Πανεπιστήμιο Κρήτης
Σχολή Θετικών Επιστημών, Τμήμα Βιολογίας
Μεταπτυχιακό Πρόγραμμα Μοριακής Βιολογίας Βιοϊατρικής

Ίδρυμα Τεχνολογίας Έρευνας
Ινστιτούτο Μοριακής Βιολογίας και Βιοτεχνολογίας

Μεταπτυχιακός τίτλος ειδίκευσης

Προσδιορισμός των LTR μεταθετών στοιχείων στο γονιδίωμα του *Anopheles gambiae*

&

Σχεδιασμός της Anobase,της σχεσιακής βάσης δεδομένων των ειδών του γένους *Anopheles*

Κουτσός Αναστάσιος

Επιβλέποντες καθηγητές:
Χρήστος Λούης & Νεκτάριος Ταβερναράκης

Ηράκλειο, Σεπτέμβριος 2002

Απέλειπεν ο Θεός Αντώνιον

Σαν έξαφνα, ώρα μεσάνυχτ' ακουσθεί,
αόρατος θίασος να περνά,
με μουσικές εξαίσιες με φωνές –
την τύχη σου που ενδίδει πια, τα έργα σου
που απέτυχαν, τα σχέδια της ζωής σου
που βγήκαν όλα πλάνες, μη ανωφέλετα θρηνήσεις.
Σαν έτοιμος από καιρο, σα θαραλέος,
αποχαιρέτα την, την Αλεξάνδρεια που φεύγει.
Προ πάντων να μη γελασθείς, μην πεις πως ήταν
ένα όνειρο, πως απατήθηκεν η ακοή σου·
μάταιες ελπίδες τέτοιες να μην καταδεχθείς.
Σαν έτοιμος από καιρό, σα θαραλέος
σαν που ταιριάζει σε που αξιώθηκες μια τέτοια πόλι,
πλησίασε σταθερά προς το παράθυρο,
κι άκουσε με συγκίνησιν, αλλ' όχι,
με των δειλων τα παρακάλια και παράπονα,
ως τελευταία απόλαυσι τους ήχους,
τα εξαίσια όργανα του μυστικού θιάσου,
κι αποχαιρέτα την, την Αλεξάνδρεια που χάνεις.

Κ.Π Καβάφης, 1911

Το ποίημα αυτό περιγράφει με τον πιο χαρακτηριστικό τρόπο πως ενιωθα κάποια περίοδο της ζωής μου για λόγους καθαρά προσωπικούς. Αυτή η περίοδος συνέπεσε με την αρχή της μεταπτυχιακής μου εργασίας και με επηρέασε τόσο, ώστε κάποια στιγμή σκέφτηκα να μην την συνεχίσω.

Η εργασία αυτή, λοιπόν, αφιερώνεται στο άτομο που με βοήθησε περισσότερο εκείνη την περίοδο.

Σ' ευχαριστώ Μαρία Μ., επειδή εκείνες τις δύσκολες στιγμές μου υπενθύμισες ότι εκτός απο «την Αλεξάνδρεια που χάνω» κάποτε θα ξαναβρώ την «Ιθακη που δε με γέλασε». Πως το χάλι γκάλι θ' αρχίσει πάλι ...

... και την επόμενη φορά θα είναι πιο δυνατά από την πρώτη.

Θα το θυμάμαι πάντα.

Περιεχόμενα

Ευχαριστίες και άλλα

Ευχαριστίες και άλλα...

Η εργασία αυτή έγινε εξ ολοκλήρου στο Ίδρυμα Τεχνολογίας Έρευνας, στο Ινστιτούτο Μοριακής Βιολογίας και Βιοτεχνολογίας, στον τομέα των εντόμων, υπό την επίβλεψη του καθηγητή κ. Λούη. Αισθάνομαι, λοιπόν, την ανάγκη να ευχαριστήσω πρώτα αυτόν, όχι μόνο επειδή μου έδωσε τη δυνατότητα να γίνω μέλος της ερευνητικής του ομάδας, αλλά επειδή και αυτός ήταν παράτολμος στο να συμφωνήσει μαζί μου σε μια εργασία στο πεδίο της βιοπληροφορικής, ένα πεδίο για το οποίο γνώριζα ελάχιστα όταν ξεκινούσα την εργασία αυτή.

Ένα ευχαριστώ αξίζει επίσης στο προσωπικό της βιβλιοθήκης του τμήματος Βιολογίας και Φυσικής, στη Μαρία Σπινθάκη και τη Γιώτα Αποστολάκη, που με βοηθούσαν πάντα στις «εξερευνήσεις» μου στη βιβλιοθήκη. Δυστυχώς ο περιορισμένος μου χρόνος κατά τη διάρκεια των 2 ετών με έκανε να επισκέπτομαι τη βιβλιοθήκη όλο και πιο σπάνια, όμως πάντα συναντούσα το χαμόγελό τους, όταν βρισκόμουν εκεί.

Θα ήθελα επίσης να ευχαριστήσω τους καθηγητές που με δέχτηκαν στα εργαστήρια τους για τις «περιστροφικές» μου ασκήσεις. Το καθηγητή κ. Λούη για την πρώτη περιστροφική μου άσκηση, και μαζί με αυτόν την Inga Siden-Kiamos, που ασχολήθηκε μαζί μου σε όλη τη διάρκεια της άσκησης μου και μου έδωσε πολύτιμες συμβουλές, τη Βάσω Μαχαιράκη, που μου κρατούσε παρέα στο εργαστήριο και υπέμενε την απαισιοδοξία μου για τα πειράματα, καθώς και όλο το εργαστήριο. Τον καθηγητή κ. Δελιδάκη, για τη δεύτερη περιστροφική μου άσκηση, και μαζί με αυτόν τον Γιαγτζόγλου Νίκο, καθώς και όλο το εργαστήριο, για τις πολύτιμες συμβουλές τους και διότι υπέμεναν σχεδόν αδιαμαρτύρητα την απαισιοδοξία μου για τα πειράματα. Τέλος, τον Δρ. Καφετζόπουλο, και μαζί με αυτόν τη Μανουέλα Καψετάκη και τη Λίλα Κουμάντου, διότι εκτός των άλλων, υπέμεναν (τι πρωτότυπο!) την απαισιοδοξία μου για τα πειράματα. Ιδιαίτερα, δεν πρόκειται να ξεχάσω ποτέ την παρομοίωση της έρευνας με το μπουκάλι της ketchup, που μου είπε κάποτε η Μανουέλα.

Η απόφαση μου να εγκαταλείψω την πιπέτα και να πιάσω το πληκτρολόγιο δεν ήταν εύκολη. Βασιζόμουν μόνο σε μια δική μου παρόρμηση να ασχοληθώ με τους υπολογιστές, χωρίς ουσιαστικά να έχω την κατάλληλη γνώση για τη βιοπληροφορική. Πολλές φορές σκέφτηκα ότι ίσως δεν θα έπρεπε να το κάνω. Θα ήθελα να ευχαριστήσω, λοιπόν, την Βάσω Μαχαιράκη και τον Ηλία Παυλόπουλο, για τις ευχάριστες συζητήσεις που έκανα μαζί τους όταν έπρεπε να πάρω την απόφαση, αλλά και διότι ουσιαστικά αυτοί ήταν που με ώθησαν στην απόφαση αυτή. Ένα ευχαριστώ επίσης και στο Νεκτάριο Ταβερναράκη, ο οποίος μου έδωσε πολύτιμες συμβουλές, και μου έκανε την τιμή να είναι ο δεύτερος διορθωτής της μεταπτυχιακής μου εργασίας.

Στα πρώτα μου βήματα στον κόσμο των υπολογιστών θα ήθελα επίσης να ευχαριστήσω την ομάδα υποστήριξης υπολογιστών του IMBB, τον Δημήτρη Αϊβαλιώτη και τον Γιάννη Κουκλινό, για την πολύτιμη βοήθεια και τεχνική υποστήριξη που παρείχαν. Επίσης, τον Χρήστο Κιάμο, που με ανέχθηκε το καλοκαίρι ως ένα μικρό αφεντικό, αλλά, επίσης, επειδή τράβηξε το βαρύ φορτίο του να φτιάξει ένα από τα προγράμματα της βάσης δεδομένων, δουλειά που διαφορετικά θα είχε βαρύνει το ήδη παραφουσκωμένο μου πρόγραμμα.

Ένα μεγάλο ευχαριστώ και στους αθέατους πρωταγωνιστές αυτής της εργασίας, τους φίλους μου. Κατά τη διάρκεια αυτής της εργασίας ένιωσα πολλούς από τους κοντινούς μου φίλους να απομακρύνονται, έκανα καινούργιους φίλους, ξαναβρήκα

άλλους παλιούς φίλους. Τους ευχαριστώ, όμως, όλους, είτε ήταν εκεί είτε με σκέφτονταν και τους εύχομαι να είναι καλά, όπου κι αν βρίσκονται.

Αναμφισβήτητα, όμως, τα άτομα στο IGL (IMBB Genome Lab) αξίζουν ένα ιδιαίτερο ευχαριστώ. Στον Λευτέρη «πλαζ πλαζ» Σπανό, ο οποίος έκανε τη δουλειά μου στο εργαστήριο λιγότερο φορτική και περισσότερο ευχάριστη, και μου έδωσε το παράδειγμα ενός ανθρώπου που πάντα θέλει να ψάχνει, να μαθαίνει, να δοκιμάζει. Στον Γιώργο «τι άχρηστο που είναι το bioinformatics team» Παπαγιαννάκη, ο οποίος με το δικό του ιδιαίτερο τρόπο μου έμαθε πως το εργαστήριο δεν είναι μόνο τα μηχανήματα, τα πειράματα και τα άψυχα αντικείμενα, αλλά περισσότερο από όλα οι άνθρωποί του.

Και τέλος, στον Παντελή Τοπάλη, στον οποίο και οφείλω όλες μου τις γνώσεις στην βιοπληροφορική μέχρι στιγμής. Η εργασία αυτή του ανήκει κατά το ήμισυ, όχι επειδή μου έμαθε τα πάντα, αλλά επειδή η συνεργασία μας ήταν σε τέτοιο βαθμό άψογη, που ορισμένα κομμάτια αυτής της εργασίας έγιναν από εμένα και ορισμένα από αυτόν, με τέτοιο τρόπο που θα ήταν δύσκολο να διαχωριστούν. Θα μπορούσα να πω περισσότερα πράγματα για τον Παντελή, αλλά δεν θα το κάνω. Ένα από τα πράγματα που έμαθα από αυτόν είναι το «ότι να κάνεις αυτό που σου αρέσει, δεν μπορεί να πληρωθεί ούτε με όλα τα χρήματα του κόσμου».

Ίσως είναι πάλι περίεργο, αλλά οι άνθρωποι που αξίζουν τη μεγαλύτερη ευγνωμοσύνη από όλους είναι και αυτοί που πέρασα ελάχιστες στιγμές μαζί τους αυτήν την περίοδο. Δεν είναι παρά η ίδια η οικογένεια μου, η οποία για μια ακόμη φορά μου εξασφάλισε ένα αρμονικό περιβάλλον για σπουδές, και η «υποτροφία» της οποίας πλήρωνε αδιαμαρτύρητα όλες τις μικρές και μεγάλες μου σπατάλες. Αν έχω φτάσει σε κάποιο σημείο μέχρι σήμερα το χρωστάω σε αυτούς.

... Και οι «καθιερωμένες» αναπάντεχες ευχαριστίες. Πρώτα, θα ήθελα να ευχαριστήσω το συγγραφέα Νίκο Δήμου, ο οποίος νομίζω ότι με παρηγόρησε με τον καλύτερο τρόπο κατά την αρχή του μεταπτυχιακού μου, όταν κάποιοι άλλοι μου είπαν ότι η βιοπληροφορική «δεν τους έχει απασχολήσει καθόλου». Τότε ήταν η πρώτη φορά που κατάλαβα πόσο δύσκολο είναι αυτό που κάνει τόσο καιρό, να προσπαθεί να πείσει όλους τους άλλους για τα μικρά θαύματα της τεχνολογίας και τον τρόπο με τον οποίο έχουν αλλάξει την καθημερινή μας ζωή. Τότε ήταν επίσης που κατάλαβα ότι εάν το περιβάλλόν μας δεν θέτει προκλήσεις, τότε θα πρέπει να το κάνουμε μόνοι μας. Τέλος θα ήθελα να ευχαριστήσω το Φράγκο Δημήτρη, για τους λόγους που ξέρει ο ίδιος.

Τελειώνοντας αυτήν την εργασία, πίσω από της σελίδες της υπάρχουν αρκετές ώρες διαβάσματος, αμέτρητες γραμμές κώδικα, πολλές περιπτώσεις μικρών καταστροφών (ενίοτε και μεγαλύτερων) και ορισμένες, αν και λίγες, περιπτώσεις απογοητεύσεων. Παρόλα αυτά, η εργασία αυτή έχει ιδιαίτερη σημασία. «Αν θέλεις κάτι πάρα πολύ ολόκληρος ο κόσμος συνωμοτεί για να σε βοηθήσει να το επιτύχεις». Για εμένα, το νόημα αυτής της φράσης κρύβεται στις επόμενες σελίδες.

Κουτσός Αναστάσιος

Γενική εισαγωγή

Γενική εισαγωγή

Η ελονοσία στον κόσμο – ιστορική αναδρομή και χαρακτηριστικά

Οι παρασιτικές ασθένειες αποτέλεσαν στο παρελθόν και αποτελούν ακόμα και σήμερα σημαντικό αίτιο καταπόνησης της ανθρωπότητας. Μια από αυτές είναι και η ελονοσία, η οποία ευθύνεται για περισσότερες από 300 εκατομμύρια περιπτώσεις το χρόνο παγκοσμίως, καθώς και για 1,5-2,7 εκατομμύρια θανάτους το χρόνο στις αφρικανικές χώρες κυρίως (Collins et al., 2000; Smyth and Wakelin, 1994).Η ελονοσία παραμένει ενδημική σε περισσότερες από 102 χώρες, συνολικού πληθυσμού 2,4 δις, δηλαδή στο 50% περίπου του πληθυσμού της γης.

Η εμφάνιση της ελονοσίας δεν είναι πρόσφατη (Bruce-Chwatt, 1993) Έχει διατυπωθεί η υπόθεση ότι η ελονοσία εμφανίστηκε κατά τον προϊστορικό χρόνο. Η πρώτη λεπτομερής περιγραφή της κλινικής εικόνας αποδίδεται στον Ιπποκράτη, τον 5 αιώνα πΧ. Παρόλα αυτά, μόλις το 1880 περιγράφτηκαν παράσιτα στο αίμα ασθενών από τον Laveran. Το 1897, ο Ronald Ross πρωτοπαρατήρησε μια αναπτυσσόμενη μορφή του παρασίτου σε κουνούπια τα οποία είχαν προηγουμένως τραφεί από μολυσμένους ανθρώπους. Είναι, όμως της περίοδο του 1898-1899 που οι μελέτες του Battista Grassi και των συναδέλφων του απέδειξαν ότι τα κουνούπια του γένους Anopheles είναι οι φορείς των παρασίτων της ελονοσίας. Στη συνέχεια, από το τέλος του 19^{ου} αιώνα και της αρχές του 20^{ου} αιώνα συντελέστηκε μεγάλη πρόοδος, όσον αφορά τη γνώση για τη μετάδοση της ασθένειας, τα συμπτώματα της ασθένειας, τις επιπτώσεις της και την προσπάθεια καταπολέμησής της.

Η ασθένεια της ελονοσίας στον άνθρωπο προκαλείται από 4 παράσιτα του γένους *Plasmodium* (φύλο *Apicomplexa (σπορόζωα),* τάξη *Hemosporida*): το *Plasmodium falciparum,* το *Plasmodium malariae,* το *Plasmodium ovale* και το *Plasmodium vivax.* Το *Plasmodium vivax* είναι το πιο διαδεδομένο γεωγραφικά και είναι υπεύθυνο για το 43% των περιπτώσεων της ελονοσίας, το *Plasmodium falciparum* για το 50% των περιπτώσεων της ελονοσίας, ενώ τα *Plasmodium ovale* και *Plasmodium malariae* συναντώνται σχετικά σπάνια. Από τα παράσιτα αυτά, το πιο επικίνδυνο είναι το *Plasmodium falciparum,* το οποίο ευθύνεται και για τους περισσότερους θανάτους.

Για να ολοκληρώσουν τον κύκλο ζωής τους τα παραπάνω 4 είδη, απαιτούν δυο ξενιστές: τα κουνούπια του γένους *Anopheles* και τον άνθρωπο Βέβαια, υπάρχουν και άλλα είδη του γένους *Plasmodium* τα οποία δεν έχουν ως ξενιστή τον άνθρωπο αλλά διάφορα ζώα. Η μετάδοση της ελονοσίας από ένα σπονδυλωτό σε ένα άλλο πραγματοποιείται μέσω αυτών των κουνουπιών, το πιο αποτελεσματικό είδος των οποίων θεωρείται το *Anopheles gambiae*. Η αποτελεσματικότητα των κουνουπιών ως φορέων μετάδοσης της ελονοσίας οφείλεται σε διάφορα χαρακτηριστικά τους, όπως είναι η διάρκεια ζωής τους, η προτίμησή τους σε ορισμένους ξενιστές για την τροφή τους (ορισμένα είδη τρέφονται με αίμα από πολλούς ξενιστές) καθώς και οι προτιμήσεις του όσον αφορά το μικροπεριβάλλον. Το είδος *Anopheles gambiae* παρουσιάζει όλα αυτά τα παραπάνω χαρακτηριστικά: έχει μεγάλη διάρκεια ζωής, εμφανίζει σχεδόν αποκλειστική προτίμηση στον άνθρωπο και παρατηρείται σε κατοικίες και σε μέρη όπου συχνάζουν οι άνθρωποι (Beaty and Marquardt, 1996). Ο *Anopheles gambiae* για αυτό το λόγο θεωρείται ο αποτελεσματικότερος φορέας από όλα τα είδη.

Το *Plasmodium* υφίσταται το σεξουαλικό κομμάτι της ζωής του στο κουνούπι. Είναι γνωστό ότι τα θηλυκά κουνούπια τρέφονται με αίμα, για να λάβουν τις

απαραίτητες θρεπτικές ουσίες για την ωρίμανση των γαμετών τους. Στο κουνούπι μεταφέρονται πολυάριθμα γαμετοκύτταρα κατά τη νύξη του δέρματος των ανθρώπων–φορέων από τα κουνούπια και την απομύζηση αίματος. Τα γαμετοκύτταρα αυτά ακολουθούν ένα πολύπλοκο μονοπάτι διαφοροποίησης στο κουνούπι (Ghosh et al., 2000): διαφοροποιούνται σε αρσενικούς και θηλυκούς γαμέτες, σε ζυγωτό και στη συνέχεια σε ωοκινέτες, ωοκύστεις και σποροζωίτες. Αυτά τα στάδια της διαφοροποίησης είναι διακριτά μεταξύ τους και συνήθως συμβαίνουν σε διαφορετικά όργανα και ιστούς του κουνουπιού, Το ζυγωτό διαφοροποιείται σε ωοκινέτες στο μεσέντερο, οι ωοκινέτες μεταπίπτουν σε ωοκύστεις στο χώρο κάτω από τη βασική μεμβράνη (basal lamina) του μεσεντέρου, ενώ οι σποροζωίτες που παράγονται εκεί διεισδύουν μέσω του αίματος στους σιελογόνους αδένες. Με επόμενο τσίμπημα του κουνουπιού, οι σποροζωίτες μεταφέρονται μέσω του αίματος στο ήπαρ του ανθρώπου, όπου υφίστανται διαδοχικές ασεξουαλικές διαιρέσεις και σχηματίζουν αρχικά σχιζοζωίδια και στη συνέχεια μεροζωίτες. Οι μεροζωίτες διαρρηγνύουν τα ηπατικά κύτταρα και εισβάλλουν μέσω του αίματος στα ερυθροκύτταρα, όπου και πολλαπλασιάζονται μονογονικά μέχρι την λύση των κυττάρων αυτών. Τα παράσιτα αυτά στη συνέχεια εισβάλλουν σε νέα αιμοσφαίρια και επαναλαμβάνουν τον κύκλο της λύσης αυτής. Με το τσίμπημα ενός κουνουπιού, ορισμένα γαμετοκύτταρα θα μεταφερθούν στο κουνούπι για να επαναλάβουν τον κύκλο.

Τα επεισόδια αυτά της λύσης των ερυθρών αιμοσφαιρίων είναι που δημιουργούν τους παροξυσμούς πυρετού στο μολυσμένο άτομο, ένα από τα χαρακτηριστικά συμπτώματα της ελονοσίας. Το *Plasmodium vivax* προκαλεί καλοήθη τριταίο πυρετό (εκδηλώνεται κάθε 48 ώρες, δηλαδή κάθε τρίτη ημέρα), το *Plasmodium ovale* προκαλεί επίσης τριταίο πυρετό, αλλά πιο ελαφριάς μορφής από αυτόν του *Plasmodium vivax*, το *Plasmodium malariae* προκαλεί τεταρταίο πυρετό (εμφανίζεται κάθε 72, δηλαδή κάθε τέταρτη ημέρα) ενώ το *Plasmodium falciparum* προκαλεί πυρετό κάθε 48 ώρες (κακοήθης τριταίος πυρετός).

Οι προσπάθειες ελέγχου και καταπολέμησης της ελονοσίας

Αρκετά νωρίς εμφανίστηκαν και οι πρώτες προσπάθειες καταπολέμησης της ασθένειας αυτής. οι οποίες εστιάστηκαν σε δυο στόχους: στο παράσιτο, είτε με την ανακάλυψη φαρμάκων για την καταπολέμηση των συμπτωμάτων είτε με την προσπάθεια δημιουργίας εμβολίου, και στον ξενιστή, το κουνούπι, είτε με το να εμποδιστεί η επαφή του κουνουπιού με τον άνθρωπο, είτε στην προσπάθεια ελέγχου ή ακόμα και εξαφάνισης των πληθυσμών με χημικές ουσίες.

Το 1955, ο Παγκόσμιος Οργανισμός Υγείας- ΠΟΥ (World Health Organization, WHO) υιοθέτησε μια στρατηγική προσπάθεια εξαφάνισης της ελονοσίας από τον πλανήτη. Είχαν προηγηθεί επιτυχημένες στρατηγικές εξαφάνισης της ελονοσίας σε ορισμένες περιοχές όπως για παράδειγμα στην Ιταλία, στην Ελλάδα και στις ΗΠΑ, κυρίως με τον έλεγχο και την εξόντωση των φυσικών πληθυσμών του γένους *Anopheles* με την ευρεία χρήση του εντομοκτόνου DDT. 15 χρόνια μετά την εφαρμογή του προγράμματος του ΠΟΥ στις αφρικανικές χώρες, τα αποτελέσματα ήταν ενθαρρυντικά. Παρόλα αυτά, στη δεκαετία 1980-1990 η ελονοσία επανεμφανίστηκε δραματικά σε ορισμένα μέρη στα οποία τα προγράμματα εξαφάνισης είχαν επιτύχει. Τις τελευταίες δεκαετίες, ο αριθμός των περιπτώσεων της ελονοσίας έχει αυξηθεί, ενώ ο αριθμός των θανάτων φαίνεται να έχει μείνει στάσιμος ή να έχει αυξηθεί σε μικρό βαθμό (Collins et al., 2000).

Η επανεμφάνιση της ελονοσίας τις τελευταίες δεκαετίες οφείλεται στην ανθεκτικότητα τόσο του παρασίτου όσο και του ξενιστή στις χημικές ουσίες. Όσον αφορά το παράσιτο, στο τέλος της δεκαετίας του 1970 εμφανίστηκαν τα πρώτα στελέχη με ανθεκτικότητα στη χλωροκινίνη (chloroquine), μια ευρέως διαδεδομένη χημική ουσία για την καταπολέμηση του παρασίτου, και ακολούθησαν και στελέχη με ανθεκτικότητα σε άλλες ουσίες. Από την άλλη μεριά, η χρήση του DDT για την εξάλειψη των πληθυσμών του *Anopheles* εγκαταλείφθηκε εξαιτίας της υψηλής τοξικότητας και πιθανής καρκινογένεσης που προκαλούσε στον άνθρωπο. Άλλες ουσίες που χρησιμοποιήθηκαν σε αντικατάσταση του DDT δεν ήταν το ίδιο αποτελεσματικές. Αναμφισβήτητα, όμως, η εμφάνιση στελεχών των πληθυσμών του *Anopheles gambiae* οι οποίοι είναι ανθεκτικοί στα αντιβιοτικά επιτείνει σημαντικά το πρόβλημα του ελέγχου των πληθυσμών του.

Τα γεγονότα αυτά σε μεγάλο βαθμό ώθησαν τον ΠΟΥ να αναθεωρήσει τη στρατηγική της ελονοσίας από την πλήρη εξάλειψη σε περιορισμό και έλεγχο της ασθένειας. Μέχρι σήμερα όλες οι προσπάθειες για τη δημιουργία εμβολίων δεν έχουν δώσει αποτέλεσμα, η προσπάθεια για περιορισμό της επαφής των ανθρώπων με τον πληθυσμό των κουνουπιών δεν έχει αποδώσει καρπούς, το σύστημα διάγνωσης της ασθένειας στα πρώτα στάδια και της αναγνώρισης των ατόμων με κίνδυνο σοβαρού περιστατικού ελονοσίας είναι ανεπαρκές, ενώ αντίθετα οι πληθυσμοί των παρασίτων και των κουνουπιών δείχνουν αυξανόμενη ανθεκτικότητα σε ουσίες που έχει ανακαλύψει ο άνθρωπος για την εξόντωσή τους. Είναι φανερό, λοιπόν, ότι οποιαδήποτε σύγχρονη στρατηγική για την αντιμετώπισης της ελονοσίας πρέπει να συνδυάσει την παραδοσιακή γνώση στον τομέα αυτό με την υιοθέτηση καινούργιων στρατηγικών και μεθόδων.

Η μοριακή βιολογία ως ένα όπλο για την καταπολέμηση της ασθένειας

Τις τελευταίες δεκαετίες, η μοριακή βιολογία έχει αναμφισβήτητα δώσει καινούργια ώθηση στην κατανόηση των μηχανισμών των ασθενειών στον άνθρωπο και σε άλλα ζώα, είτε αυτές οι ασθένειες προέρχονται από άλλους οργανισμούς είτε από βλάβες στο γενετικό υλικό των ίδιων οργανισμών. Ακόμα και στην περίπτωση της ελονοσίας, η χρησιμοποίηση κάποιων ευρέως διαδεδομένων μοριακών τεχνικών φαίνεται να έχει συμβάλλει στην προώθηση της γνώσης των φυσικών πληθυσμών του *Anopheles* όπως και στον τρόπο μετάδοσης της ασθένειας.

Αν και έχει εκφραστεί η άποψη ότι δηλαδή προσπάθεια καταπολέμησης της ασθένειας δεν θα πρέπει να εστιαστεί στη μοριακή βιολογία και τη δυνατότητα γενετικής παρέμβασης στους οργανισμούς (Curtis, 2000; Spielman, 1994), εντούτοις υπάρχει ένας ολοένα αυξανόμενος αριθμός επιστημόνων (Collins et al., 2000; Hoffman, 2000; Hoffman et al., 2002; Louis, 1999) που υποστηρίζουν ότι, αν και είναι σχετικά νωρίς να μιλήσουμε για την ολοκληρωτική εξάλειψη της αθένειας με μοριακές μεθόδους, εντούτοις η γονιδιωματική έρευνα θα μπορούσε να βοηθήσει στην καλύτερη κατονόηση της ασθένειας. Ακόμα και σήμερα υπάρχουν τέτοια παραδείγματα εφαρμογών που έχουν προκύψει από τη μορικαή έρευνα.

Αρχικά, ολοένα και περισσότεροι γενετικοί δείκτες χρησιμοποιούνται για την αναγνώριση των διαφορετικών ειδών φορέων κουνουπιών σε μια περιοχή. Τέτοιες μελέτες έχουν οδηγήσει στην καλύτερη κατανόηση των φυσικών πληθυσμών, στη μετακίνηση τους, τη γονιδιακή ροή και την ηθολογία τους. Είναι φανερό ότι η ιδέα μιας σύγχρονης στρατηγικής ελέγχου και περιορισμού της ελονοσίας πρέπει να περικλείει και την λεπτομερή γνώση των πληθυσμών των φορέων. Αντίθετα,

παλιότερα, η μορφολογική αναγνώριση των διαφορετικών ειδών είχε οδηγήσει σε πολλά λάθη, όπως επίσης και σε λανθασμένα συμπεράσματα για την ηθολογική και επιδημιολογική συμπεριφορά τους..

Παράλληλα, οι μοριακές τεχνικές στοχεύουν στην αναγνώριση γονιδίων που παίζουν ρόλο στην ασθένεια τόσο στο παράσιτο όσο και στους ξενιστές. Τέτοια γονίδια στο παράσιτο, γονίδια που συντελούν στη μετάδοση της ασθένειας, και πιο συγκεκριμένα στον κύκλο ζωής του παράσιτου, αποτελούν υποψήφιους στόχους για τη σχεδίαση εξειδικευμένων και αποτελεσματικών φαρμάκων. Γονίδια που ενέχονται σε μηχανισμούς άμυνας των ξενιστών απέναντι στο παράσιτο παρουσιάζουν αρκετό ερευνητικό ενδιαφέρον, καθώς διερευνάται η δυνατότητα εισαγωγής τέτοιων γονιδίων στους φυσικούς πληθυσμούς των εντόμων μέσω γενετικού μετασχηματισμού.

Το σπουδαιότερο, όμως, είναι ότι η μελέτη της ελονοσίας και των οργανισμών της σε μοριακό επίπεδο, ίσως οδηγήσει στη δημιουργία νέων προσεγγίσεων μελέτης και καταπολέμησης, που δεν έχουν καν εκτιμηθεί με το σημερινό επίπεδο γνώσεων για την ασθένεια.

Παρόλα αυτά, είναι φανερό ότι μέχρι πρότινος υπήρχαν πολύ λίγα πράγματα γνωστά σε μοριακό επίπεδο όσον αφορά τους 2 από τους 3 οργανισμούς και συγκεκριμένα το παράσιτο και το κουνούπι. Η αλληλούχιση του γονιδιώματος των δυο αυτών οργανισμών, σε συνδυασμό με την αλληλούχιση του ανθρώπινου γονιδιώματος θα έδινε νέες προοπτικές στη χρησιμοποίηση των μοριακών τεχνικών για την καταπολέμηση της νόσου.

Μέχρι τη στιγμή της συγγραφής της εργασίας αυτής, τα 2900 Mb του ανθρώπινου γονιδιώματος έχουν αλληλουχηθεί (Lander et al., 2001; Venter et al., 2001), ενώ στους επιστήμονες είναι διαθέσιμος ένας κατάλογος από 30000 γονίδια, που έχουν προβλεφθεί από υπολογιστικές μεθόδους, καθώς και ένας κατάλογος με πολυάριθμος πολυμορφισμούς. Η αλληλούχιση του γονιδιώματος του *Plasmodium falciparum* ξεκίνησε το 1996 (Gardner, 1999), και αυτή τη στιγμή έχει ολοκληρωθεί, ενώ έχουν ήδη γίνει γνωστές οι αλληλουχίες των χρωμοσωμάτων 2 (Gardner et al., 1998) και 3 (Bowman et al., 1999), με την επίσημη δημοσίευση ολόκληρου του γονιδιώματος να αναμένεται τον Οκτώβριο του 2002.

Είναι, όμως, φανερό ότι οποιαδήποτε σύγχρονη στρατηγική καταπολέμησης της ασθένειας αυτής δεν μπορεί να παραβλέπει τον *Anopheles gambiae* τον φορέα που ευθύνεται για τη γεωγραφική εξάπλωση της νόσου.

Το γονιδίωμα του Anopheles gambiae

Μέχρι πρότινος λίγα πράγματα ήταν γνωστά για τη γενετική και τη βιολογία του *Anopheles gambiae*. Η κατάσταση αυτή άρχισε να αλλάζει με τη δημοσίευση ενός γονιδιακού χάρτη βασισμένου σε μικροδορυφορικές αλληλουχίες και άλλους ανιχνευτές (markers)(Dimopoulos et al., 1996; Zheng et al., 1996), εργασίες που βοήθησαν στη χαρτογράφηση μορφολογικών δεικτών και γονιδίων που έχουν σχέση με την ασθένεια, οι οποίοι με τη σειρά τους βοήθησαν σε πληθυσμιακές μελέτες του *Anopheles gambiae* και των συγγενικών σε αυτά ειδών.

Σε σχέση όμως με τη *Drosophila melanogaster,* η αλληλούχιση του γονιδιώματος της οποίας ολοκληρώθηκε το 2000 (Adams et al., 2000), ο *Anopheles gambiae* υπολείπεται τόσο σε γενετική μελέτη, όσο και σε αριθμό χαρακτηρισμένων γονιδίων, ένα εμπόδιο για την περαιτέρω μοριακή ανάλυση της ελονοσίας και του τρόπου αντιμετώπισης της, μέσω ελέγχου των κουνουπιών.

Είναι λοιπόν φανερό ότι η αλληλουχία του γονιδιώματος του *Anopheles gambiae* θα έδινε μεγάλη ώθηση στην επιστημονική έρευνα. Η αλληλούχιση του γονιδιώματος του *Anopheles gambiae* πραγματοποιήθηκε χάρη στη χρηματοδότηση από το Εθνικό Ινστιτούτο Αλλεργιών και Μολυσματικών Ασθενειών των ΗΠΑ (National Institute of Allergy and Infectious Diseases) και την Γαλλική Κυβέρνηση, σε συνεργασία με τον ΠΟΥ, ο οποίος είχε και την ευθύνη του συντονισμού. Παρόλο που ορισμένες περιοχές του γονιδιώματος του *Anopheles* είχαν ήδη αλληλουχηθεί, το μεγαλύτερο μέρος του γονιδιώματος προσδιορίστηκε με τη μέθοδο whole genome shotgun από την εταιρεία Celera.

Η αλληλούχιση του γονιδιώματος του *Anopheles gambiae* έγινε με τη συνεργασία ερευνητικών ιδρυμάτων από τις ΗΠΑ, τη Γαλλία, την Ιταλία, την Αγγλία, τη Γερμανία και την Ελλάδα. Η αλληλούχιση από τη Celera βασίστηκε σε πλασμιδιακές βιβλιοθήκες με ενθέσεις κομματιών 2kb, 10 kb και 50kb, ενώ πληροφορία χρησιμοποιήθηκε από αλληλουχίες BAC, πλασμιδίων και cDNA κλώνων που είχαν προσδιοριστεί πρωτύτερα. Η μελέτη της αλληλουχίας του γονιδιώματος για τον καθορισμό ανοικτών πλαισίων διαβάσματος (open reading frames – ORFs) με υπολογιστικές μεθόδους έγινε από τη Celera και το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (EBI), ανεξάρτητα το ένα ίδρυμα από το άλλο, τα αποτελέσματα των οποίων ενοποιήθηκαν στη συνέχεια.

Η αλληλουχία του *Anopheles gambiae*[1] βρίσκεται με τη μορφή 8987 σκελετών (scaffolds), οι οποίοι αντιπροσωπεύονται στο γονιδίωμα μια μόνο φορά. Οι σκελετοί είναι ομάδες διατεταγμένων και προσανατολισμένων contigs. Λίγο παραπάνω από 100 σκελετοί (που αντιστοιχούν στο 85% του συνολικού γονιδιώματος των 280 Mb) έχουν χαρτογραφηθεί σε μια συγκεκριμένη θέση στα 4 χρωμοσώματα, ενώ επιπλέον 60,737 μικρά contigs που χρησιμοποιήθηκαν δεν μπορούν να τοποθετηθούν ακόμα σε μοναδιαία θέση στο γονιδίωμα, εξαιτίας του μικρού τους μεγέθους. Ο αριθμός των ανοικτών πλαισίων διαβάσματος που καθορίστηκαν με υπολογιστικές μεθόδους ανέρχεται στα 15189, 1509 από τα οποία πιθανολογείται ότι ανήκουν σε μεταθετά στοιχεία και 663 ότι ανήκουν σε βακτήρια (πιθανές βακτηριακές μολύνσεις). Παρόλα αυτά, ο πραγματικός αριθμός των γονιδίων του *Anopheles gambiae* είναι μεγαλύτερος από τα καθορισμένα ανοικτά πλάσια διαβάσματος, καθώς υπάρχουν γεγονότα εναλλακτικού ματίσματος εξονίων, που δημιουργούν περισσότερες από μια πρωτεΐνες.

Θα πρέπει επίσης να τονιστεί ότι σύμφωνα με ορισμένες μελέτες που πραγματοποιήθηκαν σε ορισμένες περιοχές του γονιδιώματος, υπάρχουν πολλές πολυμορφικές περιοχές, οι οποίες δυσχεραίνουν τη διαδικασία συναρμολόγησης της αλληλουχίας από τους σκελετούς. Κατά συνέπεια, άλλες περιοχές αναμένεται να αντιπροσωπεύονται πολλαπλές φορές στο γονιδίωμα ενώ άλλες να μην αντιπροσωπεύονται καθόλου, γεγονός που δημιουργεί ανακρίβειες και κενά στη συναρμολόγηση. Παρόλα αυτά, ακόμη και στην πρωταρχική του μορφή, η πληροφορία του γονιδιώματος του *Anopheles gambiae* έχει μεγάλη χρησιμότητα.

Σημαντικό είναι να αναφερθεί ότι ολόκληρη η αλληλουχία του γονιδιώματος βρίσκεται κατατεθειμένη σε βάσεις δεδομένων στις ΗΠΑ (NCBI, http://www.ncbi.nih.nlm.gov/) και στην Ευρώπη (Ensembl, EBI,

[1] Μέχρι την ολοκλήρωση της συγγραφής της παρούσας εργασίας, η επίσημη δημοσίευση της αλληλουχίας του γονιδιώματος του *Anopheles gambiae* είχε προγραμματιστεί για τον Οκτώβριο του 2002. Παρόλα αυτά, τα στοιχεία που παρατίθενται σε αυτές τις παραγράφους ήταν διαθέσιμα από στοιχεία της Ensembl στο Ευρωπαϊκό Ινστιτούτο Πληροφορικής και στο NCBI, όπως και από προσχέδια επιστημονικών εργασιών που πρόκειται να δημοσιευτούν τον Οκτώβριο.

http://www.ensembl.org/) και η πρόσβασή της είναι ελεύθερη σε όλους τους επιστήμονες.


Ειδική εισαγωγή

Ειδική εισαγωγή

Η σύγκριση μεταξύ της Drosophila melanogaster και του Anopheles gambiae

Εκτός από το γονιδίωμα του *Anopheles gambiae*, αυτή τη στιγμή υπάρχει διαθέσιμο το γονιδίωμα της *Drosophila melanogaster* (Adams et al., 2000). Η *Drosophila melanogaster* είναι ένας γνωστός πειραματικός οργανισμός, ο οποίος έχει υποβληθεί σε εξαντλητική γονιδιωματική ανάλυση, με αποτέλεσμα να υπάρχουν πολλές πληροφορίες όσον αφορά τα γονίδιά της και τις λειτουργίες τους. Το γεγονός ότι η *Drosophila* και ο *Anopheles* είναι δίπτερα, και ότι πιστεύεται ότι έχουν διαχωριστεί 250 εκατομμύρια χρόνια περίπου (Yeates and Wiegmann, 1999) κάνει ενδιαφέρουσα τη συγκριτική γονιδιωματική ανάλυση των δυο αυτών ειδών.

Η σύγκριση μεταξύ των ειδών αυτών ίσως δώσει απαντήσεις στα ερωτήματα της διαφοροποίησής τους. Αυτά τα 250 εκατομμύρια χρόνια της διαφορετικής εξέλιξης των ειδών αυτών, ο *Anopheles gambiae* εμφάνισε αλλαγές στη μορφολογία και τη φυσιολογία του ώστε να μπορεί να χρησιμοποιεί το αίμα ως τροφή, αλλαγές στο γεννητικό του σύστημα ώστε να εναποθέτει τα αβγά σε υγρά μέρη, και αλλαγές στην ηθολογία του, όσον αφορά την διαβίωσή του σε υγρά μέρη και μέρη που συχνάζουν άνθρωποι. Τέτοιες αλλαγές θα πρέπει να έχουν «αποτυπωθεί» και στο γονιδίωμά του, όσον αφορά γονίδια ελέγχου της ανάπτυξης και της διαφοροποίησης του. Επομένως, η συγκριτική ανάλυση των δυο αυτών οργανισμών ίσως δώσει περισσότερες λεπτομέρειες σχετικά με τη διαφοροποίηση του *Anopheles* στην αιματοφαγία και τους μηχανισμούς της.

Από την άλλη μεριά, ένας αριθμός χαρακτηριστικών, μορφολογικών ή και λειτουργικών των δυο ειδών είναι κοινός, οπότε αναμένει κανείς ότι ένα μεγάλο μέρος των γονιδίων των δυο ειδών είναι κοινό, ή έχει διαφοροποιηθεί σε μικρό βαθμό κατά τη διάρκεια της εξέλιξης. Μάλιστα, μελέτες σε ορισμένες περιοχές των γονιδιωμάτων των δυο ειδών έχουν δείξει παρόμοια γονιδιακή δομή (Bolshakov et al., 2002; Thomasova et al., 2002). Με δεδομένο το μέγεθος των γνώσεων όσον αφορά τη *Drosophila melanogaster,* τα γονίδια της, όπως και πληροφορίες για ορισμένες περιοχές (domains) των γονιδίων της, θα μπορούσαν να χρησιμοποιηθούν για την εύρεση των αντίστοιχων γονιδίων στον *Anopheles gambiae*. Το γεγονός αυτό είναι σημαντικό, διότι από την πρώιμη ανάλυση της αλληλουχίας του γονιδιώματος έχουν προκύψει 15189 ανοικτά πλαίσια διαβάσματος, δηλαδή πιθανά γονίδια.

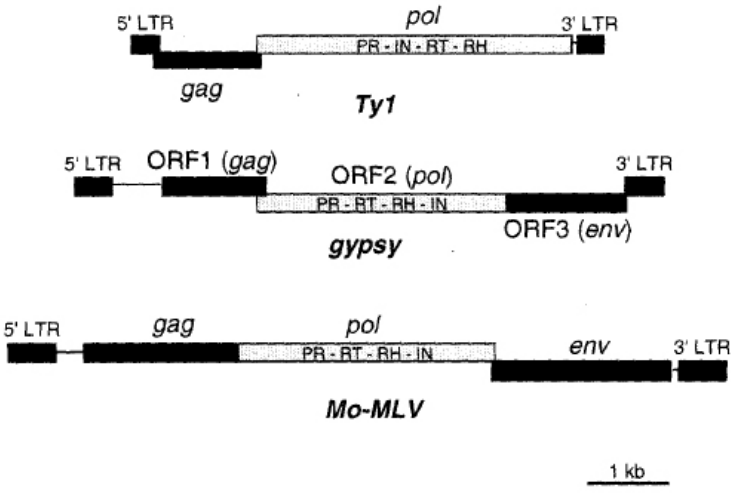
Το ότι η *Drosophila melanogaster* και ο *Anopheles gambiae* είναι εξελικτικά κοντά έχει οδηγήσει στην ιδέα ότι τεχνολογία που έχει αναπτυχθεί για τη μύγα θα μπορούσε να χρησιμοποιηθεί επίσης και στην περίπτωση του κουνουπιού. Τέτοια παραδείγματα υπάρχουν ήδη, όπως ο μετασχηματισμός γαμετικών κυττάρων στον *Aedes aegypti* (Jasinskiene et al., 1998), τον κυριότερο φορέα του κίτρινου πυρετού, και πρόσφατα και στο είδη του γένους *Anopheles* (Catteruccia et al., 2000a; Catteruccia et al., 2000b; Grossman et al., 2001), τεχνολογία που πρόκειται να βοηθήσει αρκετά στο μέλλον για λειτουργικές μελέτες των γονιδίων του *Anopheles*.

Τα μεταθετά στοιχεία και τα retrotransposons

Τα 1506 από τα 15189 ανοικτά πλαίσια διαβάσματος στο γονιδίωμα του *Anopheles gambiae* έχουν χαρακτηριστεί ως πιθανά μεταθετά στοιχεία, επειδή εντοπίστηκαν σε πολλαπλά σημεία στο γονιδίωμα του *Anopheles gmabiae*. Τα

μεταθετά στοιχεία είναι αλληλουχίες στον γονιδίωμα ευκαρυωτικών οργανισμών, οι οποίες μεταφέρονται από τη μια περιοχή στην άλλη με ένα μηχανισμό αποκοπής και ένθεσης ή ένα μηχανισμό αντιγραφής και ένθεσης, ανάλογα με το είδος του στοιχείου. Για το λόγο αυτό τα μεταθετά στοιχεία αποτελούν ένα μεγάλο κομμάτι του γονιδιώματος, σε ποσοστό που στα θηλαστικά μπορεί να φτάσει και στο 40%(Alberts, 1994).

Λόγω της ιδιότητας των μεταθετών στοιχείων να μεταφέρονται από μια περιοχή στην άλλη, έχειυποστηριχθεί ότι τα μεταθετά στοιχεία συντελούν στην εξέλιξη των γονιδιωμάτων(Alberts, 1994). Έχει παρατηρηθεί ότι η συντριπτική πλειοψηφία των αυθόρμητων μεταλλαγών στη *Drosophila melanogaster* οφείλεται σε τυχαίες ενθέσεις μεταθετών στοιχείων σε λειτουργικά γονίδια ή ρυθμιστικές περιοχές τους. Επιπλέον, διάφορες μελέτες έχουν δείξει ότι ορισμένα μεταθετά στοιχεία κατά την αποκοπή και την ένθεσή τους αφήνουν ορισμένα μικρά τμήματα από τα άκρα τους, ενώ άλλα είναι δυνατόν να μεταφέρουν ορισμένες μικρές περιοχές του γονιδιώματος σε άλλες περιοχές. Τέλος, ορισμένες έρευνες των μεταθετών στοιχείων στους πληθυσμούς της *Drosophila melanogaster* συνηγορούν υπέρ της οριζόντιας μεταφοράς μεταθετών στοιχείων, δηλαδή μεταφορά ανάμεσα σε είδη (Pinsker et al., 2001).



Εικόνα 1. Σύγκριση της δομής μεταξύ αντιπροσωπευτικών LTR retrotransposons και της δομής ρετροϊών. Το *Ty1 μεταθετό στοιχείο προέρχεται από την Ty1 οικογένεια του Saccharomyces cerevisiae,* ενώ ομόλογή οικογένεια στη *Drosiphila melanogaster* είναι η οικογένεια copia. Το μεταθετό στοιχείο gypsy προέρχεται από την οικογένεια gypsy της *Drosophila melanogaster,* ενώ ο ρετροϊός είναι ο Moloney murine leukemia virus (Mo-MLV). Στην εικόνα αυτή απεικονίζοναι τα LTR και οι 3 πρωτεΐνες gag, pol και env, καθώς και οι ξεχωριστές περιοχές της pol: PR: πρωτεάση, RT: αντίστροφη μεταγραφάση, RH: RNάση H, IN: ιντεγκράση. Προσαρμογή εικόνας από (Bucheton, 1995)

Ιδιαίτερο ενδιαφέρον παρουσιάζουν τα μεταθετά στοιχεία LTR retrotransposons. Το χαρακτηριστικό αυτών των μεταθετών στοιχείων είναι ότι παρουσιάζουν δομή παρόμοια με αυτών των ρετροϊών (εικόνα 1): αποτελούνται κυρίως από 3 ανοικτά πλαίσια διαβάσματος, τα οποία κωδικοποιούν παρόμοιες πρωτεΐνες με αυτές των ρετροϊών: την πρωτεΐνη gag, την πρωτεΐνη pol, η οποία είναι και η αντίστροφη μεταγραφάση και την πρωτεΐνη env, με την τελευταία να μην περιλαμβάνεται σε ορισμένα μεταθετά στοιχεία της οικογένειας αυτής. Χαρακτηριστικό γνώρισμα της οικογένειας αυτής είναι επίσης τα άκρα των μεταθετών στοιχείων, που περιέχουν

LTR (long terminal repeats), περιοχές όμοιες με αυτές των ρετροϊών, που όμως δεν φαίνεται να παίζουν κανένα ρόλο στη μετάθεση.

Η μετάθεση των στοιχείων αυτών γίνεται με έναν μηχανισμό ενδιάμεσου RNA μορίου, όπως ακριβώς και στους ρετροϊούς. Το DNA του μεταθετού στοιχείου δηλαδή, αντιγράφεται από την αντίστροφη μεταγραφάση σε ενδιάμεσο μόριο RNA, το οποίο στη συνέχεια μετατρέπεται πάλι σε DNA και εντίθεται. Η πολυμεράση της οικογένειας αυτής αποτελείται από πολλές λειτουργικές περιοχές: μια περιοχή με δράση πρωτεάσης, μια περιοχή με δράση αντίστροφης μεταγραφάσης, μια περιοχή με δράση RNάσης, και μια περιοχή με δράση ιντεγκράσης.

Τα μεταθετά στοιχεία αυτά έχουν γίνει αντικείμενο πολλών εξελικτικών μελετών σε αρκετούς οργανισμούς (Goodwin and Poulter, 2000; Jordan and McDonald, 1999; Marillonnet and Wessler, 1998; Springer and Britten, 1993). Επειδή η πολυμεράση των στοιχείων αυτών χρησιμεύει για τον πολλαπλασιασμό τους, και κατά συνέπεια για την μετάθεσή τους, κάθε μεταλλαγή της υπόκειται στον έλεγχο της φυσικής επιλογής. Από τις διάφορες περιοχές της πρωτεΐνης αυτής, η λιγότερο διαφοροποιημένη από πλευράς αμινοξικής αλληλουχίας είναι η αντίστροφη μεταγραφάση, γεγονός αναμενόμενο εάν αναλογιστεί κανείς ότι η περιοχή αυτή επιτελεί την πιο σπουδαία λειτουργία για το στοιχείο.

Τα LTR retrotransposons στη Drosophila melanogaster και τον *Anopheles gambiae*

Τα LTR retrotransposons στη *Drosophila melanogaster* υπολογίζεται ότι αποτελούν το 18% περίπου του γονιδιώματος (Berg and Howe, 1989). Αν και υπάρχουν πολλές οικογένειες LTR retrotransposons στη *Drosophila melanogaster*, οι μεγαλύτερες οικογένειες είναι δυο: α)η οικογένεια *copia,* η οποία είναι όμοια με την οικογένεια *Ty1* στο σακχαρομύκητα (Broach et al., 1992), στοιχεία της οποίας περιέχουν δυο γονίδια που αντιστοιχούν στα gag και pol και β)η οικογένεια gypsy, όμοια με την οικογένεια *Ty3* στο σακχαρομύκητα, αντιπροσωπευτικά στοιχεία της οποία περιέχουν 3 γονίδια που κωδικοποιούν τα gag, pol και env. Άλλες οικογένειες LTR retrotransposons στη *Drosophila* περιλαμβάνουν την οικογένεια των pao στοιχείων, που αρχικά ανακαλύφθηκε στον νηματώδη *Caenorhabditis elegans*, ενώ υπάρχει και μια σειρά στοιχείων, τα οποία σύμφωνα με τη βιβλιογραφία δεν είναι δυνατόν να ενταχθούν σε μια από τις ήδη υπάρχουσες οικογένειες, οπότε συνιστούν καινούργιες οικογένειες.

Στον *Anopheles gambiae* αντίθετα, υπάρχουν λίγα στοιχεία για την ύπαρξη LTR retrotransposons. Μια προσεκτική αναζήτηση στη νουκλεοτιδική βάση δεδομένων του EMBL δίνει μόνο δεκάδες αποτελέσματα. Με την αλληλούχιση, όμως, του γονιδιώματος του *Anopheles gambiae* περίπου 1506 υποψήφια γονίδια φαίνεται ότι ανήκουν πιθανώς σε μεταθετά στοιχεία, από τα οποία ένα ποσοστό αναμένεται να ανήκει στα LTR retrotransposons.

Σκοπός της παρούσας εργασίας

Σκοπός, λοιπόν, της παρούσας εργασίας ήταν ο προσδιορισμός των αλληλουχιών του *Anopheles gambiae* που κωδικοποιούν μεταθετά στοιχεία και ειδικότερα LTR retrotransposons. Ο προσδιορισμός των αλληλουχιών αυτών έγινε με υπολογιστικές μεθόδους, τόσο στο σύνολο των 15189 περίπου ανοικτών πλαισίων διαβάσματος,

όσον και στο συνολικό γονιδίωμα, εκμεταλλευόμενοι πληροφορίες από τα μεταθετά στοιχεία της *Drosophila melanogaster* και των υπόλοιπων οικογενειών των *Drosophilidae,* αλλά και ήδη υπάρχουσες πληροφορίες για τα μεταθετά στοιχεία στον *Anopheles gambiae*. Στη συνέχεια, δημιουργήθηκαν φυλογενετικά δέντρα με τα στοιχεία του *Anopheles gambiae* και με τα στοιχεία της *Drosophila melanogaster*.

Μέθοδοι

Μέθοδοι

Η εύρεση των LTR retrotransposons στο γονιδίωμα του *Anopheles gambiae*

Η ανάλυση για την εύρεση των αλληλουχιών στο γονιδίωμα του *Anopheles gambiae* έγινε με τη χρησιμοποίηση του αλγόριθμου (basic local alignment search tool), χρησιμοποιώντας το πρόγραμμα BLAST 2.0 (Altschul et al., 1990; Altschul et al., 1997), είτε στο NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) είτε σε έναν τοπικό υπολογιστή, όπου είχε εγκατασταθεί το αντίστοιχο πρόγραμμα, διαθέσιμο επίσης από το NCBI (ftp://ftp.ncbi.nih.gov/blast/). Στις όλες τις περιπτώσεις χρησιμοποιήθηκαν οι ισχύοντες (default) παράμετροι του προγράμματος, εκτός κι αν αναφέρεται το αντίθετο. Τα αποτελέσματα από κάθε ανάλυση αποθηκεύτηκαν με τη μορφή απλών αρχείων κειμένου για περαιτέρω επεξεργασία.

Οι αλληλουχίες που χρησιμοποιήθηκαν για να βρεθούν οι αντίστοιχες στο γονιδίωμα του *Anopheles gambiae* ήταν αρχικά νουκλεοτιδικές και πρωτεϊνικές αλληλουχίες από την οικογένεια των *Drosophilidae* (*Drosophila melanogaster* και άλλα είδη) και αργότερα νουκλεοτιδικές αλληλουχίες από κομμάτια γνωστών LTR retrotransposons του *Anopheles gambiae* (οι αλληλουχίες αναφέρονται στους πίνακες στα αποτελέσματα). Οι αλληλουχίες των διάφορων μεταθετών στοιχείων χωρίστηκαν σε οικογένειες, με βάση στοιχεία από την FlyBase και τη βιβλιογραφία. Στην περίπτωση των νουκλεοτιδικών στοιχείων, υπήρχαν αλληλουχίες που αντιστοιχούσαν σε ολόκληρο το μεταθετό στοιχείο, και αλληλουχίες που αντιστοιχούσαν σε περιοχές που κωδικοποιούσαν για κάποια πρωτεΐνη του μεταθετού στοιχείου. Αντίθετα, οι πρωτεϊνικές αλληλουχίες αντιστοιχούσαν σε ολόκληρες ή σε μέρη από τις πρωτεΐνες των μεταθετών στοιχείων.

Όσον αφορά τις νουκλεοτιδικές αναλύσεις, το γονιδίωμα του *Anopheles gambiae* που χρησιμοποιήθηκε αποτελούνταν από τους 8987 σκελετούς, οι οποίοι ήταν διαθέσιμοι στο δικτυακό τόπο του NCBI. Αντίθετα, για τις πρωτεϊνικές αναλύσεις χρησιμοποιήθηκε το πρωτέωμα του *Anopheles gambiae*, έτσι όπως είχε προβλεφθεί από τη Celera και το EBI. Στην περίπτωση αυτή, οι πληροφορίες για τα ανοικτά πλαίσια διαβάσματος αποθηκεύτηκαν σε έναν τοπικό υπολογιστή και χρησιμοποιήθηκαν για τη δημιουργία των 15189 ανοικτών πλαισίων διαβάσματος και τη μετάφρασή τους στις αντίστοιχες πρωτεΐνες.

Η περαιτέρω ανάλυση των αποτελεσμάτων, περιέλαβε τη δημιουργία των κατάλληλων προγραμμάτων, σε γλώσσα προγραμματισμού Perl, για την ανάλυση των αποτελεσμάτων καθώς και χειρωνακτικό (manual) έλεγχο των συστοιχίσεων, για την ονομασία των περιοχών του *Anopheles gambiae* με βάση την ομοιότητά τους με γνωστά μεταθετά στοιχεία.

Φυλογενετική ανάλυση και δημιουργία δέντρων

Το πρόγραμμα Clustal X, εγκατεστημένο σε ένα τοπικό υπολογιστή χρησιμοποιήθηκε για την ευθυγράμμιση (alignment) των αλληλουχιών που προέκυψαν από τα παραπάνω πειράματα και των αλληλουχιών που ήταν ήδη γνωστά από τις βάσεις δεδομένων. Τα κλαδογράμματα έγιναν με τη μέθοδο της ένωσης των γειτόνων και τα αποτελέσματα παρουσιάστηκαν με τα προγράμματα Njplot και Treeview. Στις περιπτώσεις που υπολογίστηκαν οι τιμές bootstrap, οι επαναλήψεις ήταν 1000.


Αποτελέσματα

Αποτελέσματα

Blastn ανάλυση των νουκλεοτιδικών αλληλουχιών των *Drosophilidae* σε ολόκληρο το γονιδίωμα του *Anopheles gambiae*.

Αρχικά, οι νουκλεοτιδικές αλληλουχίες που αντιστοιχούσαν σε μέρος ή σε ολόκληρα τις αλληλουχίες των μεταθετών στοιχείων των *Drosophilidae* χρησιμοποιήθηκαν με το πρόγραμμα blastn στο NCBI (το πρόγραμμα που χρησιμοποιεί νουκλεοτιδικές αλληλουχίες έναντι νουκλεοτιδικών βάσεων δεδομένων ή μιας ομάδας πρωτεϊνικών βάσεων αλληλουχιών) χρησιμοποιώντας ώς βάση δεδομένων τους 8987 σκελετούς του γονιδιώματος του *Anopheles gambiae*.

Τα αποτελέσματα από αυτές τις αναζητήσεις έδειξαν ότι δεν υπήρχε ομοιότητα ανάμεσα στις νουκλεοτιδικές αλληλουχίες των *Drosophilidae* που χρησιμοποιήθηκαν και σε όλο το γονιδίωμα του *Anopheles gambiae*, ούτε στην περίπτωση που χρησιμοποιήθηκαν ολόκληρες οι αλληλουχίες των μεταθετών στοιχείων, ούτε στην περίπτωση των κομματιών τους. (για το λόγο αυτό και δεν παρατίθενται οι αλληλουχίες των *Drosophilidae* που χρησιμοποιήθηκαν).

Blastp ανάλυση των πρωτεϊνικών αλληλουχιών των Drosophilidae στις προβλεπόμενες πρωτεϊνικές αλληλουχίες του Anopheles gambiae

Στη συνέχεια, πρωτεϊνικές αλληλουχίες των *Drosophilidae* χρησιμοποιήθηκαν με το πρόγραμμα *blastp* (το πρόγραμμα που χρησιμοποιεί πρωτεϊνικές αλληλουχίες έναντι μιας πρωτεϊνικής βάσης δεδομένων ή έναντι μιας ομάδας πρωτεϊνικών αλληλουχιών) στις προβλεπόμενες από τη Celera και το EBI πρωτεϊνικές αλληλουχίες του γονιδιώματος του *Anopheles gambiae*.

Από τα αποτελέσματα αυτά (πίνακας 1) γίνεται αντιληπτό ότι οι περισσότερες επιτυχίες (hits) προέρχονται από πρωτεΐνες που κωδικοποιούν την πολυμεράση του μεταθετού στοιχείου, επιτυχίες που φτάνουν τις 600-1000, ενώ αντίθετα σαφώς λιγότερες επιτυχίες παρουσιάζονται σε πρωτεΐνες που κωδικοποιούν την gag. Τέλος, τον μικρότερο αριθμό επιτυχιών, αλλά και το μικρότερο ποσοστό ομοιότητας μεταξύ των δυο συστοιχισμένων αλληλουχιών (οι λεπτομέρειες των συστοιχίσεων δεν παρατίθενται) έδωσαν οι πρωτεΐνες που κωδικοποιούν την πρωτεΐνη env. Μάλιστα, στην περίπτωση των env, η ομοιότητα ήταν τις περισσότερες φορές τόσο μικρή που θα μπορούσε και να αποδοθεί στην τυχαία ομοιότητα μεταξύ δυο κατά τα άλλα μη σχετιζόμενων αλληλουχιών. Όμως, αναλύσεις (λεπτομέρειες των οποίων δεν παρουσιάζονται) εντόπισαν αλληλουχίες που εμφάνιζαν ομοιότητα με πρωτεΐνες pol των *Drosophilidae* πολύ κοντά σε ορισμένες από τις περιοχές που εμφάνιζαν ομοιότητα ομοιότητα με τις env, γεγονός που συνηγορούσε ότι η περιοχή αυτή κωδικοποιούσε μεταθετό στοιχείο και ότι είχαν εντοπιστεί με ανεξάρτητες αναλύσεις οι πρωτεΐνες pol και env του στοιχείου αυτού.

Από τα αποτελέσματα αυτά, και κυρίως από τα αποτελέσματα των pol πρωτεϊνών ήταν σαφές ότι ορισμένες από τις πρωτεΐνες των *Drosophilidae* εμφάνιζαν σημαντική ομοιότητα με την ίδια πρωτεΐνη από το σύνολο των 15189 πρωτεϊνών• δηλαδή, διαφορετικές πρωτεΐνες των *Drosophilidae* είχαν «κτυπήσει» την ίδια πρωτεΐνη του *Anopheles gambiae* με ενδεχομένως διαφορετικό score και διαφορετικό ποσοστό ομοιότητας. Αντίθετα, άλλες πρωτεΐνες των *Drosophilidae* είχαν «κτυπήσει» μόνο μια πρωτεΐνη του *Anopheles gambiae*. Δημιουργήθηκαν λοιπόν τα κατάλληλα προγράμματα ώστε να κατηγοριοποιηθούν οι συστοιχίσεις των blastp αρχείων.

Κωδικός NCBI	όνομα	τύπος	Εείδος	επιτυχίες
οικογένεια copia				
OFFFCP	copia	polyprotein	D. melanogaster	633
AAA03512	copia	rev	D. melanogaster	47
PC1232	copia	polyprotein	D. simulans	283
AAF06364	copia	polyprotein	D. willistoni	139
S00954	1731	pol	D. melanogaster	420
CAA30503	1731	pol	D. melanogaster	420
S00953	1731	gag	D. melanogaster	146
CAA30502	1731	gag	D. melanogaster	146
οικογένεια gypsy				
GNFFG1	gypsy	pol	D. melanogaster	704
FOFFGY	gypsy	gag	D. melanogaster	92
S64733	gypsy	gag	D. subobscura	49
S26839	gypsy	gag	D. virilis	118
S64734	gypsy	pol	D. subobscura	684
S26840	gypsy	pol	D. virilis	653
S26841	gypsy	env	D. virilis	10
S64735	gypsy	env	D. subobscura	7
S72396	gypsy	pol	D. subobscura	690
2208454C	gypsy	env	D. subobscura	7
2208454B	gypsy	pol	D. subobscura	690
2208454A	gypsy	gag	D. subobscura	49
CAA51083	gypsy	gag	D. subobscura	49
CAA51084	gypsy	pol	D. subobscura	690
CAA51085	gypsy	env	D. subobscura	7
T13932	TV1	gag	D. virilis	54
AAC33317	TV1	gag	D. virilis	54
T13935	TV1	env	D. virilis	24
AAC33319	TV1	env	D. virilis	24
T13933	TV1	pol	D. virilis	662
AAC33318	TV1	pol	D. virilis	662
B24872	297	pol	D. melanogaster	771
CAA27159	297	gag	D. melanogaster	40
CAB57796	297	pol	D. melanogaster	771
CAB57797	297	env	D. melanogaster	15
S02021	micropia	pol	D. melanogaster	842
CAA32198	micropia	pol	D. melanogaster	833
CAA80825	tom	env	D. ananassae	13
CAA80824	tom	pol	D. ananassae	703
CAA80823	tom	gag	D. ananassae	80
T13798	mdg3	pol	D. melanogaster	859
GNFF17	17,6	pol	D. melanogaster	731
CAA25701	17,6	gag	D. melanogaster	-
CAA25703	17,6	env	D. melanogaster	-
P10394	412	pol	D. melanogaster	914
CAA27750	412	pol	D. melanogaster	914
CAA08806	idefix	gag	D. melanogaster	56
CAA08807	idefix	pol	D. melanogaster	804
CAA08808	idefix	env	D. melanogaster	16
CAA39967	ulysses	pol	D. virilis	768
S18211	ulysses	pol	D. virilis	768
οικογένεια ninja				

T31674	ninja	pol	D. simulans	845
άλλες οικογένειες				
S52564	osvaldo	pol	D. buzzatii	722
CAB39733	osvaldo	pol	D. buzzatii	813
AAC60519	osvaldo	pol	D. buzzatii	323
CAB39732	osvaldo	gag	D. buzzatii	85
AAC60518	osvaldo	pol	D. buzzatii	71
AAK01366	midline-jumper	pol	D. melanogaster	437
AAD14015	mdg1	pol	D. melanogaster	597
S70430	mdg1	pol	D. melanogaster	805
AAK53387	beagle	gag	D.melanogaster	86
AAK53386	beagle	pol	D.melanogaster	771
AAK52057	springer	gag	D.melanogaster	54
AAK52055	springer	pol	D.melanogaster	673
AAK52056	springer	env	D.melanogaster	4
AAK52060	cruiser	gag	D.melanogaster	81
AAK52058	cruiser	pol	D.melanogaster	770
AAK52059	cruiser	env	D.melanogaster	26
AAF36670	transpac	gag	D.melanogaster	51
AAF36671	transpac	pol	D.melanogaster	738
CAA04049	zam	gag	D.melanogaster	70
CAA04050	zam	pol	D.melanogaster	797
CAA04048	zam	env	D.melanogaster	30

Πίνακας 1. Οι πρωτεΐνες της οικογένειας των *Drosophilidae* που χρησιμοποιήθηκαν για την blastp ανάλυση και οι επιτυχίες που έδωσαν με τις προβλεπόμενες πρωτεΐνες του *Anopheles gambiae*. Τα μεταθετά στοιχεία έχουν διαχωριστεί στις οικογένειες τις οποίες ανήκουν. Στην οικογένεια copia, η πολυπρωτεΐνη αντιστοιχεί σε μια ενιαία μορφή πρωτεϊνών gag και pol, ενώ η rev αναφέρεται στην περιοχή αντίστροφης μεταγραφάσης της pol.

Αρχικά, με την προσεκτική παρατήρηση ενός τυχαίου δείγματος αποτελεσμάτων φάνηκε ότι συστοιχίσεις πρωτεϊνών, κυρίως των pol πρωτεϊνών με bit score μικρότερο του 200 δεν ήταν σημαντικές, με την έννοια ότι η συστοίχιση είτε γινόταν σε μικρή περιοχή μιας μεγάλης πρωτεΐνης, είτε γινόταν σε μικρό ποσοστό, ή και τα δυο, οπότε το αποτέλεσμα αυτό δεν ήταν ενδεικτικό της ομοιότητας μεταξύ των πρωτεϊνών αυτών. Όλες οι συστοιχίσεις μεταξύ 2 πρωτεϊνών που παρουσίαζαν score μεγαλύτερο ή ίσο με 200 καταγράφτηκαν από ένα ειδικό πρόγραμμα που δημιουργήθηκε, με αποτέλεσμα να δημιουργηθεί ένας κατάλογος με τις πρωτεΐνες του *Anopheles gambiae* σε κάθε μια από τις οποίες εμφανιζόταν μια σειρά πρωτεϊνών των *Drosophilidae* με τις οποίες εμφάνισε ομοιότητα. Ο κατάλογος αυτός δηλαδή έδειχνε ποιες πρωτεΐνες του *Anopheles gambiae* «κτυπήθηκαν» από τις πρωτεΐνες των *Drosophilidae* που χρησιμοποιήθηκαν για την ανάλυση.

Μολαταύτα, έγινε αντιληπτό ότι το όριο 200 του bit score απέκλειε από την ανάλυση όλες τις πρωτεΐνες που είχαν εμφανίσει ομοιότητα με τις πρωτεΐνες gag και env των *Drosophilidae*. Για το λόγο αυτό, τα προγράμματα τροποποιήθηκαν ώστε να λαμβάνουν υπόψη το είδος της πρωτεΐνης των *Drosophilidae*, και χρησιμοποιήθηκαν διαφορετικά bit score ως ουδός για τις διαφορετικές πρωτεΐνες. Ύστερα από προσεκτική παρατήρηση των αρχείων με τα αποτελέσματα από τα blastp των πρωτεϊνών gag και env, χρησιμοποιήθηκε ο ουδός του bit score μεγαλύτερου ή ίσου με 110 για την πρωτεΐνη gag και 70 για την πρωτεΐνη env.

Στη συνέχεια, τα προγράμματα αυτά μετατράπηκαν ώστε να διακρίνουν ανάμεσα στις 4 διαφορετικές οικογένειες των μεταθετών στοιχείων και στον τύπο της πρωτεΐνης. Οι κατάλογοι με τα γονίδια από όλα αυτά τα αρχεία ενοποιήθηκαν κατά τέτοιο τρόπο, ώστε κάθε πρωτεΐνη του *Anopheles gambiae* να αναφέρεται με ποιο

τύπο πρωτεΐνης και με ποια μεταθετά στοιχεία από κάθε οικογένεια εμφάνιζε σημαντική ομοιότητα

Τα αποτέλεσμα από αυτήν την ανάλυση ήταν ο εντοπισμός 813 πρωτεϊνών του *Anopheles gambiae*. Στην περίπτωση που η πρωτεΐνη του *Anopheles gambiae* εμφάνιζε ομοιότητα με μόνο μια πρωτεΐνη των *Drosophilidae*, τότε το όνομα της πρωτεΐνης αυτής αποδίδονται και στην πρωτεΐνη του *Anopheles gambiae*. Η συντιπτική πλειοψηφία, όμως, των πρωτεϊνών του *Anopheles gambiae* εμφάνιζαν ομοιότητα με περισσότερες από μια πρωτεΐνες των *Drosophilidae,* με αποτέλεσμα να εξετάζονται χειρωνακτικά όλες οι συστοιχίσεις για το χαρακτηρισμό της πρωτεΐνης.

Από την ανάλυση των αποτελεσμάτων έγινε σαφές ότι η πλειοψηφία των περιπτώσεων (715) εμφάνιζαν ομοιότητα με πρωτεΐνες pol των *Drosophilidae*, γεγονός που ήταν αναμενόμενο, καθώς η pol πρωτεΐνη εμφανίζει την μεγαλύτερη συντήρηση στα LTR retrotransposons, λόγω της λειτουργίας της. Ενώ ορισμένες πρωτεΐνες εμφάνισαν ομοιότητα μόνο μια τις pol πρωτεΐνες των *Drosophilidae* που ανήκαν σε μια οικογένεια, όπως έγινε με τις copia και pao, ε΄νας μεγάλος αριθμό πρωτεϊνών εμφάνιζε ταυτόχρονα ομοιότητα με πρωτεΐνες pol που ανήκαν στην οικογένεια gypsy και σε όλες τις άλλες οικογένειας. Σε αυτές τις περιπτώσεις μάλιστα, η μελέτη των συστοιχίσεων έδειχνε σχεδόν ίδιο ποσοστό ομοιότητας, γεγονός που δυσκόλευσε αρκετά τον χαρακτηρισμό των πρωτεϊνών αυτών.

Τα αποτελέσματα της ανάλυσης, όσον αφορά την ονομασία των 813 πρωτεϊνών φαίνονται στον πίνακα 2. Με βάση αυτά τα δεδομένα, οι περισσότερες πρωτεΐνες που ανιχνεύθηκαν ανήκουν στην οικογένεια των pao μεταθετών στοιχείων, ενώ ακολουθούν στη συνέχεια οι πρωτεΐνες στην οικογένεια gypsy, οι πρωτεΐνες που ανήκουν σε όλες τις άλλες οικογένειες και τέλος οι πρωτεΐνες της οικογένειας copia. Είναι αξιοσημείωτο να αναφερθεί ότι από τις 813 πρωτεΐνες αυτές, οι 713 είχαν ήδη χαρακτηριστεί ως πιθανά μεταθετά στοιχεία, καθώς εμφανίζονταν περισσότερο από μια φορές στο γονιδίωμα του *Anopheles gambiae*, ενώ μια από αυτές τις πρωτεΐνες είχε χαρακτηριστεί ως αλληλουχία από βακτηριακή μόλυνση και οι άλλες 99 αλληλουχίες δεν είχαν χαρακτηριστεί καθόλου.

	Pol	gag+env	σύνολο
οικογένεια copia	**106**	**43**	**149**
copia	74	-	74
1731	32	43	75
οικογένεια gypsy	**210**	**39**	**249**
gypsy	17	9	26
tom	7	-	7
zam	26	16	42
17.6	19	-	19
297	1	5	6
412	32	-	32
tv1	7	1	8
mdg3	76	-	76
micropia	14	-	14
ulysses	1	-	1
idefix	10	8	18
οικογένεια pao	**232**	**0**	**232**
ninja	232	-	232
άλλες οικογένειες	**165**	**16**	**181**

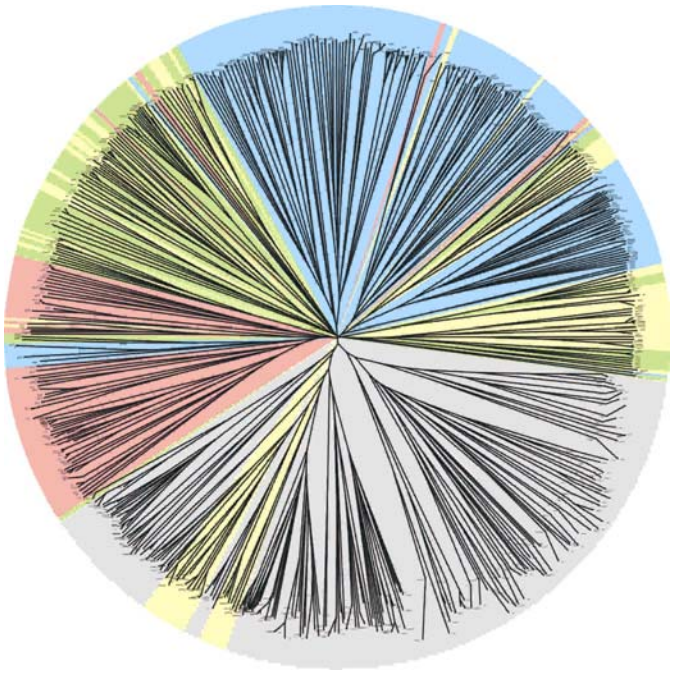
osvaldo	94	7	101
mdg1	18	-	18
transpac	10	-	10
midline-jumper	2	-	2
spinger	1	-	1
cruiser	22	3	25
beagle	18	6	24
Γενικό σύνολο	**713**	**98**	**813**

Πίνακας 2. Τα συγκεντρωτικά αποτελέσματα από την ονομασία των 813 προβλεπόμενων πρωτεϊνών, με βάση τα αποτελέσματα των blastp αναλύσεων.

Χαρακτηριστικό είναι επίσης το γεγονός ότι από αυτές τις 813 διαφορετικές πρωτεΐνες που μελετήθηκαν, μόνο οι 430 βρίσκονται σε σκελετούς που έχουν χαρτογραφηθεί πάνω στο γονιδίωμα του *Anopheles gambiae* ενώ όλες οι άλλες πρωτεΐνες ανήκουν σε σκελετούς που δεν είναι δυνατό να τοποθετηθούν σε μοναδιαίες θέσεις στο γονιδίωμα.

Οι αλληλουχίες από τις 715 πρωτεΐνες που κωδικοποιούσαν pol συστοιχήθηκαν με το πρόγραμμα CLUSTAL X, και δημιουργήθηκε ένα κλαδόγραμμα με τη μέθοδο της ένωσης των γειτόνων, απλοποιημένη έκδοση του οποίου εικονίζεται στην εικόνα 2[2]. Από την εικόνα αυτή είναι φανερό ότι τα μεταθετά στοιχεία της οικογένειας copia και της οικογένειας pao βρίσκοναι στην πλειοψηφία τους σε συγκεκριμένες περιοχές. Όσον αφορά τις πρωτεΐνες της οικογένειας gypsy και όλων των άλλων οικογενειών, αυτές εμφανίζουν μικρότερη ομαδοποίηση, αν και στην περίπτωση των άλλων οικογενειών, η ομαδοποίηση αναφέρεται σε διαφορετικά στοιχεία και όχι στο ίδιο στοιχείο. Το χαρακτηριστικό, όμως, είναι ότι υπάρχει μια μεγάλη περιοχή, στην οποία εμφανίζονται μεταθετά στοιχεία και από τις 4 ομάδες οικογενειών, αν και σε αυτήν την περιοχή παρατηρείται περισσότερο μια εναλλαγή στοιχείων που ανήκουν στην ομάδα gypsy και στοιχείων που ανήκουν σε όλες τις άλλες οικογένειες.

[2] Η εικόνα με την απεικόνιση των 715 διαφορετικών κλαδιών είναι διαθέσιμη σε μέγεθος A0 σε ένα ειδικό αρχείο στο συνοδευτικό CD με αυτήν την εργασία. Στην απλοποιημένη έκδοση, που δεν είναι διακριτά τα ονόματα των διαφορετικών κλαδιών, έχουν χρωματιστεί οι περιοχές στο δέντρο που αντιστοιχούν σε κλαδιά μεταθετών στοιχείων της ίδιας οικογένειας.



Εικόνα 2. Απλοποιημένη έκδοση του κλαδογράμματος των 715 ανοικτών πλαισίων διαβάσματος του *Anopheles gambiae*, οι οποίες χαρακτηρίστηκαν ως pol LTR retrotransposons κατά την ανάλυση. Τα διαφορετικά χρώματα απεικονίζουν περιοχές με συγκέντρωση στοιχείων που ανήκουν στην ίδια οικογένεια. Με γαλάζιο απεικονίζεται η οικογένεια pao, με ρόζ χρώμα η οικογένεια copia, με πράσινο χρώμα η οικογένεια gypsy, με κίτρινο χρώμα οι άλλες οικογένειες. Με γκρι χρώμα απιεκονιζεται η περιοχή που περιέχει στοιχεία και από τις 4 διαφορετικές ομάδες οικογενειών. Το πραγματικό κλαδόγραμμα είναι διαθέσιμο με τη μορφή αρχείου εικόνας .jpeg sto συνοδευτικό CD.

Blastn ανάλυση των νουκλεοτιδικών αλληλουχιών των ειδών του γένους Anopheles στο γονιδίωμα του Anopheles gambiae

Τα παραπάνω αποτελέσματα έδειξαν ότι στο γονιδίωμα του *Anopheles gambiae* περιέχονται και αλληλουχίες μεταθετών στοιχείων, οι οποίες σε πρωτεϊνικό επίπεδο παρουσιάζουν ομοιότητες στην αλληλουχία με αντίστοιχες αλληλουχίες των *Drosophilidae*. Επομένως, έγινε μια προσπάθεια για την εύρεση αυτών των αλληλουχιών σε όλο το γονιδίωμα του *Anopheles gambiae*, στη νουκλεοτιδική του μορφή, με τη χρησιμοποίηση νουκλεοτιδικών αλληλουχιών μεταθετών στοιχείων του *Anopheles*.. Για το σκοπό αυτό χρησιμοποιήθηκε για μια ακόμη φορά το πρόγραμμα blastn στο NCBI, χρησιμοποιώντας ως βάση δεδομένων τους 8987 σκελετούς του γονιδιώματος. Η ανάλυση στην περίπτωση αυτή περιορίστηκε στις πρωτεΐνες χαρακτηρισμένες ως pol για όλα τα μεταθετά στοιχεία, ενώ ως ουδός χρησιμοποιήθηκε το e-value μεγαλύτερο ή ίσο με 10 e^{-10}, για λόγους συνάφειας με παρόμοιες μελέτες που γίνονταν με άλλες οικογένειες μεταθετών στοιχείων για τον *Anopheles gambiae*.

Ως αλληλουχίες για την αναζήτηση χρησιμοποιήθηκαν οι νουκλεοτιδικές αλληλουχίες των μεταθετών στοιχείων του γένους *Anopheles* που υπήρχαν ήδη στις βάσεις δεδομένων και παρουσιάζονται στον πίνακα 3, και αλληλουχίες που προέκυψαν από την πιο πάνω ανάλυση. Συγκεκριμένα, χρησιμοποιήθηκε η νουκλεοτιδική αλληλουχία της μεγαλύτερης σε μέγεθος πρωτεΐνης κάθε μεταθετού στοιχείου, όπως καθορίστηκε από την παραπάνω blastp ανάλυση.

	Κωδικός NCBI
οικογένεια copia	
copia	X02599
1731	X07656
copia	AF295691
mtanga	AF387853
amer3	AJ006554
amer7	AJ006562
οικογένεια gypsy	
gypsy (D.melanogaster)	M12927
gypsy (D. subobscura)	X72390
gypsy (D.virilis)	S26840
tom	Z24451
zam	AJ000387
17.6	X01472
297	X03431
412	X04132
tv1	AF056940
mdg3	X95908
micropia	X14037
idefix	AJ009736
οικογένεια pao	
Aara5	AJ006564
Agam10	AJ006552
άλλες οικογένειες	
osvaldo	AJ133521
transpac	AF222049
midline-jumper	AF315785
spinger	AF364549
cruiser	AF364550
beagle	AJ365402
moose	AF060859

Πίνακας 3. Οι νουκλεοτιδικές αλληλουχίες των μεταθετών στοιχείων του *Anopheles gambiae* που χρησιμοποιήθηκαν κατά την blastn ανάλυση σημειώνονται με μπλε χρώμα και οι νουκλεοτιδικές αλληλουχίες των *Drosophilidae* που χρησιμοποιήθηκαν στη δημιουργία του κλοαδογράμματος σημειώνονται με μαύρο χρώμα.

Η προσέγγιση που ακολουθήθηκε ήταν παρόμοια με αυτήν που περιγράφτηκε παραπάνω. Το αποτέλεσμα ήταν ο χαρακτηρισμός 935 νουκλεοτιδικών, που περιγράφονται στον πίνακα 4. Από την εξέταση του παρακάτω πίνακα, είναι φανερό ότι πολλές από τις νουκλεοτιδικές περιοχές αποδίδονται στις γνωστές αλληλουχίες μεταθετών στοιχείων του *Anopheles*, ενώ όσον αφορά τα στοιχεία που χρησιμοποιήθηκαν και στη προηγούμενη ανάλυση, αυτά εμφανίζουν διαφορετική εικόνα, με μόνο τα μεταθετά στοιχεία mdg1, transpac, beagle, cruiser, gypsy, mdg3 και micropia να εμφανίζουν περίπου ίδιο αριθμό στοιχείων και στις 2 περιπτώσεις.

οικογένεια copia	**92**
copia	22
1731	1
amer3	22
amer7	11
mtanga	36
οικογένεια gypsy	**259**
gypsy	20
tom	26
zam	10
17.6	23
297	40
412	4
tv1	-
mdg3	58
micropia	22
ulysses	11
idefix	45
οικογένεια pao	**151**
ninja	105
Agam10	38
Aara5	8
άλλες οικογένειες	**433**
osvaldo	21
mdg1	15
transpac	10
midline-jumper	14
spinger	127
cruiser	13
beagle	23
Moose	210
Γενικό συνολο	**935**

Πίνακας 4. Τα συγκεντρωτικά αποτελέσματα από την ονομασία των 935 νουκλεοτιδικών περιοχών, με βάση τα αποτελέσματα των blastn αναλύσεων.

Από αυτές τις νουκλεοτιδικές περιοχές, μόνο οι 249 βρίσκονται σε σκελετούς που έχουν χαρτογραφηθεί στο γονιδίωμα του *Anopheles gambiae*, ενώ 238 περιοχές βρίσκονται σε σκελετούς, οι οποίοι δεν μπορούν να αποδοθούν σε μια μοναδική θέση πάνω στο γονιδίωμα. Τα υπόλοιπα στοιχεία αποτελούν τμήματα από contigs, τα οποία δεν έχουν ακόμα τοποθετηθεί σε ένα συγκεκριμένο σκελετό.

Οι νουκλεοτιδικές αλληλουχίες των pol των μεταθετών στοιχείων του *Anopheles gambiae* που δημιουργήθηκαν, συστοιχήθηκαν με τις αντίστοιχες νουκλεοτιδικές αλληλουχίες των *Drosophilidae* (που βρίσκονται επίσης στον πίνακα 3) με το πρόγραμμα CLUSTAL X, και δημιουργήθηκε ένα φυλογενετικό δέντρο με τη μέθοδο της ένωσης των γειτόνων, το οποίο απεικονίζεται στην εικόνα 3. Από το δέντρο αυτό, ενδιαφέρον παρουσιάζουν οι περιπτώσεις των pol αλληλουχιών από τα μεταθετά στοιχεία gypsy και 412 (και τα δυο της οικογένειας gypsy, όπου οι αλληλουχίες από τα δυο ήδη βρίσκονται αρκετά κοντά Αντίθετα, στο υπόλοιπο δέντρο φαίνεται ότι οι αλληλουχίες από τα είδη του γένους *Anopheles* και οι αλληλουχίες των *Drosophilidae* εμφανίζουν σχεδόν διακριτή ομαδοποίηση.

Σύγκριση μεταξύ των αποτελεσμάτων των δυο μεθόδων, των blastp kai blastn

Η διαφορά των αποτελεσμάτων των δυο αναλύσεων ήταν η αιτία για τη σύγκριση των μεθόδων αυτών, ώστε να γίνει αντιληπτό ποια από τα στοιχεία που μελετήθηκαν και στις δυο περιπτώσεις ήταν ίδια και ποια ήταν διαφορετικά. Με άλλα λόγια, αποφασίστηκε η ένωση στοιχείων από τις παραπάνω αναλύσεις ώστε να μελετηθεί το ποσοστό των όμοιων στοιχείων και στις δυο αναλύσεις.

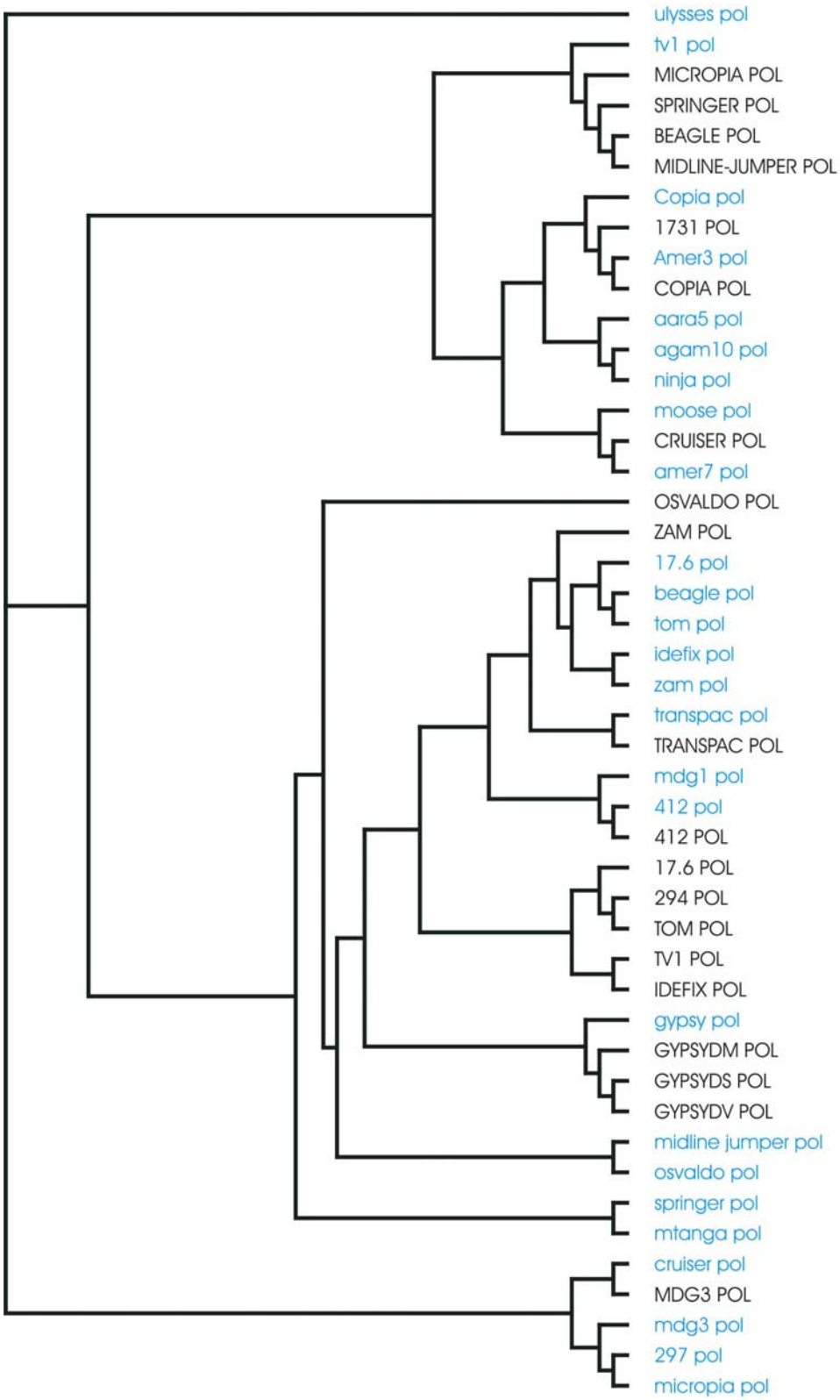
Η μελέτη αυτή πραγματοποιήθηκε με την καταγραφή των συντεταγμένων των στοιχείων από τις δυο μελέτες πάνω στους σκελετούς του *Anopheles gambiae*. Στην περίπτωση των νουκλεοτιδικών αναλύσεων, οι πληροφορίες αυτές βρίσκονται στα αρχεία με τα blastn αποτελέσματα, ενώ στην πρίπτωση των πρωτεϊνικών αναλύσεων, οι πληροφορίες για τις συντεταγμένες των πρωτεϊνών ήταν διαθέσιμες από τα στοιχεία της Celera και του EBI.

Αυτές οι πληροφορίες για τις συντεταγμένες των στοιχείων, όπως είχαν χαρακτηριστεί από τις δυο αναλύσεις συνενώθηκαν, και στη συνέχεια κατηγοριοποιήθηκαν με βάση κάθε σκελετό. Με χειρωνακτική εξέταση διαπιστώθηκε, λοιπόν, εάν υπήρχαν στοιχεία που βρίσκονται στην ίδια ακριβως θέση στο γονιίωμα ή σε γειτονικές θέσεις, και εάν τα στοιχεία αυτά είχαν χαρακτηριστεί με το ίδιο όνομα, με ονόματα μεταθετών στοιχείων που ανήκαν στην ίδια οικογένεια ή με ονόματα που ανήκαν σε διαφορετικές οικογένειες.

γειτονικά στοιχεία	**46**
στοιχεία στην ίδια περιοχή	**81**
στοιχεία στην ίδια περιοχή με το ίδιο όνομα	*37*
στοιχεία στην ίδια περιοχή που ανήκουν στην ίδια οικογένεια	*15*
στοιχεία στην ίδια περιοχή που ανήκουν σε διαφορετικές οικγένειες	*29*
σύνολο	**127**

Πίνακας 5. Πίνακας αποτελεσμάτων κοινών στιχείων από τη σύγκριση των δυο μεθόδων. Στην περίπτβση αυτήν καταγράφτηκαν στοιχεία σε γειτονικές θέσεις πάνω στους σεκλετούς και στοιχεία που βρίσκονταν στις ίδιες θέσεις

Τα αποτελέσματα από την ανάλυση αυτή φαίνονται στον πίνακα 5. Είναι φανερό ότι μόνο ένα μικρός αριθμός στοιχείων και από τις 2 αναλύσεις συμπίπτουν ή βρίσκονται στην ίδια περιοχή του γονιδιώματος του *Anopheles gambiae*, ενώ από αυτά που βρίσκονται στην ίδια περιοχή, το 50% περίπου έχει χαρακτηριστεί με το ίδιο όνομα και στις 2 αναλύσεις.



Εικόνα 3. Το κλαδόγραμμα των νουκλεοτιδικών αλληλουχιών. Με κεφαλαία γράμματα και μαύρο χρώμα απεικονίζονται οι αλληλουχίες προερχόμενες από την οικογένεια των *Drosophilidae*, ενώ με μικρά γράμματα και μπλε χρώμα απεικονίζονται οι αλληλουχίες των ειδών του γένους *Anopheles*. Στην περίπτωση των νουκλεοτιδικών αλληλουχιών gypsy των *Drosophilidae,* οι δείκτες καταδεικνύουν το είδος προέλευσης (DM: *Drosophila melanogaster*, DS: *Drosophila subobscura*, DV: *Drosophila virilis*)

Συζήτηση

Συζήτηση

Η διαδικασία εντοπισμού των αλληλουχιών των LTR retrotransposons στο γονιδίωμα του *Anopheles gambiae* περιέλαβε μια σειρά από υπολογιστικές μεθόδους, και βασίστηκε σε αλληλουχίες από την οικογένεια των *Drosophilidae*. Ο *Anopheles gambiae* αλλά και η *Drosophila melanogaster* είναι δίπτερα, και μάλιστα υπολογίζεται χωρίστηκαν κατά τη διάρκεια της εξέλιξης πριν από 250 εκατομμύρια χρόνια (Yeates and Wiegmann, 1999), κατά τη διάρκεια των οποίων το κουνούπι ανέπτυξε νέα μορφολογικά, φυσιολογικά και ηθολογικά χαρακτηριστικά ώστε να χρησιμοποιεί το αίμα ως τροφή και ως συστατικό ωρίμανσης των αυγών του.

Ανεξάρτητα, όμως, από αυτές τις διαφοροποιήσεις, πολλές είναι εκείνες οι μελέτες που καταδεικνύουν ότι το κουνούπι και η μύγα διατηρούν ορισμένα κοινά χαρακτηριστικά, τόσο στη μορφολογία, όσο και στη γονιδιακή δομή (Bolshakov et al., 2002; Thomasova et al., 2002) αλλά και στην αλληλουχία των πρωτεϊνών τους. Κατά συνέπεια, με βάση την στενή συγγένεια μεταξύ των δυο ειδών αλλά και τις ομοιότητές του, επιλέχθηκαν οι αλληλουχίες από την οικογένεια των *Drosophilidae* για την ανεύρεση των LTR retrotransposons στον *Anopheles gambiae*.

Χαρακτηριστικό είναι το γεγονός ότι σύμφωνα με τα αποτελέσματα αυτής της μελέτης, αλλά και άλλων μελετών για τα άλλα είδη των μεταθετών στοιχείων που έγιναν ανεξάρτητα[3], τα LTR retrotransposons αποτελούν ένα μεγάλο ποσοστό των μεταθετών στοιχείων και περιλαμβάνουν οικογένειες με ένα μεγάλο αριθμό διαφορετικών μεταθετών στοιχείων.

Στην αρχική ανάλυση blastn με τη χρησιμοποίηση νουκλεοτιδικών αλληλουχιών των μεταθετών στοιχείων των *Drosophilidae* δεν υπήρξε κανένα αποτέλεσμα, δηλαδή καμιά ομοιότητα αλληλουχιών. Το γεγονός αυτό, σε συνδυασμό με την περαιτέρω ανάλυση, οδηγεί στο συμπέρασμα ότι τα είδη των *Drosophilidae* διαφοροποιήθηκαν αρκετά από τον *Anopheles gambiae* σε νουκλεοτιδικό επίπεδο, τόσο ώστε να μην εντοπίζεται καμιά ομοιότητα μεταξύ των αλληλουχιών των μεταθετών στοιχείων. Από την άλλη μεριά, τα αποτελέσματα των blastp αναλύσεων καταδεικνύουν ότι αν και οι νουκλεοτιδικές αλληλουχίες έχουν διαφοροποιηθεί σε ένα μεγάλο βαθμό, εντούτοις οι πρωτεϊνικές αλληλουχίες εμφανίζονται διατηρημένες, γεγονός αναμενόμενο, αφού οι νουκλεοτιδικές αλληλουχίες εμφανίζουν μεγαλύτερη ποικιλομορφία, λόγω εκφυλισμού του γενετικού κώδικα. Από την ανάλυση των συστοιχίσεων, οι περιπτώσεις των pol πρωτεϊνών των *Drosophilidae* έδωσαν ομοιότητες σε ένα μεγάλο ποσοστό και οι ομοιτότητες αυτές εντοπίζονταν σε μια μεγάλη περιοχή, ενώ αντίθετα οι ομοιότητες των gag και env ήταν περιορισμένες. Εντούτοις, η εύρεση περιοχών στον *Anopheles gambiae* που εμφάνιζαν ομοιότητα με τις env και η εύρεση γειτονικών περιοχών που εμφάνιζαν ομοιότητα με τις pol, ενίσχυε την άποψη ότι παρά την μικρή ομοιότητα, οι περιοχές αυτές στον *Anopheles gambiae* πράγματι κωδικοποιούσαν για αυτές τις πρωτεΐνες.

Αξίζει να σημειωθεί ότι από τις πρωτεΐνες των LTR retrotransposons, μόνο η πρωτεΐνη pol έχει την πρωταρχική λειτουργία για την μετάθεση του στοιχείου, ενώ οι άλλες δυο ο πρωτεΐνες δε φαίνεται να παίζουν σημαντικό ρόλο. Για το λόγο αυτό, η πρωτεΐνη pol θα πρέπει να υφίσταται αυστηρότερη πίεση της φυσικής επιλογής με αποτέλεσμα να μην διαφοροποιείται με μεγάλο ρυθμό ανάμεσα στα είδη, ενώ αντίθετα οι πρωτεΐνες gag και env θα πρέπει να διαφοροποιούνται ταχύτερα. Μια

[3] Η παρούσα μελέτη των LTR retrotransposons αποτελεί μέρος μιας παγκόσμιας συνεργασίας για τη τον εντοπισμό και τη μελέτη όλων των ομάδων των μεταθετών στοιχείων στο γονιδίωμα του *Anopheles gambiae*.

τέτοια εικόνα παρουσιάζεται και με τα αποτελέσματα των blastp αναλύσεων, από τη στιγμή που οι pol πρωτεΐνες έδωσαν καλύτερες ομοιότητες και περισσότερες επιτυχίες από τις gag και env.

Χαρακτηριστικό είναι, όμως, ότι σε αυτήν την ανάλυση ορισμένες πρωτεΐνες του *Anopheles gambiae* παρουσίαζαν ομοιότητες ταυτόχρονα και με πρωτεΐνες pol της οικεγένεας gypsy και με πρωτεΐνες pol από μεταθετά στοιχεία άλλων οικογενειών. Τα μεταθετά στοιχεία που σε αυτήν την ανάλυση έχουν τοποθετηθεί στην κατηγορία των άλλων οικογενειών είναι στοιχεία για τα οποία, είτε στη βιβλιογραφία δεν υπάρχουν στοιχεία σχετικά με ποια οικογένεια ανήκουν, είτε αναφέρεται στη βιβλιογραφία ότι αποτελούν μια καινούργια οικογένεια. Τις περισσότερες φορές μάλιστα, η ίδια πρωτεΐνη του *Anopheles gambiae* εμφάνιζε σχεδόν το ίδιο ποσοστό ομοιότητας με πρωτεΐνες των δυο αυτών κατηγοριών. Το φαινόμενο αυτό παρατηρείται και στο κλαδόγραμμα των 715 πρωτεϊνών pol που χαρακτηρίστηκαν σε αυτήν την μελέτη. Ενώ οι οικογένειας copia και pao εμφανίζουν μεγάλη συγκέντρωση, δηλαδή συστοιχίζονται μαζί, αυτό συμβαίνει σε μικρότερο βαθμό με τις gypsy και τις άλλες οικογένειες, με στοιχεία από τη μια περίπτωση να ομαδοποιούνται με στοιχεία της άλλης..

Τα αποτελέσματα από την δική μας ανάλυση μας ενδεικνυουν ότι ορισμένα από τα μεταθετά στοιχεία που στη μελέτη μας ανήκουν στις άλλες οικογένειες θα πρέπει να τοποθετηθούν ορθότερα στην οικογένεια gypsy, ή θα να δημιουργηθεί μια καινούργια υπεροικογένεια που να περιλαμβάνει την οικογένεια gypsy και ορισμένα στοιχεία από τις άλλες οικογένειες.

Από τη στιγμή, όμως που η ανάλυση αυτή έγινε στο σύνολο των προβλεπόμενων εκφραζόμενων πρωτεϊνών του *Anopheles gambiae*, δηλαδή σε ένα μέρος του συνολικού γονιδιώματος, ενδεχομένως να υπήρχαν και άλλες αλληλουχίες που να κωδικοποιούσαν για μεταθετά στοιχεία και να μην είχαν περιληφθεί στις προβλεπόμενες πρωτεΐνες. Για το σκοπό αυτό πραγματοποιήθηκε επίσης η blastn ανάλυση, με βάση τη νουκλεοτιδική αλληλουχία των μεγαλύτερων σε μήκος pol πρωτεϊνών από κάθε μεταθετό στοιχείο.

Η ανάλυση blastn περιέλαβε και αλληλουχίες ειδών του γένους *Anopheles,* και έδωσε περισσότερα αποτελέσματα από την ανάλυση των blastp. Η σύγκριση των δυο αναλύσεων έδωσε μεγάλες διαφορές όσον αφορά τον αριθμό των χαρακτηρισμένων στοιχείων. Ένα αίτιο για αυτήν τη διαφορά μπορεί να αναζητηθεί στον τρόπο με τον οποίο έγινε η blastn ανάλυση: η πλειοψηφία των νουκλεοτιδικών αλληλουχιών που χρησιμοποιήθηκαν είχαν προέλθει από τις αντίστοιχες χαρακτηρισμένες μεγαλύτερες σε μέγεθος πρωτεΐνες για κάθε στοιχείο. Αν, λοιπόν, αυτες οι συγκεκριμένες αλληλουχίες αποδόθηκαν σε «λάθος» στοιχείο, τότε όλες οι νουκλεοτιδικές αλληλουχίες που χαρακτηρίστηκαν με βάση αυτήν την αλληλουχία, αποδόθηκαν σε άλλο στοιχείο. Στη σύγκριση των δυο αναλύσεων αυτών επομένως, περισσότερη σημασία παίζει ο συνολικός αριθμός των στοιχείων που εντοπίστηκαν στη μια και την άλλη φορά.

Σημαντικό, όμως είναι το γεγονός ότι από τη σύγκριση των δυο αυτών αναλύσεων μόνο ένας μικρός αριθμός στοιχείων εντοπίστηκε από τις δυο αυτές μεθόδους, ενώ ένας μεγάλος αριθμός στοιχείων δεν ήταν κοινός. Τα αίτια αυτά μπορούν να αναζητηθούν αρχικά στο διαφορετικό τρόπο των αναλύσεων αυτών, δηλαδή στα διαφορετικά προγράμματα, στις διαφορετικές ομάδες αλληλουχιών (data sets) και στις διαφορετικές παραμέτρους ως ουδοί. Στην περίπτωση των πρωτεϊνών, ουσιαστικά χρησιμοποιήθηκε ένα μέρος του γονιδιώματος, αυτό που περιέχει ανοικτά πλαίσια διαβάσματος. Κατά συνέπεια, τμήματα μεταθετών στοιχείων που δεν χαρακτηρίστηκαν ως ανοικτά πλαίσια διαβάσματος δεν εντοπίστηκαν. Από την άλλη

μεριά, πολλές από τις πρωτεΐνες που χαρακτηρίστηκαν από την blastp ανάλυση δεν ανιχνεύθηκαν στην blastn ανάλυση, γεγονός που ενδεχομένως να μπορεί να εξηγηθεί με το μη αντιπροσωπευτικό δείγμα των πρωτεϊνικών αλληλουχιών, από τις οποίες προήλθαν οι αντίστοιχες νουκλεοτιδικές αλληλουχίες για κάθε μεταθετό στοιχείο. Μια μελλοντική ανάλυση θα μπορούσε να χρησιμοποιήσε ένα πλήθος χαρακτηρισμένων πρωτεϊνών από κάθε μεταθετό στοιχείο για να δημιουργήσει τις αντίστοιχες νουκλεοτιδικές αλληλουχίες των στοιχείων αυ΄των και να πραγματοποιήσει την ανάλυση blastn.

Εκτός όμως από τις πρωτεΐνες, ένα από τα χαρακτηριστικά των μεταθετών στοιχείων της ανάλυσης μας είναι και η ύπαρξη των long terminal repeats-LTR στις άκρες τους. Το γεγονός ότι αυτά τα στοιχεία δεν κωδικοποιούν για καμιά πρωτεΐνη εξηγεί το ότι δεν κατέστη δυνατό να καθοριστούν LTR στον *Anopheles gambiae* με βάση αυτά των *Drosophilidae*, αφού η αρχική νουκλεοτιδική ανάλυση blastn δεν έδωσε καμιά επιτυχία. Η ανίχνευση των LTR αυτών μπορεί να γίνει είτε έμμεσα, με την εξέταση των περιοχών δίπλα από τις περιοχές που αναλύθηκαν με τις παραπάνω μεθόδους ή με τη χρησιμοποίηση κατάλληλων προγραμμάτων και ειδικά σχεδιασμένων αλγορίθμων για την ανίχνευση επαναλήψεων.

Η παραπάνω μελέτη, λοιπόν, δείχνει μια αρχική εικόνα για την ύπαρξη των LTR retrotransposons στον *Anopheles gambiae*. Ένας μεγάλος αριθμός των στοιχείων αυτών ανήκει σε σκελετούς που έχουν τοποθετηθεί στο γονιδίωμα του *Anopheles gambiae*, δηλαδή σε μια θέση στο γενετικό χάρτη του *Anopheles gambiae*, ενώ ένας μεγάλος αριθμός έχει αποδοθεί σε contigs ή σε σκελετούς που δεν έχουν ακόμα τοποθετηθεί στο γονιδίωμα του *Anopheles* σε μοναδιαία θέση. Επομένως, η πληροφορία αυτής της μελέτης, σε συνδυασμό με χαρτογραφήσεις in situ των μεταθετών αυτών στοιχείων, αλλά και την πληροφορία από ήδη δημιουργημένους γενετικούς χάρτες θα μπορούσαν βοηθήσουν στην καλύτερη χαρτογράφηση των σκελετών πάνω στο γονιδίωμα. Αναμφισβήτητα, όμως, οι πληροφορίες αυτές θα αποτελέσουν τη βάση για μελλοντικές μελέτες σχετικά με την εξέλιξη των μεταθετών στοιχείων, και την πιθανή οριζόντια γονιδιακή μεταφορά μεταξύ των ειδών.


Ειδική εισαγωγή

Ειδική εισαγωγή

Η ανάγκη για εξειδικευμένες γονιδιακές βάσεις δεδομένων

Οι μαζικές πληροφορίες που έχουν προκύψει από την αλληλούχιση των γονιδιωμάτων την τελευταία δεκαετία φαίνεται πρόκειται να αποτελέσουν το θεμέλιο, στο οποίο θα βασιστεί η επιστημονική έρευνα στο μέλλον. Η ύπαρξη μιας πλούσιας πηγής πληροφοριών θα φανεί πολύτιμη για τους ερευνητές, των οποίων τα αποτελέσματα θα οδηγήσουν σε βελτιωμένες στρατηγικές για τη διάγνωση, θεραπεία και πρόληψη των ασθενειών σε γενετική βάση. Το μέγεθος της πληροφορίας αυτής καθιστά απαραίτητη την οργάνωση της πληροφορίας με τέτοιο τρόπο ώστε να είναι εύκολη η αποθήκευση αλλά και η αναζήτηση• ένα σύστημα βάσης δεδομένων είναι ουσιαστικά ένας τρόπος διαχείρισης της πληροφορίας αυτής. Η πληροφορία από την αλληλούχιση των γονιδιωμάτων, εκτός από ελάχιστες εξαιρέσεις, είναι άμεσα προσβάσιμη στον επιστήμονα από διάφορους οργανισμούς, για παράδειγμα το NCBI (http://www.ncbi.nlm.nih.gov/) και η Ensembl (http://www.ensembl.org/).

Αυτές οι βάσεις δεδομένων, όμως, δεν εμφανίζουν εξειδικευμένες πληροφορίες, για έναν οργανισμό, δηλαδή πληροφορίες πλήν των αλληλουχιών που θα ήταν επίσης σημαντικές για έναν επιστήμονα. Πολλές βάσεις δεδομένων, λοιπόν, δημιουργήθηκαν για να καλύψουν αυτό το κενό, για να προσφέρουν όχι μόνο πληροφορία για τις αλληλουχίες, αλλά και εξειδικευμένες πληροφορίες που αφορούν φαινοτύπους, πειραματικές συνθήκες, στελέχη ορισμένων οργανισμών, και άλλες πληροφορίες που δεν μπορούν να περιληφθούν σε μια αυστηρά βάση δεδομένων αλληλουχιών, όπως είναι οι παραπάνω. Επιπλέον, αυτές οι εξειδικευμένες βάσεις δεδομένων συνήθως διαχειρίζονται από ερευνητές του συγκεκριμένου χώρου, προσθέτοντας επιστημονική αξία στην πληροφορία που εμφανίζεται σε αυτές.Τέτοιες βάσεις δεδομένων, οι οποίες αυξάνονται διαρκώς (Baxevanis, 2002), μπορούν να χρησιμοποιηθούν σε συνδυασμό με τις παραδοσιακές βάσεις δεδομένων για αλληλουχίες για αναλυτικότερη επιστημονική έρευνα.

Το παράδειγμα της FlyBase, της βάσης δεδομένων των Drosophilidae

Ένα από τα πιο επιτυχημένα παραδείγματα εξειδικευμένης βάσης δεδομένων είναι η FlyBase (http://FlyBase.bio.indiana.edu/), η βάση δεδομένων για την οικογένεια των *Drosophilidae*, με περισσότερες πληροφορίες διαθέσιμες για τη *Drosophila melanogaster*.

Οι πληροφορίες που είναι αποθηκευμένες στη FlyBase προέρχονται από μια ποικιλία πηγών: από ορισμένες γονιδιακές και πρωτεϊνικές βάσεις δεδομένων, από την αλληλούχιση του γονιδιώματος της *Drosophila melanogaster* και το annotation των γονιδίων της, από την επιστημονική βιβλιογραφία, από επικοινωνία με μέλη της επιστημονικής κοινότητας. Κάθε πληροφορία που φτάνει στη FlyBase, εισέρχεται μέσα στη βάση δεδομένων συνδεδεμένη με μια βιβλιογραφική αναφορά, είτε πρόκειται για επιστημονικό δημοσιευμένο άρθρο, είτε για αλληλουχία από άλλη βάση δεδομένων είτε ακόμα για ένα απλό άρθρο που έχει δημοσιευτεί σε μια κοινή εφημερίδα. Για αυτό το λόγο η FlyBase θα μπορούσε να θεωρηθεί ως η περιεκτικότερη βάση δεδομένων της οικογένειας των *Drosophilidae*.

Η γονιδιακή πληροφορία που παρέχει η FlyBase περιλαμβάνει πληροφορία για το annotation του γονιδιώματος της *Drosophila melanogaster*, και μάλιστα είναι

επικεφαλής της προσπάθειας που γίνεται από επιστήμονες ανά τον κόσμο για την ανάθεση λειτουργίας σε ανοικτά πλαίσια διαβάσματος που προέκυψαν από τη χαρτογράφηση. Στη συνέχεια, πληροφορία από τη χαρτογράφηση του γονιδιώματος, από τα γνωστά γονίδια, όπως και από άλλες πηγές συνδυάζονται για τη δημιουργία στατικών και αλληλεπιδραστικών γονιδιακών χαρτών του γονιδιώματος.

Η FlyBase περιλαμβάνει πλήθος πληροφοριών: πληροφορίες για γονίδια, αλληλόμορφα, προϊόντα μεταγραφής και μετάφρασης των γονιδίων, πληροφορίες για χρωμοσωμικές ανωμαλίες (chromosomal aberrations), πληροφορίες για ενθέσεις μεταθετών στοιχείων, πληροφορίες για την ύπαρξη στελεχών, την ύπαρξη συγκεκριμένων πλασμιδιακών κατασκευών, τεχνητών μεταθετών στοιχείων, κυτταρογενετικές πληροφορίες, αναφορές για συγκεκριμένα ανατομικά μέρη της *Drosophila melanogaster*, όπως επίσης και πληροφορίες για τα διάφορα αναπτυξιακά στάδια από την εμβρυϊκή και την ενήλικη ζωή της μύγας. Μια από τις σημαντικότερες διεργασίες στην οποία είναι επικεφαλής είναι η προσπάθεια δημιουργίας συγκεκριμένου και τυποποιημένου λεξιλογίου (controlled vocabularies), γεγονός που βοηθά πολύ στην επιστημονική κοινότητα, καθώς τυποποιούνται όροι, οι οποίοι σε αντίθετη περίπτωση θα εκφράζονταν με διαφορετικό τρόπο μεταξύ των επιστημόνων.

Όλη αυτή η πληροφορία της FlyBase είναι αποθηκευμένη μέσα σε ένα αποδοτικό σύστημα σχεσιακής βάσης δεδομένων[4] και παρουσιάζεται στον ερευνητή μέσω σύνδεσης με το διαδίκτυο, με ιστοσελίδες στον παγκόσμιο ιστό (world wide web). Οι πληροφορίες αυτές είναι διαχωρισμένες σε κατηγορίες, ενώ υπάρχει ένα πλήθος προγραμμάτων και δικτυακών εργαλείων που μπορεί να βοηθήσει τον ερευνητή σε ταχύτερη αναζήτηση της πληροφορίας από επιθυμεί να δει, όπως επίσης και την πιο κατανοητή απεικόνιση των πληροφοριών με τη μορφή χαρτών, πινάκων, σχεδιαγραμμάτων.

Η AnoDB, η βάση δεδομένων των ειδών του γένους Anopheles

Ακολουθώντας το παράδειγμα της FlyBase και άλλων εξειδικευμένων βάσεων δεδομένων για οργανισμούς, δημιουργήθηκε η AnoDB, μια βάση δεδομένων για όλους οργανισμούς του γένους *Anopheles*.

Οι περισσότερες πληροφορίες που περιλαμβάνει η AnoDB αφορούν τον *Anopheles gambiae*, καθώς αποτελεί το είδος με το μεγαλύτερο ερευνητικό ενδιαφέρον. Η AnoDB, σε σύνδεση με άλλες βάσεις δεδομένων, περιέχει πληροφορίες σχετικά με τα γονίδια και τα προϊόντα των γονιδίων τους, πληροφορίες για μεταλλαγές αλληλομόρφων ορισμένων γονιδίων, για ύπαρξη μοριακών μαρτύρων, όπως STSs, ESTs πληροφορίες σχετικά με κυτταρογενετικά στοιχεία (in situ hybridisations) και κυτταρογενετικούς χάρτες.

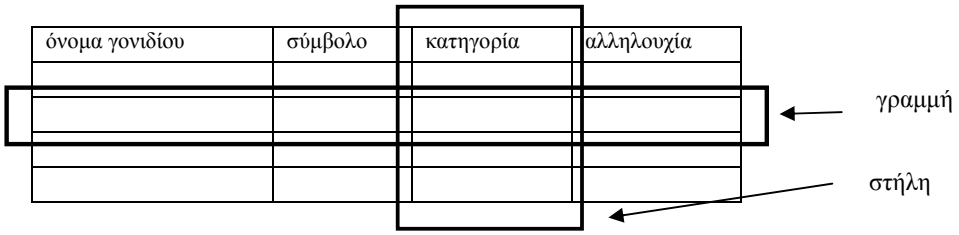
Η πληροφορία της AnoDB είναι προσβάσιμη στο διαδίκτυο, μέσω ιστοσελίδων στον παγκόσμιο ιστό. Οι πληροφορίες αυτές είναι επίσης κατηγοριοποιημένες, ενώ υπάρχουν κατάλληλα εργαλεία για την ταχύτερη αναζήτηση των πληροφοριών που υπάρχουν μέσα στην βάση δεδομένων.

Εισαγωγή στο σύστημα της σχεσιακής βάσης δεδομένων

[4] Περισσότερες πληροφορίες για τη σχεσιακή βάση δεδομένων αναφέρονται στη συνέχεια

Ο τρόπος με τον οποίο αποθηκεύονται τα δεδομένα μέσα στην AnoDB μέχρι αυτή τη στιγμή είναι σε μεγάλα αρχεία κειμένου, όπου με κατάλληλες εντολές γίνεται η αναζήτησή τους. Σε τακτά χρονικά διαστήματα, τα αρχεία αυτά επαναδημιουργούνται με ειδικά προγράμματα, ώστε να προστεθούν σε αυτά όλες οι νέες πληροφορίες που παρουσιάζονται στις βάσεις δεδομένων. Σχετικά πρόσφατα υιοθετήθηκε και το σύστημα της AceDB, το οποίο αναπτύχθηκε και χρησιμοποιήθηκε με επιτυχία στη βάση δεδομένων για τον νηματώδη *Caenorhabditis elegans*. Παρόλα αυτά, η τεράστια πληροφορία που επρόκειτο να παραχθεί από την αλληλούχιση του γονιδιώματος του *Anopheles gambiae* σε συνδυασμό με τη συσσωρευμένη πληροφορία που έχει δημοσιευτεί σε επιστημονικά περιοδικά θα καθιστούσε τη λειτουργία της AnoDB με τη σημερινή της μορφή σχετικά προβληματική, για λόγους που θα αναπτυχθούν παρακάτω. Η διατήρηση της αποτελεσματικότητας της βάσης αυτής, σε συνδυασμό με την καλύτερη διαχείριση των δεδομένων θα μπορούσε να επιτευχθεί μόνο με την υιοθέτησης μιας δομής σχεσιακής βάσης δεδομένων.

Μια σχεσιακή βάση δεδομένων (relational database) είναι η βάση δεδομένων που απεικονίζει και αποθηκεύει τα στοιχεία σε δισδιάστατους πίνακες, πίνακες που περιέχουν γραμμές και στήλες (εικόνα 4). Κάθε στήλη περιέχει επιμέρους στοιχεία τα οποία ανήκουν στην ίδια κατηγορία, ενώ κάθε γραμμή αποτελεί μια ενότητα στοιχείων, τα οποία ανήκουν φυσικά σε διαφορετικές κατηγορίες. Οι σχεσιακές βάσεις δεδομένων δεν αποτελούνται σχεδόν ποτέ από έναν πίνακα, αλλά από ένα σύνολο τέτοιων πινάκων. Η πραγματική ισχύς των σχεσιακών βάσεων δεδομένων στηρίζεται στο γεγονός ότι δεδομένα από ένα πίνακα μπορούν να συνδυαστούν με δεδομένα ενός άλλου πίνακα, υπό κατάλληλες προϋποθέσεις, ώστε να δημιουργήσουν έναν καινούργιο προσωρινό πίνακα. Στην περίπτωση των βιολογικών δεδομένων αυτό μπορεί να γίνει κατανοητό με το ακόλουθο «απλουστευτικό» παράδειγμα: πληροφορίες από έναν πίνακα που περιέχει δεδομένα για γονίδια θα μπορούσαν να συνδυαστούν με πληροφορίες από έναν άλλο πίνακα που περιέχει δεδομένα για τα αντίστοιχα αλληλόμορφά τους.



Εικόνα 4. Σχηματική παράσταση ενός πίνακα μιας σχεσιακής βάσης δεδομένων αποτελούμενος από τις γραμμές και στήλες

Το γεγονός αυτό επιτυγχάνει καλύτερη οργάνωση των στοιχείων αυτών αλλά, επίσης, αποφεύγει κατά πολύ την πλεονάζουσα πληροφορία, δηλαδή ίδια πληροφορία που θα έπρεπε να γραφτεί περισσότερες από μια φορές στη βάση των δεδομένων, εάν δεν ήταν δυνατή η σύνδεση των πινάκων αυτών. Το αποτέλεσμα θα ήταν η κατακόρυφη αύξηση του μεγέθους των αρχείων που θα ήταν αποθηκευμένες οι πληροφορίες στη βάση δεδομένων, όπως επίσης η ελάττωση της ταχύτητας και της αποτελεσματικότητας των εργαλείων αναζήτησης των δεδομένων, αφού αυτά εξαρτώνται από το μέγεθος των αρχείων.

Η σύνδεση των πινάκων μεταξύ τους επιτυγχάνεται με την ύπαρξη όμοιων στηλών στους πίνακες, όπου ο ένας παραπέμπει σε στοιχεία του άλλου (εικόνα 5). Στο προηγούμενο παράδειγμα των γονιδίων και των αλληλομόρφων, η ύπαρξη μιας

στήλης με τα ονόματα των γονιδίων στον πίνακα των αντίστοιχων αλληλομόρφων θα ήταν αρκετή για τη σύνδεση-συσχέτιση των δυο πινάκων. Οι σχέσεις μεταξύ των πινάκων μπορεί να είναι τριών ειδών: α) ένας προς ένα, κατά την οποία η κάθε γραμμή του ένα πίνακα αντιστοιχεί σε κάθε μια γραμμή του άλλου πίνακα και αντίστροφα β) ένας προς πολλά, όπου η κάθε γραμμή του ένα πίνακα αντιστοιχεί σε μια γραμμή του άλλου, αλλά ο δεύτερος μπορεί να περιέχει περισσότερες από μια αντίστοιχες γραμμές του πρώτου και γ)πολλά προς πολλά, όπου όλες οι γραμμές του πρώτου πίνακα αντιστοιχίζονται με όλες τις γραμμές του δευτέρου πίνακα.

όνομα γονίδίου	σύμβολο γονιδίου	κατηγορία
Notch	N	κανονικό γονιδιο
armadillo	arm	κανονικό γονίδιο
delta	D	κανονικό γονίδιο

όνομα πρωτεΐνης	όνομα προερχόμενου γονιδίου	σύμβολο πρωτεΐνης
-	en	$en^{+}P552$
-	armadillo	arm-P1

Εικόνα 5. Σχηματική παράσταση του συσχετισμού δυο πινάκων μέσα στη βάση δεδομένων. Ο συσχετισμός της μιας γραμμής από τον ένα πίνακα με την γραμμή του άλλου πίνακα επιτυγχάνεται με την ύπαρξη όμιοων στηλών στους δυο πίνακδες, και με την ύπαρξη όμοιων στοιχείων μέσα σε αυτούς.

Η αποθήκευση αλλά και η ανάκτηση των δεδομένων από τους πίνακες αυτούς γίνεται με τη βοήθεια ενός συνόλου εκφράσεων, ενός τυποποιημένου λεξιλογίου που στην περίπτωσή μας ονομάζεται SQL (structured query language). Με τη χρήση αυτής της γλώσσας, ο χρήστης απαλλάσσεται από τη χρονοβόρα σύνταξη πολύπλοκων εκφράσεων για την ανάκτηση δεδομένων, γεγονός που δεν θα μπορούσε να αποφύγει στην περίπτωση της μη χρησιμοποίησης μιας σχεσιακής βάσης δεδομένων. Παράλληλα, οι εσωτερικές διεργασίες από τη σχεσιακή βάση δεδομένων για την ανάκτηση τους είναι δραματικά λιγότερο χρονοβόρες από αυτές που θα χρησιμοποιούσε κανείς στην περίπτωση της ύπαρξης τεράστιων αρχείων κειμένου που θα είχαν αποθηκεύσει τις πληροφορίες.

Ιδιαίτερα στην περίπτωση των βιολογικών δεδομένων, η εφαρμογή των βάσεων δεδομένων θα είχε ιδιαίτερο νόημα, αν αναλογιστεί κανείς το πλήθος των σχετιζόμενων αντικειμένων: η συσχέτιση γονιδίων, αλληλομόρφων, μεταγράφων, πρωτεϊνών, η ένθεση μεταθετών στοιχείων με την ύπαρξη χρωμοσωμικών ανωμαλιών, ο φαινότυπος σε συνδυασμό με το γονότυπο, την ύπαρξη μεταλλάξεων κλπ. Η χρήση μιας τέτοια βάσης δεδομένων θα συντελούσε στην ορθότερη αποθήκευση των δεδομένων, με τη λιγότερη δυνατή πλεονάζουσα πληροφορία, την ευκολότερη διαχείριση ενός μεγάλου όγκου δεδομένων, όπως τέλος στην ταχύτερη ανάκτησή τους, εργασίες που στην περίπτωση των μη σχεσιακών δεδομένων, όπως της AnoDB με τη σημερινή μορφή, θα ήταν αρκετά δύσκολο να επιτευχθούν.

Σκοπός της παρούσας εργασίας

Η επιτυχημένη εφαρμογή της σχεσιακής βάσης δεδομένων στην περίπτωση της *Drosophila melanogaster* με τη FlyBase, όπως επίσης και ο μεγάλος όγκος της πληροφορίας που έχει προκύψει από την αλληλούχιση του γονιδιώματος του *Anopheles gambiae* έκανε επιτακτική την υιοθέτησή της στην περίπτωση της AnoDB. Παράλληλα, επιτακτική ήταν και η ανάγκη ενσωμάτωσης της πληροφορίας από επιστημονικές δημοσιεύσεις στη ήδη υπάρχουσα πληροφορία στην AnoDB.

Σκοπός, λοιπόν, αυτής της εργασίας είναι η περιγραφή της εσωτερικής δομής της σχεσιακής βάσης δεδομένων AnoBase (όπως έχει ήδη ονομαστεί σε άμεση αντιστοιχία με τη FlyBase), όπως και των προγραμμάτων για την εισαγωγή των δεδομένων μέσα στη βάση και των προγραμμάτων ανάκτησης των δεδομένων και παρουσίασης τους στον ερευνητή, μέσα από τις ιστοσελίδες του διαδικτύου.


εσωτερική δομή της Anobase

Εσωτερική δομή της AnoBase

Η δημιουργία της σχεσιακής βάσης δεδομένων AnoBase στηρίχθηκε κατά πολύ στη δομή της FlyBase, επειδή η τεχνογνωσία που δημιουργήθηκε κατά την δημιουργία και διαχείριση της FlyBase θα μπορούσε να εφαρμοστεί και σε συγγενικούς οργανισμούς όπως είναι τα είδη του γένος *Anopheles*. Κατά συνέπεια, σε όλα τα στάδια της δημιουργίας της AnoBase ακολουθήθηκε η δομή της FlyBase, αν και έγιναν ορισμένες διαφοροποιήσεις, όπου αυτό κρίθηκε αναγκαίο.

Επιλογή του σχεσιακού συστήματος διαχείρισης της AnoBase

Για την δημιουργία της σχεσιακής βάσης δεδομένων AnoBase, το απαραίτητο πρώτο βήμα ήταν η επιλογή του κατάλληλου συστήματος διαχείρισης. Για λόγους που είχαν να κάνουν με την αξιοπιστία των βάσεων δεδομένων, με την ευκολία διαχείρισης, την ταχύτητα αναζήτησης, την ασφάλεια και το κόστος επιλέχθηκε η MySQL (MySQL Consortium) ως σύστημα διαχείρισης της βάσης δεδομένων. Ας σημειωθεί ότι ορισμένά άλλα εμπορικά συστήματα διαχείρισης, όπως αυτό της Oracle (Sun Microsystems) ήταν επίσης αποτελεσματικά, αλλά το κόστος απόκτησης τους ήταν απαγορευτικό για το για το εγχείρημα αυτό. Η MySQL ανήκει στο προϊόντα ανοικτού κώδικα (Open Source products), που σημαίνει ότι η διάθεσή τους είναι χωρίς κόστος, αλλά, και το σημαντικότερο, οι χρήστες μπορούν ελεύθερα να μετατρέψουν τον πηγαίο κώδικα του προγράμματος ώστε να προσαρμόσουν της λειτουργίες της MySQL στις ανάγκες τους.

Τη στιγμή που γραφόταν αυτή η εργασία, η FlyBase είχε επιλέξει ως σύστημα διαχείρισης τη Sybase (Sybase Inc), αλλά όσον αφορά την παρουσίαση των στοιχείων που είχαν σχέση με το annotation της *Drosophila melanogaster* χρησιμοποιούνταν η MySQL. Μελλοντικά σχέδια της FlyBase περιλαμβάνουν την αναδιοργάνωση της δομής της βάσης δεδομένων και την υιοθέτηση της PostgreSQL(PostgreSQL Consortium), ως πρόγραμμα διαχείρισης της βάσης δεδομένων (προσωπική επικοινωνία).

Στην περίπτωση της AnoBase, η χρησιμοποίηση της MySQL εμφάνιζε πολλά πλεονεκτήματα. Εκτός από τη δυνατότητα τροποποίησης του πηγαίου κώδικα του προγράμματος, η MySQL δεν παρέχει προγράμματα για την είσοδο των δεδομένων, όπως επίσης και για την παρουσίαση των δεδομένων με τη μορφή ιστοσελίδων στον παγκόσμιο ιστό. Το γεγονός αυτό παρείχε μια ευελιξία στην περίπτωσή μας, διότι μπορούσαν να χρησιμοποιηθούν διάφορες άλλες γλώσσες προγραμματισμού, ώστε να φτιαχτούν προγράμματα επικοινωνίας με τη βάση δεδομένων, όπως επίσης προγράμματα για την είσοδο των δεδομένων και προγράμματα για την παρουσίασή τους στο διαδίκτυο.

Η έκδοση της MySQL που χρησιμοποιήθηκε κατά την ανάπτυξη της AnoBase ήταν η έκδοση 2.23 και χρησιμοποιήθηκε τόσο σε περιβάλλον Linux (έκδοση της Red Hat Inc.) και σε περιβάλλον Unix (Solaris 5.8 της Sun Microsystems). Για την δημιουργία των προγραμμάτων εισαγωγής και παρουσίασης των δεδομένων, χρησιμοποιήθηκε πληθώρα προγραμμάτων σε γλώσσες προγραμματισμού Perl (έκδοση 5.6.1), HTML (έκδοση 4.0 ή μεταγενέστερες), Java (έκδοση 4.0) και PHP (έκδοση 4.0). Η ανάπτυξη των προγραμμάτων σε αυτές τις γλώσσες προγραμματισμού έγινε με τη βοήθεια αρκετών συγγραμμάτων (Christiansen and Torkington, 1998; Descartes and Bunce, 2000; DuBois, 1999; Friedl, 1997; Orwant et al., 1999; Schwartz and Christiansen, 1997; Srinivasan, 1997; Wall et al., 1996;

Welling and Thomson, 2001; Yarger et al., 1999) και με πολύτιμες πληροφορίες από ομάδες συζητήσεων στο διαδίκτυο.

Γενική δομή της AnoBase

Η βάση δεδομένων AnoBase, όπως ακριβώς και η FlyBase, βασίζεται σε δημοσιεύσεις στοιχείων, όπου κάθε δημοσίευση αποτελεί μια καταχώριση μέσα στη βάση δεδομένων. Ο όρος δημοσίευση, όμως, δεν περιορίζεται στα στενά όρια της επιστημονικής δημοσίευσης σε βιβλία και περιοδικά, αλλά επεκτείνεται για να συμπεριλάβει όλα εκείνα τα στοιχεία που παράγονται από την επιστημονική έρευνα• περιλαμβάνει δηλαδή στοιχεία από αποτελέσματα συνεδρίων, από οπτικοακουστικά μέσα, από πανεπιστημιακές εργασίες (διδακτορικές διατριβές κ.α.) και πληροφορίες από προσωπική επικοινωνία. Όλες αυτές οι πληροφορίες αιχμαλωτίζονται από ειδικές λίστες που εμφανίζουν όλα εκείνα τα πεδία που πρέπει να συμπληρωθούν για τη συγκεκριμένη δημοσίευση, οι οποίες ονομάζονται «προφόρμες». Στην προφόρμα που υπάρχει για την καταγραφή των πληροφοριών σχετικά με την δημοσίευση προστίθενται στη συνέχεια ένας αριθμός διαφορετικών προφορμών, οι οποίες δημιουργήθηκαν για να καταγράψουν πληροφορίες σχετικά με τα γονίδια, τα αλληλόμορφα, τα μετάγραφα, τις πρωτεΐνες, τις χρωμοσωμικές ανωμαλίες, τα μεταθετά στοιχεία, τις ενθέσεις των μεταθετών στοιχείων στο γονιδίωμα, τη δημιουργία πλασμιδιακών κατασκευών, την ύπαρξη κυτταρογενετικών δεδομένων και πλήθος άλλων πληροφοριών. Μέχρι στιγμής, είχαν δημιουργηθεί οι ακόλουθες προφόρμες για την εισαγωγή δεδομένων στη βάση[5,6]:

- publication(publn) proforma: για την εισαγωγή πληροφοριών που έχουν να κάνουν με τη δημοσίευση, την προέλευση των πληροφοριών που πρόκειται να χρησιμοποιηθούν
- gene proforma: για πληροφορίες που έχουν να κάνουν με την ύπαρξη γονιδίων, όπου στην περίπτωση αυτή γονίδιο δεν θεωρείται μόνο ένα κομμάτι DNA που κωδικοποιεί για μια πρωτεΐνη, αλλά και μεταθετά στοιχεία, χιμαιρικά γονίδια, ιικά παθογόνα, μικροδορυφορικές αλληλουχίες, ψευδογονίδια, μιτοχονδριακά, πυρηνικά γονίδια και επαναλαμβανόμενες αλληλουχίες
- gmods proforma: για πληροφορίες που έχουν να κάνουν με ειδικές λεπτομέρειες ορισμένων γονιδίων, όπως πληροφορίες σχετική αφθονία των μεταθετών στοιχείων στο γονιδίωμα, και όρους γενετικής οντολογίας (gene ontology terms (Ashburner et al., 2000))
- allele proforma: για πληροφορίες αλληλομόρφων συγκεκριμένων γονιδίων
- transcript(transcr) proforma: για πληροφορίες προϊόντων RNA.
- protein proforma: για πληροφορίες πρωτεϊνικών προϊόντων γονιδίων.
- Aberration(aberr) proforma: για πληροφορίες σχετικά με την ύπαρξη χρωμοσωμικών ανωμαλιών

[5] Οι ονομασίες των προφορμών αναφέρονται στην αγγλική γλώσσα, καθώς παρουσιάζουν άμεση σχέση με τους πίνακες στη βάση δεδομένων, οι οποίο θα αναφερθούν αργότερα.

[6] Σε αυτήν την εργασία αναφέρεται περιληπτικά το είδος τον πληροφοριών που εισάγονται με κάθε proforma. Για περισσότερες πληροφορίες σχετικά με το είδος των στοιχείων που αιχμαλωτίζονται σε μια προφόρμα, υπάρχει διαθέσιμο το βιβλίο οδηγιών χρήσης σχετικά με την είσοδο δεδομένων της AnoBase.

- balancer proforma: για πληροφορίες σχετικά με την ύπαρξη ισορροπιστών (balancers)
- construct/transposable element(contra) proforma: για πληροφορίες πλασμιδιακών και άλλων κατασκευών, όπως επίσης και για τεχνητά μεταθετά στοιχεία.
- molecular segment(moseg) proforma: για πληροφορίες σχετικά με τα διάφορα στοιχεία, διάφορες περιοχές σε μια κατασκευή ή ένα τεχνητό μεταθετό στοιχείο
- junction segment(juncseg) proforma: για πληροφορίες σχετικά με περιοχές συμβολής (junctions) δυο διαφορετικών στοιχείων σε μια κατασκευή ή ένα τεχνητό μεταθετό στοιχείο, για παράδειγμα την ύπαρξη περιοριστικών θέσεων κλπ.
- molecular segment kilobase(mkilobase) proforma: για πληροφορίες σχετικά με μοριακούς χάρτες κατασκευών και τεχνητών μεταθετών στοιχείων
- kilobase proforma: για πληροφορίες που αφορούν μοριακούς χάρτες γονιδίων και άλλων στοιχείων.
- transposon insertion(ti) proforma: για πληροφορίες σχετικά με την ένθεση μεταθετών στοιχείων στο γονιδίωμα.
- in-situ proforma: για πληροφορίες σχετικά με κυτταρογενετικά στοιχεία και αναλύσεις in situ.
- species/organelle/transposable element (specorte) proforma: για πληροφορίες σχετικά με άλλα είδη, με οργανίδια και μεταθετά στοιχεία άλλων ειδών.
- Populational(pop) proforma: για πληροφορίες σχετικά με τις σχετικές αφθονίες γονιδίων και χρωμοσωμικών δεικτών σε συγκεκριμένες γεωγραφικές περιοχές

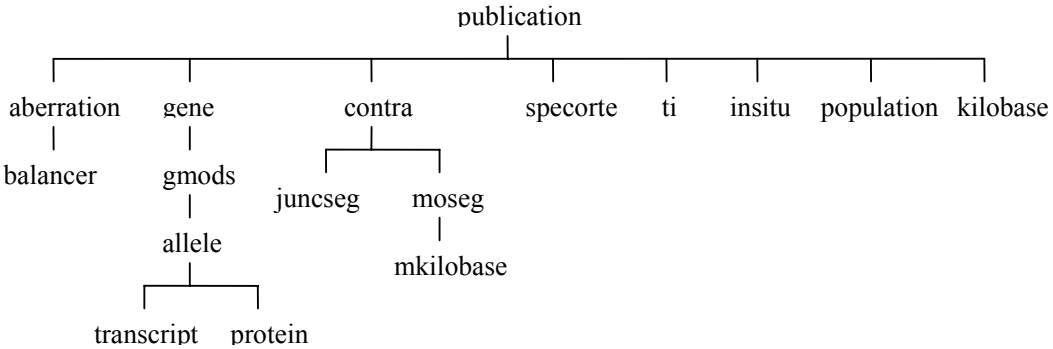
Η οργάνωση των προφορμών αυτών ακολουθεί ένα ιεραρχικό σχήμα, το οποίο εικονίζεται στην εικόνα 6. Η προφόρμα για τη δημοσίευση αποτελεί την πρώτη προφόρμα που χρησιμοποιείται για τη δημιουργία μιας νέας καταχώρισης μέσα στη βάση δεδομένων. Στη συνέχεια, σε αυτήν προστίθονται διάφορες προφόρμες ανάλογα με το είδος τον δεδομένων. Στην περίπτωσή, όμως, που υπάρχουν πληροροφορίες για ένα αλληλόμορφο, η πληροφορία αυτή δεν έχει νόημα εάν δεν περιληφθεί οι πληροφορίες που αναφέρονται στο γονίδιο του αλληλομόρφου. Κατά συνέπεια, όπως είναι φανερό και από τον πίνακα, για να χρησιμοποιηθούν ορισμένες προφόρμες είναι απαραίτητο να έχει προηγηθεί η συμπλήρωση πληροφοριών σε άλλες προφόρμες, οι οποίες προπορεύονται στην ιεραρχία.

Με τον τρόπο αυτό είναι δυνατό επίσης να αιχμαλωτιστούν πληροφορίες για πολλαπλά στοιχεία, χωρίς να υπάρξει κίνδυνος ανάμειξης πληροφοριών από το ένα στοιχείο στο άλλο. Στην περίπτωση αυτή, όπου κριθεί αναγκαίο χρησιμοποιούνται περισσότερες από μια προφόρμες, οι οποίες διατηρούν καθαρά ιεραρχικη δομή. Για παράδειγμα εάν σε μια δημοσίευση υπάρχουν πληροφορίες σχετικά με την ύπαρξη δυο γονιδίων, και το κάθε ένα από αυτά έχει δυο αλληλόμορφα, οι πληροφορίες θα αιχμαλωτιστούν από 7 προφόρμες, με την προφόρμα της δημοσίευσης να ακολουθείται από τη μια προφόρμα του γονίδιου και τις δυο αντίστοιχες προφόρμες των αλληλομόρφων του, και στη συνέχεια την προφόρμα του άλλου γονιδίου και των αλληλομόρφων του.

Ο τρόπος με τον οποίο οργανώνεται εσωτερικά η βάση δεδομένων είναι παρόμοιος με αυτόν της οργάνωσης εισόδου της πληροφορίας. Υπάρχει σχεδόν ένα

πίνακας για κάθε διαφορετική προφόρμα,ο οποίο έχει ονομαστεί με βάση την προφόρμα. (εικόνα 7). Βέβαια υπάρχουν και οι εξαιρέσεις σε αυτόν τον κανόνα. Στοιχεία από τις προφόρμες gene και gmods αποθηκεύονται στον πίνακα gene στη βάση δεδομένων, ενώ τα στοιχεία της mkilobase αποθηκεύονται στον πίνακα kilobase. Επίσης, υπάρχουν 4 πίνακες, οι οποίοι δεν αντιστοιχούν σε κάποια προφόρμα· οι πίνακες journal, photomap, division και graphic. Ο πρώτος χρησιμοποιείται για την αποθήκευση λετομερειών που έχουν να κάνουν με τα περιοδικά στα οποία δημοσιεύονται οι επιστημονικές εργασίες, ενώ οι τρεσι τελευταίοι σχετίζονται με εικόνας χαρτών που αναφέρονται στις περιπτώσεις των υβριδοποιήσεων in situ.

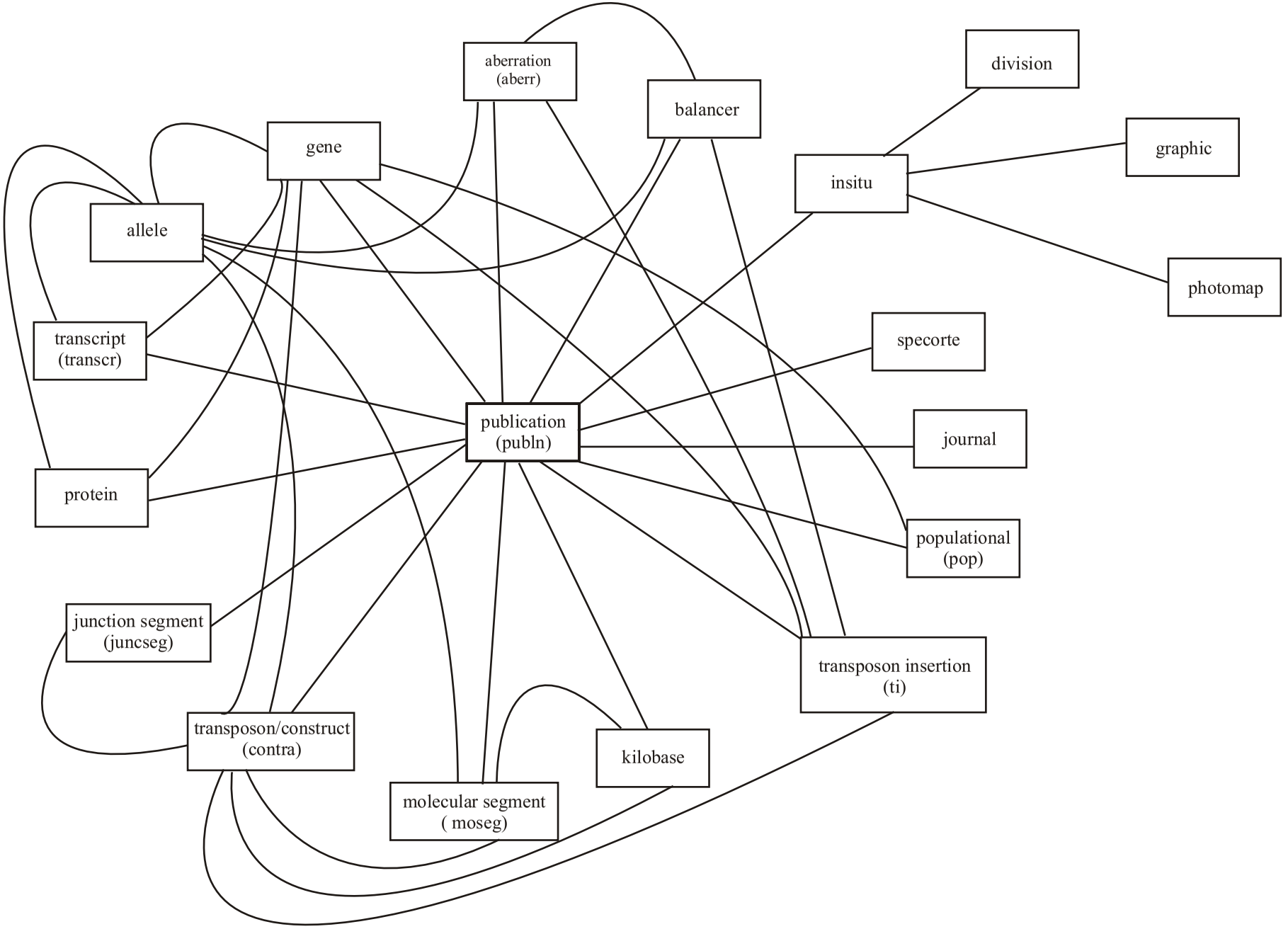
Με βάση τις παραπάνω πληροφορίες σχετικά με τις προφόρμες είναι φανερό ότι αν και τα δεδομένα καταγράφονται σε διαφορετικές προφόρμες, εντούτοις σχετίζονται μεταξύ τους. Αρχικά υπάρχει η ιεραχικη δομή ορισμένων προφορμών, ώστε ορισμένα δεδομένα σχετίζονται γραμμικά, για παράδειγμα το γονίδιο με τα αλληλόμορφά του και τα προϊόντα του, ή η πλασμιδιακή κατασκευή με τα ξεχωριστά της στοιχεία. Όμως, ο τρόπος συσχέτισης των δεδομένων μπορεί να γίνει και ανάμεσα σε διαφορετικές προφόρμες. Για παράδειγμα, πληροφορίες για ένα μεταθετό στοιχείο και τις ενθέσεις του μπορούν να καταγραφούν σε πολλές προφόρμες: στην προφόρμα των γονιδίων (gene proforma), στην προφόρμα των ενθέσεων των μεταθετών στοιχείων (transposon insertion proforma) , στην προφόρμα των χρωμοσωμικών ανωμαλιών (aberration proforma), εάν το μεταθετό στοιχείο διαταράσσει τη δομή ενός γονιδίου, στην προφόρμα των αλληλομόρφων (allele proforma) για το καινούργιοα μεταλλαγμένο αλληλόμορφο του γονιδιου που θα δώσει κλπ.



Εικόνα . Σχηματική παράσταση της ιεραρχικής δομής που ακολουθούν οι προφόρμες κατά την εισαγωγή των δεδομένων στη βάση.

Οι συσχετίσεις και ο συνδυασμός μεταξύ διαφορετικών στοιχείων γίνεται μέσα στη βάση δεδομένων, και βασίζονται στη δομή της. Ο τρόπος με τον οποίο μπορούν να συσχετιστούν ορισμένες γραμμές ενός πίνακα με ορισμένες γραμμές ενός άλλου πίνακα είναι με το να υπάρχουν κοινά στοιχεία σε αυτές τις γραμμές, στοιχεία που θα ανήκουν σε συγκεκριμένη στήλη. Με τον τρόπο αυτό, όσες γραμμές στον ένα πίνακα περιέχουν ίδια στοιχεία στην συγκεκριμένη στήλη τους, αυτές οι γραμμές μπορούν να

Εικόνα 7 (επόμενη σελίδα) Σχηματική παράσταση των πινάκων στην AnoBase. Με γραμμές εμφανίζεται η δυνατότητα σύνδεσης δυο πινάκων μεταξύ τους. Διευκρινίζεται ότι οι συσχετίσεις μεταξύ των πινάκων αυτών οφείλονται στον εσωτερικό τυος σχεδιασμό και είναι ενδεικτικές. Ανάλογα με τα στοιχεία που εισάγονται στη βάση, είναι δυνατόν να υπάρχουν και άλλες συσχετίσεις.


aberration
(aberr)
division
balancer
gene
graphic
insitu
allele
photomap
transcript
(transcr)
specorte
publication
(publn)
journal
protein
populational
(pop)
junction segment
(juncseg)
transposon insertion
(ti)
transposon/construct
(contra)
kilobase
molecular segment
(moseg)

συσχετιστούν με όσες γραμμές από τον άλλο πίνακα έχουν τα ίδια ακριβώς στοιχεία με την αντίστοιχη στήλη. Με άλλα λόγια, ο συσχετισμός και η ένωση των γραμμών επιτυγχάνεται με ίδια στοιχεία που θα έχουν και οι δυο πίνακες.

Με τον τρόπο αυτό έχουν σχεδιαστεί και οι πίνακες της δικής μας βάσης δεδομένων. Η εικόνα 7 απεικονίζει σχηματικά την εσωτερική δομή της βάσης δεδομένων και δείχνει πολλές από τις συσχετίσεις που μπορούν να γίνουν μεταξύ των πινάκων, καθώς και τις συγκεκριμένες στήλες που δημιουργούνται για το σκοπό αυτό. Αρχικά, όσον αφορά την ιεραρχική δομή ορισμένων στοιχείων, όπως για παράδειγμα το γονίδιο με τα αλληλόμορφά του και τα προϊόντα του, έχουν προστεθεί στήλες στους πίνακες που περιέχουν πληροφορίες από στοιχεία πινάκων που αντιστοιχούν σε προφόρμες πρωτύτερα στην ιεραρχία των προφορμών, δηλαδή στην περίπτωσή μας, στους πίνακες των αλληλομόρφων, των RNA προϊόντων και των πρωτεϊνών έχουν προστεθεί στήλες που περιέχουν το όνομα και το σύμβολο του γονιδίου στο οποίο αναφέρονται. Όσον αφορά τις συσχετίσεις μεταξύ των διαφορετικών πινάκων που δεν ακολουθούν τη γραμμική ιεραρχική σχέση, αυτές μπορούν να επιτευχθούν χάρη στον κατάλληλο σχεδιασμό των ίδιων των στηλών των πινάκων, ώστε στήλες που υπάρχουν σε έναν πίνακα έχουν επίσης ενσωματωθεί και σε έναν άλλο.

Αξίζει βέβαια να σημειωθεί ότι η δυνατότητα σύνδεσης μεταξύ στοιχείων του ενός πίνακα με στοιχεία του άλλου πίνακα δεν σημαίνει απαραίτητα ότι αυτές οι πληροφορίες μπορούν να συνδυαστούν, και ότι οι συνδυαζόμενες πληροφορίες έχουν βιολογική αξία. Κατά συνέπεια, κατά τη σχεδίαση της δομής, αλλά και ύστερα, κατά την παρουσίαση των δεδομένων, θα πρέπει να υπάρχει κριτική σκέψη και κατάλληλη βιολογική γνώση ώστε οι συνδέσεις μεταξύ των πινάκων να έχουν βιολογική αξία.

Το σύστημα εισαγωγής των στοιχείων στη βάση δεδομένων και τα προγράμματα διαχείρισης του (database data entering system)

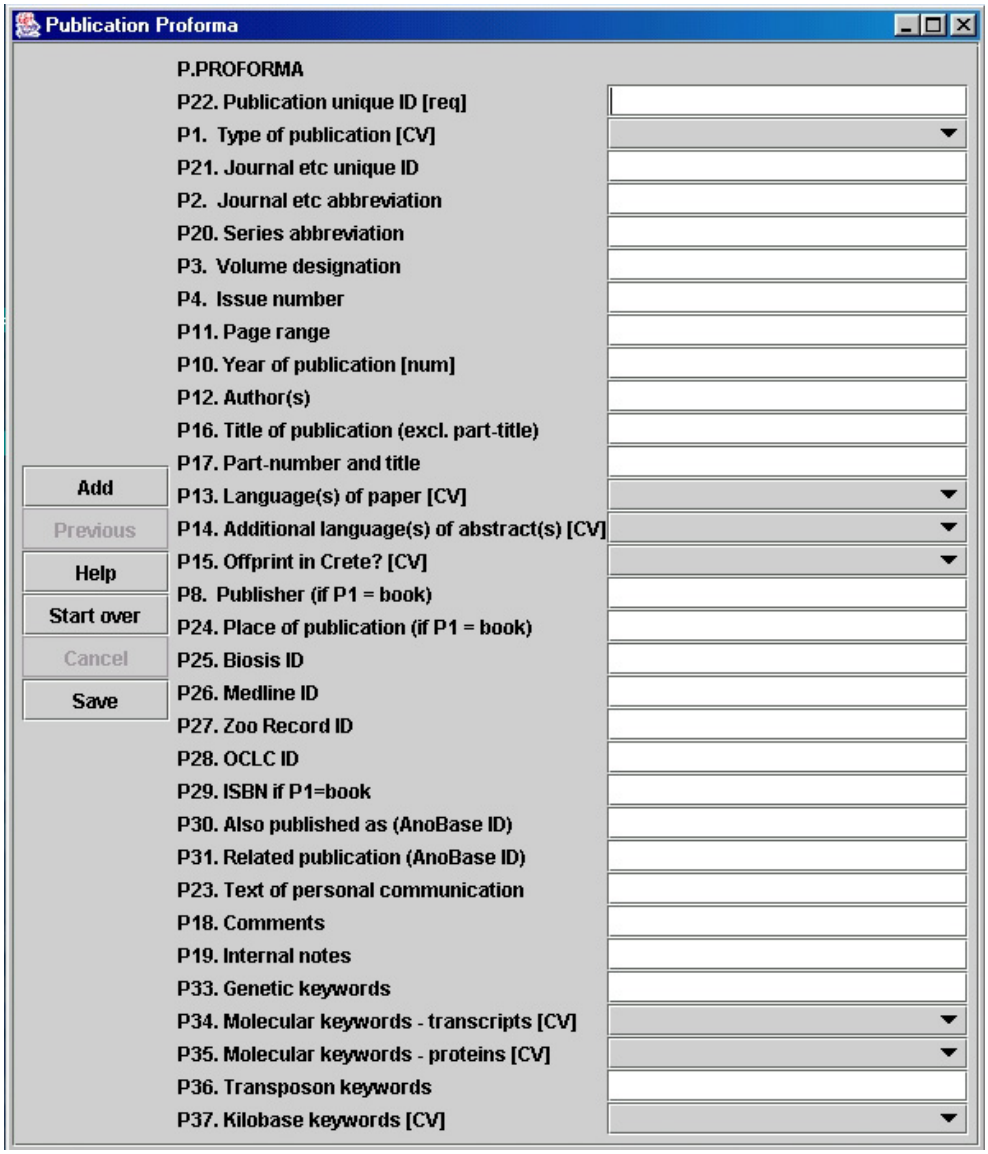
Το σύστημα εισαγωγής των στοιχείων στη βάση δεδομένων αποτελείται από 3 διαφορετικά προγράμματα, τα οποία ονομάζονται CuratorTree, simpleparse και smartparse. Με τα προγράμματα αυτά εξασφαλίζεται η σωστή διαχείριση των προφορμών, με βάση το ιεραρχικό σχήμα της εικόνας 6, η σωστή εισαγωγή των στοιχείων, η επεξεργασία τους και η είσοδός τους στις κατάλληλες στήλες και στους κατάλληλους πίνακες μέσα στη βάση δεδομένων.

Το πρόγραμμα CuratorTree[8]. Το πρόγραμμα CuratorTree είναι ένα πρόγραμμα γραφικού περιβάλλοντος, το οποίο γράφτηκε σε γλώσσα προγραμματισμού Java. Το πρόγραμμα αυτό διαχειρίζεται την εμφάνιση των προφορμών με βάση την ιεραρχική τους δομή. Είναι αυτό το πρόγραμμα που χρησιμοποιείται για τόσο για την καταχώριση νέων δημοσιεύσεων, όσο και τη διόρθωση καταχωρίσεων μέσα στη βάση δεδομένων.

Στην περίπτωση μιας νέας καταχώρισης, το πρόγραμμα αυτό θα ξεκινήσει με την εμφάνιση της προφόρμας της δημοσίευσης (publication proforma), στην οποία θα καταχωρηθούν στοιχεία σχετικά με την προέλευση, δηλαδή τη δημοσίευση αυτής της εργασίας. Η εικόνα 8 απεικονίζει το παράθυρο του γραφικού περιβάλλοντος του προγράμματος, στο οποίο υπάρχει η προφόρμα της δημοσίευσης. Όπως μπορεί να γίνει αντιληπτό, κάθε γραμμή της προφόρμας αυτή χρησιμοποιείται για την καταχώριση διαφορετικής πληροφορίας. Κάθε γραμμή αποτελείται από έναν ειδικό

[8] Η συγγραφή του πηγαίου κώδικα αυτού του προγράμματος δεν έγινε από το συγγραφέα αυτής της εργασίας αλλά από τον κ. Κιάμο Χρήστο.

κωδικό, ο οποίος αποτελείται από ένα γράμμα, χαρακτηριστικό της κάθε προφόρμας και ένα συγκεκριμένο αριθμό, ακολουθούμενο μερικές φορές και από ένα γράμμα. Ο κωδικός αυτός χρησιμοποιείται αργότερα από τα άλλα προγράμματα για την κατάλληλη αντιστοίχιση της κάθε γραμμής στη σωστή στήλη του αντίστοιχου πίνακα στη βάση δεδομένων. Εκτός από τον κωδικό υπάρχει επίσης μια μικρή σύντομη περιγραφή των στοιχείων που θα πρέπει να καταγραφούν σε αυτήν την γραμμή. Για παράδειγμα η γραμμή P12. Author(s) χρησιμοποιείται για την καταγραφή των συγγραφέων της εργασιας, ενώ η γραμμή P11. Page range χρησιμοποιείται για την καταγραφή των σελίδων της επιστημονικής εργασίας, εάν πρόκειται για δημοσιευμένο κείμενο. Δίπλα από η γραμμή υπάρχει ειδικός χώρος για την είσοδο του κειμένου (textbox).



Εικόνα 8. Η προφόρμα της δημοσίευσης (publication proforma) μέσα από το πρόγραμμα CuratorTree.

Παρόλα αυτά, σε ορισμένες περιπτώσεις κρίνεται απαραίτητο να χρησιμοποιούνται «τυποποιημένοι» όροι για να περιγράψουν μερικά στοιχεία. Πρόκειται για όρους που ανήκουν στο ελεγχόμενο λεξιλόγιο (“controlled

vocabularies") της βάσης δεδομένων, και που έχουν υιοθετηθεί για να περιγράφουν μονοσήμαντα και με σαφήνεια το συγκεκριμένο στοιχείο. Στην περίπτωση αυτή, η συμπλήρωση της συγκεκριμένης γραμμής δεν μπορεί να γίνει από ελεύθερο κείμενο και κατά συνέπεια, αντί για την εμφάνιση ειδικού χώρου για την εισαγωγή του κειμένου υπάρχει μια λίστα (popup menu) για την επιλογή του κατάλληλου όρου.

Η συμπλήρωση των στοιχείων σε κάθε προφόρμα πρέπει επίσης πρέπει να ακολοθεί κάποιους κανόνες, ώστε να εξασφαλίζεται αργότερα η διασύνδεση των πινάκων στη βάση δεδομένων. Αναλυτικές πληροφορίες για τον τρόπο εισαγωγής των στοιχείων στα πεδία των προφορμών υπάρχουν σε ένα αρχείο βοήθειας (help file), και το αρχείο είναι διαθέσιμο κατά τη συμπλήρωση των προφορμών από το κουμπί "Help" που βρίσκεται στην αριστερή μεριά του καθε παραθύρου.

Σε αυτήν την πλευρά υπάρχουν και όλα τα κουμπιά πλοήγησης του προγράμματος. Με το κουμπί "Add" εμφανίζεται μια λίστα με τις διαθέσιμες προφόρμες, δηλαδή με τις προφόρμες που μπορούν να προστεθούν στν προφόρμα που μόλις έχει συμπληρωθεί. Το κουμπί "Previous" χρησιμεύει στην εμφάνιση της προηγούμενης συμπηλρωμένης προφόρμας, και έχει προστεθεί διότι σε ορισμένες περιπτώσεις πληροφορίες που έχουν συμπληρωθεί στα πεδία της προηγούμενης προφόρμας απαιτούνται και στην παρούσα προφόρμα. Το κουμπί "Cancel" χρησιμοποιείται για την ακύρωση της παρούσας προφόρμας και επιστρέφει στην προφόρμα που είχε προηγηθεί. Το κουμπί "Start over" χρησιμοποιείται για την ακύρωση εξ ολοκλήρου της συγκεκριμένης καταχώρισης και την αρχή της δημιουργίας μιας καινούργιας καταχώρισης. Τέλος, το κουμπί "Save" χρησιμοποιείται για την αποθήκευση των στοιχείων που έχουν εισαχθεί σε όλες τις προφόρμες κατά τη διάρκεια της καταχώρισης.

Η αποθήκευση των στοιχείων που έχουν συμπληρωθεί στα πεδία των προφορμών για την καταχώριση γίνεται σε ένα αρχείο με τη μορφή απλού κειμένου (plain text file). Σε αυτό το αρχείο, για κάθε προφόρμα υπάρχει μια γραμμή, η οποία καθορίζει το όνομα της προφόρμας και ακολουθείται από τόσες γραμμές, όσες είναι και οι γραμμές της προφόρμας που έχει χρησιμοποιηθεί. Σε κάθε γραμμή υπάρχει ο κωδικός και η περιγραφή αυτής της γραμμής, ακολουθούμενη, σε όσες περιπτώσεις αυτό ισχύει, από την πληροφορία που έχει εισαχθεί.

Εκτός, όμως, από την περίπτωση της νέα καταχώρησης, το CuratorTree μπορεί να χρησιμοποιηθεί για την διόρθωση στοιχείων που έχουν ήδη αποθηκευτεί στη βάση δεδομένων. Το σύστημα διόρθωσης των προφορμών αυτών περιλαμβάνει τέσσερις διαφορετικές περιπτώσεις α) την περίπτωση της απλή διόρθωσης, όταν έχει εντοπιστεί ένα τυπογραφικό λάθος β) την περίπτωση της αφαίρεσης ενός στοιχείου, όταν το στοιχείο βρίσκεται σε μια λάθος στήλη γ) την περίπτωση της αναθεώρησης, όταν οι συγγραφείς της δημοσίευσης αναφέρουν τη αλλαγή των παλιότερων στοιχείων και δηλώνουν ότι πρόκεται για αναθεώρηση δ) την περίπτωση που οι συγγραφείς δεν αναφέρουν τη διαφορά ανάμεσα στα καινούργια και παλιότερα στοιχεία.

Και στις 4 περιπτώσεις, η διαδικασία είναι παρόμοια με αυτή που ακολουθείται κατά την νέα καταχώρηση. Οι προφόρμες που χρησιμοποιούνται, όμως, είναι ελαφρώς διαφοροποιημένες, καθώς περιλαμβάνουν και πεδία που πρόκειται να χρησιμοποιηθούν από το smartparse πρόγραμμα για τον εντοπισμό των δεδομένων που πρόκειται να διορθωθούν. Επίσης, σε αυτήν την περίπτωση δεν ακολουθείται το ιεραρχικό σχήμα της εικόνας 6, όπως γίνεται με την νέα καταχωρηση, αλλά χρησιμοποιείται απευθείας εκείνη η προφόρμα που αντιστοιχεί στον πίνακα της βάσης δεδομένων, τα στοιχεία του οποίου θα πρέπει να διορθωθούν.

Το αρχείο που αποθηκεύεται από τη διαδικασία αυτή, είναι παρόμοιο με το αρχείο που αποθηκεύεται κατά τη νέα καταχώριση, μόνο που προστίθεται ο ένας κωδικός διόρθωσης στις γραμμές της προφόρμας που πρόκεται να αλλαχτούν.

Το πρόγραμμα simpleparse. Το πρόγραμμα simpleparse είναι ένα πρόγραμμα που γράφτηκε σε γλώσσα προγραμματισμού Perl, και έχει σκοπό της επεξεργασία των δεδομένων, έτσι όπως εχουν αποθηκευτεί από το CuratorTree πρόγραμμα, και την προετοιμασία τους για την εισαγωγή τους μέσα στη βάση δεδομένων.

Το πρόγραμμα αυτό διαβάζει κάθε γραμμή της προφόρμας και διατηρεί το κωδικό της γραμμής και την πληροφορία που έχει αποθηκευτεί, αλλά απορρίπτει τη σύντομη περιγραφή της γραμμής, καθώς αυτήν δεν έχει πια καμία χρησιμότητα. Στη συνέχεια, ανιχνεύει εάν στις πληροφορίες που έχουν αποθηκευτεί περιέχουνται χαρακτήρες, οι οποίοι θεωρούνται από το πρόγραμμα της βάσης δεδομένων ως ειδικοί χαρακτήρες. Στην περίπτωση αυτή, πριν από τους χαρακτήρες αυτούς προστίθεται το σύμβολο \. Ο συνδυασμός του \ και του χαρακτήρα αφαιρεί την ειδική σημασία που έχουν αυτοί οι χαρακτήρες για τη βάση δεδομένων και τους κάνει απλούς χαρακτήρες κειμένου (κατά την είσοδο δηλαδή των στοιχείων στη βάση δεδομένων ο το σύμβολο \ δεν θα αποθηκευτεί, αλλά θα αποθηκευτεί μόνο ο συγκεκριμένος χαρακτήρας).

Το αποτέλεσμα του προγράμματος simpleparse είναι επίσης ένα αρχείο απλού κειμένου, στο οποίο έχουν αποθηκευτεί μόνο οι κωδικοί των γραμμών των προφορμών και το κείμενο που έχει εισαχθεί σε ορισμένους από αυτούς, μετά από την κατάλληλη επεξεργασία. Στην περίπτωση, βέβαια, της διόρθωσης των προφορμών, στις γραμμές που πρόκειται να διορθωθούν έχει προστεθεί και το σύμβολο της διόρθωσης.

Το πρόγραμμα smartparse. Το πρόγραμμα smartparse είναι το τελευταίο πρόγραμμα από τα προγράμματα του συστήματος εισόδου των δεδομένων, και έιναι αυτό που επικοινωνεί με τη βάση δεδομένων για την εισαγωγή των στοιχείων στις κατάλληλες θέσεις. Έχει επίσης γραφτεί σε γλώσσα προγραμματισμού Perl, και χρησιμοποιεί το αρχείο κειμένου που έχει αποθηκευτεί από το simpleparse πρόγραμμα.

Το πρόγραμμα αυτό εισάγει το δεδομένα στους κατάλληλους πίνακες και στις κατάλληλες στήλες με βάση τον κωδικό της κάθε γραμμής. Για κάθε προφόρμα που έχει συμπληρωθεί, ο κωδικός κάθε γραμμής αντιστοιχίζεται σε μια συγκεκριμένη στήλη στον κατάλληλο πίνακα μέσα στη βάση δεδομένων.

Στην περίπτωση προφορμών που ακολουθούν την ιεραρχία, το πρόγραμμα αυτό εισάγει αυτόματα στοιχεία από την τις προφόρμες που βρίσκονται πρωτύτερα στην στην ιεραρχία. Για παράδειγμα, στην περίπτωση που έχουν αποθηκευτει στοιχεία για ένα γονίδιο, το αλληλόμορφό του και την πρωτεΐνη του γονίδίου, κατά την αποθήκευση των στοιχείων της πρωτεΐνης στον κατάλληλο πίνακα, θα αποθηκευτούν στο ίδιο πίνακα, οι πληροφορίες για το γονίδιο και για το αλληλόμορφο στο οποίο αναφέρονται. Με τον τρόπο αυτό επιτυγχάνονται οι συσχετίσεις μεταξύ των πινάκων, για στοιχεία που έχουν άμεση σχέση μεταξύ τους.

Στην περίπτωση της διόρθωσης των στοιχείων που βρίσκονται ήδη μέσα στη βάση δεδομένων, το πρόγραμμα ανασύρει τις αντίστοιχες καταχωρήσεις από τον πίνακα της βάσης δεδομένων απο αντιστοιχεί στην προφόρμα και αντικαθιστά στις στήλες με τα καινούργια στοιχεία που έχουν εισαχθεί, και τα οποία έχουν το κατάλληλο κωδικό διόρθωσης στο αρχείο που έχει δημιουργηθεί από το simpleparse πρόγραμμα.

Η σχεδίαση και η εσωτερική αρχιτεκτονική των προγραμμάτων του συστήματος εισαγωγής των στοιχείων στη βάση έγινε με τέτοιο τρόπο ώστε να προσφέρουν ευελιξία στις αλλαγές, κυρίως στην πρόσθεση και αφαίρεση γραμμών στις

προφόρμες, όπως επίσης και στην προσθήκη και αφαίρεση καινούργιων προφορμών. Στην περίπτωση που χρειαστεί να προστεθούν ή να αφαιρεθούν γραμμές από τις προφόρμες, είναι απαραίτητο να γίνουν αλλαγές μόνο στον πηγαίο κώδικα του προγράμματος smartparse, ενώ στην περίπτωση που θα πρεπει να προστεθεί μια νέα προφόρμα, θα πρέπει να γίνουν αλλαγές και στο CuratorTree πρόγραμμα αλλά και στο smartparse πρόγραμμα. Αντίθετα, το simpleparse πρόγραμμα σχεδιάστηκε με τέτοιο τρόπο ώστε να μην χρειάζεται καμμιά αλλαγή στον πηγαίο κώδικα από τις παραπάνω τροποποιήσεις.

Το σύστημα παρουσίασης των στοιχείων της βάσης δεδομένων (database output system) – βασικές παράμετροι

Η λειτουργία της AnoBase δεν είναι μόνο η αποθήκευση πληροφοριών που έχουν να κάνουν με τα είδη του γένους *Anopheles* αλλά και η παρουσίαση των πληροφοριών με έναν αποδοτικό, κατανοητό και χρήσιμο τρόπο για τους ερευνητές. Το σύστημα παρουσίασης των στοιχείων της βάσης δεδομένων αποτελείται από μια σειρά προγραμμάτων, τα οποία έχουν ως στόχο την ανάσυρση των κατάλληλων δεδομένων μέσα από τους πίνακες και την παρουσίαση των δεδομένων αυτών με το μορφή ιστοσελίδων στον παγκόσμιο ιστό στο διαδίκτυο, ώστε οι πληροφορίες αυτές να είναι προσβάσιμες από οποιοδήποτε μέρος της γης.

Αν και κατα τη διάρκεια της συγγραφής αυτής της εργασίας το σύστημα παρουσίασης των δεδομένων ήταν υπό κατασκευή, εντούτοις μπορούν να περιγραφούν τα βασικά του στοιχεία. Οι γλώσσες προγραμματισμού που χρησιμοποιήθηκαν σε αυτήν την περίπτωση ήταν οι γλωσσες HTML, Perl και PHP, με τις δυο τελευταίες να είναι αυτές με τις οποίες επιτυγχάνεται η επικονωνία με τη βάση δεδομένων.

Η ανάσυρση των συγκεκριμένων δεδομένων μέσα από τη βάση, γίνεται με τη βοήθεια κατάλληλων και τυποποιημένων εντολών, που αποτελούν τη γλώσσα SQL (structured query language). Με βάση αυτήν την γλώσσα μπορούν να γίνουν αναζητήσεις συγκεκριμένων στοιχείων σε έναν πίνακα, όπως επίσης και να γίνουν οι συσχετίσεις μεταξύ των πινάκων, βασιζόμενες στα κοινά στοιχεία που περιέχουν οι πίνακες αυτοί.

Παρόλα αυτά, η SQL είναι αρκετά τεχνική γλώσσα για να χρησιμοποιηθεί από το ευρύ κοινό και προϋποθέτει όχι μόνο γνώση της εσωτερικής δομής της βάσης δεδομένων, αλλά και γνώση των κανόνων αποθήκευσης των πληροφοριών μέσα σ' αυτήν. Τα προγράμματα παρουσίασης, λοιπόν, αποτελούν τον ενδιάμσεο κρίκο ανάμεσα στον ερευνητή και τη βάση δεδομένων, όπου μετατρέπουν εντολές από απλή γλώσσα σε γλώσσα SQL για τη βάση δεδομένων.

Ο ερευνητής έχει στη διάθεσή του μια σειρά από φόρμες μέσα στις οποίες μπορεί να θέσει τα κατάλληλα ερωτήματα και να καθορίσει τις απαραίτητες ρυθμίσεις. Για παράδειγμα εάν ο ερευνητής θέλει να ανασύρει την πληροφορία σχετικά μέ ένα συγκεκριμένο όνομα γονιδίου, τότε στην περίπτωση αυτή υποδεικνύει το όνομα του γονιδίου αυτού στην κατάλληλη φόρμα, και προσθέτει, επίσης, και οποιαδήποτε πληροφορία θεωρεί απαραίτητη για την αναζήτηση

Οι διάφορες επιλογές του ερευνητή σε αυτήν την φόρμα μεταφράζονται σύμφωνα με το σχεδιασμό του προγράμματος, στην εντολή με τη μορφή της γλώσσας SQL. Το πρόγραμμα, στη συνέχεια, επικοινωνεί με τη βάση δεδομένων και θέτει την εντολη. Ανάλογα με το είδος της εντολής ανασύρονται οι συγκεκριμένες πληροφορίες από

ένα πίνακα, ή συνδυάζονται πληροφορίες από περισσότερους από έναν πίνακες και δημιουργούν ένα καινούργιο προσωρινό πίνακα.

Στη συνέχεια, όλες αυτές οι πληροφορίες παραλαμβάνονται από το πρόγραμμα και επεξεργάζονται με τέτοιο τρόπο ώστε να παρουσιάζονται με τη μορφή πινάκων και σχεδιαγραμμάτων. Στην πρώτη φάση της δημιουργίας της AnoBase, ο στόχος ήταν η παρουσίαση των βασικών πληροφοριών που έχουν σχέση με γονίδια, αλληλόμορφα, προϊόντα γονιδίων, χρωμοσωμικές ανωμαλίες, ενθέσεις μεταθετών στοιχείων, τεχνητών κατασκευών (constructs), πληροφοριών πληθυσμιακών μελετών, κυτταρογενετικών μελετών κ.α με τη μορφή απλών αναφορών (reports).

Σε μελλοντικές εκδόσεις της Anobase προγραμματίζεται η δημιουργία προγραμμάτων που θα εκμεταλλεύονται μεγάλο μέρος της αποθηκευμένης πληροφορίας και θα την παρουσιάζουν με μορφή συγκεντρωτικών χαρτών, ενός πλήθους γραφικών αναπαραστάσεων, όπως και πολύπλοκών σχεδιαγραμμάτων για την καλύτερη κατανόηση των πληροφοριών.

Βιβλιογραφία

Βιβλιογραφία

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F.*, et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Alberts, B. (1994). Molecular biology of the cell, 3rd edn (New York, Garland Pub.).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Baxevanis, A. D. (2002). The Molecular Biology Database Collection: 2002 update. Nucleic Acids Res *30*, 1-12.

Beaty, B. J., and Marquardt, W. C. (1996). The biology of disease vectors (Niwot, Colo., University Press of Colorado).

Berg, D. E., and Howe, M. M. (1989). Mobile DNA (Washington, D.C., American Society for Microbiology).

Bolshakov, V. N., Topalis, P., Blass, C., Kokoza, E., della Torre, A., Kafatos, F. C., and Louis, C. (2002). A comparative genomic analysis of two distant diptera, the fruit fly, Drosophila melanogaster, and the malaria mosquito, Anopheles gambiae. Genome Res *12*, 57-66.

Bowman, S., Lawson, D., Basham, D., Brown, D., Chillingworth, T., Churcher, C. M., Craig, A., Davies, R. M., Devlin, K., Feltwell, T.*, et al.* (1999). The complete nucleotide sequence of chromosome 3 of Plasmodium falciparum. Nature *400*, 532-538.

Broach, J. R., Pringle, J. R., and Jones, E. W. (1992). The Yeast Saccharomyces: Genome Dynamics, Protein Synthesis and Energetics (New York, Cold Spring Harbour Laboratory Press).

Bruce-Chwatt, L. J. (1993). Essential malariology, 3th edn (London ; New York, Arnold).

Bucheton, A. (1995). The relationship between the flamenco gene and gypsy in Drosophila: how to tame a retrovirus. Trends Genet *11*, 349-353.

Catteruccia, F., Nolan, T., Blass, C., Muller, H. M., Crisanti, A., Kafatos, F. C., and Loukeris, T. G. (2000a). Toward Anopheles transformation: Minos element activity in anopheline cells and embryos. Proc Natl Acad Sci U S A *97*, 2157-2162.

Catteruccia, F., Nolan, T., Loukeris, T. G., Blass, C., Savakis, C., Kafatos, F. C., and Crisanti, A. (2000b). Stable germline transformation of the malaria mosquito Anopheles stephensi. Nature *405*, 959-962.

Christiansen, T., and Torkington, N. (1998). Perl cookbook, 1st edn (Sebastopol, CA, O'Reilly).

Collins, F. H., Kamau, L., Ranson, H. A., and Vulule, J. M. (2000). Molecular entomology and prospects for malaria control. Bull World Health Organ *78*, 1412-1423.

Curtis, C. F. (2000). Infectious disease. The case for deemphasizing genomics in malaria control. Science *290*, 1508.

Descartes, A., and Bunce, T. (2000). Programming the Perl DBI (Cmabridge, MA, O'Reilly).

Dimopoulos, G., Zheng, L., Kumar, V., della Torre, A., Kafatos, F. C., and Louis, C. (1996). Integrated genetic map of Anopheles gambiae: use of RAPD polymorphisms for genetic, cytogenetic and STS landmarks. Genetics *143*, 953-960.

DuBois, P. (1999). MySQL (Indianapolis, IN, New Riders).

Friedl, J. E. F. (1997). Mastering regular expressions : powerful techniques for Perl and other tools, 1st edn (Cambridge ; Sebastopol, O'Reilly).

Gardner, M. J. (1999). The genome of the malaria parasite. Curr Opin Genet Dev *9*, 704-708.

Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C.*, et al.* (1998). Chromosome 2 sequence of the human malaria parasite Plasmodium falciparum. Science *282*, 1126-1132.

Ghosh, A., Edwards, M. J., and Jacobs-Lorena, M. (2000). The journey of the malaria parasite in the mosquito: hopes for the new century. Parasitol Today *16*, 196-201.

Goodwin, T. J., and Poulter, R. T. (2000). Multiple LTR-retrotransposon families in the asexual yeast Candida albicans. Genome Res *10*, 174-191.

Grossman, G. L., Rafferty, C. S., Clayton, J. R., Stevens, T. K., Mukabayire, O., and Benedict, M. Q. (2001). Germline transformation of the malaria vector, Anopheles gambiae, with the piggyBac transposable element. Insect Mol Biol *10*, 597-604.

Hoffman, S. L. (2000). Infectious disease. Research (genomics) is crucial to attacking malaria. Science *290*, 1509.

Hoffman, S. L., Subramanian, G. M., Collins, F. H., and Venter, J. C. (2002). Plasmodium, human and Anopheles genomics and malaria. Nature *415*, 702-709.

Jasinskiene, N., Coates, C. J., Benedict, M. Q., Cornel, A. J., Rafferty, C. S., James, A. A., and Collins, F. H. (1998). Stable transformation of the yellow fever mosquito,

Aedes aegypti, with the Hermes element from the housefly. Proc Natl Acad Sci U S A *95*, 3743-3747.

Jordan, I. K., and McDonald, J. F. (1999). Tempo and mode of Ty element evolution in Saccharomyces cerevisiae. Genetics *151*, 1341-1351.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Louis, C. (1999). Malaria entomology: can genomics help? Parassitologia *41*, 489-492.

Marillonnet, S., and Wessler, S. R. (1998). Extreme structural heterogeneity among the members of a maize retrotransposon family. Genetics *150*, 1245-1256.

Orwant, J., Hietaniemi, J., and Macdonald, J. (1999). Mastering algorithms with Perl, 1st edn (Sebastopol, CA, O'Reilly).

Pinsker, W., Haring, E., Hagemann, S., and Miller, W. J. (2001). The evolutionary life history of P transposons: from horizontal invaders to domesticated neogenes. Chromosoma *110*, 148-158.

Schwartz, R. L., and Christiansen, T. (1997). Learning Perl, 2nd edn (Sebastopol, CA, O'Reilly & Associates).

Smyth, J. D., and Wakelin, D. (1994). Introduction to animal parasitology, 3rd edn (Cambridge, Eng. ; New York, Cambridge University Press).

Spielman, A. (1994). Why entomological antimalarial research should not focus on transgenic mosquitoes. Parasitology Today *10*, 374-376.

Springer, M. S., and Britten, R. J. (1993). Phylogenetic relationships of reverse transcriptase and RNase H sequences and aspects of genome structure in the gypsy group of retrotransposons. Mol Biol Evol *10*, 1370-1379.

Srinivasan, S. (1997). Advanced Perl programming, 1st edn (Sebastopol, CA, O'Reilly).

Thomasova, D., Ton, L. Q., Copley, R. R., Zdobnov, E. M., Wang, X., Hong, Y. S., Sim, C., Bork, P., Kafatos, F. C., and Collins, F. H. (2002). Comparative genomic analysis in the region of a major Plasmodium-refractoriness locus of Anopheles gambiae. Proc Natl Acad Sci U S A *99*, 8179-8184.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A.*, et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Wall, L., Schwartz, R. L., and Christiansen, T. (1996). Programming Perl, 2nd edn (Sebastopol, CA, O'Reilly & Associates).

Welling, L., and Thomson, L. (2001). PHP and MySQL Web Development, Sams Publishing).

Yarger, R. J., Reese, G., King, T., and NetLibrary Inc. (1999). MySQL and mSQL, 1st edn (Sebastopol, Calif., O'Reilly).

Yeates, D., and Wiegmann, B. (1999). Congruence and controversy: Towards the higher-level phylogeny of Diptera. Annual Reviews in Entomology *44*, 397-428.

Zheng, L., Benedict, M. Q., Cornel, A. J., Collins, F. H., and Kafatos, F. C. (1996). An integrated genetic map of the African human malaria vector mosquito, Anopheles gambiae. Genetics *143*, 941-952.

Συνοδευτικό υλικό

Συνοδευτικό υλικό

(το συνοδευτικό υλικό υπάρχει σε CD-ROM)