετασχηματίζουν την γνώση και την εξειδίκευση τους σε κάποιο υπολογιστικό εργαλείο, κάνοντας έτσι πιο προσβάσιμη την φυλογονιδιωματική ανάλυση σε εργαστήρια και ομάδες που τους λείπει η ανάλογη ειδίκευση πιο προσιτή. Ένα εξαίρετο παραδειγμα αποτελεί το πρόγραμμα AMPHORA (a pipeline for Auto-Mated PhylogenOmic infeRence) που δοκιμάστηκε και επιτυχημένα προσέδωσε εντελώς αυτοματοποιημένα λειτουργικό σχολιασμό σε πολλά ORFs απο το τεράστιο πρόγραμμα μεταγονιδιωματικής αλληλούχησης των μικροοργανισμών της θάλασσας των Σαργασσών [Eisen & Wu 2002].

4.5 Συγκριτικές μέθοδοι που προσδιορίζουν λειτουργικές ακολουθίες.

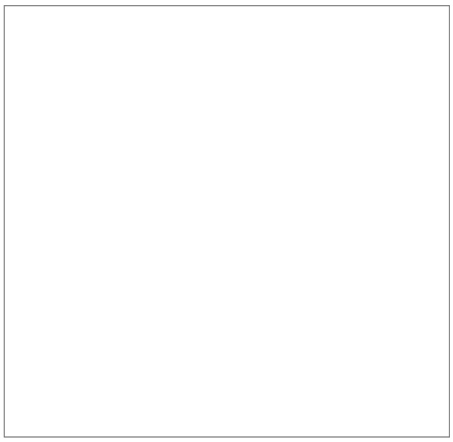
Μπορεί οι μέθοδοι που διαχωρίζουν ορθόλογες από παράλογες πρωτεϊνικές ή γονιδιακές ακολουθίες να αποτελούν ισχυρά εργαλεία για την πρόβλεψη λειτουργικών ρόλων αυτών των ακολουθιών αλλά η σύγχρονη συγκριτική γονιδιωματική έχει αναπτύξει σειρά μεθόδων που στοχεύουν συγκεκριμένα στο να προσδιορίσουν λειτουργικούς ρόλους των ακολουθιών. Στο παρακάτω κομμάτι θα παρουσιαστούν και αξιολογηθούν αυτές οι μέθοδοι που στοχεύουν στον προσδιορισμό λειτουργικών ρόλων.

4.5.1 Η μέθοδος Rosetta Stone.

Η μέθοδος αυτή (ρεφ...) αναγνωρίζει μεμονωμένες πρωτεΐνες που είναι προϊόντα σύντηξης δύο πρωτεϊνών που απαντώνται σε διαφορετικούς οργανισμούς.

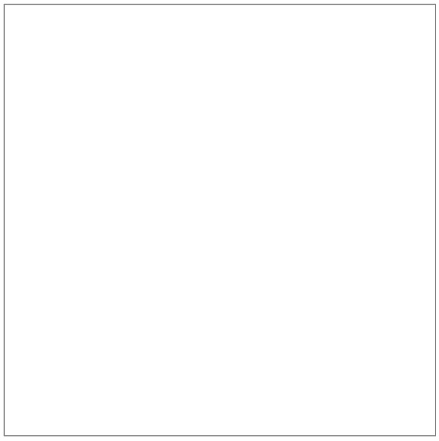
4.5.2 Η μέθοδος του φυλογενετικού προφίλ.

Η μέθοδος του φυλογενετικού προφίλ (ρεφ...) αναγνωρίζει πρωτεΐνες που απαντώνται παράλληλα ανάμεσα σε διάφορα γονιδιώματα. Η μέθοδος αυτή χρησιμοποιείται με επιτυχία και για την πρόβλεψη ρυθμιστικών περιοχών γονιδίων (ρεφ.).



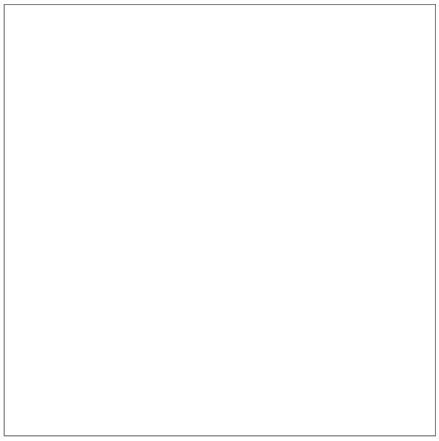
4.5.3 Η μέθοδος της συντηρημένης μικρο-συντενίας.

Η μέθοδος αυτή αναγνωρίζει γονίδια (αλλά μπορεί να χρησιμοποιηθεί και για τις πρωτεΐνες που κωδικοποιούν) που εντοπίζονται σε κοντινά χρωμοσωμικά τμήματα και μάλιστα λαμβάνει υπόψιν της και την σειρά με την οποία τα γονίδια απαντώνται στα διάφορα γονιδιώματα.



4.5.4 Η μέθοδος του οπερονίου.

Αυτή μέθοδος προσδιορίζει γονίδια που είναι πιθανόν να ανήκουν στο ίδιο οπερόνιο βασιζόμενη στην απόσταση μεταξύ παρακείμενων γονιδίων που βρίσκονται στην ίδια κατεύθυνση σε διάφορα γονιδιώματα.



4.6 Υπολογιστικές μέθοδοι συγκριτικής ανάλυσης.

Στο τμήμα αυτό θα παρουσιαστούν αναλυτικά μέθοδοι συγκριτικής ανάλυσης γονιδιωμάτων που στηρίζονται σε τεχνικές από τα διακριτά μαθηματικά αλλά και τη θεωρία υπολογιστών. Το υπόβαθρο του τμήματος αυτού είναι καθαρά αλγοριθμικό και δεν θα παρουσιαστούν εργαλεία έτοιμα προς χρήση από τους βιολόγους αλλά μια κριτική παρουσίαση των υπολογιστικών και αλγοριθμικών μεθόδων ανάλυσης των προβλημάτων στην συγκριτική γονιδιωματική.

4.7 Υπολογιστικά και αλγοριθμικά προβλήματα στην συγκριτική γονιδιωματική.

4.7.1 BLAST και BLASTphemy

Μια από τις πιο σύνηθες αλλά και πιο επικίνδυνες συγχύσεις που παράλληλα αποτελεί και το νο1 υπολογιστικό πρόβλημα σε κάθε μέθοδο ανάλυσης συγκριτικής γονιδιωματικής αποτελεί η εκτέλεση, εξιολόγηση και ερμηνία των αποτελεσμάτων των προγραμμάτων στοίχισης των ακολουθιών και δη των πιο διαδεδομένων από αυτά, των προγραμμάτων που ανήκουν στην σουίτα του NCBI BLAST. Μια πληθώρα μαθηματικών και στατιστικών περιοριστικών παραγόντων επηρεάζουν τα αποτελέσματα του BLAST και την ευαισθησία (sensitivity) σε σχέση με την εκλεκτικότητα (selectivity) του αλγόριθμου στοίχισης [Pertsemlidis & Fondon 2001]. όταν η ομοιότητα (που υπολογίζεται από το BLAST) προσπαθεί να μεταφραστεί σε βιολογική ομολογία μεταξύ ακολουθιών τότε προβλήαματ που έχουν να κάνουν με τους ευριστικούς αλγόριθμους που χρησιμοποιοεί η σουίτα BLAST, με το σχήμα σκοραρίσματος που επιλέγεται καθώς και με τις μήτρες υποκατάστασης (BLOSUM, PAM κλπ.) μπορεί να επιρρεάσουν πολύ ριζικά τα αποτελέσματα της αναζήτησης, τη στατιστική τους σημαντικότητα και άρα την βιολογική τους αξιολόγιση.

5 Τρίτο Μέρος

5.1 Σχεδιάζοντας πειράματα ελέγχου σε αναλύσεις συγκριτικής γονιδιωματικής.

Σε όλων των ειδών τις αναλύσεις συγκριτικής γονιδιωματικής υπάρχει η ανάγκη σχεδιασμού ενός πειράματος υπόβαθρου (background) ή ενός πειράματος αρνητικού κοντρόλ (negative control). Το πείραμα ελέγχου πρέπει να αποτελείτε από ακολουθίες βιομορίων (DNA/RNA, πρωτεΐνες) που θα έχουν κατασκευαστεί από μια τυχαία διαδικασία (μέσω ενός γεννήτορα τυχαίων αριθμών, σειρών κλπ.) αλλά που διατηρούν κάποια συστατικά χαρακτηριστικά των "αληθινών" ακολουθιών. Επιπλέον ο σχεδιασμός της διαδικασίας δημιουργίας του πειράματος υπόβαθρου πρέπει όσο το δυνατόν να προσομοιάζει την διαδικασία του πειράματος που μελετάται. Για παράδειγμα πολ-

λά προγράμματα υπολογιστικής εύρεσης ακολουθιών γονιδίων, χρησιμοποιούν εξελιγμένα τυχαιοποιημένα μοντέλα τεχνικά κατασκευασμένων γονιδίων για να "εκπαιδεύσουν", να τεστάρουν και να αξιολογήσουν τις υπολογιστικές μεθόδους εύρεσης γονιδίων [Picardi & Pesole 2010]. Ο σχεδιασμός και η εκτέλεση πειραμάτων τυχαιοποίησης λοιπόν είναι κομβικής σημασίας για τις μεθόδους της συγκριτικής γονιδιωματικής.

Παρακάτω θα περιγράψουμε μια προσέγγιση που ακολουθήθηκε κατά τη διάρκεια της ανάπτυξης μιας υπολογιστικής μεθόδου για την συγκριτική εύρεση στόχων μικρών μορίων RNA microRNA στο ποντίκι και στον άνθρωπο. Η μελέτη έχει δημοσιευτεί [Kiriakidou et al. 2004] και η ανάλυση έχει ως εξής.

Τα microRNA είναι μικρά μόρια RNA (21-23 νουκλοετίδια) που προέρχονται από μακρύτερες ακολουθίες RNA οι οποίες μεταγράφονται σαν πρόδρομα μόρια RNA. Κάθε microRNA δρα μέτα-μεταγραφικά πάνω σε μόρια mRNA. Με την πρόσδεση του microRNA στο 3'-UTR του mRNA δημιουργείτε ένα σύμπλοκο διπλής αλυσίδας mRNA-microRNA το οποίο αναγνωρίζεται από RNAάσες και τεμαχίζεται οδηγώντας στην απενεργοποίηση του μορίου του mRNA. Τα microRNA δρουν σαν αναστολείς της έκφρασης των γονιδίων με το να ελέγχουν τα επίπεδα του mRNA στο κυτταρόπλασμα.

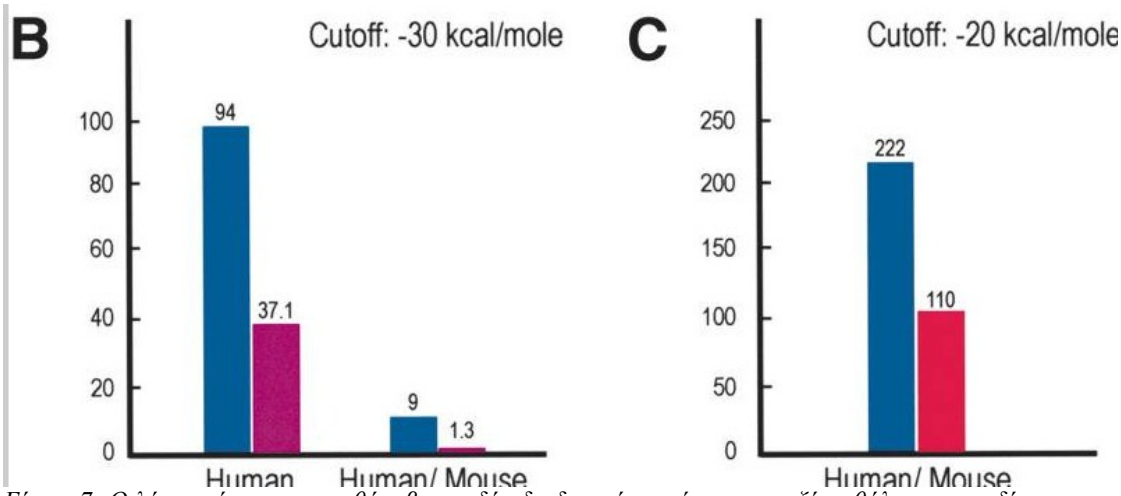
Το πρόβλημα της υπολογιστικής πρόβλεψης μορίων mRNA που αποτελούν στόχους για μόρια microRNA είναι σημαντικό καθώς πολλά microRNA έχουν βρεθεί να εμπλέκονται στη δημιουργία καρκίνων καθώς και στον αναπτυξιακό έλεγχο των βλαστοκυττάρων. Το πρόβλημα προσεγγίστηκε χρησιμοποιώντας βασικές αρχές συγκριτικής γονιδιωματικής και είχε ως εξής.

1. Πρόβλεψη ενός μεγάλου αριθμού υποθετικών στόχων mRNA για κάθε microRNA από τον άνθρωπο και από το κοντινότερο (τότε) πλήρως αποκωδικοποιημένο θηλαστικό το ποντίκι. Η πρόβλεψη των "πιθανών στόχων" έγινε με την ανάπτυξη μιας context free grammar που αναζητούμε συμπληρωματικότητα μεταξύ του microRNA και του συνόλου του πληθυσμού mRNA στον άνθρωπο και στο ποντίκι. Η διαδικασία αυτή έδωσε 2 πολύ μακριές λίστες με τους πιθανούς στόχους microRNA μια για τον άνθρωπο και μια για το ποντίκι.
2. Συγκριτική γονιδιωματική: Για κάθε στόχο ενός microRNA από τον άνθρωπο πάνω σε ένα mRNA ανθρώπινου γονιδίου αναζητήθηκε αν το ορθόλογο microRNA από το ποντίκι έχει σαν στόχο το mRNA του ορθόλογου γονιδίου

στο ποντίκι. Αν ΚΑΙ οι δύο αυτές συνθήκες εκπληρούνται τότε το mRNA-microRNA ζευγάρι από τον άνθρωπο έμπαινε στη λίστα των πιθανών στόχων.

3. Το πείραμα υποβάθρου είχε σχεδιαστεί ως εξής:
    1. Μέτρηση της συχνότητας των δίνουκλεοτιδίων στα αναγνωρισμένα ανθρώπινα microRNA από την (MirBase [Griffiths-Jones 2004]) και δημιουργία ενός πίνακα δινουκλεοτιδικών συχνοτήτων.
    2. Υπολογισμός ενός δινουκλεοτιδικού σκορ για κάθε πραγματικό microRNA.
    3. Δημιουργία με τυχαίο ανακάτεμα των υπάρχοντων βάσεων σε κάθε microRNA ενός πληθυσμού 100 τυχαίων microRNA ακολουθιών για κάθε ένα ανθρώπινο microRNA.
    4. Υπολογισμός του δινουκλεοτιδικού σκορ για όλα τα τυχαία κατασκευασμένα microRNA.
    5. Επιλογή των 5 καλύτερων σκορ από τα 100 τυχαία microRNAs. Τα 5 σκορ που ήταν πιο κοντά στο δινουκλεοτιδικό σκορ του πραγματικού microRNA συλλέχθηκαν για το επόμενο στάδιο του πειράματος.
4. Για κάθε ένα από τα 5 επιλεγμένα microRNA του πειράματος κοντρολ, όλη η διαδικασία των βημάτων 1 και 2 παραπάνω ακολουθήθηκε ξανά με ακριβώς τις ίδιες παραμέτρους εύρεσης γονιδίων στόχων.
5. Με βάση τους ομόλογους στόχους που το σύστημα προέβλεπε στο βήμα 4, ο μέσος όρος των "σωστών προβλέψεων" των τυχαιοποιημένων ακολουθιών microRNA υπολογίζονταν. Διαιρόντας το αριθμό των στόχων του πραγματικού microRNA με τον μέσο όρων των τυχαιοποιημένων υπολογίζονταν ο λόγος σήματος προς θόρυβο (signal to noise ratio).
6. Η διαδικασία επαναλαμβάνονταν με τα βήματα 1,2,4 και 5 για πολλές διαφορετικές παραμέτρους εύρεσης γονιδίων στόχων με σκοπό την μέγιστη δυνατή έκφραση του λόγου σήματος προς θόρυβο.

Η εικόνα 7 παρουσιάζει τα αποτελέσματα μετά και τη λήξη του βήματος 6 και το γράφημα με το λόγο σήματος προς θόρυβο για την υπολογιστική πρόβλεψη γονιδίων στόχων των ανθρώπινων microRNA.



Είκονα7: Ο λόγος σήματος προς θόρυβο για δύο διαδοχικά πειράματα μεταξύ ορθόλογων γονιδίων στόχων αθρώπου και ποντικιού.

Το πείραμα όπως περιγράφτηκε ήταν μια από τις πρώτες επιτυχημένες μεθόδους εύρεσης γονιδίων στόχων των microRNA στον άνθρωπο και το ποντίκι και αποτέλεσε μια πολύ επιτυχημένη μέθοδο εύρεσης στόχων που αργότερα συνετέλεσε και στη δημιουργία βάσεων δεδομένων στόχων των microRNA. Είναι ένα κλασικό παράδειγμα του πως μια καλά σχεδιασμένη μέθοδος συγκριτικής γονιδιωματικής παράλληλα με ένα στιβαρό πείραμα υποβάθρου μπορεί να συμβάλει στη δημιουργία μιας μεθόδου που να λύνει ένα δύσκολο όσο και χρήσιμο πρόβλημα.

5.2 Χρήση ροής ανάλυσης (pipeline) συγκριτικής γονιδιωματικής στη χαρτογράφηση.

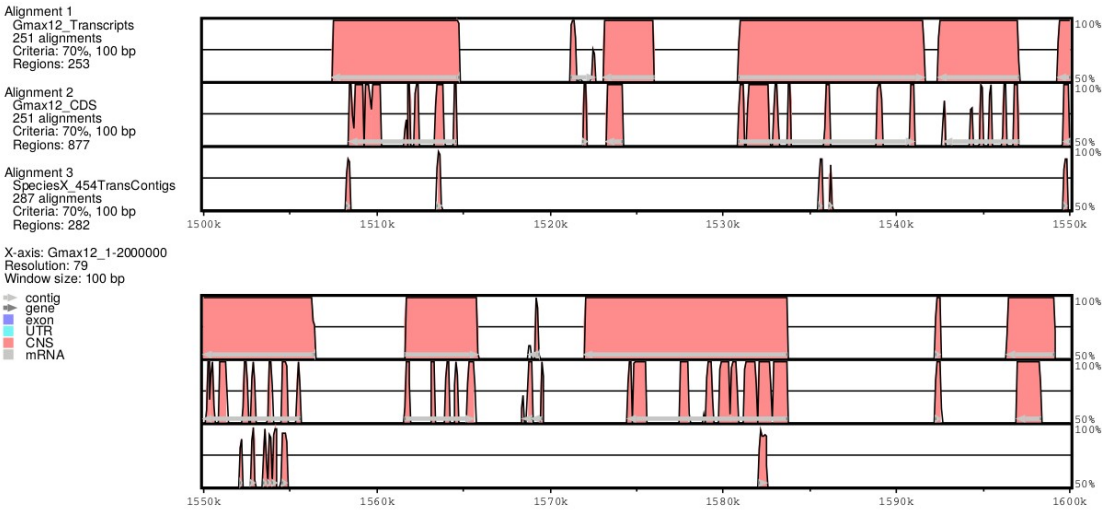
Σε αυτό το κομμάτι θα παρουσιαστεί μια ολοκληρωμένη προσέγγιση σε ένα πρόβλημα γονιδιακής χαρτογράφησης χαρακτήρων με υψηλό αγροτικά ενδιαφερόν σε φυτά με άγνωστο μη αναγνωρισμένο γονιδίωμα. Πρόκειται για μια εφαρμογή μεθόδων κλασσικής γενετικής χαρτογράφησης (λ.χ ανάλυση ανασυνδιασμών) η οποία όμως έχει συνδεθεί και δουλεύει παράλληλα με υπολογιστικές μεθόδους βιοπληροφορικής και συγκριτικής γονιδιωματικής ανάλυσης. Η μελέτη είχε σαν στόχο την εύρεση της γενετικής βάσης (ένα ή περισσότερα γονίδια της οικογένειας NB-LRRs Nucletide-Binding-Leucine-Reach-Repeats) της ανθεκτικότητας σε μια μορφή σκωριάσης της σόγιας ενός ενός φυτού άγριου μέλους της οικογένειας Fabace. Τα προϊόντα των NB-LLR γονιδίων συμμετέχουν ενεργά στην αναγνώριση και τη διαμεταγωγή σήματος για απόπτωση σε περίπτωση προσβολής από το παθογόνο. Το φυτό ήταν ένας άγριος συγγενής της σόγιας αυτοφυές μόνο στην Νότια Αμερική, με γονιδίωμα λίγο μικρότερο από τη σόγια και εντελώς άγνωστη γενετική ακολουθία. Η μελέτη είχε ξεκινήσει με την αναγνώριση μέσω μενός πρώτου περιορισμένο αριθμού ανασυνδιασμών μιας

περιοχής κοντά στο τελομερές του χρωμοσώματος 6 του φυτού. Τα δεδομένα που ήταν διαθέσιμα ήταν οι γονιδιωματικές ακολουθίες της σόγιας (*Glycine max*) και της μηδικής (*Medicago truncatula*) καθώς και μια αποκωδικοποιημένη βιβλιοθήκη του συνολικού RNA του άγριου φυτού αλλά και μια BAC βιβλιοθήκης μια καλλιεργούμενης ποικιλίας του άγριου φυτού.

Η πρώτη ενέργεια ήταν η δημιουργία μιας consensus συναρμολόγησης (assembly) των ακολουθιών από την NGS ανάλυση του μεταγραφώματος του άγριου φυτού. Η τεχνική αυτή εφαρμόζεται συχνά για την συναρμολόγηση του μεταγραφώματος στην περίπτωση που υπάρχει έλλειψη της γονιδιωματικής ακολουθίας του οργανισμού [Moreton,Dunham & Emes 2014]. Το consensus assembly συνδυάζει αρχές της συγκριτικής γονιδιωματικής και της βιοπληροφορικής καθώς προκειμένου να προχωρήσει στην συναρμολόγηση των ακολουθιών ενός άγνωστου γονιδιώματος από τις μικρού μήκους ακολουθίες του NGS εκτελεί ένα βήμα πιο πριν μια γρήγορη στοίχιση των NGS δεδομένων με ένα γονιδίωμα που είναι ήδη γνωστό και είναι συγγενές με το γονιδίωμα του οργανισμού προς μελέτη. Έτσι συνδυάζεται η δύναμη της βιοπληροφορικής να συνθέτει de-novo γονιδιώματα ή μεταγραφώματα με την αναλυτική ευαισθησία απου προσφέρει η συγκριτική γονιδιωματική. Με την στρατηγική στοίχισης και στη συνέχεια de-novo συναρμολόγησης έχουμε πρόσβαση σε μια υψηλής ποιότητας συναρμολόγησης του μεταγρσαφώματος ακόμα και με πλήρη απουσία της ακολουθίας του γονιδιώματος.

Στην περίπτωση μας χρησιμοποιήθηκε το πλήρες γονιδίωμα της σόγιας *G. max,* πάνω στο οποίο έγινε μια γρήγορη και απλή στοίχιση των δεδομένων NGS από το άγριο φυτό χρησιμοποιώντας τις βασικές παραμέτρους του προγράμματος BWA [Langmead,Trapnell,Pop & Salzberg 2009]. Στη συνέχεια οι ακολουθείες που είχαν στοιχηθεί με τη σόγια συναρμολογήθηκαν χρησιμοποιώντας τις βασικέ ςπαρμέτρους των προγραμμάτων oases και velvet για την εκ-νέου συναρμολόγηση ακολουθιών μεταγραφώματος [Zerbino & Birney 2008].

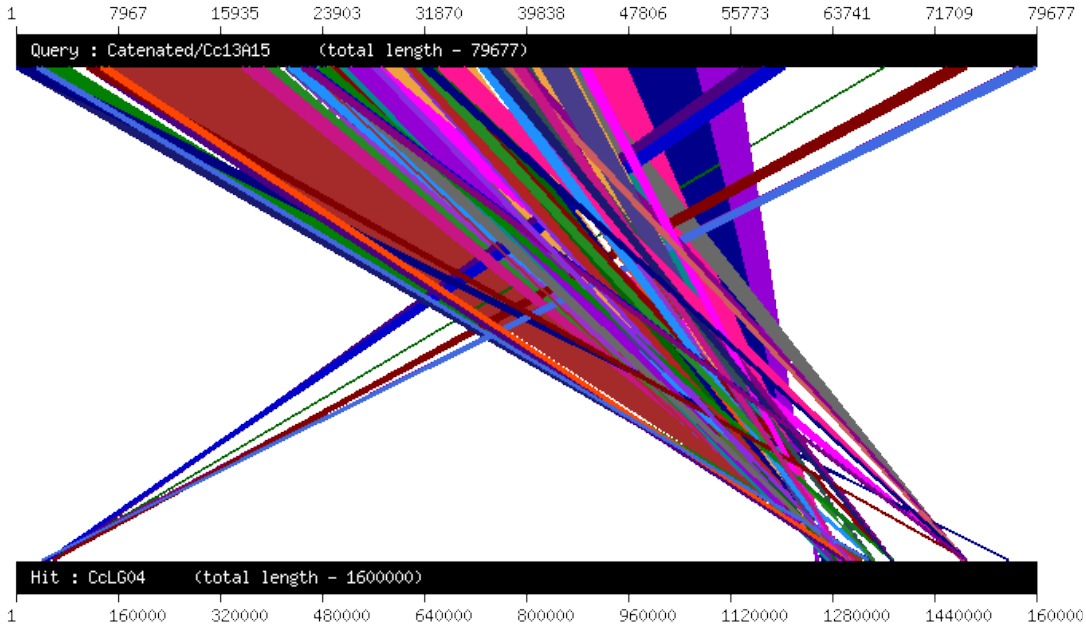
Στη συνέχεια για την αναγνώριση των εκφραζόμενων περιοχών που μπορούν να χαρτογραφηθούν στο άγριο φυτό ακολουθήθηκε η εξής στρατηγική. Οι ακολουθίες από την συναρμολόγηση του μεταγραφώματος χρησιμοποιήθηκαν στο πρόγραμμα εύρεσης συντενιάς, στοίχισης και ομοιότητας VISTA μαζί με ακολουθίες από το χρωμόσωμα 12 της σόγιας εικόνα 8.



Εικόνα 8: Η πολλαπλή συγκριτική ανάλυση του μεταγραφώματος άγριας ποικιλίας συγγενούς της σόγιας με το κομμάτι κοντά στο τελομερές του χρωμοσώματος της σόγιας.

Από την πολλαπλή συγκριτική ανάλυση έγινε δυνατό να σχεδιαστούν εκκινητές που με τη σειρά τους χρησιμοποιήθηκαν για να αλλιευθούν κλώνοι BAC από τη την βιβλιοθήκη του αγρίου φυτού και να στοιχηθούν πάλι πίσω με τις de-novo ακολουθίες του μεταγραφώματος.

Στην εικόνα 9 παρουσιάζεται η χαρτογράφηση και οι σχέσεις συντένιας ενός κομματιού BAC κλώνου του άγριου φυτού με ένα από τα contigs του μεταγραφώματος από την άγρια ποικιλία που παρουσιάζει ανθεκτικότητα στη σκωρίαση, απο την σύγκριση αυτή έγινε δυνατός ο σχεδιασμός ενός χρωμοσωμικού γεντικού δείκτη που βοήθησε στο να αυξηθεί η ανάλυση του γενετικού χάρτη και να μικρύνει σημαντικά η περιοχή αναζήτησης του γονιδίου ανθεκτικότητας.



Εικόνα 9: Διάγραμμα συντενίας ενός BAC κλώνου του καλλιεργούμενου φυτού με ένα contig απο το μεταγράφωμα του άγριου ανθεκτικού φυτού που συνετέλεσε στην βελτίωση του γενετικού χάρτη γύρω από το γονίδιο ανθεκτικότητας.

Ολοκλήρωση όλων των ανωτέρω σε μια ολοκληρωμένη προσέγγιση και ανάλυση οδήγησε στον προσδιορισμό ενός δύσκολα χαρτογραφούμενου χαρακτήρα (NB-LRR πρωτεΐνη) σε φυτό του οποίου το γονιδίωμα δεν έχει ακόμα αναγνωριστεί.

6 Συμπεράσματα

6.1 Η συγκριτική γονιδιωματική στην μετά-γονιδιωματική εποχή.

Με την έλευση της μέτα-γονιδιωματικής εποχής (post-genomics era) γη συγκριτική γονιδιωματική αποκτά ακόμα μεγαλύτερο ενδιαφέρον και πεδίο ανάπτυξης και προσφοράς. Ο αριθμός των αποκωδικοποιημένω γονιδιωμάτων αυξάνεται με πολύ γρήγορους ρυθμούς αλλά ακόμα και μέσα σε συγκεκριμένα είδη όπως π.χ. στον άνθρωπο ή στα καλλιεργούμενα φυτά οι νέες τεχνολογίες αλληλούχησης παράγουν δεδομένα με πρωτοφανείς ρυθμούς, και οι προκλήσεις είναι επίσης πρωτοφανείς. Τα προβλήματα και οι περιορισμοί που υπήρχαν στο παρελθόν φαίνεται να μειώνονται αλλά οι καινούργιες προκλήσεις όσον αφορά τη διαχείριση, την υπολογιστική ανάλυση αλλά και την εξαγωγή γνώσης από όλο αύτο τον τεράστιο όγκο δεδομένων είναι πολύ πιο απαιτητικές. Οι εποχές για την εξελικτική και συγκριτική γονιδιωματική είναι ακόμα πιο ενδιαφέρουσες.

6.1.1 Η πρωτοβουλία μοντελοποίησης του iPlant

Το iPlant consortium [Goff et al. 2011] αποτελεί μια πρωτοβουλία συγκέντρωσης και διαχείρισης του τεράστιου όγκου δεδομένων που έχει προκύψει από τις σύγχρονες μεθόδους αλληλούχησης και τα πειράματα υψηλής απόδοσης και παράλληλα μοντελοποίησης πολλών περιβαλλοντικών και εξελικτικών παραμέτρων που έχουν να κάνουν με την προσαρμοστικότητα που έχουν αναπτύξει τα καλλιεργούμενα φυτά στα διάφο-

ρα περιβάλλοντα και ασθένειες. Οι δύο μεγάλες προκλήσεις του προαγράμματος iPlant είναι το δέντρο της ζωής iPlant (iPlant Tree of Life, iPToL) και το πρόγραμμα από γονότυπο-σε-φαινότυπο (iPlant genotype-to-phenotype, iPG2P) τα οποία και τα δύο κάνουν εκτενέστατη χρήση εργαλείων και μεθόδων της συγκριτικής γονιδιωματικής, τόσο για την στοίχιση ακολουθιών και την ανακατασκευή εξελικτικών σχέσεων όσο και για το λειτουργικό σχολιασμό και πρόβλεψη του τεράστιου αριθμού νέων ακολουθιών και πολυμορφισμών που προστίθενται κάθε μέρα στην μεγάλη αυτή διαδικτυακή υποδομή που έχει αναπτύξει το iPlant.

7 Παραρτήματα

Appendix I: Heading of this appendix

8 Βιβλιογραφία

Baurain, D. and Philippe, H. (**2010**). *2*. In: (Ed.), *Current Approaches to Phylogenomic Reconstruction*, John Wiley & Sons, Inc..

Chaudhuri, R. R. and Pallen, M. J. (**2006**). *xBASE, a collection of online databases for bacterial comparative genomics.*, Nucleic Acids Research 34 : D335-D337.

Cliften, P. F.; Hillier, L. W.; Fulton, L.; Graves, T.; Miner, T.; Gish, W. R.; Waterston, R. H. and Johnston, M. (**2001**). *Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis.*, Genome Research 11 : 1175-1186.

Consortium, I. M. G. S.; Waterston, R. H.; Lindblad-Toh, K.; Birney, E.; Rogers, J.; Abril, J. F.; Agarwal, P.; Agarwala, R.; Ainscough, R.; Alexandersson, M.; An, P.; Antonarakis, S. E.; Attwood, J.; Baertsch, R.; Bailey, J.; Barlow, K.; Beck, S.; Berry, E.; Birren, B.; Bloom, T.; Bork, P.; Botcherby, M.; Bray, N.; Brent, M. R.; Brown, D. G.; Brown, S. D.; Bult, C.; Burton, J.; Butler, J.; Campbell, R. D.; Carninci, P.; Cawley, S.; Chiaromonte, F.; Chinwalla, A. T.; Church, D. M.; Clamp, M.; Clee, C.; Collins, F. S.; Cook, L. L.; Copley, R. R.; Coulson, A.; Couronne, O.; Cuff, J.; Curwen, V.; Cutts, T.; Daly, M.; David, R.; Davies, J.; Delehaunty, K. D.; Deri, J.; Dermitzakis, E. T.; Dewey, C.; Dickens, N. J.; Diekhans, M.; Dodge, S.; Dubchak, I.; Dunn, D. M.; Eddy, S. R.; Elnitski, L.; Emes, R. D.; Eswara, P.; Eyras, E.; Felsenfeld, A.; Fewell, G. A.; Flicek, P.; Foley, K.; Frankel, W. N.; Fulton, L. A.; Fulton, R. S.; Furey, T. S.; Gage, D.; Gibbs, R. A.; Glusman, G.; Gnerre, S.; Goldman, N.; Goodstadt, L.; Grafham, D.; Graves, T. A.; Green, E. D.; Gregory, S.; Guigó, R.; Guyer, M.; Hardison, R. C.; Haussler, D.; Hayashizaki, Y.; Hillier, L. W.; Hinrichs, A.; Hlavina, W.; Holzer, T.; Hsu, F.; Hua, A.; Hubbard, T.; Hunt, A.; Jackson, I.; Jaffe, D. B.; Johnson, L. S.; Jones, M.; Jones, T. A.; Joy, A.; Kamal, M.; Karlsson, E. K.; Karolchik, D.; Kasprzyk, A.; Kawai, J.; Keibler, E.; Kells, C.; Kent, W. J.; Kirby, A.; Kolbe, D. L.; Korf, I.; Kucherlapati, R. S.; Kulbokas, E. J.; Kulp, D.; Landers, T.; Leger, J. P.; Leonard, S.; Letunic, I.; Levine, R.; Li, J.; Li, M.; Lloyd, C.; Lucas, S.; Ma, B.; Maglott, D. R.; Mardis, E. R.; Matthews, L.; Mauceli, E.; Mayer, J. H.; McCarthy, M.; McCombie, W. R.; McLaren, S.; McLay, K.; McPherson, J. D.; Meldrim, J.; Meredith, B.; Mesirov, J. P.; Miller, W.; Miner, T. L.; Mongin, E.; Montgomery, K. T.; Morgan, M.; Mott, R.; Mullikin, J. C.; Muzny, D. M.; Nash, W. E.; Nelson, J. O.; Nhan, M. N.; Nicol, R.; Ning, Z.; Nusbaum, C.; O'Connor, M. J.; Okazaki, Y.; Oliver, K.; Overton-Larty, E.; Pachter, L.; Parra, G.; Pepin, K. H.; Peterson, J.; Pevzner, P.; Plumb, R.; Pohl, C. S.; Poliakov, A.; Ponce, T. C.; Ponting, C. P.; Potter, S.; Quail, M.; Reymond, A.; Roe, B. A.; Roskin, K. M.; Rubin, E. M.; Rust, A. G.; Santos, R.; Sapojnikov, V.; Schultz, B.; Schultz, Jö.; Schwartz, M. S.; Schwartz, S.; Scott, C.; Seaman, S.; Searle, S.; Sharpe, T.; Sheridan, A.; Shownkeen, R.; Sims, S.; Singer, J. B.; Slater, G.; Smit, A.; Smith, D. R.; Spencer, B.; Stabenau, A.; Stange-Thomann, N.; Sugnet, C.; Suyama, M.; Tesler, G.; Thompson, J.; Torrents, D.; Trevaskis, E.; Tromp, J.; Ucla, C.; Ureta-Vidal, A.; Vinson, J. P.; Von Niederhausern, A. C.; Wade, C. M.; Wall, M.; Weber, R. J.; Weiss, R. B.; Wendl, M. C.; West, A. P.; Wetterstrand, K.; Wheeler, R.; Whelan, S.; Wierzbowski, J.; Willey, D.; Williams, S.; Wilson, R. K.; Winter, E.; Worley, K. C.; Wyman, D.; Yang, S.; Yang, S.-P.; Zdobnov, E. M.; Zody, M. C. and

Lander, E. S. (**2002**). *Initial sequencing and comparative analysis of the mouse genome.*, Nature 420 : 520-562.

Cui, L.; Wall, P. K.; Leebens-Mack, J. H.; Lindsay, B. G.; Soltis, D. E.; Doyle, J. J.; Soltis, P. S.; Carlson, J. E.; Arumuganathan, K.; Barakat, A.; Albert, V. A.; Ma, H. and dePamphilis null, C. W. (**2006**). *Widespread genome duplications throughout the history of flowering plants.*, Genome Research 16 : 738-749.

Daubin, V.; Gouy, M. and Perrière, G. (**2002**). *A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history.*, Genome Research 12 : 1080-1090.

Duvick, J.; Fu, A.; Muppirala, U.; Sabharwal, M.; Wilkerson, M. D.; Lawrence, C. J.; Lushbough, C. and Brendel, V. (**2008**). *PlantGDB: a resource for comparative plant genomics.*, Nucleic Acids Research 36 : D959-D965.

Eddy, S. R. (**2005**). *A model of the statistical power of comparative genome sequence analysis.*, PLoS Biology 3 : e10.

Eisen, J. A. (**1998**). *Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis.*, Genome Research 8 : 163-167.

Eisen, J. A. and Fraser, C. M. (**2003**). *Phylogenomics: intersection of evolution and genomics.*, Science 300 : 1706-1707.

Eisen, J. A. and Wu, M. (**2002**). *Phylogenetic analysis and gene functional predictions: phylogenomics in action.*, Theoretical Population Biology 61 : 481-487.

Gilbert, W. (**1991**). *Towards a paradigm shift in biology.*, Nature 349 : 99.

Goff, S. A.; Vaughn, M.; McKay, S.; Lyons, E.; Stapleton, A. E.; Gessler, D.; Matasci, N.; Wang, L.; Hanlon, M.; Lenards, A.; Muir, A.; Merchant, N.; Lowry, S.; Mock, S.; Helmke, M.; Kubach, A.; Narro, M.; Hopkins, N.; Micklos, D.; Hilgert, U.; Gonzales, M.; Jordan, C.; Skidmore, E.; Dooley, R.; Cazes, J.; McLay, R.; Lu, Z.; Pasternak, S.; Koesterke, L.; Piel, W. H.; Grene, R.; Noutsos, C.; Gendler, K.; Feng, X.; Tang, C.; Lent, M.; Kim, S.-J.; Kvilekval, K.; Manjunath, B. S.; Tannen, V.; Stamatakis, A.; Sanderson, M.; Welch, S. M.; Cranston, K. A.; Soltis, P.; Soltis, D.; O'Meara, B.; Ane, C.; Brutnell, T.; Kleibenstein, D. J.; White, J. W.; Leebens-Mack, J.; Donoghue, M. J.; Spalding, E. P.; Vision, T. J.; Myers, C. R.; Lowenthal, D.; Enquist, B. J.; Boyle, B.; Akoglu, A.; Andrews, G.; Ram, S.; Ware, D.; Stein, L. and Stanzione, D. (**2011**). *The iPlant Collaborative: Cyberinfrastructure for Plant Biology.*, Frontiers in Plant Science 2 : 34.

Goodstein, D. M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R. D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N. and Rokhsar, D. S. (**2012**). *Phytozome: a comparative platform for green plant genomics.*, Nucleic Acids Research 40 : D1178-D1186.

Griffiths-Jones, S. (**2004**). *The microRNA Registry.*, Nucleic Acids Research 32 : D109-D111.

Gusfield, D., **1997**. *Algorithms on strings, trees and sequences: computer science and computational biology*. Cambridge University Press, .

Hardison, R. C. (**2003**). *Comparative genomics.*, PLoS Biology 1 : E58.

Harvey, P. H. and Pagel, M. D., **1991**. *The comparative method in evolutionary biology*. Oxford university press Oxford, .

Haubold, B. and Wiehe, T. (**2004**). *Comparative genomics: methods and applications.*, Naturwissenschaften 91 : 405-421.

Jensen, R. A. (**2001**). *Orthologs and paralogs - we need to get it right.*, Genome Biology 2 : INTERACTIONS1002.

Kiriakidou, M.; Nelson, P. T.; Kouranov, A.; Fitziev, P.; Bouyioukos, C.; Mourelatos, Z. and Hatzigeorgiou, A. (**2004**). *A combined computational-experimental approach predicts human microRNA targets.*, Genes Dev 18 : 1165-1178.

Koonin, E. V.; Aravind, L. and Kondrashov, A. S. (**2000**). *The impact of comparative genomics on our understanding of evolution.*, Cell 101 : 573-576.

Kristensen, D. M.; Kannan, L.; Coleman, M. K.; Wolf, Y. I.; Sorokin, A.; Koonin, E. V. and Mushegian, A. (**2010**). *A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches.*, Bioinformatics 26 : 1481-1487.

Kriventseva, E. V.; Tegenfeldt, F.; Petty, T. J.; Waterhouse, R. M.; Simão, F. A.; Pozdnyakov, I. A.; Ioannidis, P. and Zdobnov, E. M. (**2014**). *OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software.*, Nucleic Acids Research .

Langmead, B.; Trapnell, C.; Pop, M. and Salzberg, S. L. (**2009**). *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.*, Genome Biol 10 : R25.

Li, L.; Stoeckert Jr, C. J. and Roos, D. S. (**2003**). *OrthoMCL: identification of ortholog groups for eukaryotic genomes.*, Genome Res 13 : 2178-2189.

Magazine, S. (**2002**). *Areas to Watch in 2003*, Science 298 : 2298–2298.

Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates, T. O. and Eisenberg, D. (**1999**). *Detecting protein function and protein-protein interactions from genome sequences.*, Science 285 : 751-753.

Moreton, J.; Dunham, S. P. and Emes, R. D. (**2014**). *A consensus approach to vertebrate de novo transcriptome assembly from RNA-seq data: assembly of the duck (Anas platyrhynchos) transcriptome.*, Front Genet 5 : 190.

Pertsemlidis, A. and Fondon 3rd, J. (**2001**). *Having a BLAST with bioinformatics (and avoiding BLASTphemy).*, Genome Biology 2 : REVIEWS2002.

Picardi, E. and Pesole, G. (**2010**). *Computational methods for ab initio and comparative gene finding.*, Methods in Molecular Biology 609 : 269-284.

Rouard, M.; Guignon, V.; Aluome, C.; Laporte, M.-A.; Droc, G.; Walde, C.; Zmasek, C. M.; Périn, C. and Conte, M. G. (**2011**). *GreenPhylDB v2.0: comparative and functional genomics in plants.*, Nucleic Acids Research 39 : D1095-D1102.

Rubin, G. M.; Yandell, M. D.; Wortman, J. R.; Gabor Miklos, G. L.; Nelson, C. R.; Hariharan, I. K.; Fortini, M. E.; Li, P. W.; Apweiler, R.; Fleischmann, W.; Cherry, J. M.; Henikoff, S.; Skupski, M. P.; Misra, S.; Ashburner, M.; Birney, E.; Boguski, M. S.; Brody, T.; Brokstein, P.; Celniker, S. E.; Chervitz, S. A.; Coates, D.; Cravchik, A.; Gabrielian, A.; Galle, R. F.; Gelbart, W. M.; George, R. A.; Goldstein, L. S.; Gong, F.; Guan, P.; Harris, N. L.; Hay, B. A.; Hoskins, R. A.; Li, J.; Li, Z.; Hynes, R. O.; Jones, S. J.; Kuehl, P. M.; Lemaitre, B.; Littleton, J. T.; Morrison, D. K.; Mungall, C.; O'Farrell, P. H.; Pickeral, O. K.; Shue, C.;

Vosshall, L. B.; Zhang, J.; Zhao, Q.; Zheng, X. H. and Lewis, S. (**2000**). *Comparative genomics of the eukaryotes.*, Science 287 : 2204-2215.

Schmidt, R. (**2002**). *Plant genome evolution: lessons from comparative genomics at the DNA level.*, Plant Molecular Biology 48 : 21-37.

Tatusov, R. L.; Fedorova, N. D.; Jackson, J. D.; Jacobs, A. R.; Kiryutin, B.; Koonin, E. V.; Krylov, D. M.; Mazumder, R.; Mekhedov, S. L.; Nikolskaya, A. N.; Rao, B. S.; Smirnov, S.; Sverdlov, A. V.; Vasudevan, S.; Wolf, Y. I.; Yin, J. J. and Natale, D. A. (**2003**). *The COG database: an updated version includes eukaryotes.*, BMC Bioinformatics 4 : 41.

Tatusov, R. L.; Koonin, E. V. and Lipman, D. J. (**1997**). *A genomic perspective on protein families.*, Science 278 : 631-637.

Uchiyama, I.; Mihara, M.; Nishide, H. and Chiba, H. (**2013**). *MBGD update 2013: the microbial genome database for exploring the diversity of microbial world.*, Nucleic Acids Research 41 : D631-D635.

Yanai, I. and DeLisi, C. (**2002**). *The society of genes: networks of functional links between genes from comparative genomics.*, Genome Biology 3 : research0064.

Zerbino, D. R. and Birney, E. (**2008**). *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.*, Genome Res 18 : 821-829.