ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΠΑΝΤΕΛΗ ΤΟΠΑΛΗ

Ανάπτυξη οντολογιών για την μελέτη τροπικών ασθενειών

ΗΡΑΚΛΕΙΟ 2011

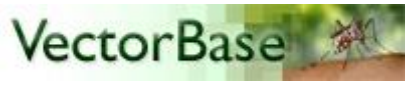
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ

ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΠΑΝΤΕΛΗ ΤΟΠΑΛΗ

Ανάπτυξη οντολογιών για την μελέτη τροπικών ασθενειών



Χρηματοδοτήθηκε από την VectorBase, ένα ερευνητικό πρόγραμμα του Εθνικού ινστιτούτου αλλεργιών και μολυσματικών νόσων των ΗΠΑ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ : Χ. ΛΟΥΗΣ

ΗΡΑΚΛΕΙΟ 2011

Ανάπτυξη οντολογιών για την μελέτη τροπικών ασθενειών

ΠΑΝΤΕΛΗΣ ΤΟΠΑΛΗΣ

Υποβλήθηκε ως μέρος των απαιτήσεων για τον τίτλο του

Διδάκτορα στο τμήμα Βιολογίας του Πανεπιστημίου Κρήτης

Επιβλέπων Καθηγητής: Χρήστος Λούης

Τριμελής Συμβουλευτική Επιτροπή

Χ. Λούης, Καθηγητής Πανεπιστημίου Κρήτης

Ν. Μοσχονάς, Καθηγητής Πανεπιστήμιου Πατρών

Δ. Αλεξανδράκη, Αν. Καθηγήτρια Πανεπιστημίου Κρήτης

Επταμελής Εξεταστική Επιτροπή

Χ. Λούης, Καθηγητής Πανεπιστημίου Κρήτης

Ν. Μοσχονάς, Καθηγητής Πανεπιστήμιου Πατρών

Δ. Αλεξανδράκη, Αν. Καθηγήτρια Πανεπιστημίου Κρήτης

Α. Χατζηγεωργίου, Καθηγήτρια Πανεπιστημίου Θεσσαλίας

Ι. Βόντας, Αν. Καθηγητής Πανεπιστημίου Κρήτης

Ν. Πουλακάκης, Επ. Καθηγητής Πανεπιστημίου Κρήτης

Ι. Ηλιόπουλος, Λέκτορας Πανεπιστημίου Κρήτης

Στους γονείς μου

ΠΡΟΛΟΓΟΣ

Η παρούσα διδακτορική διατριβή εκπονήθηκε στο εργαστήριο ελέγχου παρασίτων και ασθενειών υπό την επίβλεψη του καθηγητή Χ. Λούη. Χρηματοδοτήθηκε από Εθνικό Ινστιτούτο Αλλεργιών και Μεταδοτικών Ασθενειών των Ηνωμένων Πολιτειών στο πλαίσιο της VectorBase.

Αν η εκπόνηση μιας διατριβής είναι μια προσωπική πορεία στον χρόνο, αυτή δεν είναι ποτέ μοναχική. Πλήθος ανθρώπων συμμετέχουν άμεσα και έμμεσα συμβάλλοντας με τον τρόπο τους. Ήμουν ιδιαίτερα τυχερός, γιατί στην δική μου πορεία συνάντησα ξεχωριστούς ανθρώπους που διαμόρφωσαν την προσωπικότητά μου και ως επιστήμονα αλλά και ως ανθρώπου.

Ο Χρήστος Λούης ήταν πολλά περισσότερα πράγματα από επιβλέπων καθηγητής μου όλα αυτά τα χρόνια. Με στήριξε πολύ στις δύσκολες στιγμές μου με τρόπο άμεσο και διακριτικό, έχοντας την σοφία να με αφήνει να κάνω λάθη για να μαθαίνω από αυτά. Θέλω να ελπίζω πως δεν μετάνιωσε για την εμπιστοσύνη που μου έδειξε από την πρώτη στιγμή.

Στον Νίκο Μοσχονά χρωστώ τα πρώτα μου βήματα στον χώρο της μοριακής βιολογίας σε ερευνητικό επίπεδο, αλλά και αργότερα όταν πλέον είχα περάσει στον τομέα της βιοπληροφορικής, η υποστήριξή του σε αυτό που έκανα ήταν πολύ μεγάλη. Ήταν μια περίοδος που το περιβάλλον των τμημάτων της βιολογίας πανελλαδικά δεν καταλάβαινε ιδιαίτερα αυτούς που δούλευαν μπροστά σε μια οθόνη υπολογιστή και κατά συνέπεια δεν ήταν ιδιαίτερα φιλικό μαζί τους.

Ένα ιδιαίτερο ευχαριστώ αισθάνομαι πως οφείλω και στην Δέσποινα Αλεξανδράκη για τις συζητήσεις και τον χρόνο που μου αφιέρωσε όλα αυτά τα χρόνια. Επίσης θα ήθελα να ευχαριστήσω τους Ιωάννη Βόντα, Ιωάννη Ηλιόπουλο, Νίκο Πουλακάκη και Άρτεμη

Χατζηγεωργίου, μέλη της επταμελούς εξεταστικής μου επιτροπής για την κριτική, τα σχόλια και τις παρατηρήσεις τους.

Ο Μανώλης Διαλυνάς είναι ο μόνος πέρα από την επταμελή εξεταστική επιτροπή που διάβασε ολόκληρη τούτη εδώ την εργασία και τον ευχαριστώ πολύ γι' αυτό. Μα περισσότερο γιατί μοιραζόμαστε πέρα από το γραφείο, την καθημερινότητά μας και με το χιούμορ, τον χαρακτήρα και την προσωπικότητά του την κάνει πιο ευχάριστη.

Μιλώντας για συνεργάτες, θα ήθελα να ευχαριστήσω όλα τα μέλη του εργαστηρίου διαχρονικά, αλλά επειδή είναι μεγάλο το χρονικό διάστημα και πάρα πολλοί άνθρωποι θα αρκεστώ στους Λευτέρη Σπανό, Γιώργο Παπαγιαννάκη και Inga Siden-Kiamos που είμαστε μαζί από τότε που πρωτοβρέθηκα στο εργαστήριο σαν προπτυχιακός φοιτητής. Επίσης θα ήθελα να αναφερθώ στην Καλλιόπη Τζαβλάκη και την Ελβίρα Μητράκα που ως προπτυχιακές φοιτήτριες ανέβηκαν μαζί μου στο πλοίο των οντολογιών και ασχολήθηκαν με αυτές στο πλαίσιο της πτυχιακής τους εργασίας.

Τέλος, θα ήθελα να αναφερθώ στους γονείς μου, Αλέκο και Μαρία. Τους οφείλω πολλά περισσότερα από το «ζην» που δεν μπορούν να κλειστούν μέσα σε μια παράγραφο, ούτε να περιγραφούν με λέξεις. Χωρίς άλλο σχόλιο θα ήθελα να τους αφιερώσω την παρούσα εργασία.

Πίνακας περιεχομένων

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ ΠΙΝΑΚΩΝ

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ ΕΙΚΟΝΩΝ

Περίληψη

Δέκα χρόνια μετά την αλληλούχιση του ανθρώπινου γονιδιώματος, παραμένει ζητούμενο ο ακριβής αριθμός και η λειτουργική ανάλυση των ανθρώπινων γονιδίων. Η ανάλυση του ολοένα και μεγαλύτερου όγκου διαθέσιμων δεδομένων με την βοήθεια ηλεκτρονικών υπολογιστών επιτάχυνε την διαδικασία και υπογράμμισε την αναγκαιότητα της βιοπληροφορικής στην σύγχρονη βιολογική έρευνα. Ταυτόχρονα ανέδειξε την ανάγκη νέων εργαλείων και τρόπων οργάνωσης των βιολογικών δεδομένων. Η χρήση οντολογιών είναι κομβικό σημείο στην προσπάθεια αυτή.

Στο πλαίσιο της παρούσας διατριβής δημιουργήθηκαν οντολογίες για: την περιγραφή της ανατομίας του κουνουπιού, την περιγραφή και τον έλεγχο της ανθεκτικότητας σε εντομοκτόνα που εμφανίζουν πληθυσμοί κουνουπιών και τέλος, για την ελονοσία. Με βάση αυτές τις οντολογίες αναπτύχθηκαν εργαλεία που δίνουν την δυνατότητα προσθήκης επισημειώσεων στις υπάρχουσες βάσεις δεδομένων για τον ιστό στον οποίο εκφράζονται συγκεκριμένα γονίδια, εμπλουτίζοντάς τες. Παράλληλα, καθιστούν δυνατή την παρακολούθηση και την ταυτοποίηση της εμφάνισης ανθεκτικότητας απέναντι στα εντομοκτόνα στους πληθυσμούς κουνουπιών κάποιας περιοχής. Τέλος αποτελούν την βάση γενικότερων πληροφοριακών συστημάτων που προσπαθούν να συνδυάσουν ετερογενείς πληροφορίες μεταξύ τους και είναι γνωστότερα σαν συστήματα υποστήριξης αποφάσεων. Η οντολογία της ελονοσίας που δημιουργήθηκε στο IMBB ήδη αποτελεί την βάση του αντίστοιχου συστήματος, ενώ η οντολογία και η βάση δεδομένων για την ανθεκτικότητα σε εντομοκτόνα έχει υιοθετηθεί σαν πρότυπο και από διεθνή προγράμματα και οργανισμούς.

Summary

Ten years after the publication of the human genome, the functional analysis of human genes is still at an embryonic stage. The use of computers to analyze the increasing volume of the available data accelerated the process and underlined the necessity of bioinformatics into the modern research in the field of molecular biology. At the same time the need for new tools and new ways to semantically organize the available data it became obvious. The use of ontologies is a major step forward towards this direction.

Within the frame of this work new ontologies have been created to describe the gross anatomy of mosquitoes, to facilitate detection and monitoring of insecticide resistance often observed in mosquito populations especially in the areas where spraying is the main measure for vector control and finally to describe malaria. Based on these ontologies new tools have been developed to facilitate gene annotation with anatomical terms. Moreover a new database to detect and monitor resistance in an area has been established based on our ontology and it has been adopted as a standard from various international efforts. Finally the malaria ontology created at IMBB resides at the heart of the malaria decision support system created by the innovative vector control consortium.

ΓΕΝΙΚΗ ΕΙΣΑΓΩΓΗ

Γενικά

Είναι σχεδόν στερεότυπο, όλες οι εργασίες που αναφέρονται στις βάσεις βιολογικών δεδομένων και τη βιοπληροφορική να ξεκινούν με μια αναφορά στον τεράστιο όγκο των βιολογικών δεδομένων που είναι διαθέσιμα στις μέρες μας. Τα εργαλεία δε, που παρουσιάζουν είναι ολοένα και πιο αποδοτικά, πιο ισχυρά και υπόσχονται να αποκαλύψουν καλύτερα και ταχύτερα από οποιοδήποτε άλλο την αναζητούμενη κάθε φορά πληροφορία. Η παρούσα εργασία δεν θα μπορούσε να αποτελέσει εξαίρεση, θα ήθελε όμως να προσθέσει δύο ακόμα διαστάσεις του φαινομένου που συνεχίζει να εξελίσσεται με εντεινόμενους ρυθμούς. Η πρώτη έχει να κάνει με το γεγονός πως αυτός ο όγκος των δεδομένων που βρίσκεται αποθηκευμένος στους υπολογιστές μας, δεν είναι αρκετός για να περιγράψει με λεπτομέρεια και να αναλύσει σε ικανοποιητικό βαθμό τις σχέσεις που διέπουν τις φυσιολογικές και παθολογικές διαδικασίες της ζωής των έμβιων όντων. Η δεύτερη σχετίζεται με την ανάγκη περαιτέρω οργάνωσης των δεδομένων μια διαδικασία που περιλαμβάνει την επισημείωσή τους με πρόσθετα χαρακτηριστικά και πληροφορίες, που συνοπτικά χαρακτηρίζονται ως μετα-δεδομένα.

Η κλασσική αντιμετώπιση περιλαμβάνει την ανάπτυξη βάσεων δεδομένων που περιέχουν όλες αυτές τις πληροφορίες. Τα τελευταία χρόνια ο αριθμός των διαθέσιμων βάσεων αυξάνεται με ρυθμό ανάλογο εκείνου των δεδομένων (Galperin and Cochrane 2011) με αποτελέσματα να υπάρχουν πολυάριθμες βάσεις δεδομένων με τις ίδιες ή παρεμφερείς πληροφορίες. Για παράδειγμα υπάρχουν περισσότερες από 5 διαφορετικές βάσεις που αφορούν μοτίβα και πρότυπα που παρουσιάζονται στις πρωτεΐνες και φαίνεται να σχετίζονται με την λειτουργία τους (Hunter, Apweiler et al. 2009), (Sigrist, Cerutti et al. 2010), (Pettifer, Ison et al. 2010), (Gaudet, Bairoch et al. 2011). Γρήγορα, έγινε αντιληπτό πως αυτός ο κατακερματισμός της πληροφορίας σε επιμέρους τμήματα το οποία βρίσκονται

σε διαφορετικές βάσεις δεδομένων, συχνά σε διαφορετικά μέρη του κόσμου, υπονόμευε την αποτελεσματικότητα των ίδιων των βάσεων αλλά και των διαθέσιμων εργαλείων που χρησιμοποιούμε για να ανακτούμε τις πληροφορίες που μας ενδιαφέρουν. Με την ανάπτυξη του διαδικτύου και του παγκόσμιου ιστού οι γεωγραφικοί περιορισμοί έπαψαν να ισχύουν αλλά οι υπόλοιποι παρέμεναν. Η πρώτη απόπειρα να υπερκερασθεί το πρόβλημα ήταν η χρήση εξωτερικών συνδέσεων και διασταυρούμενων αναφορών από την μια βάση δεδομένων στην άλλη. Αυτό είχε σαν παρενέργεια το να δημιουργηθούν κυκλικές διασταυρούμενες συνδέσεις από την μια βάση δεδομένων στην άλλη χωρίς πάντοτε να είναι ξεκάθαρη η πρωτογενής πηγή της πληροφορίας. Επιπλέον, επειδή συχνά οι επιστήμονες χρησιμοποιούν συνώνυμα όρων για να περιγράψουν τα αποτέλεσματά τους οι υπολογιστές δεν μπορούν να συνδέσουν αποτελεσματικά μεταξύ τους τα διάφορα σύνολα πληροφοριών που βρίσκονται διασκορπισμένα.

Η ανάγκη για καλύτερη οργάνωση των μετα-δεδομένων δεν ήταν κάτι νέο. Από την αρχαιότητα συναντήσαμε αυτό το πρόβλημα και η πρώτη απάντηση σε αυτό ήταν η δημιουργία καταλόγων από όρους που περιέγραφαν αυτό που θέλαμε. Τα πρώτα λεξικά στην ανθρώπινη ιστορία χρονολογούνται από το 2300 π.Χ. και δεν είναι τίποτε άλλο από λίστες με όρους σαφώς ορισμένους. Τέτοιοι κατάλογοι χρησιμοποιήθηκαν και σε αυτή την περίπτωση για να διευκολυνθούν οι αναζητήσεις και περιλάμβαναν λέξεις – κλειδιά με χαρακτηριστικό παράδειγμα τους ιατρικούς θεματικούς τίτλους (Medical Subject Headings – MESH)(Aronson, Mork et al. 2004) που ήδη από το 1960 χρησιμοποιεί η Εθνική Βιβλιοθήκη της Ιατρικής (NLM) των Ηνωμένων Πολιτειών (Mary, Marquet et al. 2004). Όμως οι όροι που περιλαμβάνονται στους θεματικούς τίτλους δεν είναι ορισμένοι με σαφήνεια, δεν περιέχουν την δυνατότητα χρήσης συνωνύμων όρων και συνήθως είναι ταξινομημένοι αλφαβητικά χωρίς να περιγράφουν συγκεκριμένη θεματική περιοχή ή τομέα της επιστήμης. Σχήματα ταξινόμησης πληροφοριών και «θησαυροί» (thesauri) (Rector

1998; Serban and ten Teije 2009) καθώς και ελεγχόμενα λεξιλόγια ήρθαν να καλύψουν το κενό, αλλά η πιο εξελιγμένη μορφή οργάνωσης των μετα-δεδομένων είναι η δημιουργία και η χρήση οντολογιών (Bard 2003).

Τι είναι η οντολογία

Με τον όρο οντολογία περιγράφεται μια πληθώρα διαφορετικών πραγμάτων ανάλογα με το πεδίο αναφοράς. Σαν επιστήμη για παράδειγμα, η «*Οντολογία*» είναι κλάδος της φιλοσοφίας, παρακλάδι των μεταφυσικών, που ασχολείται με ό,τι υπάρχει. Οι απαρχές της βρίσκονται στα έργα αρχαίων Ελλήνων φιλοσόφων και τα ενδιαφέροντά της περιλαμβάνουν κυρίως την κατηγοριοποίηση όσων υπάρχουν σε βασικό και θεωρητικό επίπεδο (Griswold 2002). Οι επιστήμονες του τομέα της πληροφορικής δανείστηκαν τον όρο δίνοντας του μια διαφορετική σημασία. Η οντολογία είναι γι αυτούς η αναπαράσταση μιας περιοχής της γνώσης με δεδομένα που συνδέονται μεταξύ τους με λογικές σχέσεις κατά τέτοιο τρόπο ώστε να είναι δυνατό να αναλυθούν από ηλεκτρονικούς υπολογιστές. Βάση για αυτή την αλλαγή είναι πως για τους υπολογιστές «υπάρχουν» μόνο εκείνα τα πράγματα που μπορούν να αναπαρασταθούν (Baclawski and Niu 2006). Ο στόχος είναι κάθε συγκεκριμένη χρονική στιγμή η πραγματικότητα να περιγράφεται στην οντολογία με τον πιστότερο δυνατό τρόπο. Καθώς οι γνώσεις μας αυξάνονται, οι οντολογίες εξελίσσονται για να συμπεριλάβουν τα νέα δεδομένα ή/και να τροποποιήσουν τον τρόπο με τον οποίο συσχετίζονται οι έννοιες μεταξύ τους. Αυτό καθιστά την κατασκευή και την συντήρησή τους μια διαρκή, δυναμική διαδικασία χωρίς τέλος.

Στο χώρο της βιοπληροφορικής οι πρώτες αναφορές σε οντολογίες με την σύγχρονη έννοια της πληροφορικής εμφανίζονται στα 1996 στην εγκυκλοπαίδεια των γονιδίων και του μεταβολισμού του *E. coli*, όπου διαμορφώνουν και το σχήμα της αντίστοιχης βάσης δεδομένων (Karp, Riley et al. 1996). Όμως οι οντολογίες μπήκαν στην ζωή των βιοεπιστημόνων το 2000 όταν επιστήμονες που δούλευαν στις βάσεις δεδομένων

της δροσόφιλας, του σακχαρομύκητα και του ποντικού δημιούργησαν μια νέα βάση δεδομένων που περιείχε τα γονίδια των οργανισμών αυτών κατηγοριοποιημένα ανάλογα με την λειτουργία των προϊόντων τους σε μοριακό επίπεδο, την βιολογική διαδικασία στην οποία συμμετέχουν καθώς και σε ποιο μέρος του κυττάρου εντοπίζονται. Για τον σκοπό αυτό ανέπτυξαν την οντολογία γονιδίων (Gene Ontology, GO) που στην πραγματικότητα είναι 3 οντολογίες κάτω από την ίδια ομπρέλα και βασιζόμενοι πάνω σε αυτή δημιούργησαν την βάση δεδομένων (Ashburner, Ball et al. 2000). Είναι σημαντικός ο διαχωρισμός ανάμεσα στην οντολογία, την βάση δεδομένων, και τις επισημειώσεις που προστίθενται στα δεδομένα. Η οντολογία, όπως ήδη έχει αναφερθεί στοχεύει στην περιγραφή της γνώσης μας σε κάποιο πεδίο. Μια βάση δεδομένων είναι ένα πληροφοριακό σύστημα το οποίο περιέχει δεδομένα και οι επισημειώσεις είναι μετα-δεδομένα. Έτσι η οντολογία γονιδίων περιλαμβάνει αυτά που γνωρίζουμε για τις μοριακές λειτουργίες των προϊόντων των γονιδίων, για τις φυσιολογικές λειτουργίες στις οποίες αυτά συμμετέχουν και για το που εντοπίζονται σε κυτταρικό επίπεδο. Η βάση δεδομένων GO περιλάμβανε αρχικά τα γονίδια της δροσόφιλας, του σακχαρομύκητα και του ποντικού κατηγοριοποιημένα με βάση αυτούς τους τρεις άξονες και αυτή ακριβώς η κατηγοριοποίηση αποτελούσε επισημείωση που συμπεριλαμβανόταν πλέον και συνόδευε το αντίστοιχο γονίδιο.

Στοιχεία μιας οντολογίας

Δύο είναι τα στοιχεία τα οποία συναντά κανείς σε μια οντολογία. Οι οντότητες και οι σχέσεις (Smith and Rosse 2004). Οι οντότητες είναι εκείνες που περιγράφουν την γνώση μας για το πεδίο που καλύπτει η οντολογία, ενώ οι σχέσεις είναι εκείνες που διαμορφώνουν την δομή της (Smith, Ceusters et al. 2005).

Οι οντότητες που περιλαμβάνονται σε μια οντολογία συχνά αναφέρονται και ως *«έννοιες»* ή *«τάξεις»* καθώς επίσης σπανιότερα σαν *«είδη»*, *«τύποι»*, *«κατηγορίες»*, ονόματα τα οποία προέρχονται κυρίως από τα υπόλοιπα συστήματα οργάνωσης των μετα-δεδομένων.

Αντιπροσωπεύουν κυρίως αλλά όχι αποκλειστικά γενικότητες, όπως για παράδειγμα ο όρος άνθρωπος αναφέρεται σε κάθε άνθρωπο και όχι σε κάποιον συγκεκριμένο. Στην συνέχεια θα αναφέρω ως «*τάξεις*» εκείνες τις οντότητες που αναφέρονται σε γενικότητες ενώ ως «*άτομα*» ή «*περιπτώσεις*» εκείνες που αναφέρονται σε συγκεκριμένες περιπτώσεις. Ένας άλλος διαχωρισμός των οντοτήτων μεταξύ τους είναι εκείνος βάσει της σχέσης τους με τον χρόνο: α) Συνεχείς είναι εκείνες οι οντότητες που έχουν συνεχή παρουσία μέσα στον χρόνο, διατηρούν την οντότητά τους παρά το γεγονός πως μπορούν να μεταβάλλονται και υπάρχουν καθ' ολοκληρία όσο καιρό υπάρχουν. Για παράδειγμα ένας οργανισμός είναι μια συνεχής οντότητα, γιατί έχει συνεχή παρουσία στον χρόνο, διατηρεί την οντότητά του παρά το γεγονός πως μεταβάλλεται καθώς αναπτύσσεται και περνώντας από το ένα στάδιο στο άλλο και στο χρονικό διάστημα ανάμεσα στην γέννηση και τον θάνατό του υπάρχει καθ' ολοκληρία. β) Ασυνεχείς οντότητες ή διαδικασίες είναι εκείνες που έχουν χρονικά μέρη, εξελίσσονται σε στάδια που διαδέχεται το ένα το άλλο και υπάρχουν μόνο μέσα στα στάδια αυτά. Όλες οι φυσιολογικές διαδικασίες είναι ασυνεχείς οντότητες (Grenon, Smith et al. 2004). Κάθε μια οντότητα θα πρέπει να συνδέεται τουλάχιστον με μία άλλη μέσω κάποιας σχέσης.

Οι σχέσεις που συνδέουν τους όρους μιας οντολογίας είναι στην πραγματικότητα η ειδοποιός διαφορά ανάμεσα στις οντολογίες και τα υπόλοιπα συστήματα ταξινόμησης μεταδεδομένων κι αυτό που τις καθιστά τόσο χρήσιμες στην βιοπληροφορική. Για να το πετύχουν όμως αυτό, θα πρέπει με την σειρά τους να είναι ορισμένες με τυπικό τρόπο. Μπορούμε να διακρίνουμε τις σχέσεις αυτές ως σχέσεις ανάμεσα σε τάξεις, σχέσεις ανάμεσα σε άτομα. Η σχέση που συνδέει μια περίπτωση με μια τάξη είναι η: *instance_of*. Αφού αναφερόμαστε σε οντολογίες η πρώτη βασική σχέση που πρέπει να υπάρχει είναι η σχέση *is_a* που στην πραγματικότητα μεταφράζεται σε «*είναι υποτύπος του*». Αλλά ακόμα και αυτή η σχέση δεν είναι αξιωματική αλλά ορίζεται με τυπικό τρόπο. Μια διαδικασία Δ είναι

υποτύπος της Δ_1 αν κάθε δ που είναι περίπτωση της Δ, είναι περίπτωση της Δ_1. Όταν αναφερόμαστε σε συνεχείς οντότητες θα πρέπει να προσθέσουμε και την παράμετρο του χρόνου. Έτσι μια συνεχής οντότητα Σ είναι υποτύπος της Σ_1 την χρονική στιγμή χ, αν κάθε σ που είναι περίπτωση της Σ, στην στιγμή χ, ταυτόχρονα είναι περίπτωση του Σ_1 (Smith and Rosse 2004; Smith, Ceusters et al. 2005). Η προσθήκη του χρόνου μας επιτρέπει να αποφύγουμε λογικά λάθη όπως για παράδειγμα ο ενήλικος είναι παιδί μιας και κάθε ενήλικο άτομο υπήρξε σε κάποια χρονική στιγμή παιδί. Οι σχέσεις έχουν διάφορες ιδιότητες όπως παραδείγματος χάριν μπορεί να είναι συμμετρικές, να ισχύουν προς ορισμένη μόνο κατεύθυνση, να είναι αυτοπαθείς ή μεταβατικές (Smith and Rosse 2004; Smith, Ceusters et al. 2005). Για παράδειγμα ο άνθρωπος είναι υποτύπος του θηλαστικού δεν σημαίνει πως το θηλαστικό είναι υποτύπος του ανθρώπου αποδεικνύοντας πως η σχέση *is a* (*είναι υποτύπος του*) ισχύει προς ορισμένη μόνο κατεύθυνση. Θα διαπιστώσουμε πως η ίδια σχέση είναι μεταβατική αν επεκτείνουμε το προηγούμενο παράδειγμα προσθέτοντας πως το θηλαστικό είναι υποτύπος του σπονδυλωτού, αφού είναι αλήθεια πως ο άνθρωπος είναι υποτύπος του σπονδυλωτού. Με ανάλογο τρόπο ορίζονται και οι υπόλοιπες σχέσεις που χρησιμοποιούνται στις οντολογίες με πιο συχνά χρησιμοποιούμενη την *part of* (αποτελεί μέρος του). Αυτές οι δύο σχέσεις είναι οι θεμελιώδεις σχέσεις που υπάρχουν στις οντολογίες που περιγράφουν το πεδίο της βιοϊατρικής.

Είδη οντολογιών

Οι οντολογίες χρησιμοποιήθηκαν για την οργάνωση των μετα-δεδομένων, ώστε να καταστεί δυνατή η διασύνδεση μέσω αυτών, δεδομένων που βρίσκονται κατακερματισμένα σε διάφορες βάσεις δεδομένων. Όμως αυτό προϋποθέτει την χρήση της ίδιας οντολογίας και συχνά δεν είναι αρκετό. Το επόμενο βήμα είναι η διασύνδεση των δεδομένων από βάσεις που χρησιμοποιούν διαφορετικές οντολογίες κάτι που προϋποθέτει την διασύνδεση των ίδιων των οντολογιών μεταξύ τους. Ενώ στόχος των οντολογιών είναι η καταγραφή και η

κατηγοριοποίηση της γνώσης μας, δεν υπάρχει μια οντολογία των οντολογιών η μια σαφής ταξινόμησή τους σε κατηγορίες στην οποία να συμφωνούν όλοι. Ένας βασικός όμως διαχωρισμός τους περιλαμβάνει τις οντολογίες ανωτέρου επιπέδου (Burek, Hoehndorf et al. 2006; Raghupathi and Umar 2011), τις οντολογίες συγκεκριμένου γνωστικού πεδίου, τις οντολογίες αναφοράς (Burgun 2006) και τις οντολογίες εφαρμογών (de Clercq, Hasman et al. 2001).

Επειδή η αναπαράσταση της γνώσης μας στις οντολογίες προχωρά από το γενικό στο ειδικό, τα πρώτα επίπεδα περιγράφουν τις γενικότερες οντότητες και όσο προχωράμε συναντούμε τις περισσότερο εξειδικευμένες. Οι οντολογίες ανώτερου επιπέδου αντανακλούν διαφορετικές φιλοσοφικές θεωρήσεις της οντολογικής κατηγοριοποίησης χωρίς να έχουν αναφορές σε συγκεκριμένους επιστημονικούς τομείς ή εφαρμογές. Δημιουργούνται στηριγμένες στην υπόθεση πως όλες οι επιμέρους οντολογίες που αφορούν συγκεκριμένους τομείς και εφαρμογές θα ξεκινούν να αναφέρονται από μια γενική οντότητα και προσπαθούν να την κατηγοριοποιήσουν. Για παράδειγμα, μια από τις οντολογίες που περιλαμβάνει η GO ξεκινά από το σε ποιο συνιστόν μέρος του κυττάρου εντοπίζεται το προϊόν ενός γονιδίου. Κάποιος θα μπορούσε να συνεχίσει να γενικεύει την ταξινόμηση, ορίζοντας πως το συνιστόν μέρος του κυττάρου *part of* κυττάρου το οποίο *is a* αντικείμενο που με την σειρά του *is a* συνεχής οντότητα. Οντολογίες ανωτέρου επιπέδου είναι η βασική τυπική οντολογία (BFO) (Grenon, Smith et al. 2004), η γενική τυπική οντολογία (GFO)(Simon, Dos Santos et al. 2006) και η προτεινόμενη ενοποιημένη τυπική οντολογία (SUMO) (Soldatova and King 2006). Ο χαρακτηρισμός «τυπική» στο όνομά τους σημαίνει πως χρησιμοποιούν αυστηρά ορισμένες λογικές σχέσεις για να συνδέσουν τους όρους τους μεταξύ τους. Το μεγαλύτερο πρόβλημα στην διασύνδεση των δεδομένων που έχουν κατηγοριοποιηθεί με βάση διαφορετικές οντολογίες ανώτερου επιπέδου είναι ακριβώς η ίδια η ύπαρξη πολλών οντολογιών ανωτέρου επιπέδου, κάτι που δεν φαίνεται πως θα ξεπεραστεί σύντομα, αφού η

φιλοσοφία μετρά περισσότερα από 2000 χρόνια ύπαρξης και δεν έχει καταφέρει ακόμα να συμφιλιώσει τις διαφορετικές απόψεις.

Οι οντολογίες συγκεκριμένου γνωστικού πεδίου, όπως δηλώνει και το όνομά τους αναφέρονται και στοχεύουν στο να περιγράψουν ένα συγκεκριμένο επιστημονικό χώρο όπως για παράδειγμα τις βιοϊατρικές επιστήμες, τη φυσική, τη χημεία κτλ. Αυτή τη στιγμή υπάρχουν λίγες τέτοιες οντολογίες στον χώρο των βιοϊατρικών επιστημών. Η πλήρης κάλυψη του πεδίου παραμένει ζητούμενο και αποτελεί τον στόχο της OBO Foundry (Smith, Ashburner et al. 2007), μιας κοινοπραξίας επιστημόνων και εργαστηρίων στην οποία θα αναφερθώ παρακάτω. Στην πραγματικότητα οι οντολογίες που υπάρχουν περιγράφουν τμήμα μόνο κάποιου γνωστικού πεδίου όπως για παράδειγμα η ανατομία των οργανισμών, ή οι μολυσματικές ασθένειες και είναι στενά συνδεδεμένες με οντολογίες ανώτερου επιπέδου οι οποίες παρέχουν την φιλοσοφική βάση πάνω στην οποία στηρίζονται. Δημιουργούνται για να περιγράψουν από μια συγκεκριμένη οπτική γωνία το πεδίο στο οποίο ανήκουν χωρίς πρόβλεψη για την ανάπτυξη κάποιας εφαρμογής και συχνά αναφερόμαστε σε αυτές ως οντολογίες αναφοράς. Δεν έχουν πολλούς όρους, αλλά όσους διαθέτουν πρέπει να είναι ορισμένοι με τέτοιο τρόπο ώστε να καλύπτουν τις διαφορετικές περιπτώσεις με τους οποίους παρουσιάζονται αυτοί σε επιμέρους εφαρμογές. Για παράδειγμα οι όροι «μολυσματικός παράγοντας» και «παράσιτο» ανήκουν σε μια οντολογία αναφοράς, ενώ συγκεκριμένοι μολυσματικοί παράγοντες όπως το *Plasmodium falciparum* που ευθύνεται για την ελονοσία ανήκουν στην επιμέρους οντολογία που περιγράφει την ασθένεια.

Τέλος, οι οντολογίες συγκεκριμένων εφαρμογών στοχεύουν να καλύψουν την ανάγκη για οργάνωση συγκεκριμένων μετα-δεδομένων όπως για παράδειγμα της ανατομίας του κουνουπιού (Topalis, Tzavlaki et al. 2008) για να επισημειωθούν πειράματα γονιδιακής έκφρασης με την χρήση μικροσυστοιχιών DNA (Whetzel, Parkinson et al. 2006), ή της

ανθεκτικότητας σε εντομοκτόνα (Dialynas, Topalis et al. 2009), που εμφανίζουν πληθυσμοί εντόμων δυσχεραίνοντας τις προσπάθειες παρέμβασης και ελέγχου των ασθενειών που μεταδίδονται στον άνθρωπο από τους πληθυσμούς αυτούς. Οι οντολογίες που δημιουργούνται για συγκεκριμένες εφαρμογές βασίζονται σε μία ή περισσότερες οντολογίες αναφοράς. Χρησιμοποιούν όρους από αυτές και προχωρούν σε πιο λεπτομερή περιγραφή για να ικανοποιήσουν τις ανάγκες της εφαρμογής.

Έτσι για παράδειγμα, η οντολογία για την ελονοσία περιλαμβάνει στοιχεία τόσο από την ChEBI (Degtyarenko, de Matos et al. 2008) που περιέχει πληροφορίες για χημικές ουσίες που παρουσιάζουν βιολογικό ενδιαφέρον μιας και τα εντομοκτόνα που χρησιμοποιούνται για τον έλεγχο των πληθυσμών των κουνουπιών που μεταδίδουν ελονοσία, αλλά και τα αντιβιοτικά που χρησιμοποιούνται για την πρόληψη και θεραπεία της, εμπίπτουν σε αυτή την κατηγορία. Επίσης χρησιμοποιεί δεδομένα τόσο από την ENVO την οντολογία του περιβάλλοντος για να περιγράψει τους θώκους στους οποίους αναπτύσσονται τα κουνούπια αλλά και από την GO για να περιγράψει τις φυσιολογικές διαδικασίες των κουνουπιών. Επειδή δε, η GO προσπαθεί να παραμείνει γενική χωρίς ιδιαίτερες αναφορές σε συγκεκριμένα είδη στην IDOMAL δημιουργήθηκαν και χρησιμοποιούνται όροι για να περιγράψουν την διαδικασία αναζήτησης κατάλληλου ξενιστή από πλευράς των κουνουπιών.

Βασική Τυπική Οντολογία (BFO)

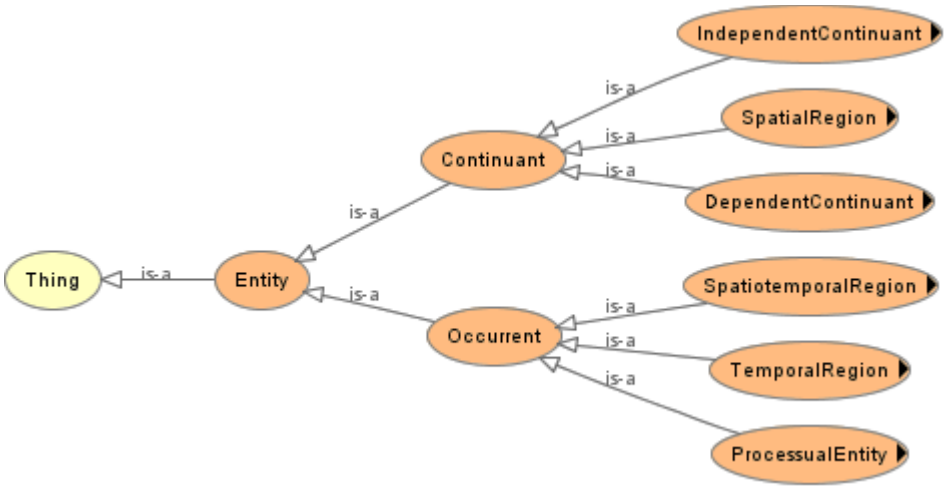
Ήδη έχουμε αναφέρει πως οι οντολογίες ανώτερου επιπέδου αντανακλούν διαφορετικές φιλοσοφικές θεωρήσεις της οντολογικής κατηγοριοποίησης χωρίς να έχουν αναφορές σε συγκεκριμένους επιστημονικούς τομείς ή εφαρμογές. Χρησιμεύουν για την διασύνδεση των οντολογιών που τις υιοθετούν και στηρίζονται στο γεγονός πως η πραγματικότητα είναι μία και αυτή περιγράφεται δια μέσου των οντολογιών. Έτσι για παράδειγμα ένας οργανισμός, ένας συγκεκριμένος φαινότυπος που εμφανίζει, κάποιο

αναπτυξιακό του στάδιο και η φυσιολογική διαδικασία του μεταβολισμού που αυτός επιτελεί για να επιβιώσει μπορούν όλα να χαρακτηριστούν σαν οντότητες. Ο τρόπος οργάνωσής τους στο ανώτερο (δηλαδή στο πιο γενικό) επίπεδο θα αποτελέσει μια οντολογία ανώτερου επιπέδου. Η βασική τυπική οντολογία (BFO) (Guarino 1998) είναι μια οντολογία ανώτερου επιπέδου (δεν είναι η μόνη που υπάρχει) και κατηγοριοποιεί τις διάφορες οντότητες με βάση την σχέση τους με τον χρόνο, διακρίνοντάς τες σε συνεχείς και ασυνεχείς (διαδικασίες). Το τμήμα που αφορά τις συνεχείς οντότητες ονομάζεται SNAP (Grenon and Smith 2004) και μπορεί να θεωρήσει κανείς πως αποτελεί μια αναπαράσταση της πραγματικότητας σε μια συγκεκριμένη χρονική στιγμή, όπως ακριβώς μια φωτογραφία και ακριβώς σε αυτή την ιδιότητα οφείλει και το όνομά της (SNAPshot). Το δεύτερο τμήμα περιλαμβάνει τις ασυνεχείς οντότητες που μεταβάλλονται στο χρόνο και ονομάζεται SPAN ακριβώς γιατί οι οντότητες που περιέχει διατρέχουν το χρόνο. Αν κανείς θα ήθελε να χρησιμοποιήσει το ανάλογο του προηγουμένου παραδείγματος της φωτογραφίας, η SPAN είναι ένα βίντεο ή μια ταινία που εξελίσσεται στον χρόνο (Rosse, Kumar et al. 2005).

Οι οντότητες που περιλαμβάνονται στην SNAP και στην SPAN (Εικόνα 1) μπορεί να διακριθούν σε ανεξάρτητες που είναι αυθύπαρκτες και στις εξαρτημένες που μπορούν να υπάρξουν μόνο όταν αποδοθούν σε κάποια αυθύπαρκτη οντότητα. Για παράδειγμα ένα έντομο ή γενικότερα οποιοσδήποτε οργανισμός είναι μια αυθύπαρκτη οντότητα, αλλά η ανθεκτικότητα σε εντομοκτόνα που εμφανίζει, ή γενικότερα ένας οποιοσδήποτε φαινότυπος δεν μπορεί να νοηθεί χωρίς την ύπαρξη αυτού που εμφανίζει τον δεδομένο φαινότυπο. Ακολουθώντας αυτήν την προσέγγιση όλες οι διαδικασίες ανήκουν στις εξαρτημένες οντότητες μιας και πραγματοποιούνται στα πλαίσια μιας ανεξάρτητης οντότητας.

Επίσης οι SNAP και SPAN (Grenon and Smith 2004) μπορούν να χρησιμοποιηθούν σε οντολογίες ανεξαρτήτως του επιπέδου λεπτομέρειας στο οποίο εκείνες

αναφέρονται. Για παράδειγμα σε μοριακό επίπεδο τα διάφορα μόρια μπορούν να θεωρηθούν αντικείμενα και το κύτταρο στο οποίο βρίσκονται η συνένωση αυτών των αντικειμένων. Σε επίπεδο οργανισμού τα κύτταρα μπορούν να θεωρηθούν αντικείμενα και οι πολυκύτταροι οργανισμοί ως η συνένωσή τους. Τέλος σε πληθυσμιακό επίπεδο οι πληθυσμοί μπορούν να θεωρηθούν σαν συνενώσεις ατόμων που αποτελούν αντικείμενα. Όσο οι οντότητες περιγράφονται και συνδέονται με λογικές σχέσεις με τις υπόλοιπες οντότητες της κάθε οντολογίας η διασύνδεση των μετα-δεδομένων είναι εξασφαλισμένη.



Εικόνα 1: Διαγραμματική απεικόνιση των ανώτερων επιπέδων της βασικής τυπικής οντολογίας

Μια αναλυτική παρουσίαση της βασικής τυπικής οντολογίας και η σύγκρισή της με τις υπόλοιπες οντολογίες ανώτερου επιπέδου ξεπερνά κατά πολύ τα πλαίσια αυτής της εισαγωγής η οποία σκοπό έχει την συνοπτική παρουσίασή της γιατί η βασική τυπική οντολογία αποτελεί την βάση πάνω στην οποία θα στηρίζονται οι οντολογίες που δημιουργούνται στα πλαίσια της OBO Foundry.

OBO Foundry

Η OBO Foundry (Smith, Ashburner et al. 2007) (OBO = Open Biological and Biomedical Ontologies - Βιολογικές και βιοϊατρικές οντολογίες ανοιχτού κώδικα) είναι μια συνεργατική προσπάθεια επιστημόνων και εργαστηρίων που αναπτύσσουν οντολογίες στον

χώρο των βιοϊατρικών επιστημών ώστε να υπάρξει μια σειρά διασυνδεδεμένων μη αλληλεπικαλυπτόμενων οντολογιών αναφοράς που να περιγράφουν τις βιοϊατρικές επιστήμες. Για να επιτευχθεί ο σκοπός αυτός κατά τον καλύτερο τρόπο υιοθετήθηκαν ορισμένες βασικές αρχές, που δεσμεύουν όσους δημιουργούν οντολογίες μέσα στα πλαίσια της OBO Foundry, οι κυριότερες από τις οποίες είναι οι εξής:

1) Οι οντολογίες είναι ανοιχτές και διαθέσιμες σε όλους να τις χρησιμοποιήσουν με τον όρο να αναφέρουν την πηγή και αν τις τροποποιήσουν οι ίδιοι καθ' οιονδήποτε τρόπο να μην τις διαθέσουν με το ίδιο όνομα σε κανέναν. Ο προτεινόμενος τρόπος αλλαγών είναι να ζητηθεί από αυτούς που δημιούργησαν την οντολογία να την τροποποιήσουν ανάλογα.
2) Οι οντολογίες πρέπει να έχουν αναπτυχθεί χρησιμοποιώντας ένα κοινό συντακτικό πρότυπο. Εκείνα της OBO και της OWL (Ontology Web Language) είναι αποδεκτά.
3) Στα πλαίσια της OBO Foundry οι οντολογίες διαθέτουν ένα μοναδικό αναγνωριστικό στοιχείο το οποίο κληρονομείται σε κάθε όρο της οντολογίας και συνήθως αποτελείται από ένα πρόθεμα κι έναν αριθμό. Το αναγνωριστικό στοιχείο κάθε όρου, αφορά τον ορισμό και όχι το όνομα του όρου.
4) Υπάρχει σαφής τρόπος διάκρισης ανάμεσα στις διαφορετικές εκδόσεις/ενημερώσεις της ίδιας οντολογίας.
5) Το περιεχόμενο της οντολογίας είναι διατυπωμένο με σαφήνεια και με τέτοιο ιεραρχικό τρόπο ώστε να αποφεύγονται οι κυκλικές συνδέσεις ανάμεσα στους όρους της.
6) Η οντολογία δεν πρέπει να αλληλεπικαλύπτεται με άλλες οντολογίες που ήδη υπάρχουν στην OBO Foundry. Αν αυτό συμβαίνει, τότε πρέπει να υπάρξει συνένωση των επιμέρους σε μια οντολογία αναφοράς.

7) Η οντολογία πρέπει να παρέχει ορισμούς για όλους τους όρους που περιέχει οι οποίοι κατά περίπτωση μπορεί να συνοδεύονται και από τυπικούς ορισμούς.

8) Η οντολογία πρέπει να χρησιμοποιεί σχέσεις που είναι μονοσήμαντα ορισμένες κατ' αναλογία αυτών που υπάρχουν στην οντολογία σχέσεων που συντηρεί η OBO Foundry (Smith, Ceusters et al. 2005).

9) Όλοι οι όροι πρέπει να διαθέτουν ένα μόνο γονικό όρο με τον οποίο να συνδέονται με την σχέση *is a.*

10) Όλοι οι όροι μιας οντολογίας πρέπει να είναι οντότητες που συναντώνται στην πραγματικότητα.

11) Όπου είναι δυνατόν, οι νέοι όροι πρέπει να προκύπτουν από ήδη υπάρχοντες όρους και σχέσεις που ήδη υπάρχουν σε άλλες οντολογίες της OBO Foundry. Για να προστεθούν νέοι όροι σε κάποια άλλη οντολογία προτείνεται μια διαδικασία ανάλογη με αυτή των αλλαγών που περιγράφεται στην πρώτη αρχή.

12) Οι οντολογίες θα αναπτύσσονται με την χρήση της βασικής τυπικής οντολογίας σαν οντολογία ανώτερου επιπέδου.

Σχέσεις που επιτρέπονται από την OBO Foundry	
is_a	has_agent
part_of	instance_of
integral_part_of	realizes
proper_part_of	inheres_in
located_in	bearer_of
contained_in	has_quality
adjacent_to	has_function
transformation_of	has_role
derives_from	has_disposition
preceded_by	has_participant

Πίνακας 1: Σχέσεις που είναι απαραίτητες για τις OBO οντολογίες

Τροπικές ασθένειες

Με τον όρο τροπικές ασθένειες περιγράφουμε ένα σύνολο ασθενειών που παρατηρούνται κυρίως ή αποκλειστικά στις τροπικές περιοχές του πλανήτη και οφείλονται σε μια πληθώρα μολυσματικών παραγόντων (πρωτόζωων, βακτηρίων, ιών) που μεταδίδονται στους ανθρώπους μέσω φορέων που τις περισσότερες φορές ανήκουν στην τάξη των αρθροπόδων. Ακριβώς αυτή η ιδιότητά τους έστρεψε το ενδιαφέρον στη μελέτη των φυσιολογικών διαδικασιών των φορέων που εμπλέκονται άμεσα ή έμμεσα στην μετάδοση της ασθένειας ως πιθανά σημεία ελέγχου, περιορισμού και τελικά αντιμετώπισής της. Η ολοκλήρωση της αλληλούχισης του γονδιώματος αρχικά του *Anopheles gambiae* (Holt, Subramanian et al. 2002) κυριότερου φορέα της ελονοσίας και στην συνέχεια μιας σειράς άλλων φορέων, όπως οι *Aedes aegypti* (Nene, Wortman et al. 2007*), Culex quinquefaciatus* (Arensburger, Megy et al. 2010)*, Ixodes scapularis* (Hill and Wikel 2005), *Pediculus humanus* (Pittendrigh, Clark et al. 2006)*, Rhodnius prolixus* (Huebner 2007)*, Glossina morsitans morsitans* έδωσε μεγάλη ώθηση στην σχετική έρευνα (Topalis, Lawson et al. 2008) (Πίνακας 2). Τα γονιδιώματα που αναλύθηκαν καθώς και οι σχετικές με αυτά πληροφορίες (επισημειώσεις κλπ) για τους 5 πρώτους οργανισμούς αποθηκεύθηκαν στην VectorBase (Lawson, Arensburger et al. 2007; Lawson, Arensburger et al. 2009)(Megy, Emrich et al. 2011), μια ενοποιημένη βάση δεδομένων που χρησιμοποιεί την αυτοματοποιημένη σουίτα ανάλυσης γονιδιωμάτων της ENSEMBL (Flicek, Amode et al. 2011) για την εύρεση ανοιχτών πλαισίων ανάγνωσης, πιθανών γονιδίων και την επισημείωση σε αυτών χαρακτηριστικών διαμέσου ανάλυσης μέσω υπολογιστών, ενώ παράλληλα με πολύ πιο αργούς ρυθμούς επιμελείται και την επισημείωση χαρακτηριστικών με μη αυτόματο τρόπο από επιμελητές-αναλυτές των διαθέσιμων αλληλουχιών. Σε αυτήν ενσωματώθηκαν και οι γενετικές πληροφορίες που υπήρχαν συγκεντρωμένες στην AnoBase (Topalis, Koutsos et al. 2005) που προϋπήρχε.

Είδος	Προσέγγιση	Κατάσταση (Αύγουστος 2011)
Aedes aegypti	Αλληλούχιση γονιδιώματος , EST	Ολοκληρώθηκε
Anopheles gambiae s.s. PEST	Αλληλούχιση γονιδιώματος , EST	Ολοκληρώθηκε
Anopheles gambiae s.s. M str	Αλληλούχιση γονιδιώματος	Ολοκληρώθηκε
Anopheles gambiae s.s. S str	Αλληλούχιση γονιδιώματος	Ολοκληρώθηκε
Culex quincefaciatus	Αλληλούχιση γονιδιώματος , EST	Ολοκληρώθηκε
Glossina morsitans morsitans	Αλληλούχιση γονιδιώματος , EST	Ολοκληρώθηκε
Glossina palpalis palpalis		Υπό σχεδιασμό
Ixodes scapularis	Αλληλούχιση γονιδιώματος , EST	Ολοκληρώθηκε
Lutzomyia longipalpis		Υπό σχεδιασμό
Musca domestica		Υπό σχεδιασμό
Pediculus humanus	Αλληλούχιση γονιδιώματος , EST	Ολοκληρώθηκε
Phlebotomus papatasi		Υπό σχεδιασμό
Rhodnius prolixus	Αλληλούχιση γονιδιώματος , EST	Ολοκληρώθηκε

Πίνακας 2: Κατάσταση προγραμμάτων αλληλούχισης των αρθροπόδων φορέων ασθενειών

Αντικείμενο της διατριβής

Αυτή ακριβώς η συγχώνευση της AnoBase στην VectorBase και η δημιουργία μιας πιο σύνθετης βάσης δεδομένων αποκάλυψε και την ανάγκη ανάπτυξης πιο σύνθετων εργαλείων για την καταχώρηση και την ανάκτηση δεδομένων από αυτήν. Από την άλλη πλευρά, πέρα από την VectorBase (Lawson, Arensburger et al. 2007; Lawson, Arensburger et al. 2009) που αυτή την στιγμή εξακολουθεί να είναι κυρίως μια βάση δεδομένων που περιέχει γενωμικές αλληλουχίες, υπάρχουν άλλες βάσεις δεδομένων και ερευνητικά προγράμματα που εστιάζουν στην καταγραφή και την ανάλυση της ποικιλομορφίας των πληθυσμών των φορέων, της ανθεκτικότητάς τους σε εντομοκτόνα και

της ικανότητάς τους να μεταδώσουν ασθένειες. Παρόλες αυτές τις προσπάθειες ο μεγαλύτερος όγκος των δεδομένων δεν υπάρχει καν καταχωρημένος σε υπολογιστές αλλά παραμένει σε έντυπη μορφή ή ακόμα και στις περιπτώσεις που είναι καταχωρημένος ηλεκτρονικά, κάθε ερευνητική ομάδα, πρόγραμμα ή εμπλεκόμενος φορέας χρησιμοποίησε δικά του πρότυπα. Αποτέλεσμα αυτού είναι κάθε σύνολο δεδομένων να είναι αποκομμένο από τα υπόλοιπα που ακολουθούν ένα διαφορετικό πρότυπο και να μην μπορεί να συνδυασθεί εύκολα με αυτά. Προφανής είναι η ανάγκη δημιουργίας κοινών προτύπων που θα αποτελέσουν τη βάση ώστε όλες οι πληροφορίες να καταστούν διαθέσιμες σ' όποιον επιθυμεί να τις χρησιμοποιήσει και για τον σκοπό αυτό είναι απαραίτητη η δημιουργία οντολογιών που θα καλύπτουν το πεδίο των τροπικών ασθενειών, η διασύνδεσή τους με τις υπόλοιπες οντολογίες του βιοϊατρικού χώρου και η επισημείωση των υπαρχόντων δεδομένων στις αντίστοιχες βάσεις με τους κατάλληλους όρους.

Αυτό ακριβώς επιχειρήθηκε με αυτήν τη διατριβή με την ανάπτυξη οντολογιών ανατομίας κουνουπιών και τσιμπουριών που αποτελούν τους φορείς αρκετών τροπικών νόσων, με την ανάπτυξη οντολογίας που αφορά την ανθεκτικότητα που εμφανίζουν ορισμένοι πληθυσμοί των ειδών αυτών στα εντομοκτόνα και τέλος μια ευρύτερη οντολογία που αφορά την ελονοσία. Βασισμένη στην οντολογία της ανθεκτικότητας σε εντομοκτόνα, δημιουργήθηκε μια βάση δεδομένων ως απόδειξη των δυνατοτήτων που παρέχουν οι οντολογίες στην δημιουργία σύγχρονων βάσεων δεδομένων.

ΚΕΦΑΛΑΙΟ ΠΡΩΤΟ: ΟΝΤΟΛΟΓΙΕΣ ΑΝΑΤΟΜΙΑΣ

Εισαγωγή

Η ολοκλήρωση της αλληλούχισης του γονιδιώματος ενός οργανισμού, δεν σηματοδοτεί το τέλος αλλά την αρχή μιας προσπάθειας. Η λειτουργική γονιδιωματική με τη βοήθεια πειραματικών προσεγγίσεων όπως οι μικροσυστοιχίες DNA, προσπαθεί να αναλύσει την δράση του συνόλου των μεταγράφων, των πρωτεϊνών, καθώς και τις αλληλεπιδράσεις των γονιδίων του υπό μελέτη οργανισμού (Rastan and Beeley 1997). Οι τεχνικές που χρησιμοποιούνται παράγουν μεγάλους όγκους δεδομένων για την ανάλυση των οποίων επιστρατεύονται εργαλεία βιοπληροφορικής. Βάσεις δεδομένων (Ashburner and Drysdale 1994; Christie, Weng et al. 2004; Blake, Bult et al. 2011) αποθηκεύουν τα αποτελέσματα η διαχείριση, η ανάλυση όσο και η παρουσίασή των οποίων προϋποθέτει την οργάνωσή τους με την χρήση των κατάλληλων επισημειώσεων και την προσθήκη μετα-δεδομένων.

Οι οντολογίες είναι από τα πιο χρήσιμα εργαλεία τα οποία διαθέτουμε για την οργάνωση μετα-δεδομένων. Όμως, ενώ η αλληλούχιση του *Anopheles gambiae* είχε ολοκληρωθεί δεν υπήρχε διαθέσιμη καμιά οντολογία πέρα από την GO (Gene Ontology Consortium 2006; Gene Ontology Consortium 2010) (η οποία δεν αναφέρεται σε κάποιο συγκεκριμένο είδος) που θα μπορούσε να χρησιμοποιηθεί για την επισημείωση σχετικών δεδομένων. Ταυτόχρονα είχαν ξεκινήσει οι προσπάθειες αποκωδικοποίησης της λειτουργίας των γονιδίων του κουνουπιού με την μελέτη του προφίλ του RNA κάθε ιστού και η ανάλυση του συνόλου των πρωτεϊνών του σαν πρώτα βήματα λειτουργικής γονιδιωματικής, ακολουθώντας το παράδειγμα και άλλων οργανισμών (Aitken 2005). Οι δύο πρώτες οντολογίες που κατασκευάσαμε στα πλαίσια της VectorBase περιέγραψαν την ανατομία/μορφολογία των κουνουπιών στην αρχή και στην συνέχεια των τσιμπουριών αφού

τόσο ο *Anopheles gambiae,* όσο και ο *Ixodes scapularis* συμπεριλαμβάνονται στους οργανισμούς που ενδιαφέρουν την VectorBase και μεταδίδουν μεταξύ άλλων, ασθένειες όπως η ελονοσία, η φιλαρίαση, ο δάγκειος πυρετός καθώς και διάφορα είδη εγκεφαλίτιδας. Η οντολογία του κουνουπιού κατασκευάστηκε πριν να υπάρξει κάποια οντολογία αναφοράς για την ανατομία των οργανισμών. Αυτή η οντολογία αναφοράς δημιουργήθηκε στην συνέχεια και ονομάστηκε CARO (Common Anatomy Reference Ontology) (Burger, Davidson et al. 2008). Η ύπαρξη μιας κοινής οντολογίας αναφοράς επιτρέπει την ανάπτυξη κοινών εργαλείων αναζήτησης δεδομένων συνδεδεμένων με ανατομικές δομές σε όλες τις βάσεις που περιέχουν ανάλογα δεδομένα ανεξαρτήτως οργανισμού. Επίσης συντελεί στη διασύνδεση οντολογιών ανατομίας ανάμεσα σε διαφορετικά είδη και μέσω αυτών στην συγκριτική ανάλυση των δεδομένων αυτών με έναν τρόπο που έχει νόημα και μπορεί να καταδείξει την ομολογία ανάμεσα στις δομές αυτές. Προϋπόθεση για τα παραπάνω είναι οι επιμέρους ανατομικές οντολογίες που αφορούν συγκεκριμένους οργανισμούς να βασίζονται στην οντολογία αναφοράς. Γι αυτόν τον λόγο, η οντολογία του κουνουπιού τροποποιήθηκε ώστε να ακολουθήσει τους κανόνες και τα πρότυπα που επέβαλε η CARO. Ανάλογες πρακτικές ακολουθήθηκαν και από άλλες οντολογίες που αφορούσαν ανατομίες οργανισμών όπως η *Drosophila melanogaster*, ενώ άλλες που δημιουργήθηκαν αργότερα υιοθέτησαν την CARO όπως η οντολογία ανατομίας των υμενοπτέρων HAO (Yoder, Miko et al. 2010) ή η οντολογία της ανατομίας των τελεόστεων (TAO) (Dahdul, Lundberg et al. 2010). Τέλος, παρόμοια δομή διέθετε και η οντολογία ανατομίας του ανθρώπου, FMA (Foundation Model of Anatomy) (Rosse and Mejino 2003) που είχε αναπτυχθεί νωρίτερα ανεξάρτητα από την CARO. Έχοντας στηριχθεί στην ίδια ανατομική οντολογία αναφοράς για τις οντολογίες του κουνουπιού, του τσιμπουριού και του ανθρώπου τα δεδομένα που θα αφορούσαν νοσήματα μεταδιδόμενα από αυτούς τους φορείς θα μπορούσαν με ευκολία να επισημειωθούν και να συσχετισθούν με οποιαδήποτε ιατρική βάση δεδομένων.

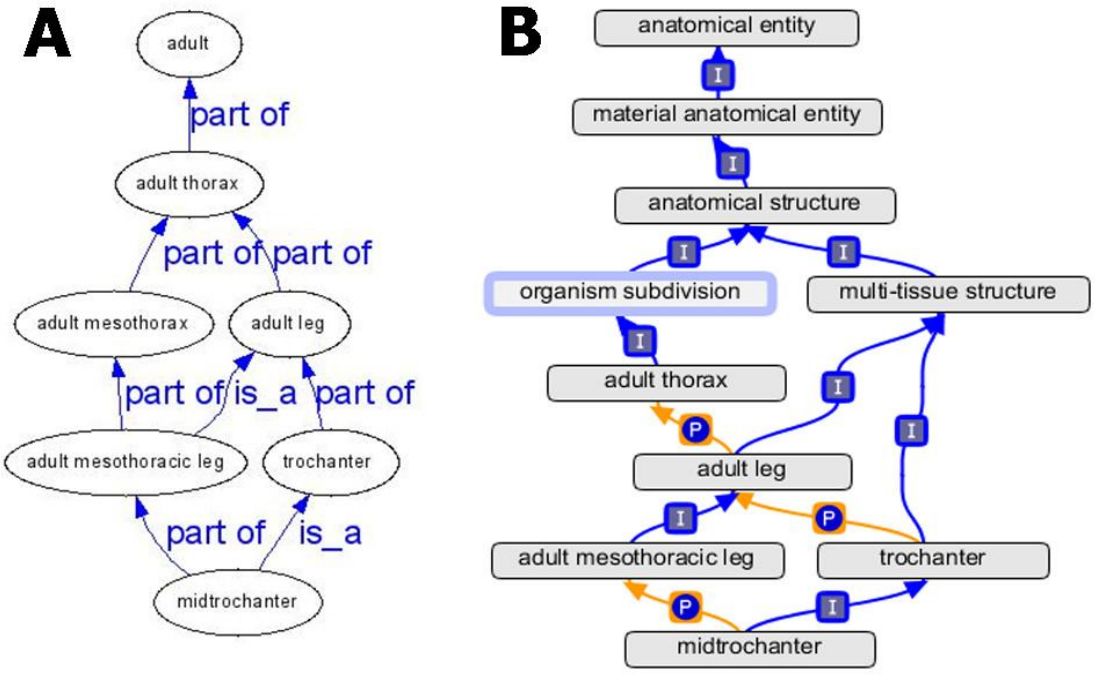
Υλικά και μέθοδοι

Για τη δημιουργία των οντολογιών χρησιμοποιήθηκε ένα πρόγραμμα ανάπτυξης οντολογιών, το OBO-Edit (Day-Richter, Harris et al. 2007) που είναι γραμμένο σε γλώσσα προγραμματισμού Java και δημιουργεί την οντολογία σαν ένα αρχείο απλού κειμένου το οποίο όμως ακολουθεί το πρότυπο obo που αναπτύχθηκε και υποστηρίχθηκε από την OBO Foundry ακριβώς για τον σκοπό αυτό (http://berkeleybop.org/~cjm/obo2owl/obo-syntax.html). Η επιλογή αυτή δημιούργησε προβλήματα συμβατότητας με οντολογίες εκτός του βιοϊατρικού χώρου οι οποίες εκφραζόταν σε μια άλλη γλώσσα γνωστή ως OWL (Ontology Web Language).

Η δυνατότητα επισκόπησης των οντολογιών ανατομίας μέσα από το διαδίκτυο, καθώς και η παρουσίαση των αντίστοιχων φωτογραφιών και σχηματικών παραστάσεων, όπου αυτό ήταν δυνατό, πραγματοποιήθηκε με τη δημιουργία νέων εξειδικευμένων μικροεφαρμογών γραμμένων αποκλειστικά για τον σκοπό αυτό σε γλώσσα προγραμματισμού PERL (Siever, Spainhour et al. 2004). Όπου χρειάσθηκε να δημιουργηθεί μια βάση δεδομένων αυτή ήταν βασισμένη στην MySQL (Vaswani 2010) ενώ επίσης χρησιμοποιήθηκε και η γλώσσα προγραμματισμού PHP (Welling and Thomson 2008) για τη διαχείριση κυρίως της βάσης δεδομένων. Ο εξυπηρετητής του διαδικτύου που χρησιμοποιήθηκε ήταν ο Apache (Aulds 2002). Όλα τα παραπάνω προγράμματα μπορούν να χρησιμοποιηθούν ελεύθερα και το ίδιο ισχύει και για τις εφαρμογές που δημιουργήθηκαν στα πλαίσια αυτής της διατριβής για τον χειρισμό και την επισκόπηση των οντολογιών. Η χρήση ελεύθερου λογισμικού συμβατού με όλα τα σύγχρονα λειτουργικά συστήματα (Windows, MacOSX, Linux) συντελεί στην ανεμπόδιστη διάδοση και χρήση των οντολογιών και των εργαλείων που αναπτύχθηκαν σε αυτή τη διατριβή.

Αποτελέσματα και συζήτηση

Στην πρώτη της μορφή η οντολογία της ανατομίας του κουνουπιού περιλάμβανε 1362 όρους κυρίως εξωτερικής μορφολογίας χωρισμένους σε 4 κατηγορίες ανάλογα με το αναπτυξιακό στάδιο στο οποίο ανήκαν (έμβρυο, προνύμφη, νύμφη, ενήλικο άτομο) και στην συνέχεια χωρισμένους ανά τμήμα οργανισμού (κεφάλι, θώρακας, κοιλιά για το ενήλικο άτομο, κεφαλοθώρακας, κοιλιά για την νύμφη). Οι όροι αυτοί αφορούσαν κυρίως την εξωτερική μορφολογία του κουνουπιού, η οποία και περιγράφεται πλήρως ενώ οι αναφορές σε εσωτερικές ανατομικές δομές και όργανα είναι περιορισμένες και αφορούν κυρίως το πεπτικό σύστημα και τους σιελογόνους αδένες του ενήλικου ατόμου καθώς έχουν ιδιαίτερο ρόλο στην μετάδοση παρασίτων και της διάδοση τροπικών ασθενειών. Στην πορεία του χρόνου και ανάλογα με τις ανάγκες επισημείωσης δεδομένων και με άλλους ανατομικούς όρους, αυτοί θα προστίθενται στην οντολογία. Όσον αφορά τις σχέσεις που είχαν χρησιμοποιηθεί για να συνδέσουν τους όρους στην αρχική μορφή της οντολογίας, κυριαρχούσε η σχέση *part_of* (Εικόνα 2Α). Η εικόνα αυτή έχει αλλάξει στην σημερινή της μορφή (Topalis, Tzavlaki et al. 2008). Περιλαμβάνει 1861 όρους που διαθέτουν 5178 συνώνυμα. Οι όροι που έχουν προστεθεί αναφέρονται κυρίως σε εσωτερικές ανατομικές δομές που προστέθηκαν για να είναι δυνατή η καλύτερη επισημείωση των γονιδίων. Χωρίς να έχουν αφαιρεθεί οι διασυνδέσεις μέσω της σχέσης *part_of* έχουν προστεθεί διασυνδέσεις μέσω της σχέσης *is_a* όπως απαιτεί η χρήση της CARO, που περιέχει και τους πιο γενικούς όρους της οντολογίας, όπως είναι αναμενόμενο για μια οντολογία αναφοράς (Εικόνα 2Β).



Εικόνα 2: Α. Στην αρχική οντολογία της ανατομίας του κουνουπιού δεν υπήρχε ολοκληρωμένη *is_a* ιεραρχία για κάθε όρο. Β. Μετά την υιοθέτηση της CARO η *is_a* ιεραρχία κατέστη πλήρης

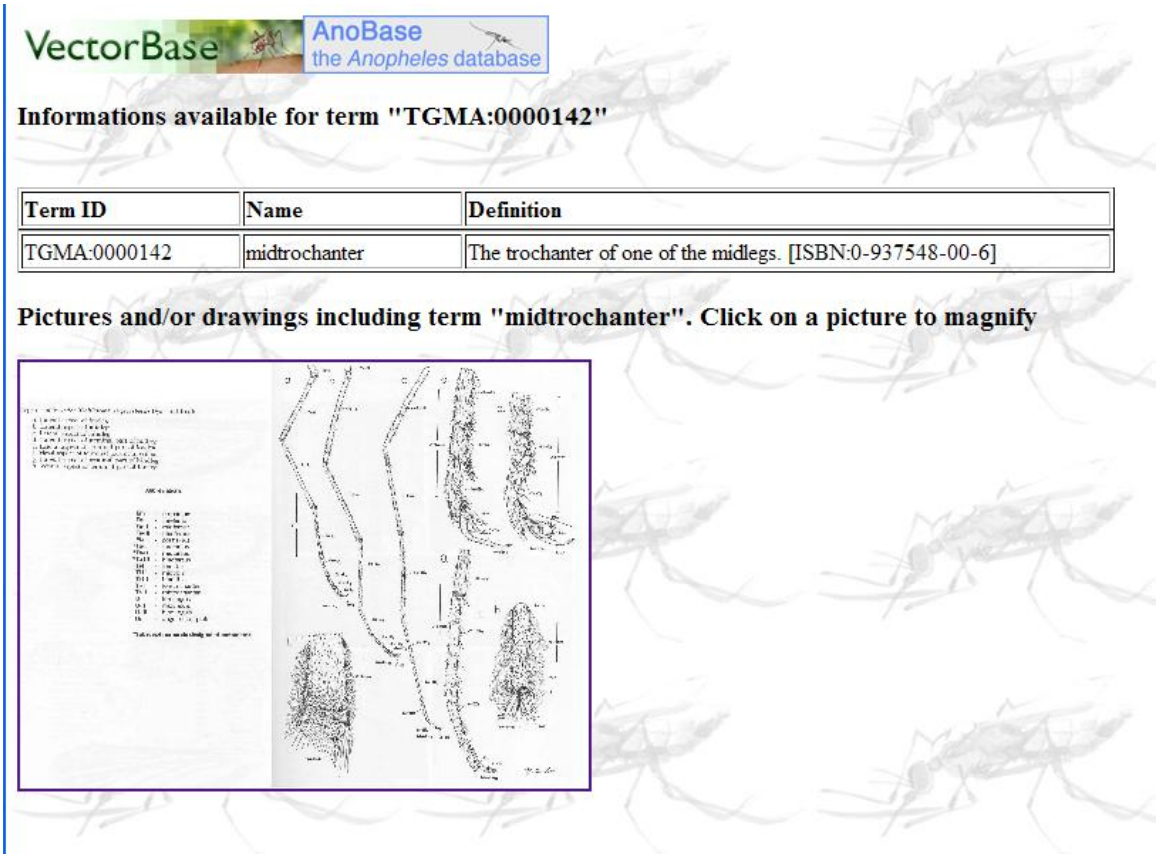
Μια ακόμα αλλαγή είναι πως το σύνολο των όρων διαθέτουν πλέον ορισμό ο οποίος συνήθως προέρχεται από το βιβλίο των Harbach και Knight "Taxonomist's Glosary of Mosquito Anatomy" (Harbach and Knight 1980). Αξίζει να τονιστεί πως η οντολογία αναφέρεται σε κουνούπια ευρύτερα και όχι μόνο στο είδος *Anopheles gambiae*. Συνεπώς κάποιες από τις ανατομικές δομές που περιγράφονται και αφορούν κυρίως το αναπαραγωγικό σύστημα του ενήλικου ατόμου καθώς και κάποια από τα στοματικά μέρη δεν υπάρχουν στον *Anopheles gambiae*. Η οντολογία της ανατομίας των τσιμπουριών χρονολογικά έπεται της CARO κι έτσι από την αρχή βασίστηκε σε εκείνη για να δημιουργηθεί. Αποτελείται από 628 όρους, 46 εκ των οποίων προέρχονται από την CARO, και 89 συνώνυμα. Όλοι οι όροι συνοδεύονται από ορισμό ο οποίος προέρχεται από το βιβλίο του καθ. Daniel Sonenshine "Biology of Ticks" (Sonenshine 1991) σε συνεργασία με τον οποίο δημιουργήθηκε η οντολογία.

Χρήση των οντολογιών ανατομίας

Η κύρια χρήση αυτών των οντολογιών είναι η επισημείωση πειραμάτων που περιλαμβάνονται στην VectorBase. Η οντολογία του κουνουπιού χρησιμοποιείται για αυτό τον σκοπό, η έλλειψη όμως των σχετικών δεδομένων για τα τσιμπούρια καθιστά τη χρήση της οντολογίας των τσιμπουριών πολύ περιορισμένη. Και οι δύο αυτές οντολογίες θα μπορούσαν να χρησιμοποιηθούν για την παρουσίαση και την περιγραφή των ανατομικών δομών σε ηλεκτρονική μορφή με την προϋπόθεση πως θα μπορούσε να υπάρξει ένα σύστημα να συνδέσει την περιγραφή τους με εικόνες. Άλλωστε ο διαχωρισμός σε ανατομικές δομές είναι μια τεχνητή διαδικασία που μας διευκολύνει στην περιγραφή και την κατανόηση της δομής και κατά συνέπεια της λειτουργίας του οργανισμού, δημιουργεί όμως μια ασάφεια ως προς τα λεπτομερή όρια των δομών μεταξύ τους. Η χρήση της CARO ως οντολογίας αναφοράς είχε ως αποτέλεσμα την χρήση καθαρά δομικών κριτηρίων και όχι λειτουργικών για την διάκριση και τον ορισμό των ανατομικών δομών του κουνουπιού και του τσιμπουριού. Για να γίνει αυτό περισσότερο σαφές, πέρα από τους ορισμούς σε κάθε όρο προσθέσαμε αναφορές για απεικονίσεις των δομών αυτών σαν σχόλιο στους όρους κάθε οντολογίας.

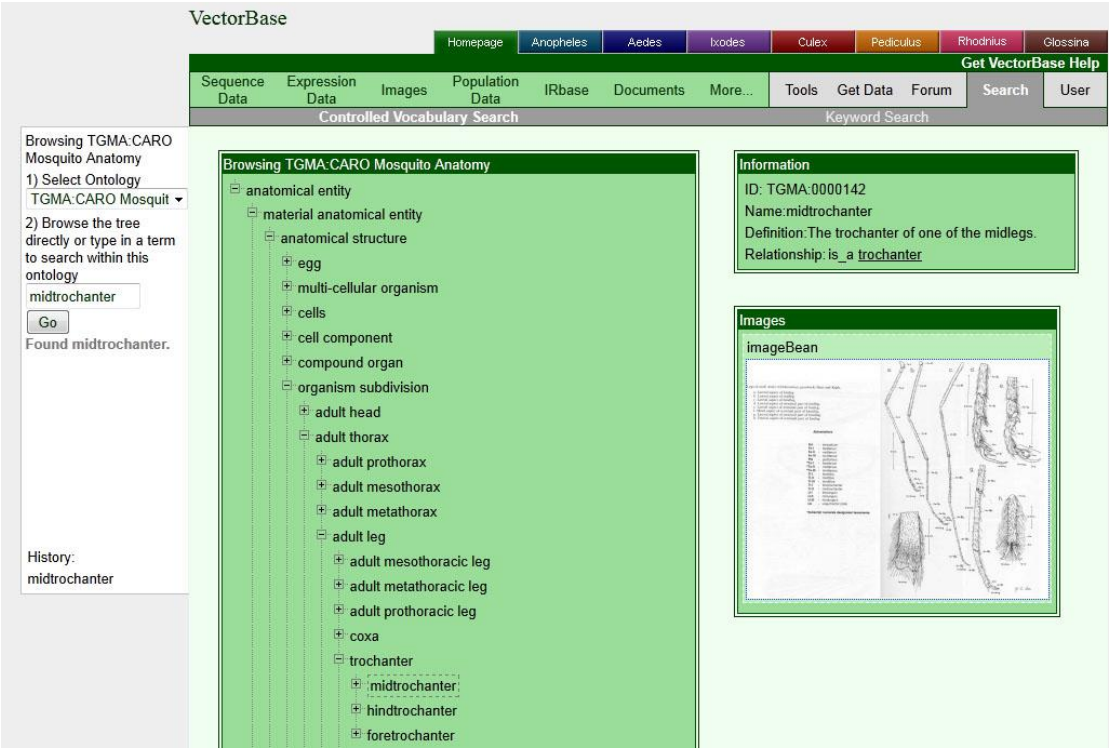
Τα γενικής χρήσης προγράμματα και εργαλεία που υπάρχουν για την παρουσίαση και την απεικόνιση των οντολογιών δεν έχουν προβλέψει τη δυνατότητα προβολής εικόνων, ούτε είναι ένα χαρακτηριστικό που ενδιαφέρει να προστεθεί (Noy, Shah et al. 2009; Srivastava and Sahni 2011). Η έμφαση δίνεται στην απεικόνιση της ίδιας της δομής της οντολογίας σαν ιεραρχία που διέπεται από ένα σύνολο σχέσεων. Για να μπορέσουμε να παρουσιάσουμε τις εικόνες αυτές που προέρχονται από τις πηγές που χρησιμοποιήθηκαν και για την δημιουργία της οντολογίας (με την άδεια αυτών που έχουν τα πνευματικά τους δικαιώματα) αναπτύχθηκε μια χωριστή εφαρμογή σε γλώσσα προγραμματισμού PERL η οποία επιτρέπει στον χρήστη να ψάξει στις συγκεκριμένες οντολογίες για ανατομικές δομές.

Για τον σκοπό αυτό οι όροι της οντολογίας μεταφορτώθηκαν σε μια βάση δεδομένων, με τέτοιο τρόπο ώστε πέρα από τους ορισμούς και τυχόν φωτογραφίες και σχηματικές αναπαραστάσεις να είναι καταχωρημένες και οι σχέσεις μεταξύ τους. Έτσι κάποιος μπορεί να υποβάλλει ερωτήματα στη βάση δεδομένων και να αναζητήσει όρους με βάση το όνομά τους, τα συνώνυμά τους, ή στοιχεία που βρίσκονται στον ορισμό του κάθε όρου. Το αποτέλεσμα της αναζήτησης είναι μια λίστα οντοτήτων μαζί με τους ορισμούς και τα συνώνυμά τους όπου περιέχεται αυτό το οποίο αναζήτησε ο χρήστης. Επιλέγοντας κάποια οντότητα ο χρήστης μπορεί να δει όλες τις φωτογραφίες/απεικονίσεις που περιέχουν τον συγκεκριμένο όρο. Έτσι ο συνδυασμός του ορισμού και της εικόνας μπορεί με απόλυτη σαφήνεια να περιγράψει την κάθε δομή που περιλαμβάνεται στις οντολογίες. Το εργαλείο όμως αυτό, δεν επιτρέπει την πλοήγηση μέσα στην οντολογία ακολουθώντας τις σχέσεις που την περιγράφουν (Εικόνα 3).



Εικόνα 3: Τα αποτελέσματα και η διαθέσιμη σχηματική αναπαράσταση του ανατομικού όρου "midtrochanter".

Η VectorBase ανέπτυξε ένα ανάλογο εργαλείο που επιτρέπει την εμφάνιση εικόνων και την πλοήγηση στους όρους της ανατομίας. Το μειονέκτημά του είναι πως δεν επιτρέπει την αναζήτηση μέσω συνωνύμων κι έτσι θεωρούμε πως είναι συμπληρωματικό πρώτου (Εικόνα 4).



Εικόνα 4: Τα αποτελέσματα της αναζήτησης για "midtrochanter" στο εργαλείο που αναπτύχθηκε από την VectorBase.

Το μέτρο της επιτυχίας μιας οντολογίας είναι η χρήση της. Κάτω από αυτό το πρίσμα η οντολογία του κουνουπιού χρησιμοποιείται για την επισημείωση των πειραμάτων γονιδιακής έκφρασης, ενώ ταυτόχρονα και οι δύο οντολογίες έχουν γίνει δεκτές από την OBO Foundry σαν υποψήφιες οντολογίες και έχουν χρησιμοποιηθεί και από άλλα ερευνητικά προγράμματα που προσπαθούν να φτιάξουν μια συνολική οντολογία ανατομίας μεταζώων χωρίς να λαμβάνουν υπόψη τους σχέσεις ομολογίας ανάμεσα στις δομές σε μια προσπάθεια να ταυτοποιήσουν περισσότερους φαινοτύπους αλλά και να βοηθήσουν την GO στην ανάπτυξη του τμήματος της που περιγράφει βιολογικές διαδικασίες. Μια τέτοια οντολογία είναι η UBERON στην οποία συμμετέχουν και οι δυο οντολογίες που δημιουργήθηκαν από εμάς (Haendel, Gkoutos et al. 2009).

ΚΕΦΑΛΑΙΟ ΔΕΥΤΕΡΟ: ΟΝΤΟΛΟΓΙΑ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ ΣΕ ΕΝΤΟΜΟΚΤΟΝΑ

Εισαγωγή

Από την στιγμή που έγιναν κατανοητά τα αίτια και ο τρόπος μετάδοσης αρκετών τροπικών νόσων και ο ρόλος που είχαν σε αυτόν τα έντομα-φορείς ήταν προφανές πως αν μπορούσε να ελεγχθεί ο πληθυσμός των εντόμων-φορέων αυτό θα σήμαινε και τον έλεγχο της ασθένειας. Εξαιτίας αυτού του φαινομένου όταν ανακαλύφθηκε η εντομοκτόνος δράση ορισμένων συνθετικών χημικών ουσιών, όπως το DDT στις αρχές της δεκαετίας του 1940 (Davidson 1951; Wright, Fritz et al. 1972), αυτές χρησιμοποιήθηκαν εκτενώς σε ψεκασμούς τόσο των καταφυγίων των κουνουπιών όσο και των ανθρώπινων κατοικιών. Αυτό συνέβη και στον ελλαδικό χώρο πειραματικά το 1945 (Vine 1947) και συστηματικά στην εξαετία 1946-1952 που η ελονοσία αποτελούσε πρόβλημα και μεταδιδόταν κυρίως από πληθυσμούς των *A. sacharovi* και *A. superpictus*. Η αποδοτικότητα των ψεκασμών ήταν εξαιρετική στην αρχή του προγράμματος και μάλιστα είχε σαν αποτέλεσμα την εξαφάνιση και άλλων εντόμων όπως η οικιακή μύγα (*Musca domestica*) από τις περιοχές όπου εφαρμόζονταν οι ψεκασμοί. Όμως ήδη από το 1947 οι πληθυσμοί της οικιακής μύγας όχι μόνο ξαναεμφανίζονται, αλλά πλέον δεν φαίνεται να επηρεάζονται από τους ψεκασμούς. Η αποτελεσματικότητα των ψεκασμών σε κάποιες περιοχές μειώνεται κατά πολύ τα τελευταία χρόνια του προγράμματος κάτι που αποδίδεται στην ανάπτυξη ανθεκτικότητας απέναντι στο DDT σε πληθυσμούς του *A. sacharovi* (Livadas and Georgopoulos 1953).

Ανάλογα φαινόμενα παρατηρήθηκαν και στις υπόλοιπες χώρες, οποιοδήποτε και αν ήταν το εντομοκτόνο το οποίο χρησιμοποιούνταν είτε για ψεκασμούς καταφυγίων κουνουπιών και σπιτιών, είτε για τον εμποτισμό υφασμάτων που στην συνέχεια χρησιμοποιούνταν ως κουνουπιέρες (Curtis 1991; Dabire, Diabate et al. 2006). Η εφαρμογή των μέτρων αυτών σε συνδυασμό με τα αντι-παρασιτικά φάρμακα αλλά και με τις

περιβαλλοντικές παρεμβάσεις (πχ. αποξηράνσεις ελών) οδήγησε στην εξαφάνιση της ελονοσίας από τις μη τροπικές περιοχές, αλλά όχι και από τις τροπικές (de Zulueta 1973). Μελέτες του Παγκόσμιου Οργανισμού Υγείας έδειξαν πως μέσα σε ένα διάστημα που ποικίλλει από 2 – 20 χρόνια για όλα τα γνωστά και ευρέως χρησιμοποιούμενα εντομοκτόνα εμφανίζεται ανθεκτικότητα (WHO Expert Committee on Vector Biology and Control. 1992). Σε αυτές τις περιπτώσεις η συνιστώμενη αντίδραση είναι η χρήση κάποιου άλλου εντομοκτόνου ή συνδυασμού εντομοκτόνων με διαφορετικό μηχανισμό δράσης (Brown 1958). Παρατηρήθηκε πως μετά από κάποιο χρονικό διάστημα από την αλλαγή χρήσης εντομοκτόνου η ανθεκτικότητα μπορεί να εξαφανιστεί από κάποια περιοχή (Keiding 1963). Κρίνεται λοιπόν αναγκαία η περιοδική παρακολούθηση των πληθυσμών των εντόμων που μεταδίδουν ασθένειες, ώστε να αξιολογείται η αποδοτικότητα της χρήσης εντομοκτόνων και να εντοπίζεται η εμφάνιση τυχόν ανθεκτικότητας όσο το δυνατόν γρηγορότερα (Curtis, Maxwell et al. 2006; Kulkarni, Malima et al. 2007). Επίσης απαραίτητος είναι και ο συντονισμός των αντιστοίχων προγραμμάτων παρέμβασης ανάμεσα σε γειτονικές χώρες στις ενδημικές περιοχές μιας και οι πληθυσμοί των κουνουπιών έχουν δυνατότητα μετανάστευσης.

Τα δεδομένα που υπάρχουν στη βιβλιογραφία και τα σχετικά πειράματα που γίνονται σήμερα σχετικά με την ανθεκτικότητα σ' εντομοκτόνα περιγράφουν τους πληθυσμούς των κουνουπιών και το αν εμφανίζουν ανθεκτικότητα σε κάποιο εντομοκτόνο στον χώρο και τον χρόνο (Chareonviriyahpap, Aum-aung et al. 1999; Briet, Galappaththy et al. 2005). Όμως, οι σχετικές πληροφορίες βρίσκονται συχνά αποθηκευμένες σε κάποιες βιβλιοθήκες μακριά από τους ανθρώπους που τις χρειάζονται. Κάποιες άλλες πληροφορίες δεν βρίσκονται ούτε στις δημόσιες βιβλιοθήκες παρά μόνο σε υπηρεσιακές αναφορές υπηρεσιών δημόσιας υγείας που δεν είναι εύκολα προσπελάσιμες από μη κυβερνητικούς παράγοντες. Ακόμα και σε περιπτώσεις που υπάρχουν σε ηλεκτρονική μορφή είναι

περιορισμένες οι δυνατότητες αναζήτησης και ανάκτησης τους. Όλα αυτά αναδεικνύουν την υπαρκτή ανάγκη για έναν ενιαίο τρόπο καταγραφής, παρουσίασης αλλά και δυνατότητας αναζήτησης των σχετικών δεδομένων, ανά περιοχή, ανά έτος και ανά εντομοκτόνο. Αυτό ακριβώς το κενό επιχειρήθηκε να καλυφθεί αρχικά στη βιολογική βάση δεδομένων του κουνουπιού (AnoBase) που διατηρούσαμε στην Κρήτη και στη συνέχεια μέσα από την VectorBase που απορρόφησε την Anobase (Topalis, Koutsos et al. 2005) με την ανάπτυξη εξειδικευμένων υπολογιστικών εργαλείων.

Αποτελέσματα – Συζήτηση

Προφόρμες καταγραφής δεδομένων

Για να συλλεχθούν και να καταγραφούν με ενιαίο τρόπο τα δεδομένα που αφορούν ανθεκτικότητα σε εντομοκτόνα έπρεπε οι συλλογείς δεδομένων να διαθέτουν κάποιον οδηγό για το πως γίνεται αυτό και να υπάρχουν κάποιοι γενικοί κανόνες που να διέπουν την διαδικασία. Ακολουθώντας το παράδειγμα της Flybase για τον χειρισμό των βιβλιογραφικών δεδομένων προκρίθηκε η λύση της δημιουργίας ενός συνόλου από προφόρμες – προκαθορισμένους πίνακες – τις οποίες οι συλλογείς που θα σταχυολογούσαν και θα αποδελτίωναν τις δημοσιευμένες πληροφορίες θα έπρεπε να συμπληρώσουν. Στην συνέχεια τα δεδομένα που υπήρχαν στις προφόρμες θα επεξεργάζονταν ηλεκτρονικά για να αποθηκευθούν στο σχετικό τμήμα της AnoBase που είχε αναπτυχθεί για τον σκοπό αυτό. Για την αποφυγή της άσκοπης εισαγωγής στοιχείων που είχαν ήδη καταχωρηθεί προκρίθηκε ένα σύστημα με ξεχωριστές αυτοτελείς προφόρμες όπου καθεμία αφορούσε και διαφορετικό είδος δεδομένων. Οι προφόρμες αυτές θα συνδέονταν μεταξύ τους με τη χρήση κάποιου κοινού πεδίου γεγονός που θα επέτρεπε την διασύνδεση του συνόλου των στοιχείων που περιείχαν. Το τελικό αρχείο καταγραφής δεδομένων θα ήταν το σύνολο των προφορμών που απαιτούνταν για την περιγραφή τους και η σειρά με την οποία βρίσκονταν οι προφόρμες σε αυτό ήταν σημαντική.

Δημιουργήθηκαν 2 διαφορετικές προφόρμες από την αρχή για να καλύψουν τα δεδομένα της ανθεκτικότητας σε εντομοκτόνα: Η πρώτη περιέγραφε λεπτομερώς τον πληθυσμό των εντόμων ο οποίος αναλυόταν στα σχετικά πειράματα, ενώ η δεύτερη θα έβρισκε εφαρμογή στην αυτή καθεαυτή περιγραφή των δεδομένων της ανθεκτικότητας. Με αυτόν τον σχεδιασμό στην συνήθη περίπτωση όπου ένας πληθυσμός είχε ελεγχθεί για την ανθεκτικότητα που τυχόν παρουσίαζε σε πολλαπλά εντομοκτόνα τα στοιχεία που τον αφορούσαν έπρεπε να εισαχθούν μόνο μία φορά. Στο τελικό αρχείο αυτή η μοναδική προφόρμα που θα περιέγραφε τον πληθυσμό θα ακολουθούνταν από πολλαπλές προφόρμες καταγραφής των πειραμάτων ανθεκτικότητας και το σύστημα θα αναγνώριζε πως οι τελευταίες αναφέρονται στην πρώτη από την σειρά τους στο αρχείο καταγραφής.

Κάθε πεδίο της προφόρμας είχε και ένα μοναδικό αντιπροσωπευτικό κωδικό. Τα περισσότερα πεδία έπρεπε να έχουν μοναδικές τιμές (πχ. το γεωγραφικό μήκος και πλάτος του τόπου από όπου προέρχεται ο πληθυσμός των εντόμων), ενώ υπήρχαν άλλα πεδία που μπορούσαν να δεχθούν πολλαπλές τιμές (όπως για παράδειγμα το ποια ήταν τα είδη που συμπεριλαμβάνονταν στην μελέτη). Η προφόρμα που περιέγραφε τους πληθυσμούς των εντόμων (Πίνακας 3) θα μπορούσε να χωριστεί σε διάφορα τμήματα: Τα αρχικά πεδία αφορούσαν τα είδη που περιελάμβανε η έρευνα και στην συνέχεια υπήρχε μια λεπτομερής αναφορά της γεωγραφικής θέσης του πληθυσμού καθώς και των περιβαλλοντολογικών ιδιαιτεροτήτων της. Πρόσθετες πληροφορίες που αφορούσαν τον χρόνο διεξαγωγής της έρευνας και των κλιματολογικών συνθηκών επίσης συλλέγονταν, ενώ ένα σημαντικό τμήμα της προφόρμας αφορούσε τη γενετική περιγραφή του πληθυσμού καθώς και αλληλομόρφων που είναι γνωστό πως σχετίζονται με ανθεκτικότητα στα εντομοκτόνα. Τέλος, υπήρχαν πεδία για να περιγράψουν αν υπήρχε DNA διαθέσιμο από άτομα του συγκεκριμένου πληθυσμού, ώστε τυχόν ενδιαφερόμενοι να μπορέσουν να το χρησιμοποιήσουν για δικά τους πειράματα.

Κωδικός	**Περιγραφή**
PO01	Κύρια είδη πληθυσμοί των οποίων συμμετείχαν στην μελέτη
PO02	Πρόσθετα είδη πληθυσμοί των οποίων συμμετείχαν στην μελέτη
PO03	Γεωγραφικοί ήπειροι όπου η μελέτη έλαβε χώρα [Africa, Europe, etc.]
PO04	Κράτος όπου η μελέτη έλαβε χώρα
PO05	Ζώνη ή περιοχή μελέτης [Eastern, dry, savanna, forest, mangrove swamp]
PO06	Πόλη ή χωριό μελέτης
PO07	Γεωγραφικό μήκος
PO08	Γεωγραφικό πλάτος
PO09	Υψόμετρο
PO10	Έτος μελέτης
PO11	Χρονική περίοδος μελέτης
PO12	Εποχή μελέτης
PO13	Ώρα ημερήσιας συλλογής δειγμάτων
PO14	Θερμοκρασία
PO15	Σχετική υγρασία
PO16	Μέσο ύψος βροχής
PO17	Αλατότητα υδάτων
PO18	Μέσο pH υδάτων
PO19	Μέθοδος ταυτοποίησης του είδους [morphological, cytological, etc.]
PO20	Φύλο εντόμων [male, female, both, unspecified]
PO21	Αναπτυξιακό στάδιο [larvae, pupae, adults, preimarginal states]
PO22	Δραστηριότητα [resting adults, endo- & exophilic, endo- & exophagy]
PO23	Κατάσταση ως προς το φαγητό [teneral, post-teneral, unfed, blooded, gravid]

PO24	Εγκαθιδρύθηκε εργαστηριακό στέλεχος ; [ΝΑΙ , ΟΧΙ]
PO25	Όνομα εργαστηριακού στελέχους
PO25α	Αν εγκαθιδρύθηκε εργαστηριακό στέλεχος, έτος εγκαθίδρυσης
PO25β	Αν εγκαθιδρύθηκε εργαστηριακό στέλεχος, τόπος προέλευσης
PO26	Μέθοδος συλλογής [indoor, outdoor, human, animal, pyrethrum, traps]
PO27	Αριθμός κουνουπιών που συλλέχτηκαν
PO28	Τα κουνούπια που συλλέχτηκαν αποθηκεύθηκαν;
PO29	Αύξων αριθμός ή όνομα στελεχών που συλλέχτηκαν
PO30	Είδη δεικτών που χρησιμοποιήθηκαν για την περιγραφή του πληθυσμού
PO31	Ονόματα αλληλομόρφων των δεικτών που χρησιμοποιήθηκαν
PO32	Τύπος στατιστικής ανάλυσης που χρησιμοποιήθηκε
PO33	Σχετική πυκνότητα κουνουπιών
PO34	Αναμενόμενη συχνότητα ετεροζυγωτίας για τους δείκτες που αναλύθηκαν
PO35	Παρατηρούμενη συχνότητα ετεροζυγωτίας για τους δείκτες που αναλύθηκαν
PO36	Συσχέτιση γονιδιακής ποικιλομορφίας
PO37	Επίπεδα σημαντικότητας στατιστικών τεστ που χρησιμοποιήθηκαν
PO38	Άλλα δεδομένα που συνδέονται με την μελέτη
PO39	Σχόλια σχετικά με τον πληθυσμό των κουνουπιών
PO40	Σχόλια συγκρίσεις με άλλους πληθυσμούς (ανταγωνισμός για πόρους)
PO41	Σχόλια σχετικά με χρήση και ανάλυση άλλων δεικτών
PO42	Σχόλια γενικού περιεχομένου
PO43	Εσωτερικές σημειώσεις αόρατες στον τελικό χρήστη

Πίνακας 3: Η αρχική προφόρμα συλλογής πληθυσμιακών δεδομένων.

Η δεύτερη προφόρμα (Πίνακας 4) αφορούσε αυτή καθεαυτή την συλλογή δεδομένων από πειράματα με σκοπό την παρακολούθηση και τον εντοπισμό της ύπαρξης

ανθεκτικότητας σε κάποιο εντομοκτόνο οπότε και πάλι υπήρχε καταγραφή του είδους που χρησιμοποιήθηκε στο πείραμα (και αποτελούσε το σημείο διασύνδεσης με την προηγούμενη προφόρμα) σε συνδυασμό με λεπτομέρειες για την ενεργή εντομοκτόνο ουσία που δοκιμάστηκε, την συγκέντρωσή της, τον τρόπο χορήγησής της, καθώς επίσης και τα αποτελέσματα του πειράματος.

Κωδικός	**Περιγραφή**
IR01	Είδος κουνουπιού που χρησιμοποιήθηκε στην μελέτη (υποχρεωτικό πεδίο)
IR02	Εντομοκτόνο που χρησιμοποιήθηκε (χημικό όνομα) (υποχρεωτικό πεδίο)
IR03	Εμπορικό όνομα εντομοκτόνου που χρησιμοποιήθηκε
IR04	Αριθμός OMS
IR05	Μίγμα της ενεργής εντομοκτόνου ουσίας με κάποιο μέσο, έτοιμο προς χρήση
IR06	Συγκέντρωση του εντομοκτόνου που χρησιμοποιήθηκε
IR07	Η μελέτη αναφέρεται σε πληθυσμό πεδίου
IR08a	Η μελέτη αναφέρεται σε εργαστηριακό πληθυσμό
IR08b	Ηλικία πληθυσμού
IR09	Είδος επιλεκτικής πίεσης
IR10	Ποσοστό θνησιμότητας
IR11	LD50
IR12	Είδος γονιδίου που επιφέρει ανθεκτικότητα στο εντομοκτόνο
IR13	Συχνότητα αλληλόμορφου ανθεκτικότητας
IR14	Σχόλια – Γενικά σχόλια ορατά σε όλους
IR15	Εσωτερικές σημειώσεις αόρατες στους τελικούς χρήστες
IR16	Κατηγορία εντομοκτόνου
IR17	KDT50 – Χρόνος που απαιτείται για την αναισθητοποίηση του 50%
IR18	KDT95 – Χρόνος που απαιτείται για την αναισθητοποίηση του 95%
IR19	Αριθμός CAS εντομοκτόνου

Πίνακας 4: Η προφόρμα για την συλλογή δεδομένων ανθεκτικότητας σε εντομοκτόνα.

Επειδή ο χαρακτηρισμός ενός πληθυσμού ως «ανθεκτικού» απέναντι σε κάποιο εντομοκτόνο εμπεριέχει στοιχεία υποκειμενικότητας και ο σκοπός μιας βάσης δεδομένων είναι η παρουσίασή τους και όχι η επεξεργασία τους, η προφόρμα κατέγραφε τα πρωτογενή

αποτελέσματα είτε σε μορφή ποσοστού θνησιμότητας, είτε σε μορφή θανατηφόρου δόσης, είτε σε μορφή ποσοστού αναισθητοποίησης και τα παρουσίαζε ως τέτοια, αφήνοντας τα όποια συμπεράσματα περί ανθεκτικότητας στον χρήστη του συστήματος. Τέλος στις προφόρμες είχε προβλεφθεί χώρος για σχόλια που θα επεξηγούσαν κάποια πεδία και θα ήταν ορατά σαν ελεύθερο κείμενο στους τελικούς χρήστες, όσο και για εσωτερικά σχόλια που αφορούσαν τη διαδικασία επιμέλειας και εισόδου των στοιχείων στο σύστημα.

Οι δύο προφόρμες που αναφέρθηκαν είναι προφανές πως δεν αρκούν για την πλήρη καταγραφή των δεδομένων. Υπάρχει ανάγκη για παράδειγμα να συλλεχθούν πληροφορίες για τυχόν δημοσιεύσεις στις οποίες αναφέρονται τα στοιχεία που έχουν προκύψει (στην πραγματικότητα στην αρχική τουλάχιστον περίοδο όλα τα δεδομένα προερχόταν από την βιβλιογραφία) καθώς επίσης και πληροφορίες για τους ερευνητές που διεξήγαγαν τις έρευνες ώστε ο ενδιαφερόμενος να μπορεί να έρθει σε επαφή μαζί τους αν θα το επιθυμούσε. Όλα αυτά τα δεδομένα καταγράφονταν σε διαφορετικές προφόρμες που αναπτύχθηκαν από την Flybase και υιοθετήθηκαν από εμάς και γι αυτό τον λόγο δεν πρόκειται να αναφερθώ στην παρούσα εργασία.

Ο τρόπος με τον οποίο γινόταν η εισαγωγή των δεδομένων απαιτούσε την δημιουργία χωριστών αρχείων καταγραφής δεδομένων για κάθε μελέτη. Στην αρχή θα έπρεπε να υπάρχουν τα δεδομένα για βιβλιογραφικές αναφορές, τους ερευνητές που διεξήγαγαν την έρευνα και στην συνέχεια η πρώτη προφόρμα που περιέγραφε αναλυτικά τον πληθυσμό ακολουθούμενη από τα πειραματικά δεδομένα της δεύτερης προφόρμας. Στην πλειοψηφία των περιπτώσεων όπου δείγματα του αρχικού πληθυσμού είχαν χρησιμοποιηθεί για να διακριβωθεί η ανθεκτικότητα σε πολλαπλά εντομοκτόνα, τα αποτελέσματα για κάθε εντομοκτόνο χωριστά (δεύτερη προφόρμα) θα συνόδευαν την αρχική περιγραφή. Στην συνηθισμένη περίπτωση όπου μια δημοσίευση περιελάμβανε δεδομένα για πολλούς

πληθυσμούς, τη μία βιβλιογραφική προφόρμα ακολουθούσαν πολλές που περιέγραφαν τους διαφορετικούς πληθυσμούς κ.ο.κ. Το τελικό αρχείο καταγραφής ήταν ένα μεγάλο αρχείο κειμένου όπου με ένα σειριακό τρόπο η μία κάτω από την άλλη συνδυαζόταν όσες προφόρμες χρειαζόντουσαν για να περιγραφούν τα δεδομένα με σαφή ιεραρχική σειρά. Αυτό θα μπορούσε να γίνει χειροκίνητα από τον συλλογέα/καταγραφέα των δεδομένων ενώ υπήρχε η δυνατότητα να γίνεται και ημιαυτοματοποιημένα με την χρήση μιας εφαρμογής που αναπτύξαμε για τον σκοπό αυτό. Η προσέγγιση αυτή εξασφάλισε την ομοιομορφία και την δυνατότητα αυτοματοποίησης της εισαγωγής δεδομένων, όμως αυτό δεν ήταν αρκετό καθώς επέτρεπε παντού την εισαγωγή ελεύθερου κειμένου. Για όσα πεδία ήταν δυνατό και αφορούσαν για παράδειγμα ονόματα ειδών οργανισμών, εντομοκτόνων, κατηγορίες εντομοκτόνων κλπ δημιουργήθηκαν λίστες επιτρεπτών τιμών και αυτός που εισήγαγε τα δεδομένα στο σύστημα αντί να πληκτρολογήσει, έπρεπε απλά να διαλέξει την κατάλληλη τιμή, μειώνοντας την πιθανότητα να υπάρχουν λάθη πληκτρολόγησης και επιβάλλοντας την ενιαία χρήση κάποιων όρων (πχ. «*An. gambiae s.s.*» στην θέση του «*Anopheles gambiae* sensu stricto»). Το σύστημα ολοκληρωνόταν με μια σειρά από προγράμματα επεξεργασίας τα οποία εισήγαγαν τα δεδομένα που είχαν συλλεχθεί στις προφόρμες σε μια σχεσιακή βάση δεδομένων. Ο τελικός χρήστης αποκτούσε πρόσβαση σε αυτά μέσω της ιστοσελίδας της AnoBase που του επέτρεπε να δει και να αναζητήσει τις πληροφορίες για τα δεδομένα ανθεκτικότητας που τον ενδιέφεραν ανά είδος οργανισμού, περιοχή, χρονική περίοδο και εντομοκτόνο. Με αυτό τον τρόπο η πληροφορία που βρισκόταν στις βιβλιοθήκες θα έφτανε στους υπολογιστές των ανθρώπων που τις χρειαζόντουσαν ακόμα κι αν αυτοί δεν είχαν πρόσβαση στις βιβλιοθήκες, κάτι πολύ συνηθισμένο στις χώρες του Τρίτου κόσμου που είναι και αυτές που μαστίζονται από τις τροπικές νόσους.

Οντολογία ανθεκτικότητας σε εντομοκτόνα

Το σύστημα με τις προφόρμες ήταν ένα βήμα προς την σωστή κατεύθυνση αλλά σύντομα αποκάλυψε τις αδυναμίες του. Η βασικότερη αδυναμία ήταν πως οι προφόρμες παρόλο που χρησιμοποιούσαν λίστες όρων για ορισμένα πεδία έλειπε η συσχέτιση μεταξύ τους και υπήρχε αδυναμίας χρήσης συνώνυμων όρων. Για παράδειγμα το γεγονός πως τα είδη *Anopheles arabiensis* , *Anopheles bwambe, Anopheles melas, Anopheles merus* και *Anopheles quadriannulatus* μαζί με τα *Anopheles gambiae* sensu stricto αναγνωρίστηκαν σαν σύμπλοκο είδος με το όνομα *Anopheles gambiae* sensu lato ήταν πηγή προβλημάτων ακόμα και με την χρήση της προφόρμας. Οι πληθυσμοί των κουνουπιών δεν μπορούσαν να συσχετισθούν με τους υπόλοιπους καθώς αντιμετωπίζονταν σαν διαφορετικές οντότητες μεταξύ τους και δεν υπήρχε δυνατότητα συσχέτισης. Έχοντας την εμπειρία των οντολογιών της ανατομίας που περιγράφηκαν στο προηγούμενο κεφάλαιο και των πλεονεκτημάτων τους, γνωρίζαμε πως η λύση ήταν η δημιουργία μιας οντολογίας που θα περιγράφει την ανθεκτικότητα σε εντομοκτόνα. Η οντολογία αυτή θα ήταν η βάση των εφαρμογών που θα αναπτύσσονταν για την μελέτη της ανθεκτικότητας σε εντομοκτόνα. Με πυρήνα τις υπάρχουσες προφόρμες προχωρήσαμε στη δημιουργία της.

Όπως και στο σύστημα με τις προφόρμες κεντρικό σημείο της οντολογίας αποτελούν τα πειράματα ταυτοποίησης και παρακολούθησης της ανθεκτικότητας που εμφάνιζε ένας πληθυσμός απέναντι στα εντομοκτόνα. Σύντομα όμως έγινε αντιληπτό πως μια πληθώρα διαφορετικών οντοτήτων πρέπει να χρησιμοποιηθούν για να περιγράψουν κατάλληλα τα υπάρχοντα δεδομένα. Συχνά δε αυτές οι οντότητες αλληλεπικαλύπτονταν μεταξύ τους. Για παράδειγμα πειραματικά εξακριβώνεται και το σε ποιο είδος ανήκει ένας οργανισμός αλλά και το αν παρουσιάζει ανθεκτικότητα σε κάποιο εντομοκτόνο. Τα πειράματα έχουν διαφορετικό στόχο, αλλά από οντολογικής άποψης εξακολουθούν να είναι πειράματα άρα ανήκουν στο ίδιο κλαδί της υπό ανάπτυξη οντολογίας. Αυτό σηματοδοτούσε

την υποχρέωση της οντολογίας να οργανώσει τα δεδομένα με διαφορετικό τρόπο σε σύγκριση με τις προφόρμες αναδιοργανώνοντάς την.

Η πρώτη απόφαση αφορούσε το αν θα χρησιμοποιηθεί μια οντολογία ανώτερου επιπέδου ως βάση για την δημιουργία της οντολογίας. Αρχικά επιλέξαμε να αποφύγουμε κάτι τέτοιο καθώς επιζητούσαμε τη μέγιστη δυνατή συμμετοχή της επιστημονικής κοινότητας στη διαμόρφωση της οντολογίας. Αυτή θα ήταν πιο αποδοτική με μια ευθεία προσέγγιση όπου οι γνώστες του πεδίου θα αντιμετώπιζαν άμεσα τους οικείους όρους που το περιγράφουν παρά η χρήση ενός θεωρητικού οικοδομήματος που θα καθιστούσε την οντολογία δυσνόητη για όσους δεν ήταν εξοικειωμένοι με οντολογίες. Κατά τα άλλα υιοθετήθηκαν όλοι οι κανόνες που έθεσε η OBO Foundry και οι σχέσεις περιορίστηκαν σε *is*_a , *part*_of και *agent_in* (πίνακας με τις σχέσεις και τους ορισμούς τους). Όροι που ανήκαν σε άλλες οντολογίες αναφοράς αποφασίστηκε να χρησιμοποιηθούν ως έχουν. Η οντολογία που δημιουργήθηκε ονομάστηκε MIRO (Mosquito Insecticide Resistance Ontology) (Dialynas, Topalis et al. 2009) αντανακλώντας τη συσχέτιση και την εστίασή της σε θέματα τροπικών νόσων που μεταδίδονται από κουνούπια. Διέθετε 4 αρχικές κατηγορίες :

α) «Βιολογικό υλικό» που περιέγραφε τους οργανισμούς και τους πληθυσμούς που ελεγχόντουσαν για το αν παρουσιάζουν ανθεκτικότητα ή όχι.

β) «Εντομοκτόνες ουσίες» που περιλάμβανε όλα τα εντομοκτόνα που χρησιμοποιούνταν στις αντίστοιχες έρευνες.

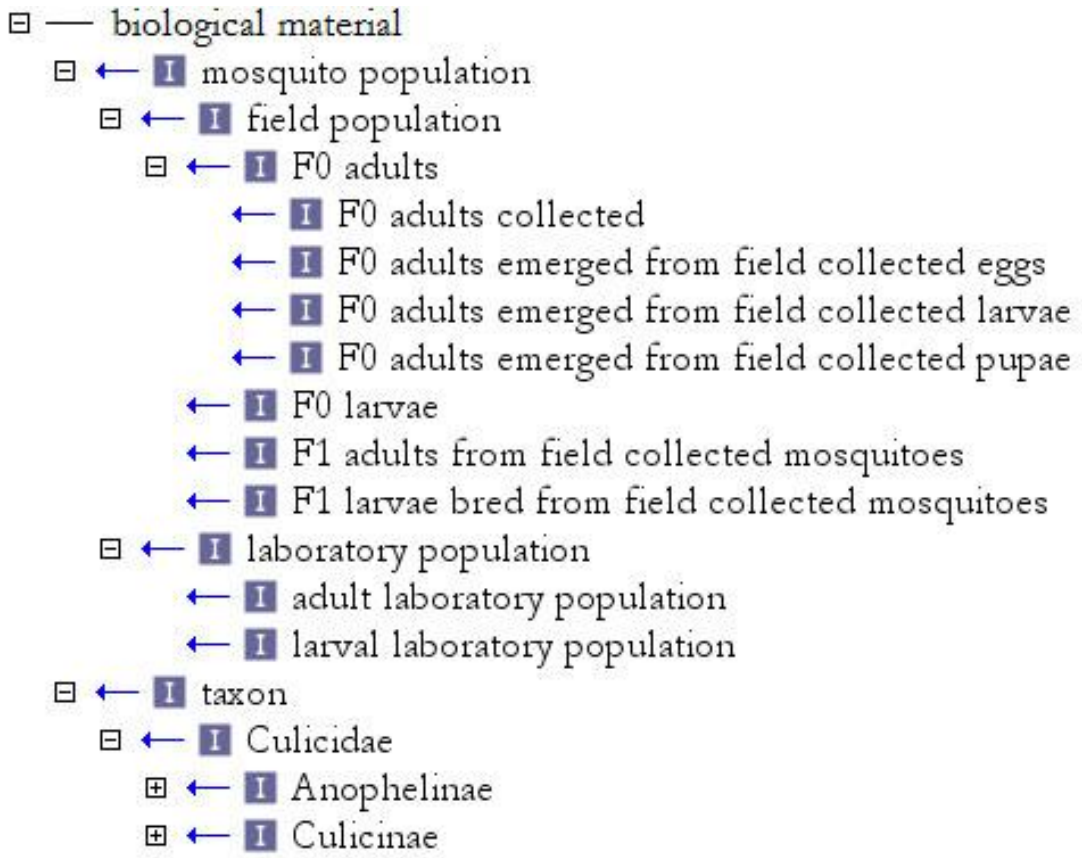
γ) «Μέθοδοι» που περιλάμβανε συγκεντρωμένες όλες τις πειραματικές μεθόδους.

δ) «Ανθεκτικότητα» που ταξινομούνται όλοι οι γνωστοί μηχανισμοί μέσω των οποίων επιτυγχάνεται η ανθεκτικότητα σε κάποιο εντομοκτόνο.

Μια πέμπτη κατηγορία που αφορούσε τις γεωγραφικές τοποθεσίες δεν πρόκειται να συζητηθεί καθώς αποτελεί μια ξεχωριστή οντολογία άλλων ερευνητών και δεν αποτελεί μέρος της εργασίας αυτής. Χρησιμοποιήθηκε όμως αυτούσια παρέχοντας τους κατάλληλους γεωγραφικούς όρους που χρειαζόταν η MIRO. Είναι αυτονόητο πως όλες οι προαναφερθείσες κατηγορίες δεν είναι στατικές, αλλά ανανεώνονται όταν παραστεί ανάγκη.

Α) ΒΙΟΛΟΓΙΚΟ ΥΛΙΚΟ

Η κατηγορία περιλαμβάνει όρους σχετιζόμενους με την περιγραφή των πληθυσμών που συμμετέχουν στα πειράματα ανθεκτικότητας και την ταξινόμησή τους. Η περιγραφή αφορά την προέλευσή τους (αν είναι κάποιο γνωστό στέλεχος εργαστηρίου ή φυσικός πληθυσμός που αιχμαλωτίστηκε πρόσφατα για τους σκοπούς του πειράματος) καθώς και το αναπτυξιακό στάδιο στο οποίο ανήκαν τα άτομα του πληθυσμού (Εικόνα 5).



Εικόνα 5: Τα τρία πρώτα επίπεδα όρων της MIRO που αναφέρονται στο βιολογικό υλικό.

Για την ταξινόμηση των οργανισμών περιοριστήκαμε στην οικογένεια *Culicidea* καθότι το πρωταρχικό ενδιαφέρον μας ήταν να μπορέσουμε να συμπεριλάβουμε δεδομένα από προγράμματα παρατήρησης ανθεκτικότητας σε πληθυσμούς κουνουπιών που μεταδίδουν τροπικά νοσήματα, αναγνωρίζοντας πως όλα τα *Culicidea* δεν είναι φορείς κάποιας νόσου, αλλά και πως υπάρχουν και άλλα αρθρόποδα πέρα από τα κουνούπια (πχ ακάρεα) που μεταφέρουν ασθένειες. Η διαθέσιμη ταξινομική οντολογία από το NCBI δεν κάλυπτε τις ανάγκες της οντολογίας μας καθώς υπήρχε περιορισμένη κάλυψη των *Culicidea* με αναφορά κυρίως σε αυτά για τα οποία υπήρχαν δεδομένα στις βάσεις δεδομένων αλληλουχιών γονιδίων, όπως η Genbank. Έτσι χρησιμοποιήθηκε ο συστηματικός κατάλογος που υπήρχε στη Βιοσυστηματική μονάδα Walter Reed (Gaffigan, Wilkerson et al.) που όχι μόνο παρείχε πληρέστερη κάλυψη του συγκεκριμένου τάξου, αλλά και τα συνώνυμα που περιείχε για τα συγκεκριμένα είδη ήταν κατά πολύ περισσότερα. Για παράδειγμα μόνο 312 είδη του γένους *Anopheles* υπήρχαν στην ταξινομική οντολογία του NCBI τον καιρό δημιουργίας της οντολογίας, ενώ στην λίστα της βιοσυστηματικής μονάδας Walter Reed φτάνουν τα 470 χωρίς να συμπεριλαμβάνονται τα συνώνυμά τους. Το μειονέκτημά της τελευταίας ήταν πως από την φύση της βάσης δεδομένων του Walter Reed οι όροι που αντλήθηκαν από εκεί δεν συμπεριλάμβαναν κάποιο ορισμό, ο οποίος προστέθηκε από εμάς στην MIRO. Τα πλεονεκτήματα της χρήσης της MIRO έγιναν εμφανή από την πρώτη στιγμή, γιατί τώρα υπήρχε τρόπος να συσχετισθούν με την σχέση *is_a*, οι διάφορες χρωμοσωμικές και μοριακές μορφές που εντοπίστηκαν στο *Anopheles gambiae* sensu stricto (della Torre, Costantini et al. 2002). Έτσι οι μοριακές μορφές (M και S) αλλά καθώς επίσης και οι χρωμοσωμικές Bamako, Mopti, Bissau, Forest, Savanna συνδέονται με σχέση *is_a* με το *Anopheles gambiae* sensu stricto που επίσης συνδέεται με την ίδια σχέση με το *Anopheles gambiae* sensu lato. Ανάλογη (αν και λίγο απλούστερη) είναι και η κατάσταση με το *Anopheles funestus* sensu lato (δεν υπάρχουν χρωμοσωμικές και μοριακές

μορφές). Με αυτόν τον τρόπο εργαλεία που θα βασιστούν στην MIRO θα μπορούν να συμπεριλάβουν αυτόματα στα αποτελέσματα κάποιας αναζήτησης για δεδομένα που αφορούν *Anopheles gambiae* και δεδομένα που έχουν καταχωρηθεί χρησιμοποιώντας την αναφορά σε συγκεκριμένη μορφή.

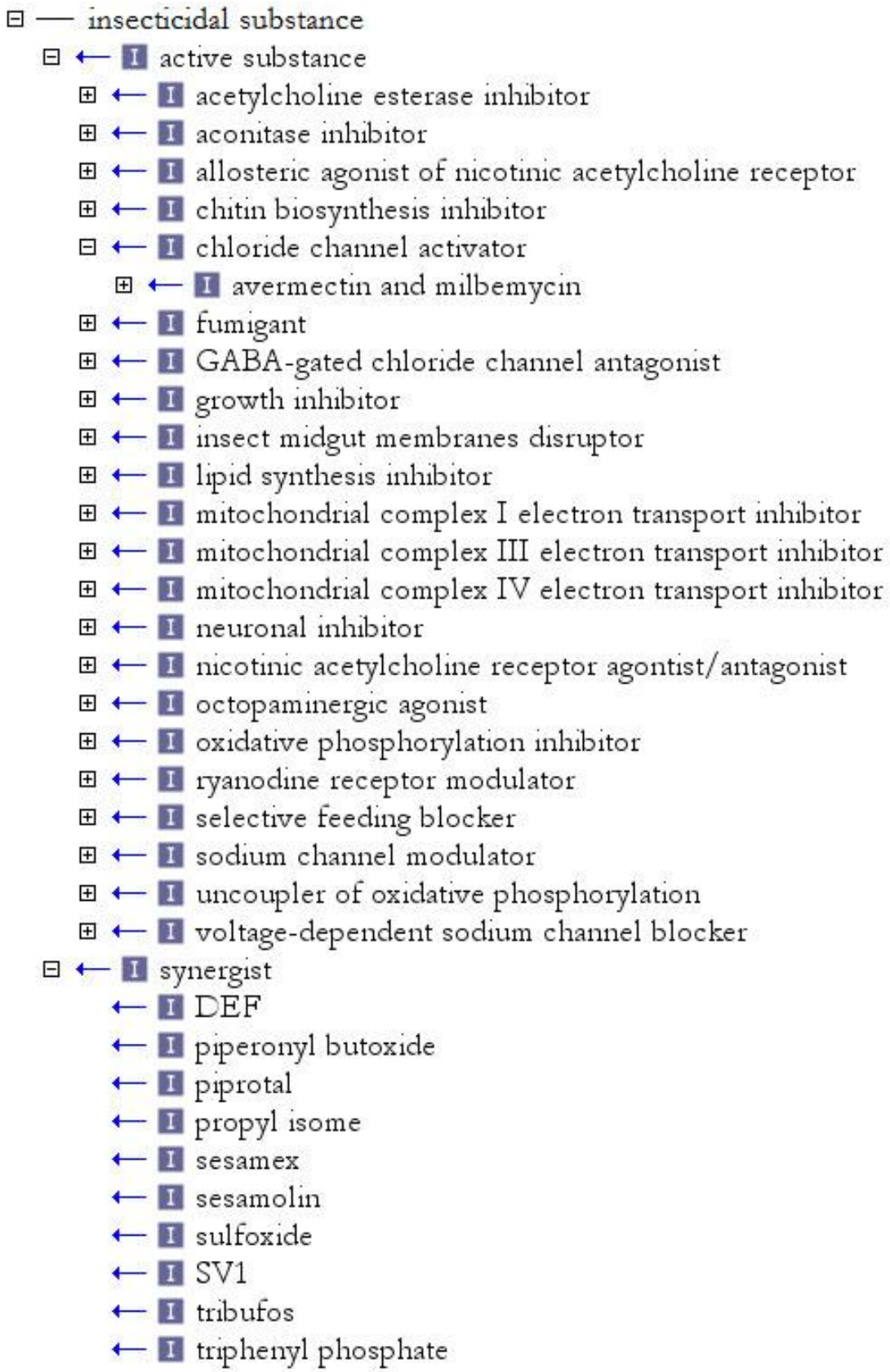
Β) ΕΝΤΟΜΟΚΤΟΝΕΣ ΟΥΣΙΕΣ

Η κατηγορία αυτή είναι σχετικά προβληματική μιας και υπάρχει μια οντολογία που περιγράφει χημικές ενώσεις που παρουσιάζουν ενδιαφέρον από βιολογική σκοπιά (ChEBI), ενώ το μεγαλύτερο μέρος της επιστημονικής κοινότητας χρησιμοποιεί σαν σημείο αναφοράς την επιτροπή δράσης για την ανθεκτικότητα σε εντομοκτόνα (IRAC) όπου τα ενδιαφέροντά της δεν περιορίζονται μόνο στην αντιμετώπιση προβλημάτων από αρθρόποδα - φορείς ασθενειών ανθεκτικά σε εντομοκτόνα, αλλά και από έντομα που επηρεάζουν την αγροτική παραγωγή. Το δίλημμα να ακολουθήσει κανείς την ChEBI (Degtyarenko, de Matos et al. 2008; Degtyarenko, Hastings et al. 2009; de Matos, Alcantara et al. 2010) ή την ταξινόμηση του IRAC γίνεται ακόμα πιο πολύπλοκο γιατί και οι δύο προσεγγίσεις παρουσίαζαν σημαντικά αρνητικά στοιχεία. Η ChEBI ανέφερε πολύ λίγα εντομοκτόνα σε σχέση με το IRAC. Η ταξινόμηση του IRAC, εστιασμένη στον τρόπο δράσης των εντομοκτόνων, περιείχε οντολογικά προβληματικές κατηγορίες όπως εντομοκτόνα με άγνωστο τρόπο δράσης ενώ επιπλέον υπήρχαν και αντιφάσεις όπου ουσίες αναφερόμενες ως εντομοκτόνα κατά το IRAC στην ChEBI παρουσιάζονταν μόνο ως ακαρεοκτόνα.

Στην MIRO οι εντομοκτόνες ουσίες χωρίστηκαν σε ενεργές εντομοκτόνες ουσίες και συνεργατικές με τις τελευταίες απλά να χρησιμοποιούνται σε συνδυασμό με τις πρώτες και να διευκολύνουν την δράση τους (Εικόνα 6). Υιοθετήθηκε σε γενικές γραμμές η προσέγγιση του IRAC με κριτήριο (ακριβώς όπως και στην περίπτωση των τάξων της οικογένειας Culicidea) τηn εξοικείωση των χρηστών με την ταξινόμηση του IRAC. Η κατηγορία των εντομοκτόνων με άγνωστο τρόπο δράσης δεν παρουσιάστηκε ως τέτοια,

απλά τα εντομοκτόνα που ανήκαν σ' αυτή συμπεριλήφθηκαν γενικά στις ουσίες με εντομοκτόνο δράση. Παράλληλα σε συνεργασία με την ChEBI λύσαμε όλες τις υπάρχουσες αντιφάσεις και προσθέσαμε όσα εντομοκτόνα δεν αναφέρονταν ως τέτοια. Στην MIRO κρατήσαμε τα εντομοκτόνα αυτά με τον δικό τους κωδικό προσθέτοντας ως εξωτερική αναφορά και τον κωδικό της ChEBI. Έτσι όσοι είχαν χρησιμοποιήσει τους όρους της MIRO πριν τη διαδικασία κανονικοποίησής της δεν θα χρειαστεί να αλλάξουν κάτι στα δικά τους εργαλεία. Ταυτόχρονα η MIRO είχε διασυνδεθεί με την ChEBI και μέσω αυτής με όλες τις οντολογίες που θα χρησιμοποιούν την ChEBI ως οντολογία αναφοράς και θα ενδιαφέρονται για παρόμοια δεδομένα.

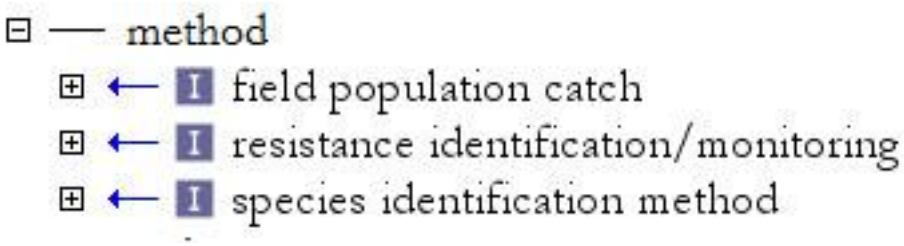
Επειδή ο στόχος ήταν να υπάρξει μια γενική παρουσίαση των εντομοκτόνων με παγκόσμια εμβέλεια χρησιμοποιήσαμε σαν σημεία αναφοράς τις ενεργές ουσίες με εντομοκτόνο δράση και όχι τα εμπορικά ονόματα με τα οποία είναι διαθέσιμες. Τα τελευταία αλλάζουν από χώρα σε χώρα και από εταιρεία σε εταιρεία κι ενώ είναι απαραίτητα από πρακτικής πλευράς η χρησιμοποίησή τους παραπέμφθηκε στο επίπεδο της ανάπτυξης βάσης δεδομένων που θα στηρίζεται στην οντολογία που αναπτύξαμε. Άλλωστε το σύνηθες είναι στην επιστημονική βιβλιογραφία να αναφέρεται η ενεργός ουσία και όχι το εμπορικό της όνομα. Παρόλα αυτά όμως δεν αποκλείεται στο μέλλον η προσθήκη αυτών των ονομάτων ως συνώνυμα.



Εικόνα 6: Η ταξινόμηση των εντομοκτόνων στην MIRO.

γ) ΜΕΘΟΔΟΙ

Είναι φανερό πως η συγκεκριμένη είναι μια ετερογενής κατηγορία που περιέχει όλες τις πειραματικές μεθόδους και διαδικασίες που περιλαμβάνονται στις σχετικές αναφορές και σχετίζονται όχι μόνο με την ταυτοποίηση της ανθεκτικότητας, αλλά και την ανάλυση του μηχανισμού της (αν είναι γνωστός), τις διαδικασίες που χρησιμοποιήθηκαν για την συλλογή του πληθυσμού από την φύση καθώς και εκείνες για την εξακρίβωση του σε ποιο είδος ανήκει ο πληθυσμός (Εικόνα 7). Το εύρος αυτών των τεχνικών είναι πολύ μεγάλο μιας και εκτείνεται από την χρήση εξειδικευμένων παγίδων για την συλλογή εντόμων μέχρι γενικές βιοχημικές, γενετικές μεθόδους καθώς και μεθόδους μοριακής βιολογίας που απαιτούνται για την ταυτοποίηση συγκεκριμένων μεταλλαγών για παράδειγμα, που αποδεδειγμένα έχουν σαν αποτέλεσμα την ανθεκτικότητα σε εντομοκτόνα.

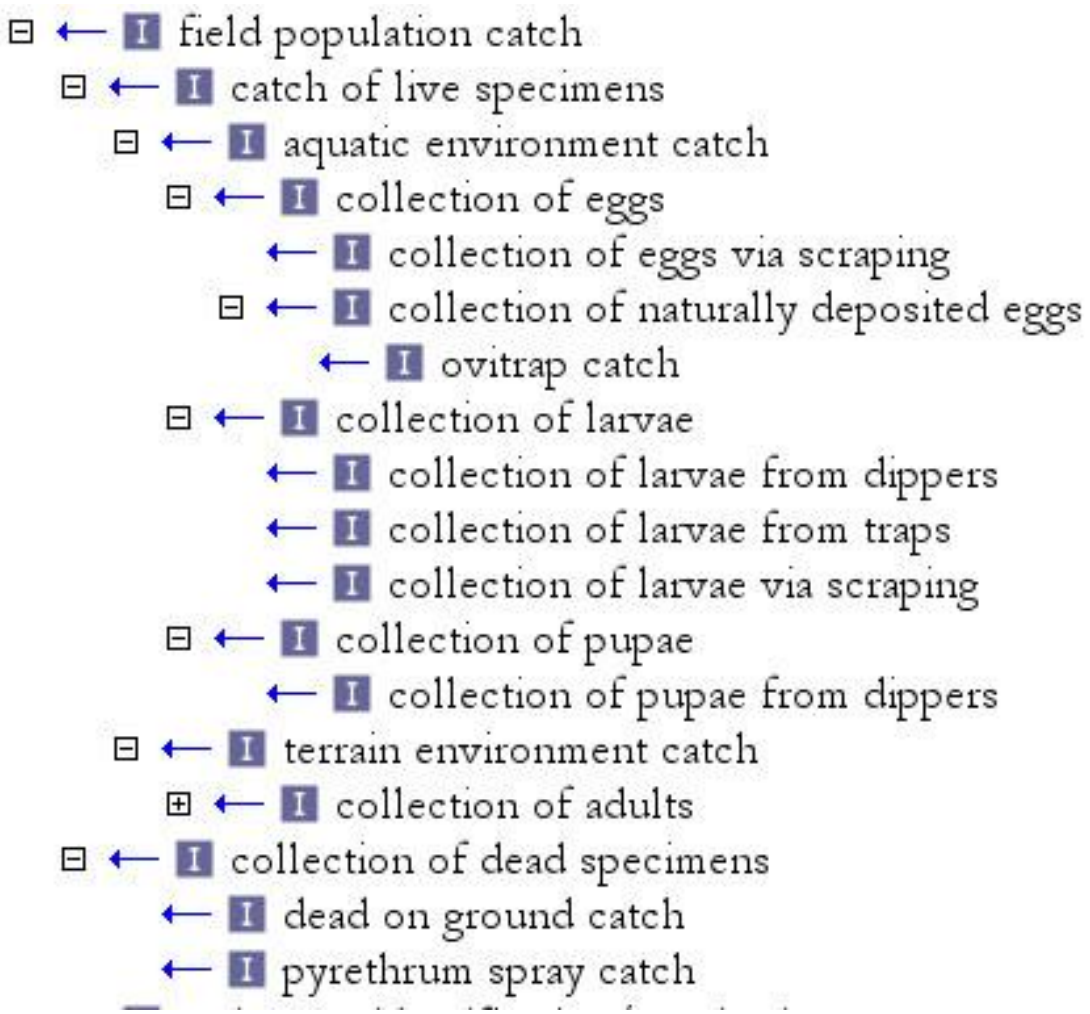


Εικόνα 7: Τα διαφορετικά είδη μεθόδων που οντολογικά ταξινομούνται μαζί στην MIRO.

Η υπάρχουσα οντολογία βιολογικών ερευνών (OBI) (Brinkman, Courtot et al. 2010) αναπτύσσεται παράλληλα με την MIRO, αλλά το εύρος της είναι τέτοιο που δεν μπορούσε να ανταποκριθεί άμεσα και να μας παράσχει τους απαιτούμενους όρους. Γι αυτό όπως κάναμε και με τα εντομοκτόνα συμπεριλάβαμε στη δική μας οντολογία τους όρους που είχαμε ανάγκη αφήνοντας για το μέλλον μια αντίστοιχη διαδικασία κανονικοποίησης αντίστοιχη με εκείνη που έγινε ανάμεσα στην MIRO και την ChEBI, ώστε να μην υπάρχουν αλληλεπικαλύψεις στα πλαίσια της OBO Foundry. Τρεις είναι οι κύριες κατηγορίες των

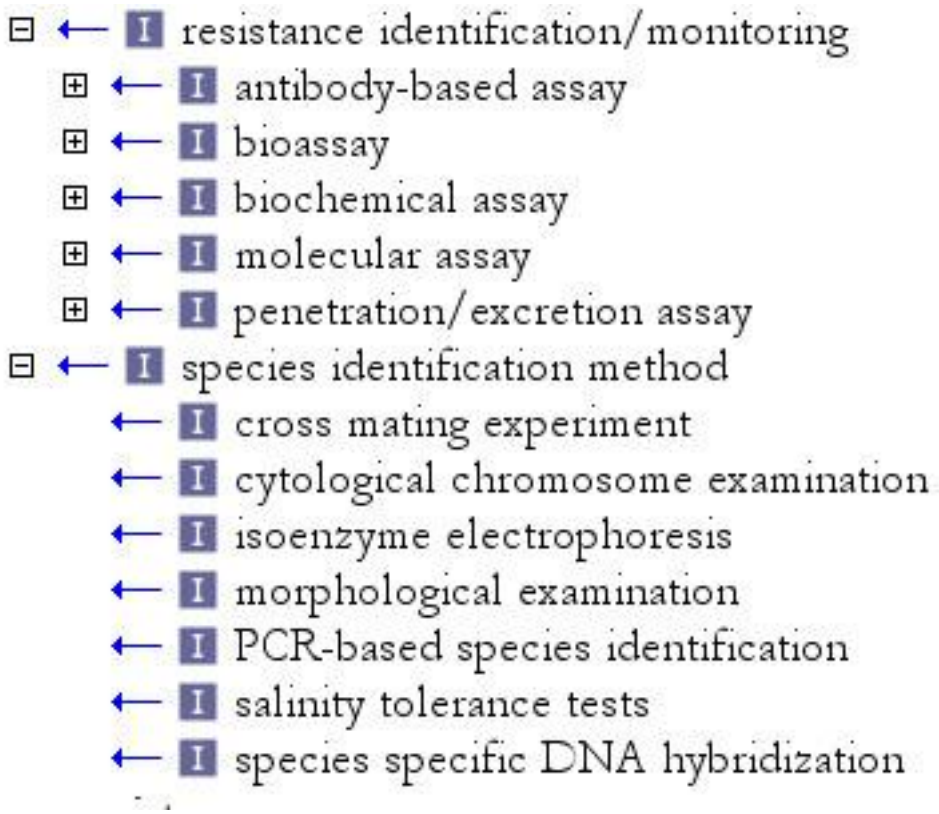
μεθόδων στις οποίες καταλήξαμε: Μέθοδοι για την συλλογή βιολογικού υλικού από την φύση, μέθοδοι για την ταυτοποίηση των οργανισμών που συλλέχθηκαν και τέλος μέθοδοι εντοπισμού και ανάλυσης της τυχόν ανθεκτικότητας σε εντομοκτόνα. (Εικόνα 7).

Στις μεθόδους συλλογής κουνουπιών συγκαταλέγονται μέθοδοι συλλογής κάθε αναπτυξιακού σταδίου (αυγά, προνύμφες, νύμφες, ενήλικα άτομα) στο αντίστοιχο μικροπεριβάλλον όπου απαντάται κάθε στάδιο (υδάτινο ή χερσαίο). Ειδικότερα για τα ενήλικα άτομα τα κυριότερα είδη παγίδων που υπάρχουν περιγράφονται και όπως σε όλα τα τμήματα μιας οντολογίας υπάρχει πάντα η δυνατότητα τροποποίησης και προσθήκης και άλλων μεθόδων αν υπάρξει η σχετική ανάγκη. Τέλος, έχουμε συμπεριλάβει και την συλλογή νεκρών δειγμάτων γιατί και αυτά χρησιμοποιούνται συχνά για περαιτέρω ανάλυση (Εικ 8).



Εικόνα 8: Οι τρόποι συλλογής κουνουπιών που περιλαμβάνονται στην MIRO

Οι μέθοδοι ταυτοποίησης του είδους είναι οι πιο συχνά χρησιμοποιούμενες για τον σκοπό αυτό από την μορφολογική εξέταση, τις διασταυρώσεις μεταξύ των ατόμων και την αντοχή τους σε διαφορετικές αλατότητες μέχρι κυτταρολογικές (εξέταση χρωμοσωμάτων), βιοχημικές (ανάλυση ισοενζύμων) και μοριακές (υβριδοποίηση DNA και PCR) (Εικόνα 9).



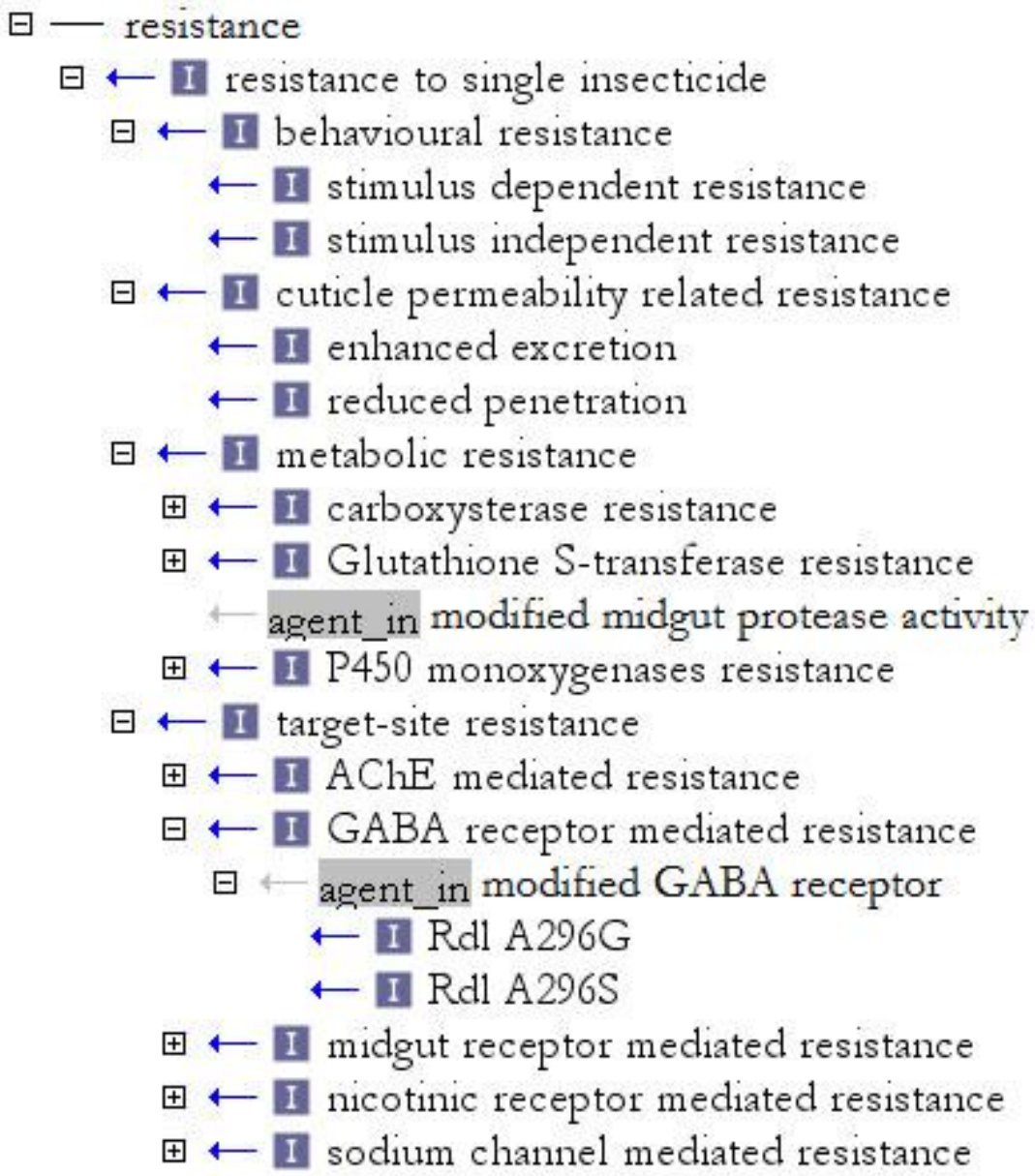
Εικόνα 9: Οι μέθοδοι παρακολούθησης επιπέδων ανθεκτικότητας σε εντομοκτόνα ενός πληθυσμού κουνουπιών καθώς κι εκείνες που χρησιμοποιούνται για την ταυτοποίηση του είδους των ατόμων που τον αποτελούν.

Τέλος, όπως είναι φυσικό το μεγαλύτερο μέρος του τμήματος αυτού αφορά μεθόδους και διαδικασίες για την ταυτοποίηση των επιπέδων ανθεκτικότητας σε εντομοκτόνα που παρουσιάζει κάποιος πληθυσμός. Πρόκειται για βιοδοκιμές που γίνονται είτε παρουσία, είτε απουσία κάποιας συνεργατικής ουσίας και σκοπό έχουν την ανίχνευση ανθεκτικότητας σε κάποιο εντομοκτόνο (διαγνωστικοί έλεγχοι). Μια άλλη σειρά μεθόδων εστιάζεται στην απόκριση του ελεγχόμενου πληθυσμού σε συγκεκριμένες δόσεις εντομοκτόνου καθώς επίσης και στον χρόνο ο οποίος απαιτείται για να έχει αποτέλεσμα η εφαρμογή συγκεκριμένης δόσης εντομοκτόνου στον ελεγχόμενο πληθυσμό. Εκτός από τις βιοδοκιμές μια πληθώρα βιοχημικών, μοριακών και τεχνικών ανοσοεντοπισμού, χρησιμοποιούνται για τον έλεγχο της παρουσίας συγκεκριμένων οικογενειών ενζύμων καθώς και την ταυτοποίηση των γονιδίων που τα κωδικοποιούν και την λειτουργική τους ανάλυση

τόσο σε επίπεδο γονιδιακής έκφρασης όσο και βιοχημικής λειτουργίας. Οικογένειες ενζύμων ευθύνονται για την αδρανοποίηση τοξικών για τον οργανισμό ουσιών. Όποια μεταλλαγή τους, αυξάνει την αποδοτικότητα αυτών των συστημάτων είναι πιθανό να επιφέρει και ανθεκτικότητα. Αυτό όμως μας οδηγεί στην επόμενη ενότητα της οντολογίας την

δ) ΑΝΘΕΚΤΙΚΟΤΗΤΑ

Η ενότητα αυτή αντανακλά την τρέχουσα γνώση για τους μηχανισμούς με τους οποίους επιτυγχάνεται ανθεκτικότητα απέναντι σε κάποιο εντομοκτόνο. Αυτή μπορεί να επιτευχθεί με την αποφυγή από μεριάς των κουνουπιών των περιοχών στις οποίες είναι παρόν το εντομοκτόνο (ανθεκτικότητα λόγω συμπεριφοράς). Επίσης μειωμένη δυνατότητα προσρόφησης του εντομοκτόνου από το έντομο είτε λόγω αλλαγών στη δομή της επιδερμίδας, είτε λόγω αυξημένης απεκκριτικής λειτουργίας μπορεί να οδηγήσει σε ανάλογα αποτελέσματα. Ακόμα μηχανισμοί γενικευμένης αποτοξίνωσης του οργανισμού μπορούν να οδηγήσουν σε ανθεκτικότητα απέναντι σε κάποια εντομοκτόνα αν έχουν μεταλλαχθεί κατά τέτοιο τρόπο ώστε να αυξηθεί η αποδοτικότητά τους. Γνωστοί μηχανισμοί αδρανοποίησης τοξικών ουσιών περιλαμβάνουν την συμμετοχή γνωστών οικογενειών ενζύμων όπως οι οικογένειες των καρβοξυλεστερασών (CoE), των γλουταθιόνη S-τρανσφερασών (GST) και των P450 μονοξυγενασών. Τέλος έχει παρατηρηθεί ανθεκτικότητα εξαιτίας της παρουσίας μεταλλαγμένων αλληλομόρφων σε συγκεκριμένους γενετικούς τόπους όπως τα κανάλια ιόντων νατρίου, ο υποδοχέας του νευροδιαβιβαστή γ-αμινοβουτυρικό οξύ (GABA) κλπ. (Εικόνα 10).

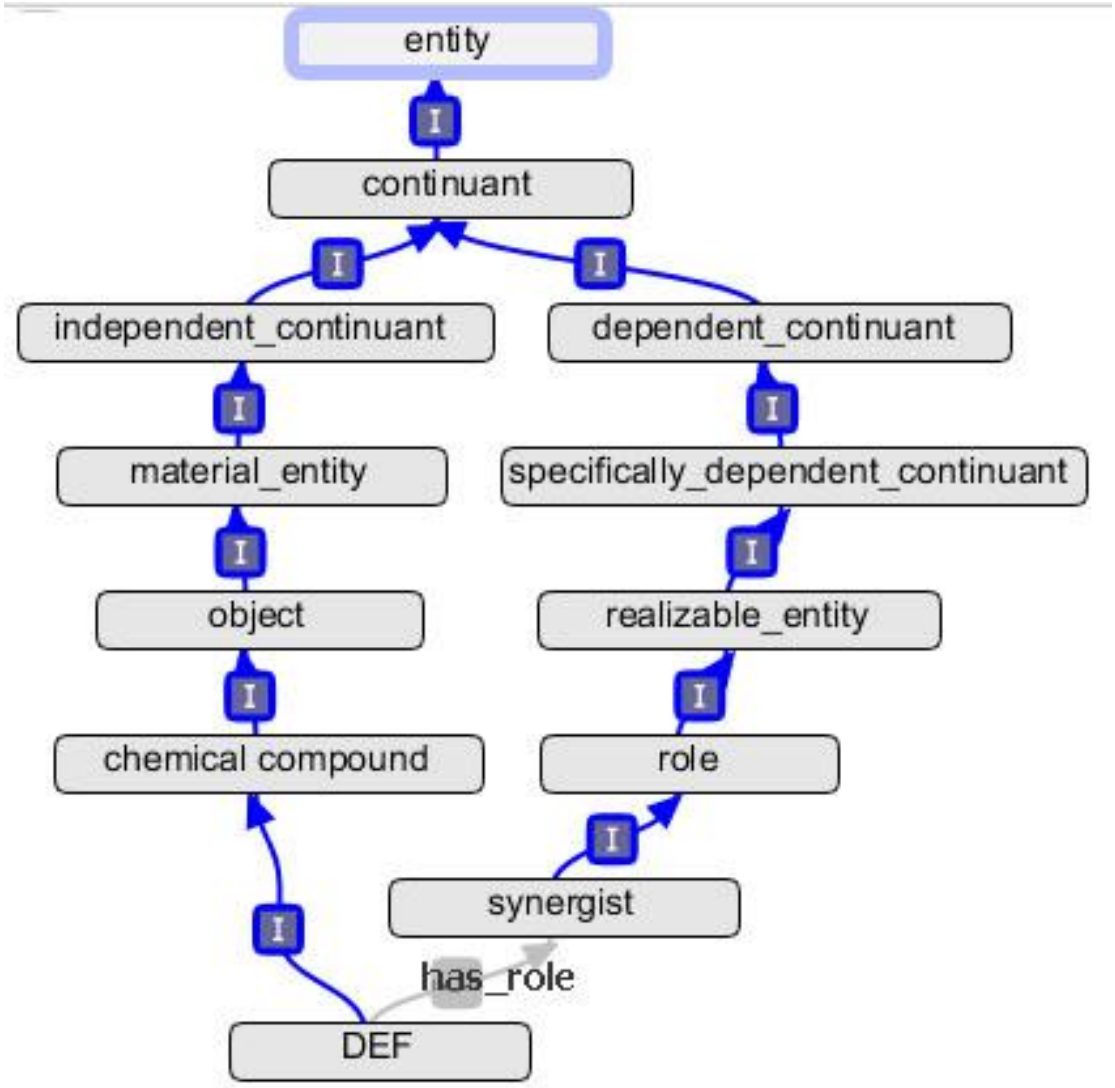


Εικόνα 10: Το τμήμα της MIRO που περιγράφει τους μηχανισμούς της ανθεκτικότητας σε εντομοκτόνα.

Προσαρμογή της MIRO στο πρότυπο BFO

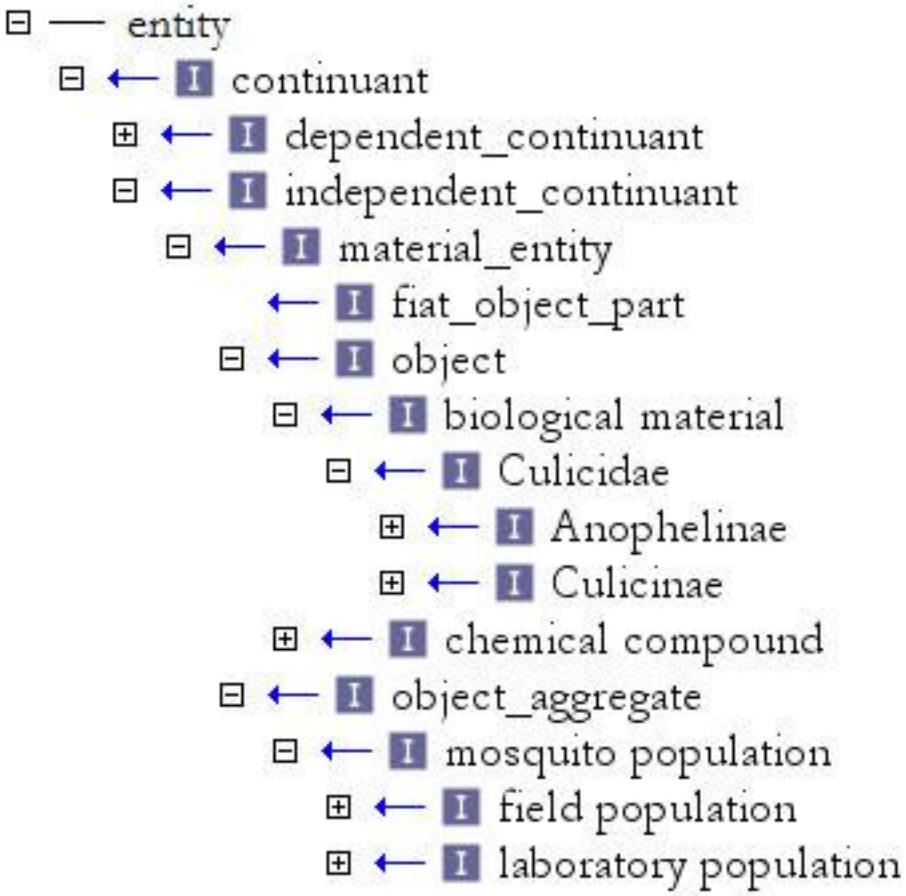
Η επιλογή να μην ακολουθηθεί η BFO διευκόλυνε κατά πολύ την παρουσίαση, την κατανόηση και την αποδοχή της MIRO από τους μη εξοικειωμένους με οντολογίες επιστήμονες. Ταυτόχρονα όμως δημιούργησε και προβλήματα που αφορούσαν την ακρίβεια της περιγραφής του γνωστικού πεδίου. Από τα πιο χαρακτηριστικά παραδείγματα του προβλήματος είναι πως οι συνεργατικές ουσίες θεωρούνται πως έχουν εντομοκτόνο δράση ενώ στην πραγματικότητα κάτι τέτοιο είναι ανακριβές. Συνήθως η δράση των συνεργατικών ουσιών σχετίζεται με την παρεμπόδιση των φυσιολογικών μηχανισμών αποτοξίνωσης με

συνέπεια την αποτελεσματικότερη δράση του εντομοκτόνου. Όμως στην MIRO υπάρχει η συσχέτιση πως οι συνεργατικές ουσίες είναι εντομοκτόνες ουσίες. Έτσι η περιγραφή της MIRO πως «η *DEF* είναι μια *συνεργατική ουσία* και πως *οι συνεργατικές ουσίες* είναι *εντομοκτόνες ουσίες*», οδηγεί στο λανθασμένο συμπέρασμα πως η DEF έχει εντομοκτόνο δράση. Η σωστή οντολογικά περιγραφή θα ήταν πως: «*Η DEF* είναι *μια χημική ένωση*. *Οι χημικές ενώσεις* είναι *αβιοτικά αντικείμενα*. *Τα αβιοτικά αντικείμενα* είναι *αντικείμενα*. *Τα αντικείμενα* είναι *ανεξάρτητες οντότητες συνεχείς στον χρόνο*. Η *DEF* παίζει τον ρόλο *της συνεργατικής ουσίας ενός εντομοκτόνου*. *Η συνεργατική ουσία ενός εντομοκτόνου* είναι *ρόλος*». (Στις περιγραφές με πλάγια γράμματα παρουσιάζονται οι όροι της οντολογίας) (Εικόνα 11). Είναι προφανές πως η δεύτερη περιγραφή ανταποκρίνεται πολύ καλύτερα στην πραγματικότητα και θεωρήθηκε σκόπιμο να υιοθετηθεί από την στιγμή που η οντολογία είχε αποκτήσει μια σταθερή δομή αποδεκτή από την κοινότητα στην οποία απευθυνόταν. Ο μετασχηματισμός έπρεπε να γίνει με τέτοιο τρόπο ώστε να είναι συμβατός με τα εργαλεία που είχαν αναπτυχθεί στο μεταξύ και βασίζονταν στην MIRO και θα δούμε παρακάτω. Επιπρόσθετα αυτή η αλλαγή δομής θα επέτρεπε στην MIRO να είναι ακόμα περισσότερο συμβατή με τις υπόλοιπες οντολογίες της OBO Foundry.



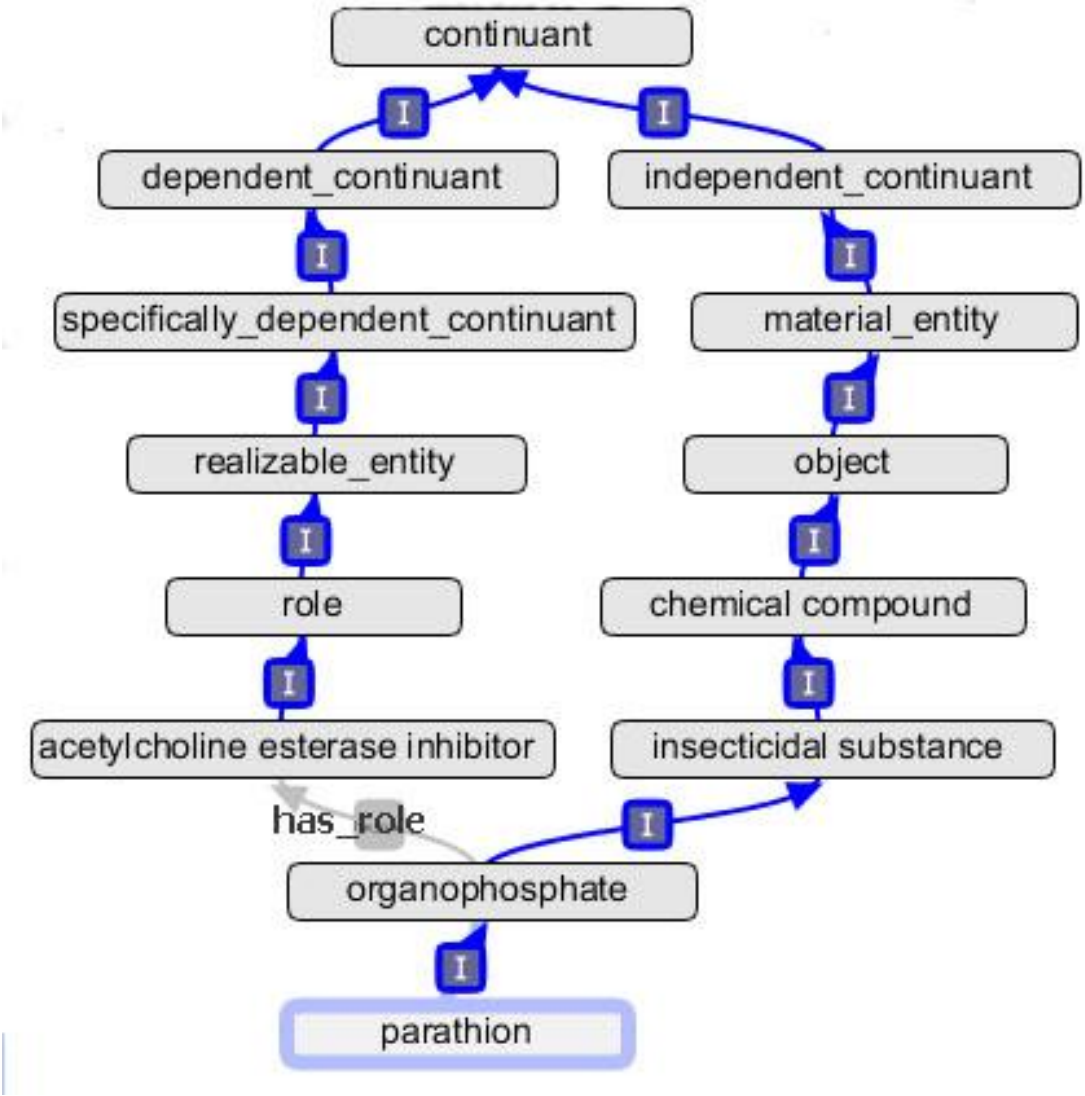
Εικόνα 11: Η οντολογική θέση της συνεργατικής ουσίας DEF σύμφωνα με την βασική τυπική οντολογία.

Κατά συνέπεια ολόκληρα τμήματα της MIRO μετακινήθηκαν χωρίς να σβηστούν όροι ή να αλλάξουν οι κωδικοί τους. Στην περιγραφή του βιολογικού υλικού η ταξινόμηση σε τάξα βρίσκεται στα αντικείμενα που είναι ανεξάρτητες οντότητες (Εικόνα 12), ενώ όλοι οι πληθυσμοί βρέθηκαν στα σύνολα αντικειμένων μιας και είναι ακριβώς σύνολα οργανισμών που κατοικούν σε μια περιοχή (Εικόνα 12). Η θεώρηση του πληθυσμού σαν σύνολο ομοειδών πραγμάτων (οργανισμών) διευκολύνει και την εισαγωγή εννοιών όπως παρατηρούμενες συχνότητες αλληλομόρφων ή μεταλλαγών που θα χρειαστούν για την περιγραφή του πληθυσμού.



Εικόνα 12: Σύμφωνα με την βασική τυπική οντολογία το βιολογικό υλικό είναι ανεξάρτητη ουσία με συνεχή παρουσία στον χρόνο, ενώ ένας πληθυσμός άθροισμα ατόμων.

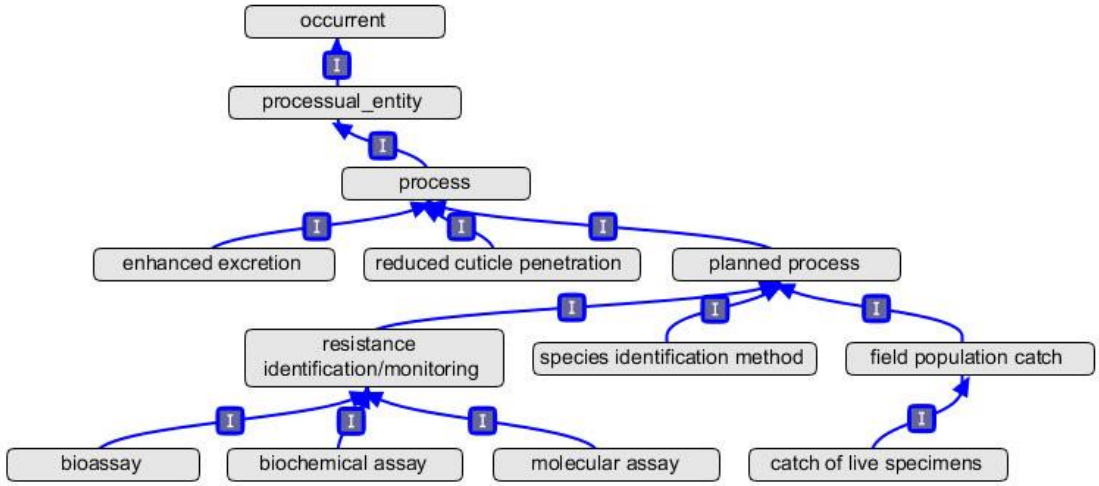
Τα εντομοκτόνα αποτελούσαν μια περισσότερο πολύπλοκη κατηγορία, όπως αναμενόταν και από το γεγονός πως τα περισσότερα λογικά προβλήματα με τη μη χρήση της BFO εμφανίζονταν σε αυτή την κατηγορία. Το πρόβλημα επιτείνονταν και από την υιοθέτηση των κατηγοριών του IRAC όπου στην πραγματικότητα περιγράφουν τον τρόπο δράσης των ουσιών αυτών. Όλα τα εντομοκτόνα είναι χημικές ουσίες, (Εικόνα 13) ενώ ο τρόπος δράσης τους αναφέρεται στον ρόλο που παίζουν σε συγκεκριμένες φυσιολογικές διαδικασίες. Για παράδειγμα το παραθείο είναι η οργανοφωσφορική ένωση η οποία παίζει τον ρόλο του αναστολέα της εστεράσης της ακετυλχολίνης (Εικόνα 13).



Εικόνα 13: Η αναπαράσταση των εντομοκτόνων και του μηχανισμού δράσης τους αναδεικνύεται από την βασική τυπική οντολογία.

Όλες οι πειραματικές μέθοδοι αποτελούν διαδικασίες καθώς μεταβάλλονται στο πέρασμα του χρόνου και εκτυλίσσονται σε στάδια και σαν τέτοιες έχουν καταγραφεί. Σε αντιδιαστολή όμως με τις φυσιολογικές διαδικασίες του οργανισμού, αυτές είναι προσχεδιασμένες από τους επιστήμονες (Εικόνα 14). Τέλος μεταβολές στις ιδιότητες των οργανισμών όπως παραδείγματος χάριν μια ποσοτική αλλαγή στα επίπεδα των καρβοξυλεστερασών έχουν σαν αποτέλεσμα τη μεταβολή φυσιολογικών διαδικασιών όπως της αδρανοποίησης των τοξικών ουσιών. Η αποδοτικότερη αδρανοποίηση τοξικών ουσιών δημιουργεί την προδιάθεση για εμφάνιση ανθεκτικότητας σε κάποιο εντομοκτόνο. Στις

περιπτώσεις που το περιβάλλον είναι τέτοιο ώστε να εκφραστεί η συγκεκριμένη προδιάθεση, τότε και ο οργανισμός θα εμφανίσει τον φαινότυπο της ανθεκτικότητας, που δεν είναι τίποτε άλλο από μια πρόσθετη ιδιότητα του οργανισμού αυτού.



Εικόνα 14: Οι πειραματικές μέθοδοι είναι προσχεδιασμένες διαδικασίες.

Οι αλλαγές αυτές δημιούργησαν ένα πιο στέρεο πλέγμα σχέσεων ανάμεσα στους όρους της MIRO, που όμως δεν επηρέασαν καθόλου την χρηστική της αξία. Άλλωστε μια οντολογία συνήθως κατασκευάζεται για να παράσχει τους απαραίτητους όρους για την περιγραφή κάποιων δεδομένων και όχι για να προσεγγίζεται άμεσα και απευθείας από τους τελικούς χρήστες. Αυτοί θα ωφεληθούν από την αποδοτικότερη χρήση των βάσεων δεδομένων που στηρίζονται στις οντολογίες. Η τρέχουσα έκδοση της MIRO περιλαμβάνει 4317 όρους που διαθέτουν όλοι ορισμούς και 1768 συνώνυμους όρους. Η υιοθέτηση της BFO στα ανώτερα επίπεδα οδήγησε στο να υπάρχει μία και μόνη αρχική κατηγορία.

Χρήση της MIRO

Η δημιουργία της MIRO, διαμόρφωσε τελείως διαφορετικά δεδομένα στην AnoBase και τον τρόπο παρουσίασης δεδομένων ανθεκτικότητας σε σχέση με όσα ίσχυαν την περίοδο όπου τα πάντα βασίζονταν στις προφόρμες που είχαμε αναπτύξει. Πλέον ήταν δυνατή η δημιουργία μιας εξειδικευμένης βάσης δεδομένων όπου θα μπορούσε να

αποθηκεύσει δεδομένα ανθεκτικότητας για περισσότερους οργανισμούς, με πιο εξελιγμένους τρόπους αναζήτησης δεδομένων και χωρίς την ανάγκη της ανασύστασης βήμα προς βήμα μιας δημοσίευσης όπως ήταν απαραίτητο παλαιότερα σύμφωνα με το σύστημα με τις προφόρμες. Μια τέτοια προσέγγιση δεν αποκλειόταν, απλά δεν αποτελούσε την μοναδική δυνατότητα. Έτσι υπήρχε η δυνατότητα ανάπτυξης ενός εργαλείου που δεν θα ήταν απλά η ηλεκτρονική μορφή των δημοσιεύσεων που κάλυπταν το πεδίο, αλλά ένα δυναμικό σύστημα το οποίο θα διατηρούσε τη δυνατότητά του να επεκταθεί και σε άλλους τομείς της ανθεκτικότητας σε εντομοκτόνα πέρα από τον αναφερόμενο στις τροπικές νόσους και την δημόσια υγεία.

Η βάση δεδομένων, ονομάστηκε IRBase και σχεδιάστηκε από την αρχή, ώστε να αρθούν περιορισμοί του προηγούμενου συστήματος και για να καταστεί δυνατή τυχόν επέκτασή της στο μέλλον. Όλοι οι όροι της MIRO, αλλά και εκείνοι της GAZ (gazetteer), της οντολογίας που χρησιμοποιήθηκε για τους γεωγραφικούς όρους συμπεριλαμβανομένων και των μεταξύ τους συσχετίσεων αποθηκεύτηκαν στη βάση δεδομένων. Αποτέλεσμα αυτού ήταν η δυνατότητα αυτόματης επέκτασης ερωτημάτων που υπέβαλλαν οι χρήστες ώστε να καλύψουν οντότητες που περιλαμβάνονταν στην αρχική. Έτσι για παράδειγμα αν κάποιος ήθελε να βρει δεδομένα για την ανθεκτικότητα απέναντι στα οργανοφωσφορικά εντομοκτόνα θα έπρεπε να καταφύγει σε πολλαπλές αναζητήσεις, αναζητώντας κάθε φορά δεδομένα για συγκεκριμένο εντομοκτόνο. Με την ύπαρξη της οντολογίας και εκμεταλλευόμενοι την σχέση *is_a* κάποιος με ένα μόνο ερώτημα για το σύνολο των οργανοφωσφορικών μπορεί να ανακτήσει όλα τα σχετικά δεδομένα. Επίσης η αναζήτηση περιλαμβάνει όχι μόνο τους όρους της οντολογίας αλλά και συνώνυμά τους αυξάνοντας το ποσοστό των επιθυμητών αποτελεσμάτων που ανακτώνται με μία και μόνη αναζήτηση.

Τα αποτελέσματα μπορούν είτε να εμφανιστούν με μορφή πίνακα στην οθόνη του υπολογιστή, δίνοντας πάντα την δυνατότητα για καθένα χωριστά να υπάρξει πλήρης αναφορά όλων των αποθηκευμένων στοιχείων (Εικόνα 15 πάνω), είτε να αποσταλούν στον χρήστη με την μορφή αρχείου, ώστε αυτός να τα επεξεργαστεί όπως θέλει. Τέλος με την βοήθεια του Google Earth τα δεδομένα μπορούν να απεικονισθούν πάνω σε χάρτες με την προϋπόθεση πως τα δεδομένα που υπάρχουν στη βάση περιλαμβάνουν και τις γεωγραφικές συντεταγμένες της περιοχής στην οποία αναφέρονται (Εικόνα 15 κάτω).

VectorBase

Insecticide Resistance

Insecticide Resistance Assays

	Location	Year	Species	Assay type	Insecticide	Resistance mechanism	Full Report
☐	Ipokia Local Government Area	2003	Anopheles gambiae		deltamethrin	modified sodium channel	
☐	Ipokia Local Government Area	2003	Anopheles gambiae		permethrin	modified sodium channel	
☐	Ipokia Local Government Area	2003	Anopheles gambiae		DDT	modified sodium channel	
☐	Ivory Coast	2003	Anopheles gambiae		Control		
☐	Ivory Coast	2003	Anopheles gambiae		carbosulfan		
☐	Ivory Coast	2003	Anopheles gambiae		carbosulfan		
☐	Ivory Coast	2003	Anopheles gambiae		carbosulfan		
☐	Ivory Coast	2003	Anopheles gambiae		carbosulfan		
☐	Ivory Coast	2003	Anopheles gambiae		carbosulfan		

Εικόνα 15: Πάνω η απεικόνιση σε μορφή πίνακα μέρους των αποτελεσμάτων αναζήτησης για δεδομένα ανθεκτικότητας σε εντομοκτόνα που αφορούν πληθυσμούς του *A. gambiae*. Κάτω η απεικόνιση των περιοχών για τις οποία υπάρχουν σχετικά δεδομένα.

Η MIRO ελαφρώς παραλλαγμένη με την προσθήκη των εμπορικών ονομάτων των εντομοκτόνων και την αφαίρεση των φορέων που δεν μεταδίδουν ελονοσία, χρησιμοποιείται στον πυρήνα του εντομολογικό τμήματος του συστήματος υποστήριξης αποφάσεων για την ελονοσία (MDSS = Malaria Decision Support System) όπως αυτό αναπτύσσεται στα

πλαίσια του IVCC (Innovative Vector Control Consortium) τόσο για την ανίχνευση όσο και για την παρακολούθηση της ανθεκτικότητας σε εντομοκτόνα στις Αφρικανικές χώρες.

Συμπεράσματα

Η MIRO συνοδευόμενη από την IRBase που την υλοποιεί χρησιμοποιήθηκαν για την περιγραφή και την ανάλυση της ανθεκτικότητας σε εντομοκτόνα που εμφανίζουν πληθυσμοί κουνουπιών σε ολόκληρο τον κόσμο. Είναι η πρώτη απόδειξη στην πράξη πως ο συνδυασμός μιας εφαρμοσμένης οντολογίας με μια βάση δεδομένων μπορεί να αποτελέσει τη βάση πάνω στην οποία θα οικοδομηθεί ένα ευρύτερο σύστημα παρακολούθησης της ανθεκτικότητας σε εντομοκτόνα. Μέχρι σήμερα τα αντίστοιχα συστήματα ήταν κατακερματισμένα ανά περιοχές ή κράτη ή ανά ερευνητικό πρόγραμμα. Η χρήση της MIRO τόσο από την VectorBase στα πλαίσια της οποίας δημιουργήθηκε, αλλά και από το υποστηρικτικό σύστημα αποφάσεων για την ελονοσία του IVCC, καθώς επίσης και η υιοθέτησή της από το αφρικανικό τμήμα του Παγκόσμιου Οργανισμού Υγείας στα πλαίσια του αφρικανικού δικτύου για την ανθεκτικότητα σε εντομοκτόνα φορέων ασθενειών (ANVR) αποτελεί εγγύηση πως όλα τα δεδομένα που συλλέγονται από αυτούς τους φορείς θα έχουν τη δυνατότητα να μεταφέρονται από το ένα σύστημα στο άλλο χωρίς ιδιαίτερα προβλήματα. Μοιραία αυτό θα οδηγήσει και στην περαιτέρω ανάπτυξη εργαλείων που θα χειρίζονται με κοινό τρόπο τα δεδομένα αυτά ενοποιώντας τα. Με αυτό τον τρόπο όχι μόνο θα μπορούμε να οδηγούμαστε σε αποτελεσματικότερες παρεμβάσεις για τον έλεγχο των πληθυσμών των φορέων που μεταδίδουν ασθένειες αλλά και να αξιολογούμε τα αποτελέσματά τους, τροποποιώντας τις όπου αυτό κρίνεται απαραίτητο.

Με την επέκταση της MIRO ώστε να καλύψει την παρακολούθηση της ανάπτυξης ανθεκτικότητας σε εντομοκτόνα από έντομα τα οποία παρουσιάζουν ενδιαφέρον από την

σκοπιά της «αγροτικής παραγωγής» και όχι εκείνη της «δημόσιας υγείας» στην οποία έχουμε επικεντρωθεί μέχρι στιγμής, θα ήταν δυνατόν να μελετηθούν και οι επιβαρυντικές συνέπειες της αλόγιστης χρήσης εντομοκτόνων στις καλλιέργειες στην διευκόλυνση της μετάδοσης τροπικών νόσων από ανθεκτικούς φορείς.

ΚΕΦΑΛΑΙΟ ΤΡΙΤΟ: ΟΝΤΟΛΟΓΙΑ ΤΗΣ ΕΛΟΝΟΣΙΑΣ

Εισαγωγή

Μετά την επιτυχημένη προσπάθεια με την οντολογία και τη βάση δεδομένων που περιείχαν στοιχεία για την ανθεκτικότητα σε εντομοκτόνα που εμφανίζουν πληθυσμοί κουνουπιών, το επόμενο λογικό βήμα ήταν η απόπειρα κατασκευής μιας οντολογίας που θα περιγράφει και θα παρέχει όρους για την επισημείωση δεδομένων που αφορούν την ελονοσία, αντιμετωπίζοντάς την ολιστικά και όχι μόνο από την ιατρική ή βιολογική πλευρά της και συμπεριλαμβάνοντας όρους που έχουν να κάνουν με τις προσπάθειες ελέγχου της ασθένειας σε όλα τα επίπεδα.

Βέβαια προσπάθειες ταξινόμησης και περιγραφής των ασθενειών έχουν υπάρξει αρκετές στον χώρο της ιατρικής και μάλιστα πολύ πριν υπάρξουν οι ηλεκτρονικοί υπολογιστές και η μαζική παραγωγή δεδομένων που οδήγησαν στην ανάγκη των οντολογιών. Για παράδειγμα ο ίδιος ο Λινναίος πέρα από την ταξινόμηση των ειδών επιχείρησε ήδη από το 1768 να ταξινομήσει σε κατηγορίες και τις ασθένειες (Egdahl 1907). Αλλά και άλλα προγράμματα οργάνωσης δεδομένων κυρίως βιβλιογραφικών αναφορών όπως οι όροι MESH (Medical Subject Headings) της εθνικής βιβλιοθήκης Ιατρικής των ΗΠΑ (NLM) από το 1960 (Nelson, Schopen et al. 2004) και το τμήμα κλινικών όρων της SNOMED CT (Systematic Nomenclature of MEDicine Clinical Terms) (Ruch, Gobeill et al. 2008) επιχειρούν να περιγράψουν το πεδίο των ασθενειών και να παρέχουν ένα σύνολο κλινικών όρων. Όμως και οι δυο αυτές προσπάθειες δεν επιχειρούν να συσχετίσουν τους όρους που περιλαμβάνουν μεταξύ τους με λογικές σχέσεις. Αυτό έχει σαν αποτέλεσμα να καθίσταται προβληματική η διασύνδεση μεταξύ των όρων που περιέχουν μεταξύ τους, αλλά και με άλλες υπάρχουσες οντολογίες, αφού δεν είναι ξεκάθαρη κάθε φορά η έννοια της κάθε

οντότητας. Επιπλέον η έμφαση δίνεται κυρίως σε κλινικά ευρήματα χωρίς να καλύπτονται επαρκώς τομείς όπως η ανοσοβιολογία και οι παθογενετικοί μηχανισμοί που οδηγούν σε μολυσματικές νόσους και αυτό επιτείνει τη δυσκολία διασύνδεσης με ήδη υπάρχουσες οντολογίες που καλύπτουν τομείς βιολογικού ενδιαφέροντος.

Την ανάγκη αυτή στο χώρο της ελονοσίας ήρθε να την καλύψει η IDOMAL (Topalis, Mitraka et al. 2010), μια οντολογία που αναπτύχθηκε ακριβώς γι' αυτόν τον σκοπό, σύμφωνα με τα κριτήρια της OBO Foundry που έχουν περιγραφεί παραπάνω. Όμως για να αποφευχθεί το λάθος της SNOMED, η IDOMAL έπρεπε να ενταχθεί και να ακολουθήσει κάποια οντολογία ανώτερου επιπέδου και κάποια οντολογία αναφοράς, ακριβώς όπως η οντολογία της ανατομίας του κουνουπιού ακολουθεί το πρότυπο της βασικής τυπικής οντολογίας (BFO) και την κοινή ανατομική οντολογία αναφοράς (CARO) ώστε μέσω αυτών να επιτυγχάνεται η διασύνδεση με τις υπόλοιπες οντολογίες του πεδίου. Στην περίπτωση της IDOMAL και πάλι η βασική τυπική οντολογία θα παρείχε τους όρους του ανώτερου επιπέδου, αλλά έλειπε η οντολογία αναφοράς. Για το σκοπό αυτό συμπράξαμε με άλλα εργαστήρια ώστε να συμβάλλουμε στην ανάπτυξη μιας οντολογίας αναφοράς για τις μολυσματικές ασθένειες (Infectious Disease Ontology – IDO) (Sintchenko 2010) αποδεχόμενοι να αναπτύξουμε την οντολογία της ελονοσίας σαν επέκτασή της. Από εκεί προέκυψε και το όνομα IDOMAL (Infectious Disease Ontology – MALaria). Η απόφαση αυτή εξασφαλίζει πως όλες οι οντολογίες στο πεδίο των μολυσματικών ασθενειών θα περιγράφονται με ενιαίο τρόπο διευκολύνοντας τα μέγιστα για την διασύνδεσή τους. Ο τρόπος με τον οποίο πραγματοποιείται αυτό είναι πως η IDO παρέχει μόνο έναν πυρήνα βασικών όρων, οι οποίοι εξειδικεύονται κατά περίσταση στις επιμέρους επεκτάσεις που αφορούν συγκεκριμένες ασθένειες. Δεν επιλύει όμως αυτόματα όλα τα ζητήματα σχετικά με τη δομή της καινούργιας οντολογίας μιας και η πλήρη κάλυψη μιας ασθένειας απαιτεί δεδομένα και όρους ετερογενείς μεταξύ τους που θα αφορούν τα κλινικά δεδομένα, τα

βιολογικά δεδομένα, τους τρόπους αντιμετώπισης της ασθένειας, είτε πρόκειται για θεραπεία ή πρόληψή της, που μπορεί με την σειρά της να περιλαμβάνει μια πληθώρα παρεμβάσεων.

Αποτελέσματα και συζήτηση

Δομή της οντολογίας της ελονοσίας

Η ελονοσία σαν ασθένεια, αποτελεί τυπικό παράδειγμα μιας υποομάδας των μολυσματικών ασθενειών όπου το κλασσικό μοντέλο πως κάποιο παθογόνο προκαλεί την ασθένεια στον ξενιστή του είναι πιο πολύπλοκο μιας και περιλαμβάνει και την ύπαρξη του φορέα ή ενδιάμεσου ξενιστή ο οποίος είναι απαραίτητος για την συμπλήρωση του κύκλου ζωής του παθογόνου και συνεπώς και για την μετάδοση της ασθένειας. Η IDO εν πολλοίς αντιμετωπίζει το συγκεκριμένο φαινόμενο σαν μια ιδιαιτερότητα ορισμένων μόνο μολυσματικών ασθενειών αφήνοντας τον χώρο σε εκείνες για να τον καλύψουν. Η απόφαση για το πως θα γίνει αυτό με τον καλύτερο δυνατό τρόπο είναι από αυτές που πρέπει να παρθούν από την αρχή της ανάπτυξης μιας οντολογίας και στη συγκεκριμένη περίπτωση συνδυάστηκε με την απάντηση στο ερώτημα ποια είναι η πραγματική ανάγκη για τη δημιουργία της οντολογίας της ελονοσίας και το πως αυτή θα χρησιμοποιηθεί. Η προφανής απάντηση είναι πως η οντολογία θα αποτελέσει τη βάση για την ανάπτυξη πληροφοριακών εργαλείων για τον συγκερασμό της πληθώρας των κατακερματισμένων δεδομένων που υπάρχουν διαθέσιμα για την ελονοσία με στόχο την αποτελεσματικότερη παρέμβαση σε όλα τα επίπεδα ώστε να καταστεί δυνατός ο έλεγχος της ασθένειας. Κατά συνέπεια απευθύνεται σε μια πληθώρα ανθρώπων διαφορετικών ειδικοτήτων (ελονοσιολόγους, ειδικούς δημόσιας υγείας, βιολόγους, τεχνικούς υπολογιστών) και πρέπει να μεγιστοποιεί την δυνατότητα διασύνδεσης με άλλες οντολογίες και να επιτρέπει την μελλοντική της επέκταση και σε τομείς που δεν θα καλύπτει αρχικά.

Για τον σκοπό αυτό προκρίθηκε η λύση η IDOMAL να συμπεριλάβει όρους από το επίπεδο του μακρομορίου ως εκείνο του οικοσυστήματος από διαφορετικές οντολογίες ώστε

να περιγράψει με πιστότερο τρόπο τα σχετικά με την ασθένεια και τους τρόπους αντιμετώπισής της, από το να αφήσει τους όρους αυτούς στις οικείες οντολογίες, οι οποίες θα συμπεριλαμβάνονταν στα υπό ανάπτυξη εργαλεία χωριστά. Με τον τρόπο αυτό επιτρέπεται αφενός η λειτουργική διασύνδεση με λογικές σχέσεις ανάμεσα στους όρους που προέρχονται από άλλες οντολογίες και τους όρους που υπάρχουν στην IDOMAL και αφετέρου τα σημεία που ενώνουν την IDOMAL με τις υπόλοιπες οντολογίες της OBO Foundry. Η ένταξη των εξωτερικών αυτών όρων έγινε με τη διατήρηση του μοναδικού κωδικού τους από την οντολογία από την οποία προέρχονται και την προσθήκη ενός δεύτερου, μοναδικού και χαρακτηριστικού για την IDOMAL σε ορισμένες περιπτώσεις. Αυτό κυρίως συνέβη αναφορικά με όρους που δεν υπήρχαν σε άλλες οντολογίες στις αρχικές εκδόσεις της IDOMAL, αλλά είχαμε την ανάγκη να τους χρησιμοποιήσουμε, οπότε προστέθηκαν αποκτώντας ένα μοναδικό κωδικό της IDOMAL παρά το γεγονός πως αναγνωρίζαμε πως είναι όροι ευρύτερου ενδιαφέροντος και ως τέτοιοι θα έπρεπε να υπάρχουν σε κάποια οντολογία αναφοράς. Όταν προστέθηκαν σε αυτές τις οντολογίες ο κωδικός αντικαταστάθηκε με εκείνον της οντολογίας αναφοράς, αλλά παρέμεινε και ο αρχικός της IDOMAL ώστε αν κάποιος τον είχε ήδη χρησιμοποιήσει να μη χαθούν οι υπάρχουσες επισημειώσεις και ταυτόχρονα να μπορούν να διασυνδεθούν με άλλες που χρησιμοποιούν τον όρο της οντολογίας αναφοράς.

Η επιλογή να αποτελέσει η IDOMAL επέκταση της IDO, σήμαινε αυτόματα πως θα υιοθετούσε στη βάση της τους όρους της βασικής τυπικής οντολογίας, γεγονός που την καθιστά δυσκολονόητη για τους μη ειδικούς εξασφαλίζοντας την ορθότητά της από οντολογική σκοπιά. Για παράδειγμα κάποιοι όροι ταξινομήθηκαν στην κατηγορία "υποθετικό τμήμα διαδικασίας " (fiat process part) η οποία ορίζεται σαν μέρος μιας διαδικασίας χωρίς σαφώς καθορισμένη αρχή και τέλος (Smith and Grenon 2004). Το πρόβλημα ξεπερνιέται αν κατά τη δημιουργία των εργαλείων που βασίζονται στην οντολογία

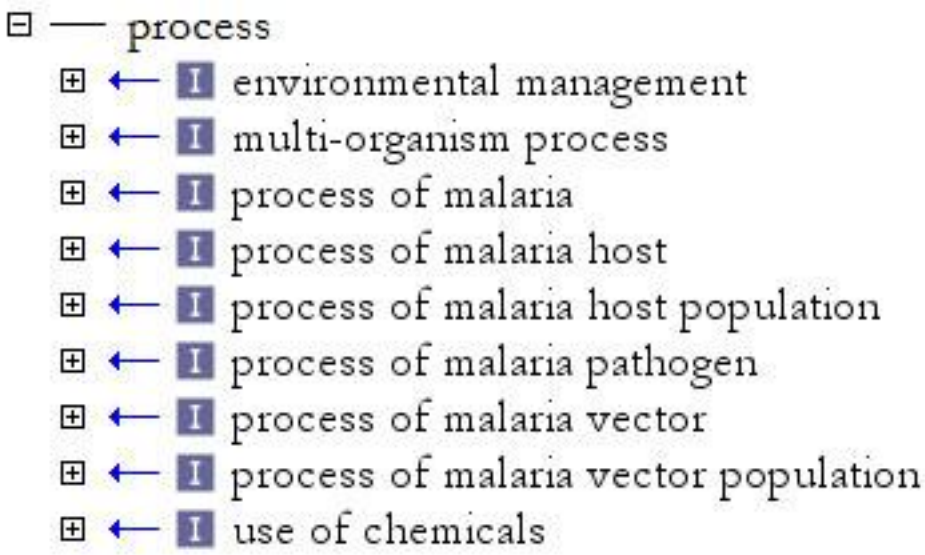
ληφθεί ειδική μέριμνα ώστε να μειώνεται η έκθεση του τελικού χρήστη στην λογική της οντολογίας. Η τελευταία θα ακολουθηθεί από τους υπολογιστές για να εξασφαλίσει την ανάκτηση των δεδομένων που επιθυμεί ο χρήστης. Ρόλος δικός του είναι να αναλύσει τα δεδομένα που θα ανακτήσει με αυτόν τον τρόπο.

Αν και η βασική τυπική οντολογία παρέχει τους όρους των αρχικών επιπέδων της IDOMAL δεν χρησιμοποίησα τα τρία πρώτα επίπεδα (Εικόνες 1 και 16). Αυτό έγινε για να μειωθεί το βάθος της οντολογίας και να είναι πιο εύκολη η επισκόπηση της από μη οντολογολόγους αλλά ειδικούς με γνώση της ελονοσίας. Άλλωστε οποιαδήποτε στιγμή μπορούν να προστεθούν οι όροι των 3 πρώτων επιπέδων.

Εικόνα 16: Από την IDOMAL απουσιάζουν τα 3 πρώτα επίπεδα της βασικής τυπικής οντολογίας.

Στην πιο απλή τους μορφή οι ασθένειες αφορούν μια βιολογική οντότητα, τον πάσχοντα. Στις μολυσματικές ασθένειες η κατάσταση περιπλέκεται γιατί πλέον υπάρχει ένα ζεύγος βιολογικών οντοτήτων που συμμετέχουν, το παράσιτο και ο ξενιστής. Στην περίπτωση της ελονοσίας αλλά και γενικότερα των ασθενειών που μεταδίδονται μέσω

φορέων, υπάρχει και μια τρίτη βιολογική οντότητα ο φορέας ή ενδιάμεσος ξενιστής που είναι απαραίτητος για να ολοκληρώσει τον βιολογικό του κύκλο το παράσιτο. Το γεγονός αυτό αλλά και το γεγονός πως υπάρχουν παρεμβάσεις για τον περιορισμό της ασθένειας που αφορούν είτε πληθυσμούς ατόμων, είτε το περιβάλλον μας οδήγησε σε ένα επιπλέον διαχωρισμό. Έτσι για παράδειγμα το τμήμα που η βασική τυπική οντολογία περιγράφει τις διαδικασίες (processes) διαιρέθηκε σε κατηγορίες που περιελάμβαναν τις διαδικασίες που αφορούσαν τον ξενιστή (με την έννοια αυτού που νοσεί), το παράσιτο και τον φορέα. Με αυτόν τον διαχωρισμό θα μπορούσαν να μείνουν συγκεντρωμένες κάτω από έναν πατρικό όρο οι διαδικασίες που θα λάμβαναν χώρα στον ίδιο οργανισμό. Γρήγορα φάνηκε πως αυτό δεν ήταν αρκετό γιατί υπήρχαν διαδικασίες που αφορούσαν πληθυσμούς ξενιστών ή φορέων, ενώ υπάρχουν και φυσιολογικές διαδικασίες οι οποίες αφορούν την συνύπαρξη και την συμμετοχή τουλάχιστον δύο διαφορετικών ειδών (παρασίτου με φορέα, παρασίτου με ξενιστή ή φορέα με ξενιστή). Τέλος η χρήση χημικών ουσιών καθώς και η διαχείριση του περιβάλλοντος αποτελούν από μόνες τους κατηγορίες διαδικασιών (Εικόνα 17).



Εικόνα 17: Η κατηγοριοποίηση των διαδικασιών στην οντολογία της ελονοσίας.

Ο ίδιος διαχωρισμός σε παράσιτο, φορέα και ξενιστή χρησιμοποιήθηκε και όπου αλλού κρίθηκε δόκιμο ώστε να αποσαφηνισθεί πλήρως το περιεχόμενο της οντολογίας και σε τι ακριβώς αναφέρεται κάθε οντότητα που περιλαμβάνεται σε αυτή. Έτσι για παράδειγμα και η κατηγορία που περιγράφει τις ιδιότητες (quality) χωρίσθηκε αναλόγως.

Η τελευταία επιλογή όσον αφορά την δομή της IDOMAL είχε να κάνει με το να μην συμπεριληφθούν σε αυτή όσοι όροι της βασικής τυπικής οντολογίας δεν επρόκειτο να χρησιμοποιηθούν σε αυτή. Αντίθετα, επιδιώχθηκε όπου ήταν δυνατό η χρήση ενός όρου περισσότερο από μία φορές με τέτοιο τρόπο ώστε να αποδίδει όσο το δυνατόν πιο πιστά την πραγματικότητα. Φυσικά στις περιπτώσεις αυτές μία και μόνο ήταν η *is_a* σχέση που συνέδεε τον συγκεκριμένο όρο με κάποιον πατρικό, ενώ στις υπόλοιπες περιπτώσεις η συσχέτιση γινόταν με διαφορετικές σχέσεις αφού σύμφωνα με τον νόμο της αντίφασης του Αριστοτέλη κάθε έννοια δεν μπορεί να αντιφάσκει με τον εαυτό της, να είναι δηλαδή συγχρόνως ίδιο και όχι ίδιο με τον εαυτό της, (το Α δεν μπορεί να είναι συγχρόνως Α και όχι Α) γιατί δύο έννοιες όπου η μία βεβαιώνει κάτι για ένα πράγμα και η άλλη αρνείται αυτό το κάτι, είναι αντιφατικές και δεν μπορεί να είναι ταυτόχρονα και οι δύο αληθινές. Σχέσεις που, όπως και στην περίπτωση της οντολογίας της ανθεκτικότητας στα εντομοκτόνα ανήκουν σε αυτές που είναι λογικά ορισμένες και έχουν εγκριθεί για χρήση σε βιοϊατρικές οντολογίες από την OBO Foundry.

Περιληπτικά η IDOMAL (έκδοση 1.22, Οκτώβριος 2010) περιλαμβάνει 2394 μοναδικούς όρους από τους οποίους οι 2379 διαθέτουν ορισμούς. Αυτοί κατανέμονται σε 12 κατηγορίες όλες προερχόμενες από την βασική τυπική οντολογία και ορισμένες από αυτή. Πολυπληθέστερη όλων είναι η κατηγορία «process» (1320 όροι) που περιλαμβάνει μαζί με τις κατηγορίες «fiat process part» (121 όροι) και «process boundary» (2 όροι) το σύνολο των διαδικασιών που αφορούν την ελονοσία. Η έμφαση έχει δοθεί κυρίως από την πλευρά

του φορέα και με κανένα τρόπο δεν υπονοείται πως αυτοί οι υπάρχοντες 1443 περιγράφουν καθ' ολοκληρία τις διαδικασίες που συμβαίνουν και σχετίζονται με την ελονοσία. Μια άλλη πολυπληθής κατηγορία είναι η κατηγορία «object» που περιλαμβάνει όλες τις ανεξάρτητες συνεχείς οντότητες όπως αυτές έχουν ορισθεί στη βασική τυπική οντολογία από το επίπεδο του μακρομορίου μέχρι το επίπεδο του οργανισμού. Όλες οι χημικές ουσίες που χρησιμοποιούνται για τη θεραπεία και την πρόληψη της ελονοσίας συμπεριλαμβάνονται με *is_a* σχέση στην κατηγορία αυτή και αυτό εξηγεί το μεγάλο της μέγεθος. Από την άλλη μεριά οι πληθυσμοί των ατόμων όπου χρειάστηκε να περιγραφούν ως τέτοιοι κατηγοριοποιήθηκαν ως «object aggregate». Όμως η ίδια κατηγορία περιλαμβάνει και τους συνδυασμούς φαρμάκων τα οποία χρησιμοποιούνται κατά περίπτωση. Οι κατηγορίες «quality» και «role» είναι επίσης πολυπληθείς με την τελευταία να περιλαμβάνει και τα είδη των παρασίτων που προκαλούν ελονοσία και τους φορείς που τα μεταφέρουν αλλά και τους τελικούς ξενιστές τους μιας και δεν είναι μόνο ο άνθρωπος που νοσεί από ελονοσία. Η γενική παρατήρηση είναι πως αν κανείς αθροίσει τους όρους που συμμετέχουν σε κάθε μια από τις αρχικές κατηγορίες θα ξεπεράσει κατά πολύ τον αριθμό των 2394 όρων και αυτό γιατί όπως έχει ήδη αναφερθεί υπάρχουν όροι που αναφέρονται περισσότερες από μια φορές στην IDOMAL. Επίσης το 27% των όρων της IDOMAL διαθέτει περισσότερους από έναν γονείς (Πίνακας 5).

Όροι με αριθμό γονέων	**Πλήθος όρων**
0	12 (< 1%)
1	1759 (73%)
2	540 (22%)
3	48 (2%)
4	20 (< 1%)
>4	15 (< 1%)

Πίνακας 5: Στατιστική ανάλυση του αριθμού των πατρικών όρων κάθε όρου της οντολογίας της ελονοσίας.

Μόνο ένας όμως από τους γονείς είναι συσχετισμένος με σχέση *is_a* καθιστώντας έτσι πλήρως γραμμικό και αποσαφηνισμένο το περιεχόμενο της οντολογίας. Τέλος, οι 12 όροι που εμφανίζονται χωρίς γονείς αντιστοιχούν στο πρώτο επίπεδο της IDOMAL και όπως ήδη αναφέρθηκε παραπάνω είναι όροι του τέταρτου επιπέδου της βασικής τυπικής οντολογίας. Στον (Πίνακα 6) μπορεί κανείς να δει περιληπτικά και ποιοτικά τα διαφορετικά είδη των σχετικών με την ελονοσία δεδομένων που μπορούν να επισημειωθούν με χρήση των όρων της IDOMAL και πως αυτά ποικίλλουν από τα κλινικά δεδομένα της ασθένειας (συμπεριλαμβάνοντας και την επιδημιολογία της), και τους τρόπους αντιμετώπισής της μέχρι την βιολογία του παρασίτου, των φορέων και των ξενιστών.

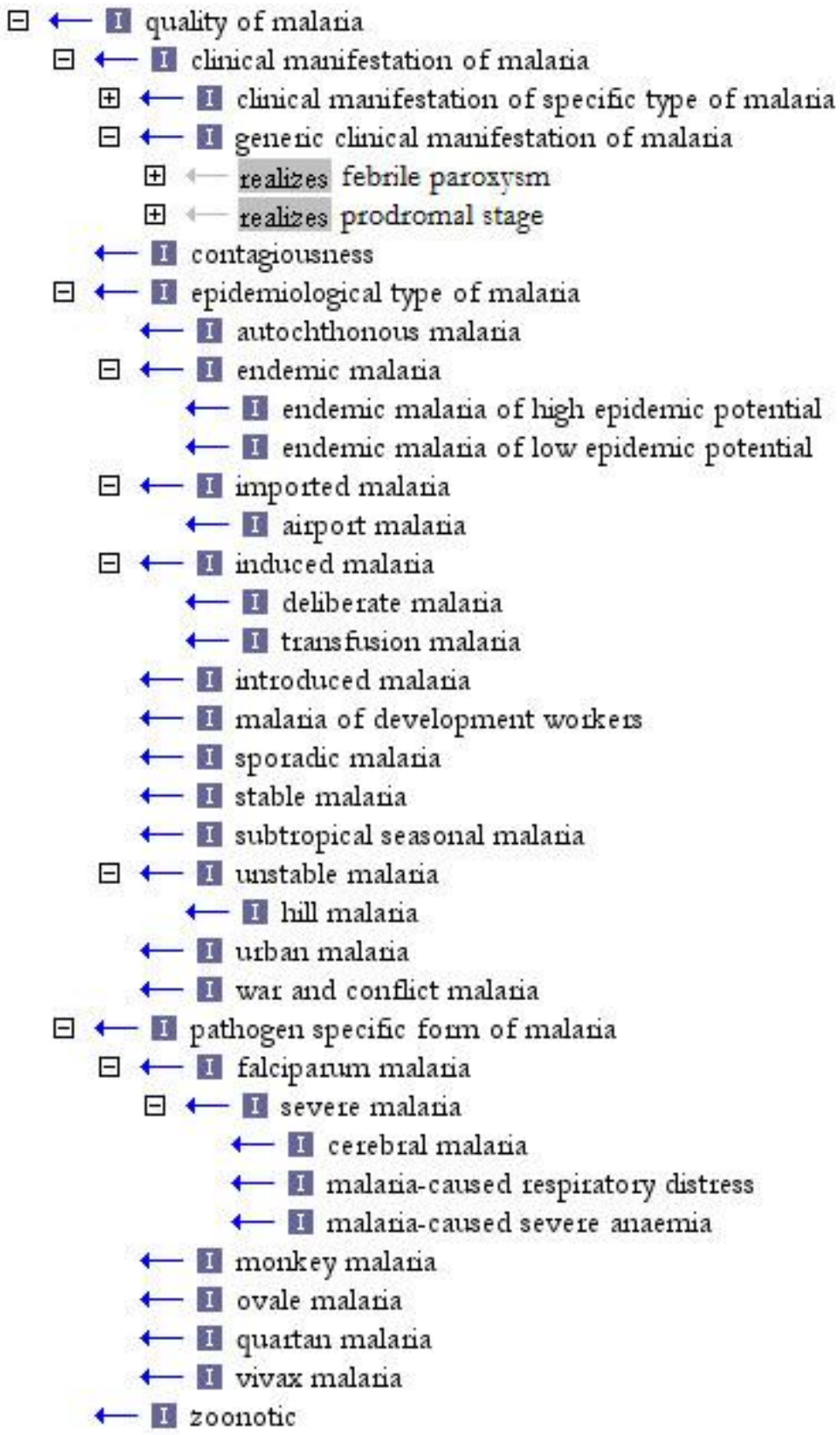
Κατηγορίες	**Όροι**	**Περίληψη περιεχομένων**
condition	45	Κλινικά δεδομένα του πάσχοντος από ελονοσία
disposition	77	Όροι σχετικοί με την ασθένεια της ελονοσίας
fiat process part	121	Φυσιολογικές διαδικασίες κυρίως του φορέα
object	1150	Χημικές ενώσεις, εντομοκτόνα, ανθελονοσιακά φάρμακα, οργανισμοί και ανατομικές δομές
object aggregate	89	Πληθυσμοί οργανισμών και σύμπλοκα πρωτεϊνών
process	1320	Φυσιολογικές διαδικασίες του ξενιστή, του φορέα, του παρασίτου αλλά και προγραμματισμένες διαδικασίες όπως η προφύλαξη ή οι προσπάθειες ελέγχου της ελονοσίας
process boundary	2	
quality	253	Φαινότυποι του ξενιστή, του φορέα και του παρασίτου
role	577	Ρόλοι των χημικών ουσιών (φάρμακα, εντομοκτόνα) αλλά και των οργανισμών (ξενιστής , φορέας, παράσιτο)
spatial region	57	Χαρακτηριστικά του περιβάλλοντος (λίμνες, οικολογικοί θώκοι)

Πίνακας 6: Συνοπτική παρουσίαση του πρώτου επιπέδου όρων της IDOMAL, το μέγεθός του και των περιεχομένων του.

Οντολογία της ελονοσίας : Οι όροι που σχετίζονται με την ασθένεια

Αναπόφευκτα στην οντολογία της ελονοσίας πολλοί όροι θα αναφέρονται στα συμπτώματά της, τις διάφορες θεραπευτικές αντιμετωπίσεις της, τις διαδικασίες που συμβάλλουν στην πρόληψή της, καθώς και σε όρους που θα παραπέμπουν στα παράσιτα που είναι οι αιτιολογικοί της παράγοντες (Εικόνα 18). Όμως είναι πέρα από τον σκοπό της οντολογίας της ελονοσίας να παρέχει όρους που θα αναφέρονται γενικά σε ασθένειες. Αυτοί θα πρέπει να προέλθουν από την οντολογία των ασθενειών (disease ontology – DO) (Osborne, Flatow et al. 2009) και την οντολογία μολυσματικών ασθενειών (infectious disease ontology – IDO). Συνεπώς, μόνο οι σχετικοί με την ελονοσία όροι προστέθηκαν, εκτός από τα σημεία που ήταν εντελώς απαραίτητοι κάποιοι γενικότεροι όροι ώστε να είναι δυνατή η επισημείωση των δεδομένων που υπάρχουν με όρους της IDOMAL.

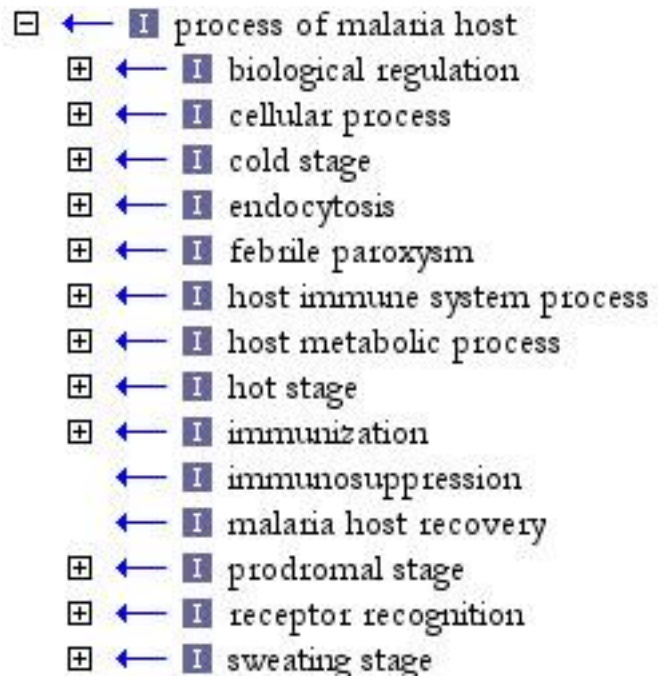
Ανάμεσα στους όρους που σχετίζονται με την ελονοσία ως ασθένεια ανήκουν και αυτοί που περιγράφουν τις κλινικές εκφάνσεις και τα συμπτώματα που την συνοδεύουν και αποτέλεσαν ένα σημείο προβληματισμού τόσο για την οντολογική τους θέση μέσα στο πλαίσιο της βασικής τυπικής οντολογίας, όσο και εξαιτίας του γεγονότος πως περιλαμβάνουν γενικότερες οντότητες οι οποίες δεν συναντώνται μόνο στην περίπτωση της οντολογίας της ελονοσίας, όπως για παράδειγμα η αναιμία, ο πυρετός, η ηπατική δυσλειτουργία, το πνευμονικό οίδημα και άλλα. Όπως και προηγουμένως έτσι και σε αυτή την περίπτωση η τελική επιλογή ήταν να συμπεριληφθούν στην οντολογία της ελονοσίας, ώστε να δοθεί η δυνατότητα επισημείωσης των αντίστοιχων κλινικών δεδομένων που υπάρχουν διαθέσιμα. Αντίστοιχη αντιμετώπιση είχαν και τα διάφορα ανθελονοσιακά σκευάσματα είτε πρόκειται για τη θεραπεία είτε για την πρόληψή της ασθένειας. Σχεδόν η πλήρης λίστα τους έχει συμπεριληφθεί και μάλιστα σε αρκετές περιπτώσεις έχουν προστεθεί και τα εμπορικά τους ονόματα σαν συνώνυμα.



Εικόνα 18: Οι ιδιότητες της ασθένειας της ελονοσίας και η κατηγοριοποίησή τους.

Οντολογία της ελονοσίας: Οι όροι που σχετίζονται με τον ξενιστή και τους πληθυσμούς του

Ένα σημαντικό τμήμα της IDOMAL περιλαμβάνει όρους που συσχετίζουν τον τελικό ξενιστή και τους πληθυσμούς του με την ελονοσία. Η πιο προφανής ομάδα τέτοιων όρων και η πολυπληθέστερη είναι εκείνη που περιγράφει τις παθοφυσιολογικές διαδικασίες που συμβαίνουν σε κάποιον πάσχοντα από ελονοσία. Αυτή η ομάδα συνδέεται άμεσα με εκείνη που περιγράψαμε προηγουμένως και αφορούσε τις κλινικές εκφάνσεις της ελονοσίας μιας και αυτές συνήθως αποτελούν το αποτέλεσμα των εν λόγω παθοφυσιολογικών διαδικασιών. Στην κατηγορία αυτή ανήκουν και οι όροι που περιγράφουν την βιολογία της ελονοσίας και πιο συγκεκριμένα την ανοσολογία της, αφού όλες οι αποκρίσεις του ανοσοποιητικού συστήματος απέναντι στα παράσιτα του γένους *Plasmodium* που προκαλούν την ελονοσία είναι στην πραγματικότητα φυσιολογικές διαδικασίες του ξενιστή. Όμως και στην περίπτωση αυτή το ενδιαφέρον είναι εστιασμένο σε όρους που περιγράφουν τις σχετικές διαδικασίες ανοσοαπόκρισης σε σχέση πάντα με την ελονοσία και όχι γενικά.



Εικόνα 19: Οι κύριες κατηγορίες όρων που περιγράφουν τις διαδικασίες του ξενιστή της ελονοσίας.

Παρόλα αυτά όμως για λόγους πληρότητας στις περιπτώσεις όπου είναι γνωστές οι πρωτεΐνες που συμμετέχουν στις διαδικασίες αυτές, προστέθηκαν και αυτές με την σειρά τους, έτσι ώστε να είναι δυνατόν να χρησιμοποιηθούν για την επισημείωση των σχετικών δεδομένων. Μια άλλη σειρά όρων σχετική με τους προηγούμενους προστέθηκε ώστε να περιγράψει τελικά αν ο ξενιστής εμφανίζει τελικά και σε ποιο ποσοστό ανοσία προς την ελονοσία σαν σύνολο ή σε συγκεκριμένες διαδικασίες της. Αυτοί οι όροι δεδομένου ότι αποτελούν φαινότυπο του ξενιστή βρίσκονται στην κατηγορία «quality of host».

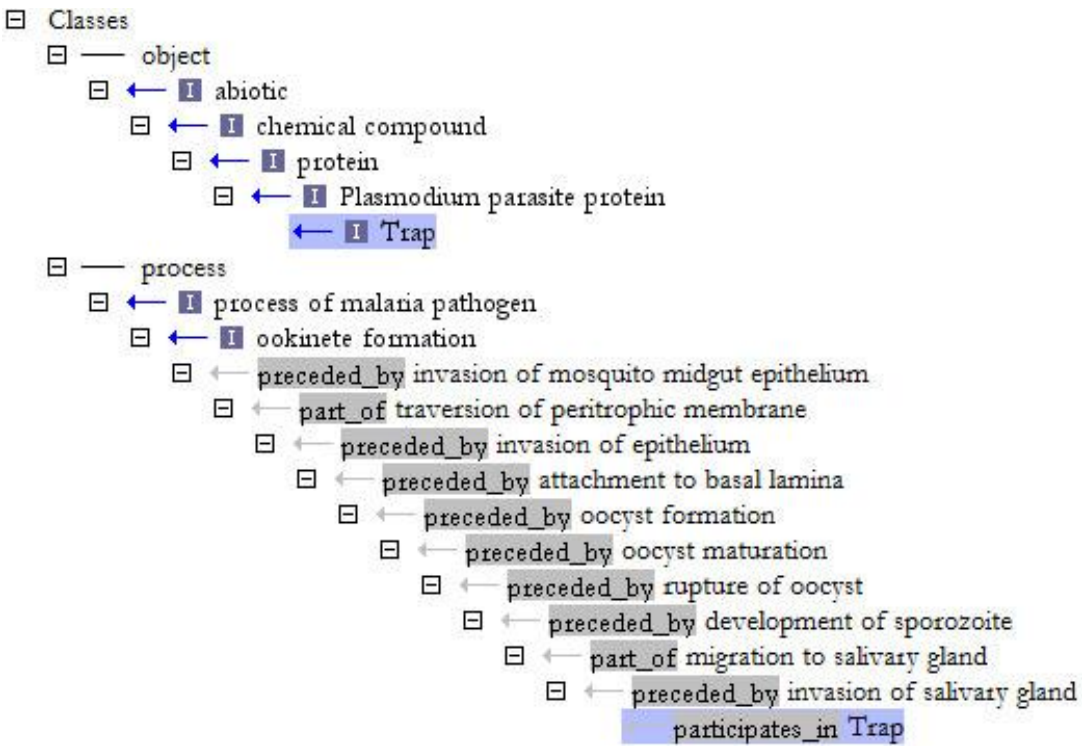
Σε ένα άλλο επίπεδο πολυπλοκότητας τώρα υπάρχει η ανάγκη να περιγραφούν διαδικασίες όπως οι απόπειρες ανοσοποίησης αλλά και φαινότυποι που εμφανίζονται σε πληθυσμούς ξενιστών όπως για παράδειγμα ο αριθμός των κρουσμάτων σε δεδομένο πληθυσμό και χρονικό διάστημα. Τέτοιου είδους περιγραφές πληθυσμών είναι και οι συχνότερα απαντούμενες στο χώρο της δημόσιας υγείας.

Οντολογία της ελονοσίας: Οι όροι που σχετίζονται με το παράσιτο

Από την IDOMAL δεν θα μπορούσαν να λείπουν οι όροι που περιγράφουν τη βιολογία και τα στάδια ζωής των παρασίτων του γένους *Plasmodium* . Ακριβώς αυτοί οι όροι μπορούν να αποτελέσουν και το σημείο διασύνδεσης της οντολογίας της ελονοσίας με εκείνες τις οντολογίες που θα περιγράφουν τις φυσιολογικές διαδικασίες του παρασίτου και κατά συνέπεια θα είναι δυνατό να διασυνδέσουν την VectorBase με την PlasmoDB (Aurrecoechea, Brestelli et al. 2009) (τη βάση δεδομένων που περιέχει πληροφορίες για τα πλασμώδια) και κατ' επέκταση την EuPathDB (Aurrecoechea, Brestelli et al. 2010) η οποία αντικατέστησε την PlasmoDB. Το δίλημμα σε αυτή την περίπτωση ήταν αν θα συμπεριλάβουμε αναλυτικά ακόμα και σε επίπεδο πρωτεϊνών ανάλογους όρους στην IDOMAL ή αν θα τους αφήσουμε εκτός οντολογίας.

Ένα παράδειγμα που μπορεί να παρουσιάσει αυτήν ακριβώς την κατάσταση είναι η σχετική με την θρομβοσπονδίνη ανώνυμη πρωτεΐνη (TRAP) του πλασμωδίου. Ταυτοποιήθηκε για πρώτη φορά πριν από 21 χρόνια στο *Plasmodium falciparum* (Robson, Hall et al. 1988) και από τότε και σε διάφορα άλλα είδη του ιδίου γένους. Η λειτουργία αυτής της διαμεμβρανικής πρωτεΐνης εντοπίστηκε στο στάδιο των σποροζωιτών, όπου αλληλεπιδρούσε με το υπόστρωμα συμβάλλοντας στην ικανότητα μετακίνησής τους (Spaccapelo, Naitza et al. 1997; Sultan, Thathy et al. 1997). Αργότερα διαπιστώθηκε και ο ενεργός της ρόλος στην διείσδυση του παρασίτου στα ηπατοκύτταρα (Akhouri, Sharma et al. 2008; Morahan, Wang et al. 2009). Θα έπρεπε λοιπόν η TRAP να αποτελέσει μέρος της IDOMAL; Υπάρχουν αρκετά είδη του γένους *Plasmodium* στα οποία δεν έχουμε καμιά πληροφορία για την ύπαρξη της ομόλογης πρωτεΐνης και του ορθόλογου γονιδίου. Μάλιστα μέσα σε αυτά συγκαταλέγονται και κάποια που προκαλούν ελονοσία στους ανθρώπους. Επιπλέον, οι πληροφορίες του είδους αποθηκεύονται σε ειδικές βάσεις δεδομένων και αποτελούν αντικείμενο επισημειώσεων με όρους της οντολογίας γονιδίων (GO). Τέλος δεν υπάρχει για την ώρα ανάγκη επιπλέον επισημειώσεων για την συγκεκριμένη πρωτεΐνη. Για όλους αυτούς τους λόγους, μάλλον η TRAP δεν θα έπρεπε να συμπεριληφθεί σαν αυτόνομη οντότητα στην IDOMAL. Από την άλλη μεριά όμως αποτελεί αντικείμενο ερευνών ως πιθανό συστατικό ενός μελλοντικού εμβολίου (Dolo, Modiano et al. 1999; Epstein, Giersing et al. 2007) και ως τέτοιο ενδεχομένως να υπάρξει η ανάγκη προσθήκης επισημειώσεων σε βάσεις δεδομένων που συγκεντρώνουν στοιχεία για τις προσπάθειες ανοσοποίησης κατά της ελονοσίας. Ανάλογη είναι η κατάσταση και με άλλες πρωτεΐνες του πλασμωδίου που είναι δυνητικοί φαρμακολογικοί στόχοι σε προσπάθειες ανεύρεσης νέων ανθελονοσιακών σκευασμάτων (Artavanis-Tsakonas, Tongren et al. 2003; Hviid 2005). Ακριβώς αυτή η πιθανή εμπλοκή τους σε προσπάθειες ελέγχου της ελονοσίας μας οδήγησε στο να τις συμπεριλάβουμε τελικά στην IDOMAL. Η προσθήκη αυτή μάλιστα έγινε με

τέτοιο τρόπο ώστε να συσχετισθούν με οντολογικά ορθό τρόπο με σχέση *is_a* με τις πρωτεΐνες του πλασμωδίου κι αυτές με την σειρά του με τις πρωτεΐνες γενικά που είναι τελικά αντικείμενα, δηλαδή ανεξάρτητες οντότητες με συνεχή παρουσία στον χρόνο. Από την άλλη μεριά προστέθηκαν και ως συμμετέχοντες παράγοντες στις αντίστοιχες βιολογικές διαδικασίες, συσχετιζόμενες με αυτές με τις σχέσεις *participates_in* ή *agent_in* ανάλογα με το αν η πρωτεΐνη απλά συμμετέχει ή είναι ο αιτιολογικός παράγοντας της διαδικασίας αντίστοιχα (Εικόνα 20). Αυτή τη στιγμή 36 πρωτεΐνες του πλασμωδίου συμπεριλαμβάνονται στην IDOMAL για αντίστοιχους λόγους και ο αριθμός τους αναμένεται να αυξηθεί με τον καιρό καθώς μεγαλώνουν οι γνώσεις μας για την μοριακή βάση της ελονοσίας.



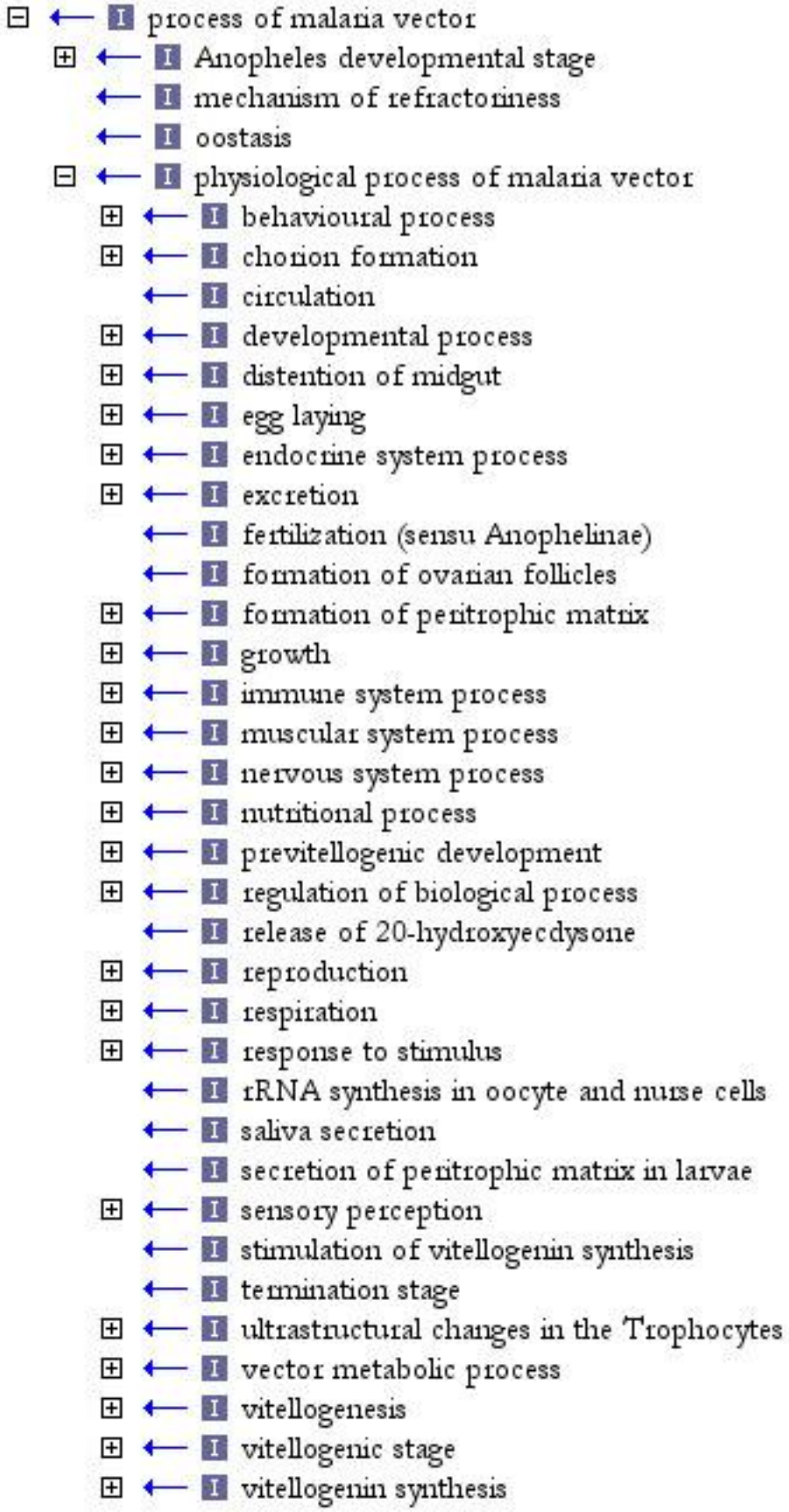
Εικόνα 20: Η θέση της πρωτεΐνης TRAP στην οντολογία της ελονοσίας.

Ένας άλλος τομέας στον οποίο θα πρέπει να επεκταθεί το τμήμα αυτό της οντολογίας είναι ο φαινότυπος της ανθεκτικότητας σε αντιβιοτικά που εμφανίζουν πολλά παράσιτα. Στην παρούσα μορφή της IDOMAL έχει απλά διαμορφωθεί η δομή και έχουν

προστεθεί ελάχιστοι σχετικοί όροι ώστε να μπορεί η επέκταση να λάβει χώρα χωρίς να δημιουργήσει προβλήματα στα υπόλοιπα τμήματα της οντολογίας.

Οντολογία της ελονοσίας: Οι όροι που σχετίζονται με τον φορέα και τους πληθυσμούς του

Αν για την ανθεκτικότητα του πλασμωδίου στα αντιβιοτικά δεν υπήρχε κάποια σχετική οντολογία, για την ανθεκτικότητα των φορέων της ελονοσίας απέναντι στα εντομοκτόνα υπήρχε η MIRO που περιγράφηκε στο προηγούμενο κεφάλαιο ως ανεξάρτητη οντολογία. Δύο ήταν οι εναλλακτικοί δρόμοι που μπορούσαμε να ακολουθήσουμε. Να παραπέμψουμε όσους σκόπευαν να δημιουργήσουν βιοπληροφορικά εργαλεία για την ελονοσία στο να χρησιμοποιήσουν ανεξάρτητα και την MIRO και την IDOMAL σαν οντολογίες, ή να συμπεριλάβουμε ένα μεγάλο μέρος ή και καθ' ολοκληρία την MIRO στην IDOMAL. Έγινε το δεύτερο και οι όροι της MIRO που χρησιμοποιήθηκαν διατήρησαν τους μοναδικούς κωδικούς που είχαν από την αρχή χωρίς να τους αποδοθούν νέοι με κανένα τρόπο. Η όλη διαδικασία πραγματοποιήθηκε ακολουθώντας τους κανόνες που είναι αποδεκτοί από την OBO Foundry και εφαρμόζονται και από άλλες οντολογίες του βιοϊατρικού χώρου χρησιμοποιώντας έναν ελάχιστον αποδεκτό όγκο πληροφορίας για κάθε όρο που προέρχεται από μια εξωτερική οντολογία και περιλαμβάνει τους μοναδικούς κωδικούς και ονόματα τόσο της οντολογίας από την οποία προέρχεται ο όρος , όσο και του ίδιου του όρου που σημαίνει πως μεταφέρεται αυτούσιος ο κωδικός, το όνομα, ο ορισμός και τα συνώνυμά του. Ένα παράδειγμα είναι ο όρος «MIRO:00000129» με όνομα «midgut receptor resistance» . Κατ' αυτόν τον τρόπο ο όρος υπάρχει μια και μόνη φορά στον χώρο των βιοϊατρικών οντολογιών, αλλά ταυτόχρονα είναι δυνατή η συσχέτισή του με άλλους όρους της IDOMAL αν υπάρχει ανάγκη για κάτι τέτοιο.


process of malaria vector
Anopheles developmental stage
mechanism of refractoriness
oostasis
physiological process of malaria vector
behavioural process
chorion formation
circulation
developmental process
distention of midgut
egg laying
endocrine system process
excretion
fertilization (sensu Anophelinae)
formation of ovarian follicles
formation of peritrophic matrix
growth
immune system process
muscular system process
nervous system process
nutritional process
previtellogenic development
regulation of biological process
release of 20-hydroxyecdysone
reproduction
respiration
response to stimulus
rRNA synthesis in oocyte and nurse cells
saliva secretion
secretion of peritrophic matrix in larvae
sensory perception
stimulation of vitellogenin synthesis
termination stage
ultrastructural changes in the Trophocytes
vector metabolic process
vitellogenesis
vitellogenic stage
vitellogenin synthesis

Εικόνα 21: Οι κατηγορίες των φυσιολογικών διαδικασιών του φορέα.

Ένα άλλο σημαντικό μέρος αυτού του τμήματος της οντολογίας αποτελούν οι όροι εκείνοι που περιγράφουν τα στάδια ζωής και τις φυσιολογικές διαδικασίες των κουνουπιών (Εικόνα 21). Στην πραγματικότητα πρόκειται για μια ξεχωριστή οντολογία που αναπτύχθηκε αμέσως μετά την οντολογία της ανατομίας του κουνουπιού για να περιγράψει τις φυσιολογικές διαδικασίες που είχαν σχέση με την ελονοσία και δεν συμπεριλήφθηκαν στο οικείο τμήμα της GO, η οποία διατηρεί ουδέτερη αντιμετώπιση απέναντι σε ειδοειδικές φυσιολογικές διαδικασίες. Αποτέλεσμα αυτού είναι η κάλυψη των διαφόρων φυσιολογικών διαδικασιών να μην είναι ισοβαρής. Για παράδειγμα πολύ λίγες αναφέρονται στα όσα συμβαίνουν στα στάδια της προνύμφης του κουνουπιού, ενώ διαδικασίες όπως η συμπεριφορά του κατά την αναζήτηση ξενιστή και όσες σχετίζονται με τα γεύματα αίματος παρουσιάζονται πολύ πιο λεπτομερειακά. Επίσης όροι που αφορούν το ανοσοποιητικό σύστημα του φορέα και την ανοσοαπόκρισή του απέναντι στο παράσιτο (Dimopoulos 2003; Alphey 2009) έχουν για την ώρα παραληφθεί από την συγκεκριμένη έκδοση της IDOMAL. Σύντομα όμως η οντολογία θα επεκταθεί και προς αυτήν την κατεύθυνση.

Τέλος, στο τμήμα που αφορά τους πληθυσμούς των φορέων περιγράφεται το σύνολο των προσπαθειών για τον βιολογικό έλεγχο και μέρος των γενικότερων προσπαθειών για έλεγχο των πληθυσμών των κουνουπιών. Μέσω αυτών μπορεί να ελεγχθεί σε κάποιο βαθμό και η ίδια η ελονοσία.

Η οντολογία της ελονοσίας και άλλες οντολογίες

Από τους 2394 μοναδικούς όρους της IDOMAL, οι 885 (περίπου 35%) προέρχονται από άλλες οντολογίες, όπως η MIRO, η ChEBI (οντολογία χημικών ουσιών που παρουσιάζουν βιολογικό ενδιαφέρον), η GO (οντολογία γονιδίων) και η CL (οντολογία κυτταρικών σειρών). Η τελευταία χρησιμοποιήθηκε ιδιαίτερα στο τμήμα που περιέγραφε την ανοσολογία της ελονοσίας. Σε όλες αυτές τις περιπτώσεις οι εξωτερικοί όροι

συμπεριλήφθηκαν στην οντολογία με την διαδικασία της χρήσης της ελάχιστης πληροφορίας αναφοράς που περιγράφηκε παραπάνω και είναι γνωστή ως MIREOT (the Minimum Information to Reference an External Ontology Term) (Courtot, Gibson et al. 2009). Έτσι δεν ήταν αναγκαίο να συμπεριλάβουμε στην IDOMAL τις εξωτερικές αυτές οντολογίες στην ολότητά τους (η τρέχουσα έκδοση της GO περιλαμβάνει 33857 όρους) που θα καθιστούσε την χρήση της προβληματική και ταυτόχρονα να διατηρήσουμε τη δυνατότητα να συσχετίσουμε τους όρους των εξωτερικών οντολογιών με την δική μας.

Στο μέλλον και καθώς οι οντολογίες αναφοράς θα εξελίσσονται και θα επεκτείνονται, όροι που τώρα βρίσκονται στην IDOMAL θα μεταφερθούν εκεί που πραγματικά ανήκουν. Η διαδικασία αυτή είναι διαρκής και θα λαμβάνει υπόψη της το ενδεχόμενο κάποια εργαλεία που θα έχουν αναπτυχθεί στο μεταξύ να χρησιμοποιούν αυτούς τους όρους με τους μοναδικούς κωδικούς ταυτότητας της IDOMAL, όπως ήδη έχει συμβεί μέχρι σήμερα.

Χρήση της οντολογίας της ελονοσίας

Η οντολογία της ελονοσίας είναι διαθέσιμη εδώ και ένα χρόνο περίπου. Φυσικός χρήστης της είναι η VectorBase η βάση δεδομένων των φορέων που μεταδίδουν ασθένειες και στην οποία συμμετέχουμε σαν εργαστήριο. Βρίσκεται λοιπόν υπό ανάπτυξη ένα εργαλείο το οποίο θα επεκτείνει την IRBase που ήδη είναι τμήμα της VectorBase και βασίζεται στην MIRO, προσθέτοντάς στην VectorBase την δυνατότητα να φιλοξενήσει δεδομένα που αφορούν τη δομή των πληθυσμών των κουνουπιών. Στο τμήμα το σχετικό με τους φορείς της ελονοσίας η IDOMAL είναι η οντολογία που βρίσκεται στο επίκεντρο. Το σχήμα και η δομή της βάσης δεδομένων στην καρδιά αυτού του εργαλείου είναι ήδη έτοιμο, ενώ μέσα στο 2012 θα ολοκληρωθεί και η διασύνδεσή του με τον παγκόσμιο ιστό και θα είναι διαθέσιμο προς γενική χρήση, ενοποιώντας δεδομένα που αυτή την στιγμή βρίσκονται κατακερματισμένα στην Κρήτη, το Ηνωμένο Βασίλειο και στις Ηνωμένες Πολιτείες. Επίσης

θα δώσει την δυνατότητα να διασυνδεθούν με την VectorBase δεδομένα για πληθυσμούς κουνουπιών τα οποία συλλέγονται από τον Παγκόσμιο Οργανισμό Υγείας. στην Αφρική. Δυστυχώς στα πλαίσια της VectorBase δεν υπάρχει κάποιο πλάνο συλλογής και παρουσίασης κλινικών δεδομένων στο άμεσο μέλλον, οπότε η σχετική δυνατότητα που παρέχει η IDOMAL αυτή την στιγμή παραμένει ανενεργή.

Υπάρχουν όμως άλλες προσπάθειες από το IVCC για την ανάπτυξη συστημάτων που διευκολύνουν την λήψη αποφάσεων για την ελονοσία τα οποία ήδη βρίσκονται σε πιλοτική φάση και χρησιμοποιούν μεγάλο μέρος της IDOMAL. Το γεγονός πως οι ίδιοι σχεδόν φορείς που μεταδίδουν ελονοσία, μεταδίδουν και άλλες ασθένειες όπως για παράδειγμα Δάγκειο πυρετό οδήγησε στο να χρησιμοποιηθεί μεγάλο μέρος της IDOMAL και στο σύστημα υποστήριξης αποφάσεων για δάγκειο πυρετό που εφαρμόζεται πιλοτικά στο Μεξικό. Αυτό αποδεικνύει στην πράξη τη δύναμη των οντολογιών στο να διασυνδέσουν διαφορετικά σετ δεδομένων μεταξύ τους.

Συμπεράσματα

Ο στόχος μας με τη δημιουργία της οντολογίας της ελονοσίας ήταν να δημιουργήσουμε ένα χρήσιμο εργαλείο που σε βάθος χρόνου να βοηθήσει στο να αντιμετωπιστεί η ελονοσία σε παγκόσμιο επίπεδο. Ο καθοριστικός όμως παράγοντας γι' αυτό είναι το να γίνει αποδεκτή από την επιστημονική κοινότητα και η οποία θα την χρησιμοποιήσει για να εμπλουτίσει τις επισημειώσεις που υπάρχουν ήδη στις βάσεις δεδομένων. Αυτό θα μεταμορφώσει τις βάσεις δεδομένων από απλούς αποθηκευτικούς χώρους σε συστήματα που μπορούν να υποστηρίξουν πολύπλοκα εργαλεία και να αποκαλύψουν διασυνδέσεις ανάμεσα σε δεδομένα που ως σήμερα δεν ήταν προφανείς. Παρά το γεγονός πως η IDOMAL διαθέτει περίπου 2400 όρους η αλήθεια είναι πως απέχει πολύ από το να είναι ολοκληρωμένη, άλλωστε καμιά οντολογία δεν μπορεί ποτέ να είναι ολοκληρωμένη καθώς περιγράφει την πραγματικότητα και η πραγματικότητα είναι πως

συνεχώς οι γνώσεις μας για την επιστήμη και τον κόσμο διαρκώς επεκτείνονται. Ελπίζουμε πως η ανάπτυξη και η καθημερινή χρήση των υπό ανάπτυξη εργαλείων που χρησιμοποιούν την IDOMAL θα αποτελέσει το έναυσμα για την εμπλοκή και την συνεισφορά της επιστημονικής κοινότητας που ασχολείται με την ελονοσία στην περαιτέρω ανάπτυξή της.

ΓΕΝΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΟΠΤΙΚΕΣ

Στο πλαίσιο αυτής της διδακτορικής διατριβής δημιουργήθηκαν για πρώτη φορά οντολογίες που περιέγραψαν την ανατομία κουνουπιών και τσιμπουριών, το φαινόμενο της ανθεκτικότητας ορισμένων πληθυσμών κουνουπιών σε εντομοκτόνα, καθώς και την ελονοσία στο σύνολό της δίνοντας τη δυνατότητα επισημείωσης κλινικών δεδομένων, δεδομένων που σχετίζονται με τη βιολογία των φορέων που μεταδίδουν ελονοσία και των παρασίτων που την προκαλούν και τέλος με τις ανθρώπινες παρεμβάσεις σε όλα τα επίπεδα για τη θεραπεία, την προφύλαξη και τον περιορισμό της ασθένειας (Πίνακας 7).

Οντολογία	Περιγραφή	Πλαίσιο
TGMA	Ανατομία κουνουπιών	Στα πλαίσια της διατριβής
TADS	Ανατομία τσιμπουριών	Σε συνεργασία με καθ. D. Sonenshine
--	Φυσιολογικές διαδικασίες κουνουπιών	Στα πλαίσια της διατριβής
MIRO	Ανθεκτικότητα σε εντομοκτόνα	Στα πλαίσια της διατριβής
IDOMAL	Ελονοσία	Στα πλαίσια της διατριβής
VBcv	Κατάλογος όρων που δεν υπάρχουν σε καμιά οντολογία ακόμα, αλλά είναι απαραίτητοι για την ανάπτυξη βιοπληροφοριακών εργαλείων	Ακολουθώντας το παράδειγμα της FlyBase

Πίνακας 7: Οντολογίες που δημιουργήθηκαν στα πλαίσια αυτής της διατριβής.

Με βάση αυτές τις οντολογίες αναπτύχθηκαν εργαλεία που αποκαλύπτουν στην πράξη τα πλεονεκτήματα αυτής της προσέγγισης ιδιαίτερα όταν έχουμε να αντιμετωπίσουμε κατακερματισμένα και ετερογενή σύνολα δεδομένων. Βρισκόμαστε σε μια εποχή που παράγουμε μεγαλύτερο όγκο βιολογικών δεδομένων από αυτόν που μπορούμε να αναλύσουμε. Είναι λοιπόν σημαντικό να διαθέτουμε εργαλεία που θα αντλούν πληροφορίες

από πολλαπλές πηγές και ανταποκρινόμενα σε σύνθετες αναζητήσεις θα μπορούν να μας τροφοδοτήσουν με τα δεδομένα που μας ενδιαφέρουν άμεσα.

Η επιτυχία, η αναγνώριση και η χρήση της GO (που σε κάποιες στιγμές πλησιάζει και τα όρια της κατάχρησης) δείχνει τον δρόμο προς το μέλλον. Την ίδια στιγμή υπάρχουν ερευνητικά προγράμματα που προσπαθούν να συνδυάσουν τις ομοιότητες και τις συσχετίσεις που έχουν οι GO επισημειώσεις που έχουν αποδοθεί σε συγκεκριμένα γονίδια, ώστε να ταυτοποιήσουν πιθανές σχέσεις μεταξύ τους, οι οποίες είναι άγνωστες μέχρι σήμερα. Όμως βάσεις δεδομένων που περιείχαν αλληλουχίες υπάρχουν από τα τέλη της δεκαετίας του 1970, ενώ η GO ξεκίνησε το 1998 μετρώντας σήμερα 13 χρόνια ζωής. Επιπλέον η τελευταία υποστηρίχθηκε ενεργά από τις κοινότητες που μελετούσαν οργανισμούς – μοντέλα για την σύγχρονη βιολογική έρευνα όπως η *D. melanogaster, M. musculus, S. cerevisiae* με αποτέλεσμα από τα αρχικά στάδια της προσπάθειας να υπάρχει μια ικανή μάζα επισημειώσεων συνδεδεμένη με τα γονίδια των οργανισμών αυτών. Η δυνατότητα ηλεκτρονικής απόδοσης αντίστοιχων επισημειώσεων (αν και με μειωμένη αξιοπιστία) βάσει της ομοιότητας σε επίπεδο αλληλουχίας με άλλους οργανισμούς μεγάλωσε ακόμα περισσότερο τον όγκο των σχετικών διαθέσιμων επισημειώσεων. Τα αντίστοιχα δεδομένα στον χώρο των τροπικών νόσων είναι πολύ διαφορετικά. Δεν υπάρχουν αυτή την στιγμή οργανωμένες κεντρικές βάσεις δεδομένων που να αποθηκεύουν τα σχετικά δεδομένα. Αυτά βρίσκονται σε συρτάρια ερευνητικών εργαστηρίων και κρατικών υπηρεσιών δημόσιας υγείας, σε φακέλους νοσοκομείων και σε τόμους που βρίσκονται σε ράφια βιβλιοθηκών. Δεν υπάρχουν άλλες οντολογίες πέραν αυτών που δημιουργήθηκαν στα πλαίσια αυτής της εργασίας και όσων είναι συνδεδεμένες με αυτές για να περιγράψουν τα σχετικά δεδομένα. Είναι λοιπόν δεδομένο πως για να μπορέσει κανείς να φτάσει στα επίπεδα της επιτυχίας της GO, θα πρέπει να πείσει την κοινότητα να χρησιμοποιήσει τις νέες δυνατότητες που της παρέχονται.

Η διαδικασία δεν είναι απλή, γιατί ο τελικός χρήστης είναι μόνο έμμεσος χρήστης της οντολογίας. Εκείνο που βλέπει στην οθόνη του υπολογιστή του είναι ένα γραφικό περιβάλλον εργασίας και προσπαθεί να ανακτήσει όσο το δυνατόν περισσότερα σε όγκο και ανταποκρινόμενα στα ενδιαφέροντά του δεδομένα γίνεται. Την ίδια στιγμή όταν καταθέτει δεδομένα στις βάσεις δεδομένων θέλει αυτό να γίνει με τον μικρότερο δυνατό κόπο, αφιερώνοντας τον λιγότερο δυνατό χρόνο. Έτσι είναι πιθανό να αγνοεί εντελώς αν μια βάση δεδομένων χρησιμοποιεί στον πυρήνα της για την διασύνδεση των δεδομένων που περιλαμβάνει οντολογίες ή όχι. Στην πραγματικότητα οι βιοπληροφορικοί που δημιουργούν εργαλεία προς χρήση είναι εκείνοι οι οποίοι πρέπει να θεωρούνται σαν άμεσοι χρήστες μιας οντολογίας και ο στόχος είναι να αναπτύξουν με τέτοιο τρόπο τα εργαλεία τους ώστε να κάνουν χρήση οντολογιών. Τα πλεονεκτήματα των τελευταίων είναι τόσο προφανή που στις μέρες μας μεγάλα γενικά σχήματα βάσεων δεδομένων που προσφέρονται έτοιμα προς χρήση χωρίς να χρειάζεται κανείς να ξαναεφεύρει τον τροχό για να φτιάξει μια βάση δεδομένων, όπως το Chado (Zhou, Emmert et al. 2006; Mungall and Emmert 2007), βασίζονται σχεδόν αποκλειστικά σε οντολογίες για να διασυνδέσουν τα δεδομένα μεταξύ τους.

Παρόλα αυτά, οι οντολογίες της ανατομίας που δημιουργήθηκαν στο πλαίσιο αυτής της διατριβής σε συνδυασμό με τις ιστοσελίδες που δίνουν την δυνατότητα αναζήτησης για συγκεκριμένες δομές και παρουσιάζουν και φωτογραφίες ή σχηματικές απεικονίσεις τους μπορούν να χρησιμοποιηθούν σαν εκπαιδευτικό εργαλείο και εργαλείο αναφοράς απευθείας από τους τελικούς χρήστες. Παράλληλα παρέχουν τους απαραίτητους εκείνους όρους ώστε να μπορούν να επισημειωθούν αποτελέσματα πειραμάτων σε ανατομικές δομές κάτι ιδιαίτερα χρήσιμο στις μέρες μας όπου η επιστήμη της λειτουργικής γονιδιωματικής παράγει μεγάλους όγκους δεδομένων και πιο συγκεκριμένα η ανάλυση μικροσυστοιχιών DNA μπορεί να αναλύσει τη γονιδιακή έκφραση σε συγκεκριμένους ιστούς και ανατομικές

δομές. Αυτό που υπήρχε διαθέσιμο από την GO μέχρι σήμερα ήταν μόνο ο εντοπισμός συγκεκριμένων προϊόντων γονιδίων σε υποκυτταρικό επίπεδο. Επιπλέον, οι οντολογίες ανατομίας μπορούν να αποτελέσουν την αφετηρία για την περιγραφή και την κωδικοποίηση μορφολογικών χαρακτηριστικών και φαινοτύπων των κουνουπιών και να συντελέσουν και με αυτό τον τρόπο στον εμπλουτισμό των πληροφοριών στις υπάρχουσες βάσεις δεδομένων. Τέλος, σε έναν κόσμο όπου οι επισημειώσεις των γονιδίων που αφορούν τις ανατομικές δομές στις οποίες εκφράζονται είναι πολυπληθείς, θα μας δινόταν η δυνατότητα εντοπισμού πιθανών ορθόλογων γονιδίων μόνο και μόνο από το γεγονός της έκφρασής τους σε ανάλογες δομές.

Η MIRO και κατ' επέκταση η IRBase σαν βάση δεδομένων για την παρακολούθηση της ανθεκτικότητας σε εντομοκτόνα δημιουργεί μια νέα κατάσταση καθώς καθιστά δυνατή την ανταλλαγή δεδομένων μεταξύ διαφορετικών ερευνητικών προγραμμάτων, με την προϋπόθεση της υιοθέτησης της οντολογίας. Με σχετικά περιορισμένες μετατροπές ανάλογα εργαλεία μπορούν να αναπτυχθούν και να συμπεριλάβουν αγροτικού ενδιαφέροντος έντομα παρέχοντας έναν ενιαίο τρόπο καταγραφής και παρακολούθησης του φαινομένου διευκολύνοντας στη λήψη αποφάσεων αλλά και στον εντοπισμό περιοχών που χρειάζεται εντονότερη παρέμβαση.

Τέλος η IDOMAL είναι αυτή που μπορεί να συνεισφέρει πάρα πολλά στην διαχείριση και την ενοποίηση όλων των δεδομένων που υπάρχουν διαθέσιμα για την ελονοσία. Η υιοθέτησή της σαν πρότυπο από το τμήμα Αφρικής του Παγκόσμιου Οργανισμού Υγείας για την παρακολούθηση όχι μόνο της ανθεκτικότητας σε εντομοκτόνα αλλά και της περιγραφής των πληθυσμών των κουνουπιών, ανοίγει νέες προοπτικές στην δημιουργία ενός διεθνούς εργαλείου που θα επιτρέπει την συσχέτιση των σχετικών δεδομένων μεταξύ τους. Η χρήση της στο σύστημα υποστήριξης αποφάσεων τόσο της

ελονοσίας όσο και του Δάγκειου πυρετού δείχνει ακόμα περισσότερο την ανάγκη για περισσότερες οντολογίες που θα καλύψουν και άλλες ασθένειες που μεταδίδονται από αρθρόποδους φορείς, αλλά και το γεγονός πως οι οντολογίες αυτές μπορούν να συνδυασθούν μεταξύ τους φτιάχνοντας ίσως μια γενικότερη οντολογία αναφοράς για μολυσματικές ασθένειες που μεταδίδονται μέσω φορέων. Η τελευταία θα βρίσκεται μεταξύ της IDO και των οντολογιών των επιμέρους ασθενειών.

Ο όγκος των δεδομένων προς διαχείριση είναι τέτοιος που δεν επιτρέπει να τεθεί το ερώτημα αν οι οντολογίες είναι χρήσιμες στην σύγχρονη βιοπληροφορική ανάλυση. Το μόνο ερώτημα που τίθεται είναι στο πόσο γρήγορα θα μπορέσουν να αναπτυχθούν τα εργαλεία και οι οντολογίες που απαιτούνται και με ποιο τρόπο θα επισημειωθούν τα δεδομένα με τους κατάλληλους όρους ώστε να μπορούμε να εκμεταλλευτούμε τα πλεονεκτήματα που μας προσφέρουν. Αυτό δεν είναι κάτι που μπορεί να απαντηθεί με παρουσιάσεις μελλοντικών σχεδίων, αλλά να αποδειχθεί στην πράξη στα επόμενα χρόνια. Το γενικότερο ενδιαφέρον και ο αριθμός των οντολογιών που δημιουργούνται τα τελευταία χρόνια, επιτρέπουν να είμαστε αισιόδοξοι

ΒΙΒΛΙΟΓΡΑΦΙΑ

Aitken, S. (2005). "Formalizing concepts of species, sex and developmental stage in anatomical ontologies." Bioinformatics **21**(11): 2773-2779.

Akhouri, R. R., A. Sharma, et al. (2008). "Role of Plasmodium falciparum thrombospondin-related anonymous protein in host-cell interactions." Malaria journal **7**: 63.

Alphey, L. (2009). "Natural and engineered mosquito immunity." Journal of biology **8**(4): 40.

Arensburger, P., K. Megy, et al. (2010). "Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics." Science **330**(6000): 86-88.

Aronson, A. R., J. G. Mork, et al. (2004). "The NLM Indexing Initiative's Medical Text Indexer." Studies in health technology and informatics **107**(Pt 1): 268-272.

Artavanis-Tsakonas, K., J. E. Tongren, et al. (2003). "The war between the malaria parasite and the immune system: immunity, immunoregulation and immunopathology." Clinical and experimental immunology **133**(2): 145-152.

Ashburner, M., C. A. Ball, et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nature genetics **25**(1): 25-29.

Ashburner, M. and R. Drysdale (1994). "FlyBase--the Drosophila genetic database." Development **120**(7): 2077-2079.

Aulds, C. (2002). Linux Apache web server administration. San Francisco, Sybex.

Aurrecoechea, C., J. Brestelli, et al. (2009). "PlasmoDB: a functional genomic database for malaria parasites." Nucleic acids research **37**(Database issue): D539-543.

Aurrecoechea, C., J. Brestelli, et al. (2010). "EuPathDB: a portal to eukaryotic pathogen databases." Nucleic acids research **38**(Database issue): D415-419.

Baclawski, K. and T. Niu (2006). Ontologies for bioinformatics. Cambridge, Mass., MIT Press.

Bard, J. (2003). "Ontologies: Formalising biological knowledge for bioinformatics." BioEssays : news and reviews in molecular, cellular and developmental biology **25**(5): 501-506.

Blake, J. A., C. J. Bult, et al. (2011). "The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics." Nucleic acids research **39**(Database issue): D842-848.

Briet, O. J., G. N. Galappaththy, et al. (2005). "Maps of the Sri Lanka malaria situation preceding the tsunami and key aspects to be considered in the emergency phase and beyond." Malaria journal **4**: 8.

Brinkman, R. R., M. Courtot, et al. (2010). "Modeling biomedical experimental processes with OBI." Journal of biomedical semantics **1 Suppl 1**: S7.

Brown, A. W. (1958). "The insecticide-resistance problem: a review of developments in 1956 and 1957." Bulletin of the World Health Organization **18**(3): 309-321.

Burek, P., R. Hoehndorf, et al. (2006). "A top-level ontology of functions and its application in the Open Biomedical Ontologies." Bioinformatics **22**(14): e66-73.

Burger, A. G., D. Davidson, et al. (2008). Anatomy ontologies for bioinformatics : principles and practice. London, Springer.

Burgun, A. (2006). "Desiderata for domain reference ontologies in biomedicine." Journal of biomedical informatics **39**(3): 307-313.

Chareonviriyahpap, T., B. Aum-aung, et al. (1999). "Current insecticide resistance patterns in mosquito vectors in Thailand." The Southeast Asian journal of tropical medicine and public health **30**(1): 184-194.

Christie, K. R., S. Weng, et al. (2004). "Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms." Nucleic acids research **32**(Database issue): D311-314.

Courtot, M., F. Gibson, et al. (2009). MIREOT: the Minimum Information to Reference an External Ontology Term. The 1st International Conference on Biomedical Ontology (ICBO 2009). Buffalo, NY.

Curtis, C. F., Ed. (1991). Impregnated bed nets and curtains against malaria mosquitoes. Control of disease vectors in the community. London, Wolfe.

Curtis, C. F., C. A. Maxwell, et al. (2006). "Insecticide-treated bed-nets for malaria mosquito control." Journal of the American Mosquito Control Association **22**(3): 501-506.

Dabire, R. K., A. Diabate, et al. (2006). "Personal protection of long lasting insecticide-treated nets in areas of Anopheles gambiae s.s. resistance to pyrethroids." Malaria journal **5**: 12.

Dahdul, W. M., J. G. Lundberg, et al. (2010). "The teleost anatomy ontology: anatomical representation for the genomics age." Systematic biology **59**(4): 369-383.

Davidson, G. (1951). "Results of recent experiments on the use of DDT and BHC against adult mosquitos at Taveta, Kenya." Bulletin of the World Health Organization **4**(3): 329-332.

Day-Richter, J., M. A. Harris, et al. (2007). "OBO-Edit--an ontology editor for biologists." Bioinformatics **23**(16): 2198-2200.

de Clercq, P. A., A. Hasman, et al. (2001). "The application of ontologies and problem-solving methods for the development of shareable guidelines." Artificial intelligence in medicine **22**(1): 1-22.

de Matos, P., R. Alcantara, et al. (2010). "Chemical Entities of Biological Interest: an update." Nucleic acids research **38**(Database issue): D249-254.

de Zulueta, J. (1973). "Malaria eradication in Europe: the achievements and the difficulties ahead." The Journal of tropical medicine and hygiene **76**(11): 279-282.

Degtyarenko, K., P. de Matos, et al. (2008). "ChEBI: a database and ontology for chemical entities of biological interest." Nucleic acids research **36**(Database issue): D344-350.

Degtyarenko, K., J. Hastings, et al. (2009). "ChEBI: an open bioinformatics and cheminformatics resource." Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] **Chapter 14**: Unit 14 19.

della Torre, A., C. Costantini, et al. (2002). "Speciation within Anopheles gambiae--the glass is half full." Science **298**(5591): 115-117.

Dialynas, E., P. Topalis, et al. (2009). "MIRO and IRbase: IT tools for the epidemiological monitoring of insecticide resistance in mosquito disease vectors." PLoS neglected tropical diseases **3**(6): e465.

Dimopoulos, G. (2003). "Insect immunity and its implication in mosquito-malaria interactions." Cellular microbiology **5**(1): 3-14.

Dolo, A., D. Modiano, et al. (1999). "Thrombospondin related adhesive protein (TRAP), a potential malaria vaccine candidate." Parassitologia **41**(1-3): 425-428.

Egdahl, A. (1907). "Linnaeus' "Genera Morborum," and Some of His Other Medical Works." Medical library and historical journal **5**(3): 185-193.

Epstein, J. E., B. Giersing, et al. (2007). "Malaria vaccines: are we getting closer?" Current opinion in molecular therapeutics **9**(1): 12-24.

Flicek, P., M. R. Amode, et al. (2011). "Ensembl 2011." Nucleic acids research **39**(Database issue): D800-806.

Gaffigan, T., R. Wilkerson, et al. (1/1/2010). "Systematic Catalog of Culicidae." 2008, from http://www.mosquitocatalog.org/.

Galperin, M. Y. and G. R. Cochrane (2011). "The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection." Nucleic acids research **39**(Database issue): D1-6.

Gaudet, P., A. Bairoch, et al. (2011). "Towards BioDBcore: a community-defined information specification for biological databases." Nucleic acids research **39**(Database issue): D7-10.

Gene Ontology Consortium (2006). "The Gene Ontology (GO) project in 2006." Nucleic acids research **34**(Database issue): D322-326.

Gene Ontology Consortium (2010). "The Gene Ontology in 2010: extensions and refinements." Nucleic acids research **38**(Database issue): D331-335.

Grenon, P. and B. Smith (2004). "SNAP and SPAN: Towards Dynamic Spatial Ontology." Spatial Cognition & Computation **4**(1): 69 - 104.

Grenon, P., B. Smith, et al. (2004). "Biodynamic ontology: applying BFO in the biomedical domain." Studies in health technology and informatics **102**: 20-38.

Griswold, C. L. (2002). Platonic writings/Platonic readings. University Park, Pennsylvania State University Press.

Guarino, N. (1998). Formal ontology in information systems : proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy. Amsterdam ; Washington, DC

Tokyo, IOS Press ;

Omsha.

Haendel, M., G. V. Gkoutos, et al. (2009). "Uberon: towards a comprehensive multi-species anatomy ontology." Nature Proceedings.

Harbach, R. and K. Knight (1980). Taxonomist's Glossary of Mosquito Anatomy. Marlton, NJ, Plexus Pub.

Hill, C. A. and S. K. Wikel (2005). "The Ixodes scapularis Genome Project: an opportunity for advancing tick research." Trends in parasitology **21**(4): 151-153.

Holt, R. A., G. M. Subramanian, et al. (2002). "The genome sequence of the malaria mosquito Anopheles gambiae." Science **298**(5591): 129-149.

Huebner, E. (2007). "The Rhodnius Genome Project: The promises and challenges it affords in our understanding of reduviid biology and their role in Chagas' transmission." Comparative Biochemistry and Physiology a-Molecular & Integrative Physiology **148**: S130-S130.

Hviid, L. (2005). "Naturally acquired immunity to Plasmodium falciparum malaria in Africa." Acta Tropica **95**(3): 270-275.

Karp, P. D., M. Riley, et al. (1996). "EcoCyc: an encyclopedia of Escherichia coli genes and metabolism." Nucleic acids research **24**(1): 32-39.

Keiding, J. (1963). "Possible reversal of resistance." Bulletin of the World Health Organization **29 Suppl**: 51-62.

Kulkarni, M. A., R. Malima, et al. (2007). "Efficacy of pyrethroid-treated nets against malaria vectors and nuisance-biting mosquitoes in Tanzania in areas with long-term insecticide-treated net use." Tropical medicine & international health : TM & IH **12**(9): 1061-1073.

Lawson, D., P. Arensburger, et al. (2007). "VectorBase: a home for invertebrate vectors of human pathogens." Nucleic acids research **35**(Database issue): D503-505.

Lawson, D., P. Arensburger, et al. (2009). "VectorBase: a data resource for invertebrate vector genomics." Nucleic acids research **37**(Database issue): D583-587.

Livadas, G. A. and G. Georgopoulos (1953). "Development of resistance to DDT by Anopheles sacharovi in Greece." Bulletin of the World Health Organization **8**(4): 497-511.

Mary, V., G. Marquet, et al. (2004). "MeSH and specialized terminologies: coverage in the field of molecular biology." Studies in health technology and informatics **107**(Pt 1): 530-534.

Megy, K., S. J. Emrich, et al. (2011). "VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics." Nucleic acids research.

Morahan, B. J., L. Wang, et al. (2009). "No TRAP, no invasion." Trends in parasitology **25**(2): 77-84.

Mungall, C. J. and D. B. Emmert (2007). "A Chado case study: an ontology-based modular schema for representing genome-associated biological information." Bioinformatics **23**(13): i337-346.

Nelson, S. J., M. Schopen, et al. (2004). "The MeSH translation maintenance system: structure, interface design, and implementation." Studies in health technology and informatics **107**(Pt 1): 67-69.

Nene, V., J. R. Wortman, et al. (2007). "Genome sequence of Aedes aegypti, a major arbovirus vector." Science **316**(5832): 1718-1723.

Noy, N. F., N. H. Shah, et al. (2009). "BioPortal: ontologies and integrated data resources at the click of a mouse." Nucleic acids research **37**(Web Server issue): W170-173.

Osborne, J. D., J. Flatow, et al. (2009). "Annotating the human genome with Disease Ontology." BMC genomics **10 Suppl 1**: S6.

Pettifer, S., J. Ison, et al. (2010). "The EMBRACE web service collection." Nucleic acids research **38**(Web Server issue): W683-688.

Pittendrigh, B. R., J. M. Clark, et al. (2006). "Sequencing of a new target genome: the Pediculus humanus humanus (Phthiraptera: Pediculidae) genome project." Journal of medical entomology **43**(6): 1103-1111.

Raghupathi, W. and A. Umar (2011). "Upper-level Ontologies for Health Information Systems. Towards an Archetype Patterns Approach." Methods of information in medicine **50**(2).

Rastan, S. and L. J. Beeley (1997). "Functional genomics: going forwards from the databases." Current opinion in genetics & development **7**(6): 777-783.

Rector, A. L. (1998). "Thesauri and formal classifications: terminologies for people and machines." Methods of information in medicine **37**(4-5): 501-509.

Robson, K. J., J. R. Hall, et al. (1988). "A highly conserved amino-acid sequence in thrombospondin, properdin and in proteins from sporozoites and blood stages of a human malaria parasite." Nature **335**(6185): 79-82.

Rosse, C., A. Kumar, et al. (2005). "A strategy for improving and integrating biomedical ontologies." AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium: 639-643.

Rosse, C. and J. L. Mejino, Jr. (2003). "A reference ontology for biomedical informatics: the Foundational Model of Anatomy." Journal of biomedical informatics **36**(6): 478-500.

Ruch, P., J. Gobeill, et al. (2008). "Automatic medical encoding with SNOMED categories." BMC medical informatics and decision making **8 Suppl 1**: S6.

Serban, R. and A. ten Teije (2009). "Exploiting thesauri knowledge in medical guideline formalization." Methods of information in medicine **48**(5): 468-474.

Siever, E., S. Spainhour, et al. (2004). The Perl CD bookshelf. Sebastopol, CA, O'Reilly,**:** 1 CD-ROM.

Sigrist, C. J., L. Cerutti, et al. (2010). "PROSITE, a protein domain database for functional characterization and annotation." Nucleic acids research **38**(Database issue): D161-166.

Simon, J., M. Dos Santos, et al. (2006). "Formal ontology for natural language processing and the integration of biomedical databases." International journal of medical informatics **75**(3-4): 224-231.

Sintchenko, V. (2010). Infectious disease informatics. New York, Springer.

Smith, B., M. Ashburner, et al. (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." Nature biotechnology **25**(11): 1251-1255.

Smith, B., W. Ceusters, et al. (2005). "Relations in biomedical ontologies." Genome biology **6**(5): R46.

Smith, B. and P. Grenon (2004). "The Cornucopia of Formal-Ontological Relations." Dialectica **58**: 17.

Smith, B. and C. Rosse (2004). "The role of foundational relations in the alignment of biomedical ontologies." Studies in health technology and informatics **107**(Pt 1): 444-448.

Soldatova, L. N. and R. D. King (2006). "An ontology of scientific experiments." Journal of the Royal Society, Interface / the Royal Society **3**(11): 795-803.

Sonenshine, D. E. (1991). Biology of ticks. New York, Oxford University Press.

Spaccapelo, R., S. Naitza, et al. (1997). "Thrombospondin-related adhesive protein (TRAP) of Plasmodium berghei and parasite motility." Lancet **350**(9074): 335.

Srivastava, A. K. and N. S. Sahni (2011). "OntoVisT: A general purpose Ontological Visualization Tool." Bioinformation **6**(7): 288-290.

Sultan, A. A., V. Thathy, et al. (1997). "TRAP is necessary for gliding motility and infectivity of plasmodium sporozoites." Cell **90**(3): 511-522.

Topalis, P., A. Koutsos, et al. (2005). "AnoBase: a genetic and biological database of anophelines." Insect molecular biology **14**(6): 591-597.

Topalis, P., D. Lawson, et al. (2008). "How can ontologies help vector biology?" Trends in parasitology **24**(6): 249-252.

Topalis, P., E. Mitraka, et al. (2010). "IDOMAL: an ontology for malaria." Malaria journal **9**: 230.

Topalis, P., C. Tzavlaki, et al. (2008). "Anatomical ontologies of mosquitoes and ticks, and their web browsers in VectorBase." Insect molecular biology **17**(1): 87-89.

Vaswani, V. (2010). MySQL database usage & administration. New York, McGraw-Hill.

Vine, J. M. (1947). "Malaria Control with D.D.T. on a National Scale-Greece, 1946 [Abridged]." Proceedings of the Royal Society of Medicine **40**(13): 841-848.

Welling, L. and L. Thomson (2008). PHP and MySQL Web development. Upper Saddle River, NJ, Addison-Wesley.

Whetzel, P. L., H. Parkinson, et al. (2006). "The MGED Ontology: a resource for semantics-based description of microarray experiments." Bioinformatics **22**(7): 866-873.

WHO Expert Committee on Vector Biology and Control. (1992). Vector resistance to pesticides : fifteenth report of the WHO Expert Committee on Vector Biology and Control. Geneva, World Health Organization.

Wright, J. W., R. F. Fritz, et al. (1972). "Changing concepts of vector control in malaria eradication." Annual review of entomology **17**: 75-102.

Yoder, M. J., I. Miko, et al. (2010). "A gross anatomy ontology for hymenoptera." PloS one **5**(12): e15991.

Zhou, P., D. Emmert, et al. (2006). "Using Chado to store genome annotation data." Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.] **Chapter 9**: Unit 9 6.

ΠΑΡΑΡΤΗΜΑ Α' : ΚΑΤΑΛΟΓΟΣ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ

Συντομογραφία	**Περιγραφή**
BFO	Basic Formal Ontology
CARO	Common Anatomy Reference Ontology
ChEBI	Chemical Entities of Biological Interest
CL	Cell Line Ontology
CoE	Carboxyesterase
DDT	Dichlorodiphenyltrichloroethane
DEF	S,S,S-Tributyl phosphorotrithioate
DO	Disease Ontology
ENVO	Environmental Ontology
FMA	Foundation Model of Anatomy
GABA	g-aminobouteric acid
GAZ	Gazetteer
GFO	Generic Formal Ontology
GO	Gene Ontology
GST	Glutathion-S-transferase
HAO	Hymenoptera Anatomy Ontology
IDO	Infectious Disease Ontology
IDOMAL	Infectious Disease Ontology – Malaria
IRAC	Insecticide Resistance Action Committee
IVCC	Innovative Vector Control Consortium
MDSS	Malaria Decision Support System
MESH	Medical Subject Headings
MIREOT	Minimal Information to Reference External Ontology Terms
MIRO	Mosquito Insecticide Resistance Ontology
NCBI	National Center
NCBO	National Center of Biomedical Ontologies
NLM	National Library of Medicine
OBI	Ontology of Biomedical Investigations
OBO	Open Biomedical Ontologies
OWL	Ontology Web Language
PERL	Practical Extraction and Report Language

PHP	PHP: Hypertext Preprocessor
SNAP	SNAP(shot) the continuant part of BFO
SNOMED	Systematized Nomenclature of Human Medicine
SPAN	The occurent part of BFO
SUMO	Suggested Upper Level Merged Ontology
TADS	Tick Anatomy Daniel Sonenshine
TAO	Teleost Anatomy Ontology
TGMA	Taxonomist's Glossary of Mosquito Anatomy
TRAP	Thrombospondin Anonymous Protein
UBERON	Uber Ontology
VBcv	VectorBase controlled vocabulary
WHO / Π.Ο.Υ.	World Health Organization / Παγκόσμιος Οργανισμός Υγείας

ΓΛΩΣΣΑΡΙ

Ασυνεχής οντότητα ή διαδικασία (occurent or processual entity): Οντότητα που εξελίσσεται σε φάσεις όπου η μια διαδέχεται την άλλη.

Αυθύπαρκτη οντότητα (independent entity): Συνεχής οντότητα που μπορεί να υπάρξει ανεξάρτητα. Για παράδειγμα ένας οργανισμός.

Διασταυρούμενη αναφορά (cross reference, x-ref): Αναφορά σε ένα κείμενο, μια οντολογία ή μια βάση δεδομένων που παραπέμπει σε σχετική πληροφορία σε διαφορετικό κείμενο, οντολογία ή βάση δεδομένων.

Ελεγχόμενο λεξιλόγιο (controlled vocabulary): Τρόπος οργάνωσης μετα-δεδομένων για την πιο εύκολη ανάκτησή τους. Περιλαμβάνει μια προεπιλεγμένη λίστα όρων από τον δημιουργό του, χωρίς όμως να περιέχει ορισμούς των όρων ή συσχετίσεις μεταξύ τους.

Εξαρτημένη οντότητα (dependent entity): Συνεχής ή ασυνεχής οντότητα η οποία δεν μπορεί να υπάρξει από μόνη της αλλά αναφέρεται σε κάποια αυθύπαρκτη οντότητα. Για παράδειγμα ο φαινότυπος ενός οργανισμού.

Εξυπηρετητής (server): Πρόγραμμα ηλεκτρονικού υπολογιστή που επιτρέπει την διασύνδεση απομακρυσμένων χρηστών και την ανάκτηση πληροφοριών από αυτούς.

Επισημείωση (annotation): Οποιαδήποτε πρόσθετη πληροφορία συνδέεται με κάποιο δεδομένο καθώς και η διαδικασία σύνδεσής της. Για παράδειγμα ο ορισμός της κωδικής περιοχής σε μια αλληλουχία DNA, η προσθήκη χαρακτηριστικών στην περιγραφή ενός πληθυσμού κουνουπιών αποτελούν επισημειώσεις.

Θησαυρός (thesaurus): Τρόπος οργάνωσης μετα-δεδομένων. Περιλαμβάνει θεματικές λίστες όρων που είναι ορισμένοι με σαφήνεια καθώς και διασταυρούμενες αναφορές με άλλα

μέρη που περιέχουν ανάλογες πληροφορίες. Οι όροι δεν είναι συσχετισμένοι μεταξύ τους με λογικές σχέσεις.

Μετα-δεδομένα (meta-data): Δεδομένα τα οποία χρησιμοποιούν για την περιγραφή άλλων δεδομένων.

Οντολογία (ontology): Τρόπος οργάνωσης μετα-δεδομένων. Περιλαμβάνει όρους που είναι πλήρως ορισμένοι και αποσαφηνισμένοι καθώς και τα συνώνυμά τους. Οι όροι είναι συσχετισμένοι μεταξύ τους με λογικές σχέσεις ορισμένες με σαφήνεια και αν είναι δυνατόν με τυπικό τρόπο. Οι συνώνυμοι όροι είναι επίσης συσχετισμένοι με τον κύριο όρο μέσω λογικών σχέσεων. Η Οντολογία σαν επιστήμη είναι παρακλάδι των μεταφυσικών που είναι κλάδος της φιλοσοφίας.

Οντολογία αναφοράς (reference ontology): Μια οντολογία που έχει σκοπό να καλύψει ένα συγκεκριμένο θεματικό πεδίο. Διαθέτει μεγάλη έκταση, αλλά μικρό βάθος και αποτελεί τον σύνδεσμο ανάμεσα στις οντολογίες ανώτερου επιπέδου και τις οντολογίες εφαρμογών.

Οντολογία ανώτερου επιπέδου (upper level ontology): Ξεχωριστή φιλοσοφική θεώρηση για την οντολογική ταξινόμηση των πραγμάτων χωρίς ειδικές αναφορές σε επιμέρους τομείς ή επιστήμες.

Οντολογία εφαρμογών (application ontology): Οντολογίες που δημιουργούνται για να καλύψουν τις ανάγκες μιας συγκεκριμένης κοινότητας προσφέροντας όρους που μπορούν να χρησιμοποιηθούν και την περιγραφή και την επισημείωση των αποτελεσμάτων που υπάρχουν και για να αποτελέσουν τον πρότυπο βάσει του οποίου θα σχεδιαστούν βάσεις δεδομένων.

Οντότητα (entity): Οτιδήποτε υπάρχει.

Περίπτωση (instance): Συγκεκριμένο μέλος μιας τάξης που φέρει την ιδιότητα που την χαρακτηρίζει.

Συνεχής οντότητα (continuant): Μια οντότητα που παραμένει σταθερή στον χρόνο για όσο υπάρχει.

Σύστημα υποστήριξης αποφάσεων (decision support system): Πληροφοριακό σύστημα που συνδυάζει δεδομένα από διαφορετικές πηγές και τα παρουσιάζει με ενιαίο τρόπο, ώστε να δίνει μια συνολική εικόνα και να διευκολύνει τους ειδικούς να λαμβάνουν αποφάσεις.

Σχήμα ταξινόμησης (classification scheme): Τρόπος οργάνωσης μετα-δεδομένων. Περιλαμβάνει λίστα θεματικών όρων που συνήθως δεν συνοδεύονται από ορισμούς.

Τάξη (class): Σύνολο ομάδων ή αντικειμένων που μπορεί να περιγραφεί πέρα πάσης αμφιβολίας μέσω μιας κοινής ιδιότητας όλων των μελών του.

Τυπική οντολογία (formal ontology): Οντολογία που χρησιμοποιεί αποκλειστικά σχέσεις που έχουν οριστεί με αυστηρό, λογικό τρόπο.

ΔΗΜΟΣΙΕΥΣΕΙΣ ΠΟΥ ΠΕΡΙΕΧΟΥΝ ΤΜΗΜΑΤΑ ΑΥΤΗΣ ΤΗΣ ΔΙΑΤΡΙΒΗΣ

Insect Molecular Biology (2005) **14**(6), 591–597

SHORT NOTE

AnoBase: a genetic and biological database of anophelines

P. Topalis*, A. Koutsos*†, E. Dialynas*, C. Kiamos*, L. K. Hope‡, C. Strode‡, J. Hemingway‡ and C. Louis*†

**Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas, Vassilika Vouton, Heraklion, Crete, Greece;* †*Department of Biology, University of Crete, Heraklion, Crete, Greece; and* ‡*Liverpool School of Tropical Medicine, Liverpool, UK*

Abstract

AnoBase (http://www.anobase.org) is an integrated, relational database of basic biological and genetic data on anopheline species, with a particular emphasis on *Anopheles gambiae*. It has been designed as an information source and research support tool for the broad vector biology community. Although AnoBase is not a primary genomic database that develops and provides tools to access the genome of the malaria mosquito, it nevertheless contains several sections that offer data of genomic interest such as *in situ* hybridization images, an integrated gene tool and direct online access to AnoXcel, the proteomic database of *An. gambiae*. Moreover, AnoBase also contains information on non-*gambiae* mosquito species and a novel section on studies related to insecticide resistance.

Keywords: *Anopheles gambiae*, bioinformatics, malaria, mosquito, genomics.

Introduction

The 'revolution' in biological research, initiated with the development of recombinant DNA technology in the mid-seventies, would not have been possible if it had not been supported by equivalent progress in the field of informatics, on the levels of both software and hardware. The beginnings of bioinformatics, i.e. the branch of computer sciences that deals with the storage, processing, analysis and mining of biological data can be traced back to the early eighties, when *Nucleic Acids Research* dedicated its entire issue of January 1984 to the description of a variety of software mostly related to nucleic acid analysis; the issue included, among others, a paper describing the *GCG* package of DNA sequence analysis (Devereux *et al*., 1984). It is this seminal paper, alongside that of Altschul *et al*. (1990), describing the BLAST analysis tools, now the fifth and third most cited 'biochemical' papers of the eighties and nineties, respectively, that best demonstrates the triumphant advance of bioinformatics.

The development of even more sophisticated software and the increase of the speed at which data, and most prominently, nucleic acid sequences can be acquired, have driven the requirement for databases that store these data and at the same time make them available for analysis and mining. The simple storing of sequences is not enough to assist researchers in the analysis of the immense wealth of information available to the end-user in the era of genome sequencing. Additional bioinformatics resources, such as those dealing with the structure and function of proteins and the analysis of gene expression, are being constantly added to the pool of data available. This huge mass of information makes it necessary to integrate databases to a point where they become manageable and easily accessible to the general research community.

FlyBase was one of the first attempts to develop such an integrated database, centred on the fruit fly *Drosophila melanogaster* (Ashburner & Drysdale, 1994; The FlyBase Consortium, 2003; Drysdale *et al*., 2005). From its inception, the FlyBase Consortium concentrated on making available to the research community the enormous genetic heritage of this model organism, primarily based on the curation of the vast *Drosophila* literature. FlyBase later joined efforts with the Berkeley *Drosophila* Genome Project in a collaboration that was crucial in the initial, and the later updated, annotation of the *D. melanogaster* genome

doi: 10.1111/j.1365-2583.2005.00596.x

Received 30 March 2005; accepted after revision 24 June 2005. Correspondence: Christos Louis, IMBB-FORTH, Vassilika Vouton, PO Box 1527, 711 10 Heraklion, Crete, Greece. Tel.: +30 281 0391119; fax: +30 281 0391104; e-mail: louis@imbb.forth.gr

(Adams *et al*., 2000; Celniker & Rubin, 2003) as well as the full integration of the related data in the fruit fly database.

Facilitated by the close links between the *Drosophila* and the *Anopheles* research community, FlyBase became the model for the development of an integrated *Anopheles* database. The decision to initiate the development of such a resource was reached in 1995, and AnoDB, the first product became accessible in early 1996. Although the dual format of the database, flat text files and AceDB (Eeckman & Durbin, 1995), could handle the data available then, it was decided to move AnoDB to a relational format to better handle the wealth of information that was to be available after the anticipated *Anopheles* genome sequencing project. Here, we describe AnoBase, the successor database, and discuss its significance in the frame of mosquito-related research.

Results and discussion

AnoDB migrated to the current format in 2002 and was renamed AnoBase to reflect this change; the move coincided with the publication of the whole genome shotgun sequence (WGS) of *An. gambiae* (Holt *et al*., 2002). The new format is based on a relational database management system implemented on an upgraded server and fully utilizing the MySQL database server engine. The vast majority of the data previously stored in AnoDB were retained in AnoBase, and additional datasets and services were incorporated.

The in situ *hybridization section*

In situ hybridization data and images have been stored in the *Anopheles* database since the launch of AnoDB. Although the completion of the WGS of *An. gambiae* made the availability of these images less important (they were ideal milestones for positional cloning strategies), their value is still substantial, especially considering the existence of a large number of natural inversions in *An. gambiae* (Coluzzi *et al*., 1979) and the importance of these chromosomal rearrangements for the understanding of the biology, including the evolution and taxonomy of the main African malaria vector (Coluzzi *et al*., 2002). At present AnoBase holds 1375 *in situ* hybridization images and this number could double if all available data worldwide are submitted for inclusion in AnoBase. To date, most of these images are derived from the mapping of sequenced ends of BAC clones (Hong *et al*., 2003) and are housed within the cytogenetics section of AnoBase, alongside the polytene maps of M. Coluzzi and his collaborators. In addition to the actual *in situ* hybridization image, every entry makes available a high-resolution photographic representation of the cytogenetic segments around the hybridization signal and a drawn map of the area, and links to the actual sequences are presented, providing a direct linkage to the WGS.

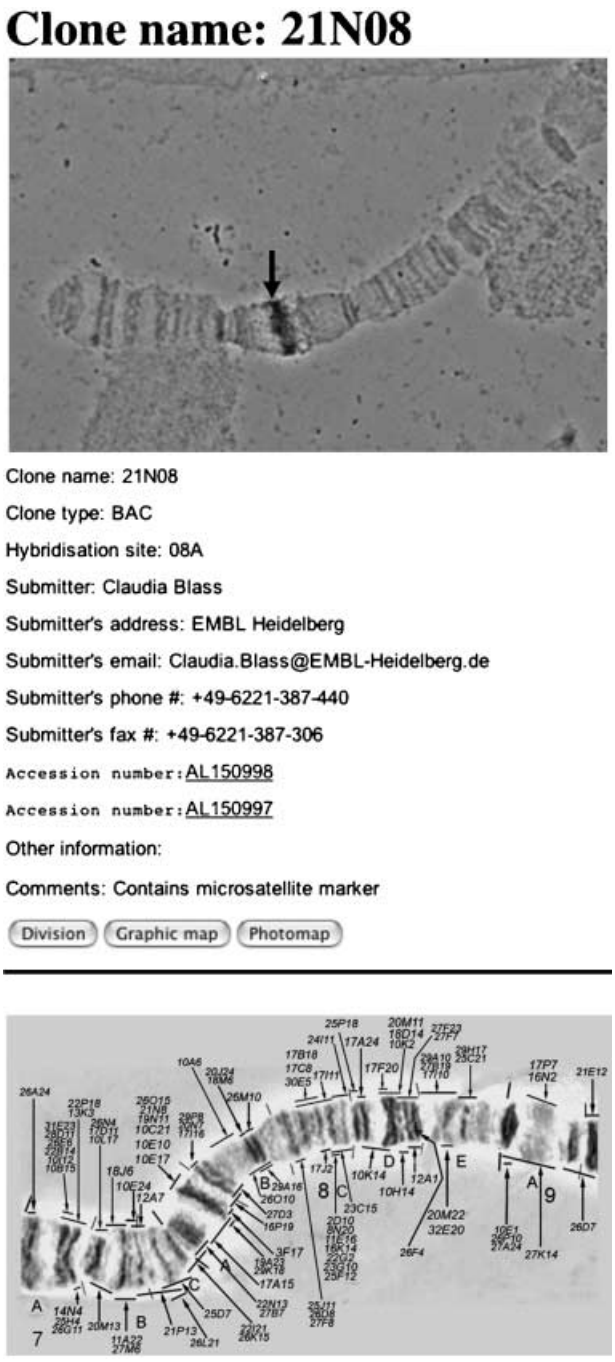


Figure 1. Output of an *in situ* hybridization entry. The upper part of the figure shows the actual image (an arrow points to the hybridization site, in this case 8A) as well as the additional details of the entry; links to the sequenced ends of the BAC clone (21N08) are provided. The lower part of the figure shows the image retrieved when clicking the 'Division' button. The hybridization sites of other clones for the particular chromosomal division are schematically shown on a photographic map.

Figure 1 shows an example of such an entry. The BAC clone 21N08 has been shown to hybridize to the cytogenetic location 8A. The upper part of the Figure shows the information stored for each image entry. When clicking on

the 'Division' button, the information shown in the bottom appears. Here, a photographic map of divisions 7 and 8 (from left to right) appears, including all information available for that division, i.e. clones found to hybridize *in situ* and extent of the hybridization signal.

Genes

The information gap on mosquito genes was recently filled through the acquisition of the WGS, which theoretically provided a complete list of the *An. gambiae* genes. However, to be precise, genes discovered by a WGS should not be called genes *sensu stricto* but, rather putative genes or gene models, as their identity only relies on analysis performed with bio-computing tools. In the particular case of *An. gambiae* the majority of the annotation of genes has been automatic, with very few gene families based on manual annotation performed by scientists outside the genome project. It should therefore be stressed that the actual genes of the mosquito may include additional genes not previously identified (see for example Gomez *et al*., 2005); or individual genes may be joined, or split or even discarded later at repeated annotation updates. All these automatically annotated genes and the related information can be viewed in the mosquito pages of the ENSEMBL database (http://www.ensembl.org/Anopheles_gambiae), which is responsible for the automatic annotation of the *An. gambiae* genome and its updating, as well as the presentation of the data. For a scientist skilled in genomics the search tool used to query ENSEMBL is easy and the information available there is comprehensive and accessible without difficulty. However, data accessibility for nonmolecular entomologists, who on a practical level are relatively distant from genomic analysis, may still be perceived as difficult. For example, ENSEMBL, correctly, differentiates between genes, transcripts and (encoded) proteins, a fact that sometimes leads to a slight confusion of uninitiated users. In addition, information can be cumbersome to retrieve as ENSEMBL identifies a gene by an ENSEMBL ID that is characterized by the prefix ENSENGG … followed by a series of twelve digits that include six to seven zeroes.

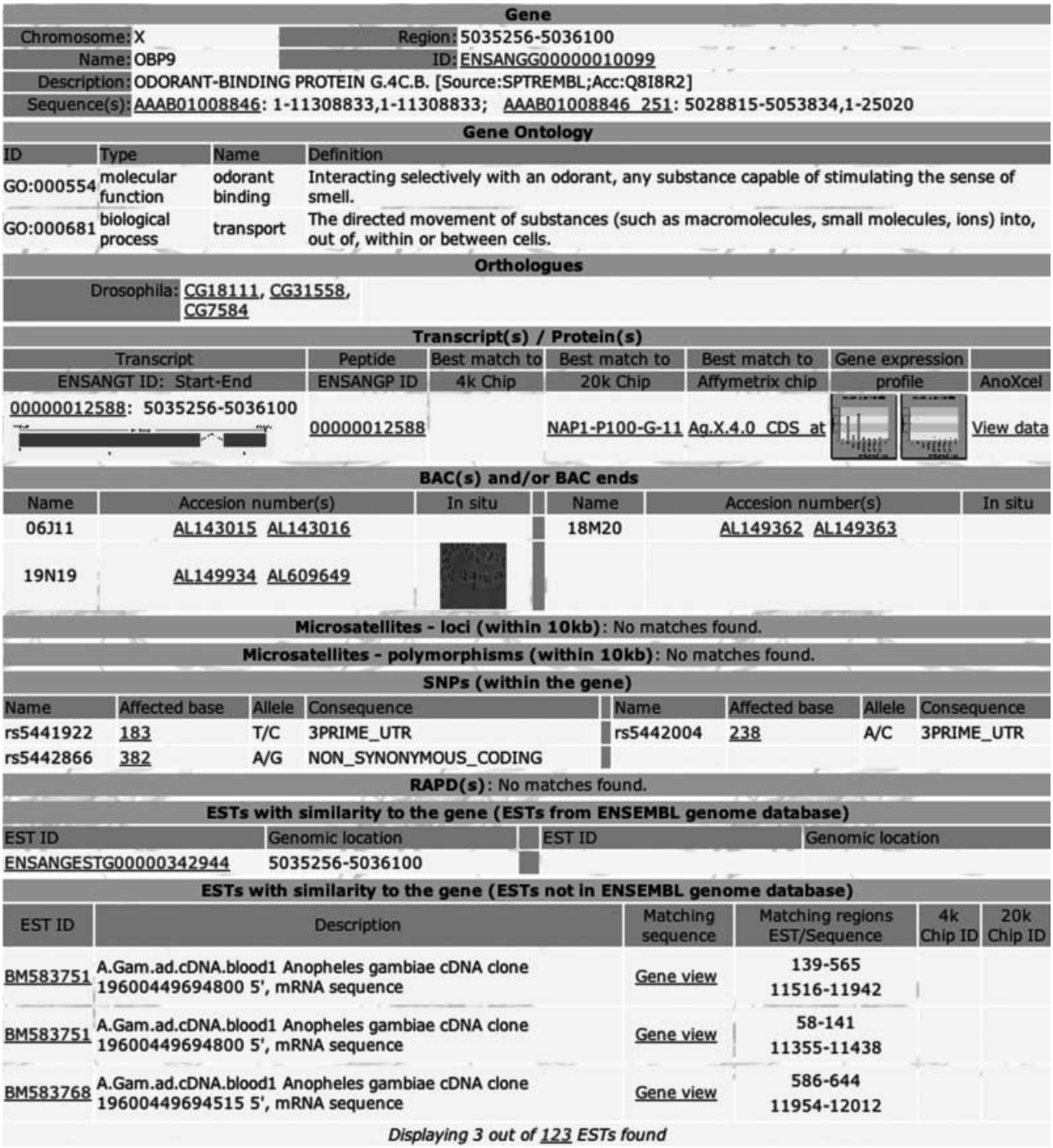
Planning, initially, to help the nongenomics-orientated scientists, we compiled the information available at ENSEMBL in a slightly different way and also incorporated additional data from AnoBase and other resources. With our present gene tool, the user can select the database or data type to be searched using a drop-down menu: gene names (name adopted by the International *Anopheles gambiae* Genome Consortium), gene functions as used in either the gene name submissions or the ENSEMBL genome data, ENSANGXn ID numbers (both protein and gene), Gene Ontology (GO) terms (The Gene Ontology Consortium, 2000), and AnoXcel (Ribeiro *et al*., 2004) data. The output of the searches contains data on additional potential items of interest related to the gene of interest. Figure 2 shows the example of the output of a search using the gene name *OBP9* (for odorant binding protein 9). In addition to GO terms where available and cross-references (e.g. FlyBase, AnoXcel, etc.), the output includes all microsatellites and random amplified polymorhpic DNAs (RAPDs) that have been mapped within 10 kb on either side of the gene queried (none in this particular case), and single nucleotide polymorphisms (SNPs) along the gene's sequence (three SNPs detected here). Moreover, if the gene searched has been included in the GeneChip® *Plasmodium/Anopheles* Genome Array (Plasmodium_datasheet.pdf, downloadable from http://www.affymetrix.com/products/arrays/specific/plasmodium.affx) and/or the '4K' and '20K' microarray chips (annotation available at: http://komar.embl-heidelberg.de/), the corresponding accession numbers are also indicated, so that the respective annotation can be viewed. In addition, where the gene of interest has been included in the *Plasmodium/Anopheles* Genome Array, the results of a genome-wide developmental profiling experiment performed by Marinotti *et al*. (2005) are also displayed in a graphical form by clicking on the thumbnails provided; the graphs are then linked to the 'mother' database at the University of California at Irvine (http://www.angagepuci.bio.uci.edu/), in order to access the full results of the analysis. The outputs of the searches also include lists of all available *An. gambiae* ESTs displaying high similarity scores in BLAST searches, along with the coordinates of the similar regions. Finally, if a gene has been mapped within a BAC whose ends have been sequenced, this is also indicated; corresponding links to *in situ* hybridization data are provided when available. Here, a thumbnail along BAC 19N19 indicates that an *in situ* hybridization image is also available for the BAC. It is planned to include in the future additional data, such as results from other microarray profiling experiments, links to pertinent literature, images, etc., as these become available. It should be stressed that the gene tool in AnoBase is not an alternative view of the data in ENSEMBL but, rather, an individual device that acquires and displays all available information relating to a given gene.

AnoXcel

AnoXcel, a proteomic database for *An. gambiae*, was presented in a previous report (Ribeiro *et al*., 2004). The data presented are updated regularly and, most importantly, the dataset is generated *de novo* with each major update of the ENSEMBL mosquito database (Mongin *et al*., 2004). Links to the AnoXcel database in AnoBase are provided in different sections such as from within any entry accessed through the Genes section.

Insecticide resistance

A survey of entomologists in disease-endemic countries conducted in cooperation with the World Health Organization Special Programme for Research and Training in Tropical



Gene

Chromosome:	X	Region:	5035256-5036100
Name:	OBP9	ID:	ENSANGG00000010099
Description:	ODORANT-BINDING PROTEIN G.4C.B. [Source:SPTREMBL;Acc:Q8I8R2]		
Sequence(s):	AAAB01008846: 1-11308833,1-11308833; AAAB01008846_251: 5028815-5053834,1-25020		

Gene Ontology

ID	Type	Name	Definition
GO:000554	molecular function	odorant binding	Interacting selectively with an odorant, any substance capable of stimulating the sense of smell.
GO:000681	biological process	transport	The directed movement of substances (such as macromolecules, small molecules, ions) into, out of, within or between cells.

Orthologues

Drosophila:	CG18111, CG31558, CG7584

Transcript(s) / Protein(s)

Transcript	Peptide	Best match to	Best match to	Best match to	Gene expression	
ENSANGT ID: Start-End	ENSANGP ID	4k Chip	20k Chip	Affymetrix chip	profile	AnoXcel
00000012588: 5035256-5036100	00000012588		NAP1-P100-G-11	Ag.X.4.0_CDS_at		View data

BAC(s) and/or BAC ends

Name	Accesion number(s)	In situ	Name	Accesion number(s)	In situ
06J11	AL143015 AL143016		18M20	AL149362 AL149363	
19N19	AL149934 AL609649				

Microsatellites - loci (within 10kb): No matches found.

Microsatellites - polymorphisms (within 10kb): No matches found.

SNPs (within the gene)

Name	Affected base	Allele	Consequence	Name	Affected base	Allele	Consequence
rs5441922	183	T/C	3PRIME_UTR	rs5442004	238	A/C	3PRIME_UTR
rs5442866	382	A/G	NON_SYNONYMOUS_CODING				

RAPD(s): No matches found.

ESTs with similarity to the gene (ESTs from ENSEMBL genome database)

EST ID	Genomic location	EST ID	Genomic location
ENSANGESTG00000342944	5035256-5036100		

ESTs with similarity to the gene (ESTs not in ENSEMBL genome database)

EST ID	Description	Matching sequence	Matching regions EST/Sequence	4k Chip ID	20k Chip ID
BM583751	A.Gam.ad.cDNA.blood1 Anopheles gambiae cDNA clone 19600449694800 5', mRNA sequence	Gene view	139-565 11516-11942		
BM583751	A.Gam.ad.cDNA.blood1 Anopheles gambiae cDNA clone 19600449694800 5', mRNA sequence	Gene view	58-141 11355-11438		
BM583768	A.Gam.ad.cDNA.blood1 Anopheles gambiae cDNA clone 19600449694515 5', mRNA sequence	Gene view	586-644 11954-12012		

Displaying 3 out of 123 ESTs found

Figure 2. Output of a search using the gene tool with the term 'OBP9'. The annotations of the gene in three different *Anopheles gambiae* microarray chips are provided (here, the specific DNA is not included in the '4K' chip). Gene Ontology (GO) terms are derived directly from the latest version of the GO database. Orthologues, where offered, are derived from ENSEMBL. A thumbnail of the transcript map of the gene is provided. All thumbnails in the output are clickable, retrieving the respective information. Molecular genetic markers are listed, where available, with their genomic coordinates.

Diseases (TDR) identified the inclusion of data on insecticide resistance as a need of the community that should be urgently addressed. For this, we developed data inclusion proformas that could be used in order to curate the results of studies on insecticide resistance conducted earlier and published in refereed journals. It should be stressed that the aim of this section is not to include *in toto* the published data but, rather, indicate the studies based on the region(s)

Table 1. Population information proforma used by AnoBase

Proforma code	Description
PO1	Main species used in paper (including multiple) [CV]
PO2	Additional species encountered in paper
PO3	Continent of study [Africa, Europe, etc.]
PO4	Country of study
PO5	Zone or region of study [Eastern, dry, savanna, forest, mangrove swamp]
PO6	City or village of study
PO7	Longitude of the collecting site
P08	Latitude of the collecting site
P09	Altitude of the collecting site
PO10	Year of studies (including multiple years)
PO11	Time frame of study
PO12	Season of study
PO13	Time of day of collection
PO14	Temperature of area
PO15	Relative humidity
PO16	Rainfall average
PO17	Salt concentration of water
PO18	Relative pH of water
PO19	Species identification [morphological, cytological, etc.]
PO20	Sex of mosquitoes [male, female, both, unspecified]
PO21	Physiological stage [larvae, pupae, adults, preimarginal states]
PO22	Activity, e.g. resting adults, endo- & exophilic, endo- & exophagy
PO23	Food state of mosquitoes [teneral, post-teneral, unfed, blooded, gravid]
PO24	Laboratory strain established?
PO25	Name of lab strain
PO25a	If yes, year of establishment
PO25b	If yes, geographical origin (if different from above)
PO26	Mode of collection [indoor, outdoor, human, animal, pyrethrum, traps]
PO27	No of mosquitoes collected
PO28	Have mosquitoes been banked?
PO29	Accession no or names of mosquito strains
PO30	Type of markers used [allozymes, microsatellites, RFLPs, RAPDs, SSCP, nucleotide sequences non-coding, -coding, TEs, RNA, DNA-DNA hybridization, morphology, chromosome (rearrangements)
PO31	Names of alleles of markers used
PO32	Type of statistical analysis performed
PO33	Relevant mosquito densities reported
PO34	Relevant heterozygote frequencies expected for markers used
PO35	Relevant heterozygote frequencies observed for markers used
PO36	Correlation of gene variation with distance variation?
PO37	Significance levels of statistical studies
PO38	Associated data present with current study
PO39	Comments – population of insect
PO40	Comments – comparison with other insect populations (competition with other species for resources)
PO41	Comments – use of different marker methods
PO42	Comments – general comments
PO43	Internal notes

[CV] indicates that special controlled vocabularies have been developed to handle the corresponding data.

Table 2. Insecticide resistance proforma used by AnoBase

Proforma code	Description
IR1	Species used [req] [CV]
IR2	Insecticide used (chemical name) [req] [CV]
IR3	Insecticide's trade name [CV]
IR4	OMS number [CV]
IR5	Insecticide formulation [CV]
IR16	Insecticide class [CV]
IR19	Insecticide CAS number [CV]
IR6	Insecticide concentration used
IR7	Study is referring to field population
IR8a	Study is referring to colony bred in lab
IR8b	Age of mosquito [one day old, mixed]
IR9	Type of selection pressure [larvicide, adulticide, unknown]
IR10	Percentage of mortality
IR11	LD50 value or resistance ratio
IR17	Knock down time 50 [KDT50]
IR18	Knock down time 95 [KDT95]
IR12	Type of resistance gene [altered AChE, kdr, esterase, P450, GST, GABA, other]
IR13	Resistance gene frequency
IR14	Comments – general comments
IR15	Internal notes

[CV] indicates that special controlled vocabularies have been developed to handle the corresponding data.
OMS, insecticide database (WHO).

performed (at the level of countries) and the insecticides tested. The data types that will be included by the curators are shown in Tables 1 and 2. Two different proformas, each dealing with parts of the aspects (i.e. Proforma 1: information on the actual populations tested and Proforma 2: the insecticide parameters) are used. Moreover, wherever [CV] is indicated, special controlled vocabularies have been developed to handle the corresponding data. Finally, a web interface is now being developed that will enable a more streamlined curation of pertinent data.

The 'Coluzzi' papers

This rather simple, new section provides a collection of 'classical' papers, in PDF format, on anophelines, written by various authors, obtained from the reprint collection of Prof. Mario Coluzzi (Rome). The abstracts of the papers are searchable, and it is hoped that the selection will ultimately be expanded to include his entire collection on malaria and mosquito-related publications, hopefully recruiting the whole *Anopheles* community. This has become imperative as several important papers on mosquito biology, and in particular reports dealing with population biology of mosquitoes were published several decades ago in journals that are no longer published or that will not make these publications available online.

Other data and resources

AnoBase contains a variety of other data and resources that are of potential value to the *Anopheles* research community. Given their relative simplicity in terms of search possibilities, etc., they will not be discussed here in detail. These include a BLAST server that can be used to quickly query different *Anopheles*-related sequence datasets, compilations of both nucleic acid and protein sequences

from non-*gambiae* anophelines (including the respective links), data on a variety of molecular markers used in population genetic studies (e.g. microsatellites, RAPD markers and mtDNA sequences from all available anopheline species), general information on non-*gambiae* species and, finally, useful information and links for the research community such as addresses of researchers and specific information on meetings or important documents.

Conclusions

Primary whole genome sequence information is best handled by specialized databases that can accommodate and handle the enormous datasets that result from a genome. In the case of *An. gambiae* the annotation of the genome and its periodic updates is handled by ENSEMBL (Mongin *et al*., 2004). AnoBase is complementary, and certainly not an alternative to the ENSEMBL *Anopheles*-WGS section. This is exemplified both by the large amount of data that are absent from either the ENSEMBL gene browser or AnoBase, but also by some common data that are presented in a completely different mode, such as genes.

Research in vector biology, and in particular *An. gambiae* has advanced rapidly since the completion of the WGS of *An. gambiae* (Holt *et al*., 2002). The increase of scientific, especially molecular, output has not been confined to *An. gambiae*, but covers a series of other insects (and other arthropods) of social/medical importance. Furthermore, a large number of new sequencing projects have been initiated. The *Aedes aegypti* WGS will soon be completed, a pregenome analysis is underway in *Culex quinquefasciatus* as a precursor of a full WGS, and an expressed sequence tag (EST)-based Tsetse project is proceeding, while other projects will undoubtedly start in the near future. The tools that will become available for the control of insect vectors are bound to improve, and their number to increase. The need for databases that hold the information and make it available for mining is therefore obvious.

AnoBase now forms part of an National Institute of Allergy and Infectious Diseases (NIAID)-funded project, VectorBase, that brings together a series of bioinformatics resources with the aim of developing a integrated scheme for insect vectors. *Anopheles* obviously acts as the 'model' for this enterprise that will encompass other disease vectors such as *Anopheles funestus*, *Aedes aegypti*, *Culex pipiens*, *Glossina* spp., and the tick *Ixodes scapularis*. Although the details of AnoBase's incorporation within VectorBase have not yet been implemented, it is planned to continue to make available all present data to the community. Moreover, AnoBase will also make available controlled vocabularies and ontologies relating to specific processes such as, for example, haematophagy and interactions between *Anopheles* and the disease agents that it carries, that are presently being constructed.

Experimental procedures

Hardware and software

AnoBase relies on MySQL database engine and runs on a Sunfire server (Sun Microsystems, Denver, CO) with 2 Gb RAM and 144 Gb storage space, half of which is in an external RAID array. Two other Pentium 4-based machines (Intel, Santa Clara, CA) with 1 Gb RAM and 80 Gb storage space, both running Linux, are used as developing stations for AnoBase. The web interface is based on both Perl and PHP and takes advantage of the SQL query language already built into MySQL. The web server supporting AnoBase is the apache server (http://www.apache.org).

Data input followed the philosophy of FlyBase requiring the curators to complete text-based proformas, then incorporating the data by parsing those files. A 'proforma' is a form containing prescribed, defined fields that are to be used for data inclusion. The core of the system is the publication proforma. All others are linked to it, each one with predefined relations forming a tree structure (i.e. a publication proforma can have several branches of population proformas which describe a specific mosquito population; each population may have other branches of data describing specific assays to test for resistance in insecticides). A newly developed standalone tool written in Java simplifies the process of defining the proper relations between the various data. Fields and data that are repeated several times and could fall in specific categories (e.g. chemical and trade names of insecticides) are transformed into controlled vocabulary lists presented as pull-down menus in the tool, decreasing the likelihood of erratic submissions. The outcome is again text-based files similar to the proformas, which have to be parsed into the database. Another version of the web-based tool is in progress. This communicates directly with a copy of the database adding or modifying entries in real time. Then the curator incorporates these modifications into the main database.

Acknowledgements

We are particularly indebted to Drs Boris Dobrokhotov and Yeya Touré for their long-standing interest in, and the support of, the *Anopheles* database project. Thanks also to Bruno Arcà, Michael Ashburner, Kate Aultman, Marc Benedict, Takis Benos, Nora Besansky, Ewan Birney, George Christofides, Frank Collins, Mario Coluzzi, Paolo Costantino, David Emmert, Bill Gelbart, Aubrey de Grey, Martin Hammond, Tony James, Fotis Kafatos, Yannis Kouklinos, Osvaldo Marinotti, Jules Milgram, Manu Mongin, Carlos Morel, Ayo Oduola, Elias Papanikolaou, José Ribeiro and David Roos for their constructive comments and active support at different stages of the project. Our apologies to those whom we simply forgot to include! The AnoDB and AnoBase projects were made possible through seed money from the John D. and Catherine T. MacArthur Foundation and generous funding from the UNICEF/UNDP/World Bank/World Health Organization Special Programme for Research and Training in Tropical Diseases (TDR), the National Institute of Allergy and Infectious Diseases (NIAID) to C.L. as well as the BIOMALPAR Network of Excellence. The help of the Fondazione Cenci-Bolognetti is also gratefully acknowledged.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G. *et al.* (2000) The genome sequence of Drosophila melanogaster. *Science* **287**: 2185–2195.

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.

Ashburner, M. and Drysdale, R. (1994) FlyBase – the Drosophila genetic database. *Development* **120**: 2077–2079.

Celniker, S.E. and Rubin, G.M. (2003) The *Drosophila melanogaster* genome. *Annu Rev Genomics Hum Genet* **4**: 89–117.

Coluzzi, M., Sabatini, A., Petrarca, V. and Di Deco, M.A. (1979) Chromosomal differentiation and adaptation to human environments in the Anopheles gambiae complex. *Trans R Soc Trop Med Hyg* **73**: 483–497.

Coluzzi, M., Sabatini, A., della Torre, A., Di Deco, M.A. and Petrarca, V. (2002) A polytene chromosome analysis of the Anopheles gambiae species complex. *Science* **298**: 1415–1418.

Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res* **12**: 387–395.

Drysdale, R.A., Crosby, M.A. and The FlyBase Consortium (2005) FlyBase: genes and gene models. *Nucleic Acids Res* **33** Database Issue: D390–395.

Eeckman, F.H. and Durbin, R. (1995) ACeDB and macace. *Methods Cell Biol* **48**: 583–605.

Gomez, S.M., Eiglmeier, K., Segurens, B., Dehoux, P., Couloux, A., Scarpelli, C. *et al.* (2005) Pilot *Anopheles gambiae* full-length cDNA study: sequencing and initial characterization of 35575 clones. *Genome Biol* **6**: R39.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Ribeiro, D.M.C. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**: 129–149.

Hong, Y.S., Hogan, J.R., Wang, X., Sarkar, A., Sim, C., Loftus, B.J. *et al.* (2003) Construction of a BAC library and generation of BAC end sequence-tagged connectors for genome sequencing of the African malaria mosquito *Anopheles gambiae*. *Mol Genet Genomics* **268**: 720–728.

Marinotti, O., Calvo, E., Nguyen, Q.K., Ribeiro, J.M.C. and James, A.A. (2005) Genome-wide analysis of gene expression in adult *Anopheles gambiae*. *Insect Mol Biol* **14**: 365–373.

Mongin, E., Louis, C., Holt, R.A., Birney, E. and Collins, F.H. (2004) The Anopheles gambiae genome: an update. *Trends Parasitol* **20**: 49–52.

Ribeiro, J.M.C., Topalis, P. and Louis, C. (2004) AnoXcel: a database of *Anopheles gambiae* proteins oriented to the bench scientist. *Insect Mol Biol* **13**: 449–457.

The FlyBase Consortium (2003) The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res* **31**: 172–175.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* **25**: 25–29.

Insect Molecular Biology (2008) **17**(1), 87–89

SHORT NOTE

Anatomical ontologies of mosquitoes and ticks, and their web browsers in VectorBase

P. Topalis*, C. Tzavlaki*†, K. Vestaki*†, E. Dialynas*, D. E. Sonenshine‡, R. Butler§, R. V. Bruggner§, E. O. Stinson§, F. H. Collins§ and C. Louis*†

**Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas, Heraklion, Crete, Greece;* †*Department of Biology, University of Crete, Heraklion, Crete, Greece;* ‡*Department of Biological Sciences, Old Dominion University, Norfolk, VA, USA; and* §*Center for Global Health and Infectious Diseases, University of Notre Dame, Notre Dame, IN, USA*

Abstract

VectorBase, an integrated, relational database that manages genomic and other genetic/biological data pertaining to arthropod vectors of disease, has recently embarked on the construction of ontologies and controlled vocabularies (CVs). It aims, thus, at providing all necessary tools for the complete annotation of vector genomes and, in particular, the annotation of functional genomic data. This task was initiated with the development of anatomical ontologies of mosquitoes and ticks, both of which were made compliant to CARO, the common anatomy reference ontology. The ontologies are complemented by the development of novel web-based browsers that can show figures for anatomical terms, something that is especially helpful for fully illustrating the controlled vocabularies of anatomy.

Keywords: arthropods, bioinformatics, databases, genome annotation, insect anatomy.

Introduction

While progress in information technology over recent decades has played a key role in both the acquisition and handling of DNA and protein sequence data, the wealth of information directly associated with them was accompanied by massive complexity in terms of management of the data in a fashion that can be understood by both computers and humans. This difficulty was especially recognized in the functional annotation of genes, which was based, to a large extent, on similarities detected between genes and gene models identified when genomes from different organisms were compared to one another. This led to the formation of the so-called Gene Ontology (GO) consortium (Ashburner *et al*., 2000), originally consisting of representatives of three model organism databases: FlyBase (Drysdale *et al*., 2005; Crosby *et al*., 2007), the Mouse Genome Database (Blake *et al*., 2006) and the *Saccharomyces* Genome Database (Christie *et al*., 2004). The consortium's aim was the construction of a controlled vocabulary to describe the roles of genes and gene products valid for any organism.

The development of GO was soon followed by the establishment of the OBO (Open Biological Ontologies) consortium (and its successor OBO Foundry (Smith *et al*., 2007); accessible at: http://www.obofoundry.org/), a loose collaboration of scientists involved in the construction of biological ontologies. The OBO Foundry has been instrumental in ensuring that biological ontologies '*... being developed are mutually compatible, expandable, and correctable in line with advances in biomedical science*'. The OBO Foundry now stores and makes available about 70 biological ontologies that are open; ie they are available to all for free and without any limitation on their use or re-distribution. The ontologies range from the GO to anatomy, to taxonomy, etc., and they also concern a variety of organisms. Most of them have been built using the OBO-Edit software that was developed for the editing of biological ontologies and controlled vocabularies (CVs; Day-Richter *et al*., 2007).

Here, we present two new open anatomical ontologies, for mosquitoes and ticks, which were developed to be included in VectorBase (Lawson *et al*., 2007). These are the first ontologies that are specific for arthropod vectors of disease and their purpose is to facilitate the annotation of genomic and biological data. Moreover, we briefly present two browsers that were also developed to handle the ontologies through VectorBase.

Received 11 September 2007; accepted after revision 29 October 2007. Correspondence: C. Louis, Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas, P.O. Box 1385, Vassilika Vouton, 71110 Heraklion, Crete, Greece. Tel.: +30 281 0391119; fax: +30 281 0391104; e-mail: louis@imbb.forth.gr

Results and discussion

The construction of anatomical and/or morphological ontologies on all levels of resolution is rapidly becoming essential, especially given the need to annotate an ever-increasing number of experiments that deal with RNA profiling and proteome analysis in a variety of organisms (Aitken, 2006). Given the complete absence of such resources for arthropod vectors of disease, VectorBase (Lawson *et al*., 2007) has embarked on the development of ontologies that will help medical entomologists handle results acquired using high-throughput molecular techniques such as whole genome sequences (Holt *et al*., 2002; Nene *et al*., 2007) and microarray analysis. An ontology is classically defined as '*the philosophical study of what exists: the study of the kinds of entities in reality, and the relationships that these entities bear to one another*' (Smith, 2003). On a practical level, we could describe it here as a controlled vocabulary that uses authorized, pre-defined terms in hierarchical order, that are linked to one another through a limited set of statements termed relations. The first two such ontologies developed in the frame of this project describe the anatomy of mosquitoes and ticks, both vectors of pathogens that cause a variety of diseases such as malaria, lymphatic filariasis, dengue and yellow fever, Lyme disease, as well as several kinds of encephalitis.

To build the anatomical ontologies, we made use of the OBO-Edit software that was developed for the purpose of constructing biological ontologies (Day-Richter *et al*., 2007). The first ontology built, that for the mosquito, was originally very grossly modelled on the anatomy ontology of *Drosophila melanogaster*. All terms and similar items were based on the descriptions and definitions of Harbach & Knight (1980). The ontology originally contained 1818 terms and 5127 synonyms. It fully described all external features of the mosquito anatomy, whereas internal features are, at this point, focused on the adult stage, and especially on the alimentary canal and salivary glands, as these are the main tissues that are 'involved' in pathogen transmission. Obviously, the ontology can (and will) be further expanded to include missing terms. Upon initial completion, it was made publicly available through its inclusion in the catalogue of ontologies that are listed and freely downloadable at the OBO foundry (http://obofoundry.org/cgi-bin/table.cgi). Following the recent development of the Common Anatomy Reference Ontology, CARO (Haendel *et al*., 2007; http://obofoundry.org/cgi-bin/detail.cgi?caro), which attempts to set rules for the unification of the schemes of anatomical ontologies throughout metazoa, we decided to completely rebuild the mosquito anatomy CV to meet the requirements of this reference ontology. The CARO-compliant version now consists of 1861 terms with 5178 synonyms. The two mosquito anatomy ontologies are quite different from each other in terms of architecture, although they do contain, to a large degree, the same anatomical entities. OBO foundry now lists the new, CARO-compliant ontology (TGMA), while the older version is available, upon request, from the authors. The tick ontology (TADS in the OBO foundry), in contrast, was made CARO-compliant *ab initio*. It contains 628 terms and 89 synonyms, and their descriptions are fully based on Sonenshine (1991) with small modifications. Both ontologies now almost uniquely use either the *is_a* or the *part_of* relations (see http://www.obofoundry.org/ro/ for the exact formal definition of these, and other relations). In addition to the OBO foundry, the ontologies have also been made publicly available, first through AnoBase (http://www.anobase.org), the *Anopheles* database, and then through VectorBase, into which AnoBase is in the process of being incorporated. As is the case for most ontologies available to the research community, although fully operative, it should be stressed that these ontologies should not be considered as 'final' because they will constantly undergo additions and modifications on a regular basis. Finally, we should like to stress the fact that these were the first anatomical ontologies to adopt the CARO rules.

A series of browsers have already been developed for viewing ontologies, such as AMIGO (Gene Ontology Consortium, 2006), DynGO (Liu *et al*., 2005) and OBO-Edit (Day-Richter *et al*., 2007), but most of them require the ontologies to be downloaded on the user's own computer. While the OBO format used for the construction of the anatomical ontologies is easy to be interpreted by computers, it is inappropriate for direct access by human users. The serial way of cataloguing terms and the compact way of storing the relations are its most profound disadvantages. Using the locally installed OBO-Edit software, which was also used to build the CVs, makes browsing easier, but it does not allow figures to be shown together with the description of anatomical entities. We therefore proceeded with the design of a simple web-based browser that would have this capacity.

In order to use this browser, the OBO file is parsed into four different tables of a database. The term table holds information on the terms (term names, term IDs and definitions). The synonym table contains the synonyms available for each term. The relations table contains data on the kind of relations and the related terms, including the parent–child relations. Finally, the pictures table brings together the file names of the pictures available with the appropriate anatomical entities. This browser, which can be accessed at AnoBase, will display both the mosquito (http://www.anobase.org/cgi-bin/anatomy.pl) and the tick (http://www.anobase.org/cgi-bin/tick_anatomy.pl) anatomical ontologies accessing the figures from Harbach & Knight (1980) and Sonenshine (1991), respectively (permission of the copyright holders is available). The user is able to query the anatomy for a known specific term name (or synonym) and/or a keyword included in the definition of a term. The query

then returns all relevant terms (with their IDs, names, description and synonyms, where available) linked to a more detailed view. The figures can be shown in higher analysis by clicking on them.

The newer VectorBase CV browser is a comprehensive system for navigating CVs in VectorBase (http://www.vectorbase.org/Search/CVSearch); it is also fully capable of showing figures attached to any ontology being searched. Hosting GO as well as both the mosquito and the tick anatomy ontologies, the VectorBase browser provides an expandable tree for user navigation. A user is given a hierarchical representation of a group of terms, making parent–child relationships the method of traversal between CV terms. Information and associated features related to a term are displayed upon navigation to it, including a term's relationship to its parent, shown in the box describing the term. Some terms are associated with other features hosted natively on VectorBase; they provide entry-points to report pages on these features hosted on the VectorBase site. As is the case with the simpler browser, some terms also have associated images, which are provided upon navigation to these terms. The CV/ontology browser utilizes a searching system as the quickest access point into the browser. Searching will either yield a direct CV term upon an exact match, or a set of search results, which can be directly investigated. These capabilities make the browser a comprehensive resource for investigating CVs or ontologies hosted on VectorBase.

Ontologies, as mentioned earlier, enable the annotation of large datasets, and especially genomes and genes, in a way that can be 'understood' and 'handled' by computers, provided, of course, that they are adopted by the community and they are widely used for that purpose. A wide adoption of ontologies by databases, then, offers the bonus advantage of enhancing searches; this happens by also utilizing synonym terms in queries submitted to a database and/or asking to be included in the search output results matching children (or parents) of the term used as a query. With this in mind, VectorBase is currently developing additional ontologies that are related directly or indirectly to disease vectors' physiology and/or their capacity to transmit pathogens. It is hoped that the full set of ontologies, once developed, will first allow for an improved possibility to compare among disease-related biological processes across species and second, will provide immediate help in decision support in matters such as insecticide resistance management.

Acknowledgements

The core VectorBase project is funded by contract HHSN266200400039C from the NIAID; this work is part of the activities of the BioMalPar European Network of Excellence supported by a European grant (LSHP-CT-2004-503578) from the Priority 1 'Life Sciences, Genomics and Biotechnology for Health' in the 6th Framework Programme. We are indebted to Dr Ralph Harbach for granting permission to use the figures from his book, Dr Michael Ashburner for his encouragement as well as his critical comments during the development of the ontologies, and Nikos Giakoumakis for his help in transferring TGMA to the CARO format.

References

Aitken, S. (2006) Formalizing concepts of species, sex and developmental stage in anatomical ontologies. *Bioinformatics* **21**: 2773–2779.

Ashburner, M. and Drysdale, R. (1994) FlyBase – the Drosophila genetic database. *Development* **120**: 2077–2079.

Ashburner, M., Ball, C.A., Blake, J.A, Botstein, D., Butler, H. Cherry, J.M. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–29.

Blake, J.A., Eppig, J.T., Bult, B.J., Kadin, J.A., Richardson, J.E. and Mouse Genome Database Group (2006) The Mouse Genome Database (MGD): updates and enhancements. *Nucleic Acids Res* **34**: D562–D5674.

Christie, K.R., Weng, S., Balakrishnan, R., Costanzo, M.C., Dolinski, Dwight, S.S. *et al*. (2004) *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res* **32**: D311–D314.

Crosby, M.A., Goodman, J.L., Strelets, V.B., Zhang, P., Gelbart, W.M. and The FlyBase Consortium (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res* **35**: D486–D491.

Day-Richter, J., Harris, M.A., Haendel, M., The Gene Ontology OBO-Edit Working Group and Lewis, S. (2007) OBO-Edit, an ontology editor for biologists. *Bioinformatics* **23**: 2198–2200.

Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res* **34**: D322–326.

Haendel, M.A., Neuhaus, F., Osumi-Sutherland, D., Mahee, P.M., Mejino J.L.V. Jr, Mungall, C.J. *et al*. (2007) CARO – The Common Anatomy Reference Ontology. In *Anatomy Ontologies for Bioinformatics: Principles and Practice* (Burger, A., Davidson, D. and Baldock, R., eds), In press. Springer, New York.

Harbach, R.E. and Knight, K.L. (1980) *Taxonomists' Glossary of Mosquito Anatomy*. Plexus Publishing Inc., Marlton, NJ.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R. *et al*. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**: 129–149.

Lawson, D. *et al*. (2007) VectorBase: a home for invertebrate vectors of human pathogens. *Nucleic Acids Res* **35**: D503–D505.

Liu, H., Hu, Z.Z. and Wu, C.H. (2005) DynGO: a tool for visualizing and mining of Gene Ontology and its associations. *BMC Bioinformatics* **6**: 201.

Nene, V., Wortman, J.R., Lawson, D., Haas, B., Kodira, C. *et al*. (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* **316**: 1718–1723.

Smith, B. (2003) Ontology. In *Blackwell Guide to the Philosophy of Computing and Information* (Floridi, L., ed.). Blackwell, Oxford.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W. *et al*. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* **25**:1251–1255.

Sonenshine, D.E. (1991) *Biology of Ticks*, Vol. 1. Oxford University Press, New York, NY.

Opinion

How can ontologies help vector biology?

Pantelis Topalis[1], Daniel Lawson[2], Frank H. Collins[3] and Christos Louis[1,4]

[1] Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas, P.O. Box 1385, Vassilika Vouton, 71110 Heraklion, Crete, Greece
[2] European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK
[3] Center for Global Health and Infectious Diseases, University of Notre Dame, PO Box 369, Notre Dame, IN 46556-0369, USA
[4] University of Crete, Department of Biology, P.O. Box 2208, 71103 Heraklion, Crete, Greece

The reach of genomics has now extended to vector biology, with three mosquito genomes already sequenced and more arthropod vector genomes in the pipeline. The availability of these genomes has paved the way for high-throughput investigations on genome-wide gene expression and proteomics in vector biology. Such investigations would not have been possible without parallel progress in bioinformatics. It is now necessary to construct specific ontologies that will enable vector biologists to achieve computer-comprehensible annotation of genes and genomes, but also of various experimental, clinical and surveillance data. This will inevitably lead to the enhanced usage of such controlled vocabularies, and to an effort to develop novel ontologies, particularly in the context of disease control.

Databases as genomic tools

Few would have predicted at the time of development of modern DNA sequencing techniques [1] that the human genome would be completely sequenced less than two decades later [2,3]. In addition to the nucleic acid sequence repositories, such as the European Molecular Biology Laboratory (EMBL) and GenBank [4,5], established in the early 1980s for storing this newly acquired data, the exponential expansion of data generation also led to the development of specialized databases. These evolved from simple electronic storage areas into complex structures that provide their users with the tools necessary for viewing, understanding, mining and manipulating the complex mass of information generated. Examples of these multi-faceted databases include FlyBase [6,7], PlasmoDB [8,9] and Ensembl [10,11]. These databases go beyond the representation of the sequence data to include, for example, extensive genetic information and literature on the fruit fly (FlyBase), analytical tools for searching sequence variation and comparison with related genomes (Ensembl) and the integration of functional data relating to genes and gene products of malaria pathogens (PlasmoDB).

Using the Internet to access allied data kept in other data repositories and making full use of these resources through cross-referencing has led to a dramatic expansion in the information available for a given organism. Starting from a gene entry, information can be found about all known genetic and molecular data relating to that gene and, potentially, links can be found to access 3D protein structures and molecular and genetic interactions, making it possible to explore the function of the gene product in a particular organism and in the wider biological realm.

Vector biology, defined as the science that studies all arthropod biological processes that are specific to, and have a direct or indirect role in, disease transmission, was relatively slow in adopting modern molecular approaches. However, this situation changed when a decision was reached to sequence the genome of *Anopheles gambiae*, the second insect to be studied at this level of analysis [12], after *Drosophila melanogaster*. When the genome sequence was completed [13] the annotation was stored in Ensembl [10,11] rather than in the existing *Anopheles* database, AnoBase [14]. The strength of Ensembl in automated genome annotation and the handling of large biological datasets was deemed important by The *Anopheles* Genome Project Consortium. In contrast to the model

Glossary of ontologies and important acronyms

Biological ontology: Biological ontologies describe concepts and relationships that can exist for a biological entity in a computer-readable fashion.
CARO: Common Anatomy Reference Ontology developed to facilitate inter-operability between existing anatomy ontologies for different species and to provide a template for building new anatomy ontologies.
Controlled Vocabulary (CV): A restricted list of terms relevant to a topic; these are words or phrases that are acceptable values for completing certain metadata fields.
Evidence code: A two- or three-letter code showing the evidence that supports a specific GO annotation (http://www.geneontology.org/GO.evidence.shtml).
Gene ontology (GO): An ontology consisting of three separate ontologies (cellular component, biological process and molecular function) that is used to describe genes and, more often, gene products in a species-independent manner.
NCBO: The National Center for Biomedical Ontology is a consortium of leading biologists, clinicians, informaticians and ontologists who develop innovative technology and methods that enable scientists to create, disseminate and manage biomedical information and knowledge in computer-processable form.
OBO Foundry: The OBO Foundry is an open community of scientists who aim to develop fully interoperable ontologies that should enable best communication of scientists and their tools. OBO Foundry is part of NCBO.
OBO: Open Biomedical Ontologies. A collection of freely available, well-structured controlled vocabularies for shared use across different biological and medical domains.
Orphan (hypothetical) gene: Putative open reading frames (ORFs), in ESTs or genomic sequences, without any resemblance to previously determined protein coding sequences.
Relations: The edges that link together terms in a given ontology. A Relations Ontology (OBO-REL) lists all relations that are accepted for use in an OBO Foundry-hosted ontology.

Corresponding author: Louis, C. (louis@imbb.forth.gr).

1471-4922/$ – see front matter doi:10.1016/j.pt.2008.03.002 Available online 24 April 2008

Table 1. Genome projects for arthropod vectors of disease

Species	Approach	Status
Aedes aegypti	WGS, EST	Finished
Anopheles gambiae Pest strain	WGS, EST	Finished
An. gambiae M strain	WGS	Ongoing (sequencing completed)
An. gambiae S strain	WGS	Ongoing (sequencing completed)
Culex pipiens	WGS, EST	Finished
Glossina m. morsitans	EST	Ongoing (WGS initiated)
Glossina p. palpalis		Planned
Ixodes scapularis	WGS, EST	Ongoing (sequencing completed)
Lutzomyia longipalpis		Planned
Musca domestica		Planned
Pediculus humanus	WGS, EST	Finished
Phlebotomus papatasi		Planned
Rhodnius prolixus	WGS, EST	Ongoing (WGS initiated)
Xenopsylla cheopis		Projected

WGS, whole genome shotgun; EST, expressed sequence tag.

organisms that had been sequenced previously, there was not a large body of published experimental data for the African malaria vector. The storage of data of the mosquito genome in different databases was not optimal, and a new project was initiated to bring together the two data sources to a single site, VectorBase [15]. VectorBase, a unified resource (not just for *Anopheles)* was created to include several other arthropod vector species (Table 1). Currently, VectorBase also contains data for the mosquitoes *Aedes aegypti* and *Culex pipiens*, for which genome sequencing efforts are completed, as well as the body louse *Pediculus humanus* and the tick *Ixodes scapularis*. Plans are in place to expand this catalogue on the basis of demand.

Biological ontologies

Tools used by most, if not all, genome databases are the Gene Ontology (GO) terms [16,17] that are key instruments for the annotation of genes and gene products of the organisms they handle (see Glossary of terms relating to ontologies). The GO currently consists of three 'sub-ontologies' that deal with cellular localization, molecular function and biological processes relating to genes and gene products. Apart from the model organism databases, in which dedicated data curators assign terms based on the literature, most organisms have terms assigned in an automatic process based on protein sequence similarity using the InterPro2GO software that utilizes a curated file associating domains with GO terms [18]. The assignment of GO terms to genes has helped to make sense of data acquired on the completion of a genome sequence and extending to other species. GO terms have proved invaluable as a common annotation of gene products in the analysis of gene clusters from microarray experiments in which other annotations are unreliable because of differences in the annotation processes. Moreover, GO is an ontology and therefore has inherent relationships between terms such that searches can be made beyond the actual keyword query to include a term that is its 'child'. As an example of the structure of an ontology, and the GO in particular, Figure 1 shows an excerpt of the GO sub-ontology of biological processes.

Although the complete concept of ontologies might have seemed complex, the GO was adopted by the research community well beyond the initial model systems; its practical success has had an impact in areas directly linked to its usage as briefly described here, and also in giving a push to the development of additional biological ontologies aimed at describing, for example, biological notions, scientific disciplines and inter-related biomedical terms [19,20].

The establishment of the Open Biomedical Ontologies (OBO) consortium (http://obo.sourceforge.net/main.html) and then the OBO Foundry (http://www.obofoundry.org) had a central role in ontology development, setting out a

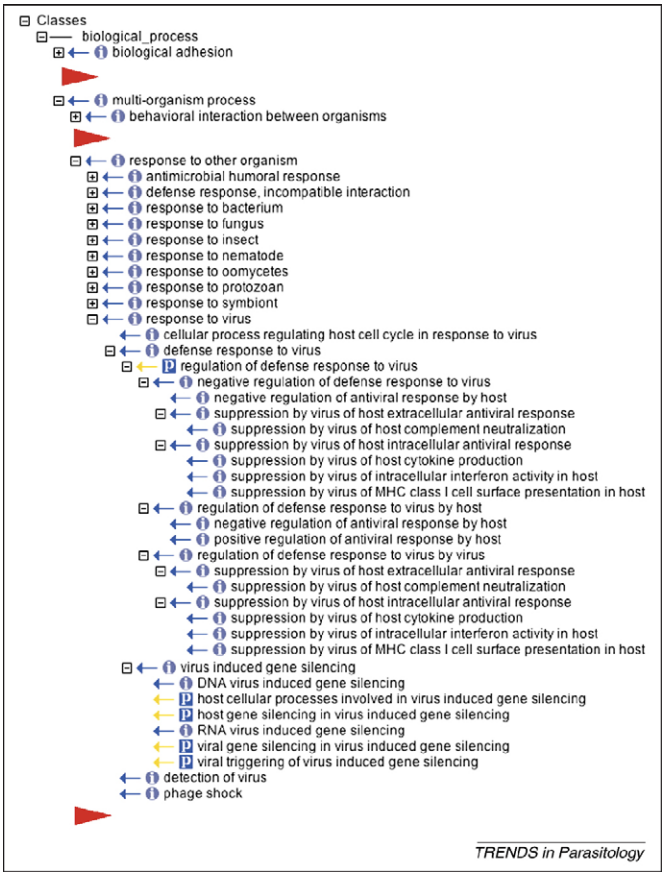


Figure 1. The figure shows an edited screen shot from the OBO Edit software into which the terms of the GO ontology were loaded. Only a part of the sub-ontology describing biological processes is shown here (the terms referring to 'response to virus'), but several terms and their respective child terms have been omitted at the positions marked by red arrowheads. Relations used in this example are *is_a* (indicated by a small grey circle labeled 'i') and *part_of* (indicated by a small blue rectangle that includes the letter 'p'). The small white rectangles indicate whether children of a given term are hidden (plus sign) or not (minus sign). Finally, all children of a given parent term are aligned left.

framework for ontology development and interoperability between biological ontologies. An example of the much needed standardization is in the area of relations [21]. What differentiates an ontology from a controlled vocabulary is mainly the fact that in an ontology terms are linked through specific relations (i.e. logical connection of terms into a hierarchy) joining 'parent' and 'child', but not 'sibling' terms. For an ontology to be comprehensible by all its users the relations must be described such that they have a clearly defined meaning understood by annotators (often the bench scientists performing the experiments). The OBO foundry currently makes available through its site a total of 70 ontologies; 19 of these deal with the anatomy of several organisms, including the mosquito and tick [22].

Ontologies and disease vectors

The mosquito (OBO prefix: TGMA) and tick (OBO prefix: TADS) gross anatomy ontologies are the first relating to arthropod disease vectors. As such, they are good examples of what ontologies can provide to a research community. These two ontologies were based on two excellent books [23,24] that described in detail the anatomy of these two taxa. Both ontologies were built to be compliant with the Common Anatomy Reference Ontology (CARO) standard [25]. The CARO compliancy enables linking to other ontologies in evolutionary studies. Despite the availability of the source books mentioned, before the TGMA ontology was developed it was not feasible to annotate fully an experiment concerning gene expression in a given mosquito because it was impossible to describe consistently (in a way that could be handled by a database) the exact anatomic location in which a given gene would have been found to be active. Moreover, the inclusion of figures showing the actual anatomical features in both vector anatomy ontologies, using two browsers developed at the same time, makes their use simple even for a novice [22]. Finally, because these ontologies can be accessed through VectorBase (http://www.vectorbase.org/Search/CVSearch/) there is no need for the uninitiated to download them from the OBO Foundry (http://www.obofoundry.org/cgi-bin/detail.cgi?id=mosquito_anatomy and http://www.obofoundry.org/cgi-bin/detail.cgi?id=tick_anatomy) and work with them locally using dedicated software such as OBO-Edit [26], which although specifically designed for the development and editing of these ontologies, can be cumbersome for the inexperienced user. The recent publication of the complete sequence of the *Aedes aegypti* genome [27] also shows how important the availability of a mosquito anatomical ontology could be in the near future because the availability of the genome sequence of this mosquito is expected to dramatically increase the number of experiments dealing with gene expression, which will need to be annotated.

Soon after the completion of the *An. gambiae* genome sequence it became apparent that in addition to the development of anatomy ontologies, GO annotation of the mosquito genes was also necessary. The manual assignment of GO terms to genes and gene products is a labour-intensive task, requiring dedicated curators. For the *An. gambiae* genome (later also adopted for the *Ae. aegypti* genome) an automated approach was developed based on the Ensembl project. Briefly, GO terms were assigned to genes based on similarity searches at the level of InterPro [28] domains or by transitive annotation via orthology from a high-quality manual set of GO annotations. InterPro domains are calculated for the proteome using InterProScan [29] and then GO terms are associated with the gene based on a mapping file that associates GO terms with domains [18]. Finally, the InterPro2GO methodology was supplemented by the *D. melanogaster* GO annotations from FlyBase to project high-quality orthologous pairs between the two dipterans. According to the GO convention, terms that were annotated solely on the basis of results of similarity searches, such as BLAST, carry evidence codes reflecting this (IEA – inferred through electronic annotation or ISS – inferred through sequence or structural similarity). Notwithstanding, the VectorBase project decided that it would be better to attach low confidence GO terms to the genes of the genomes it handles, instead of none at all.

VectorBase is currently about to launch a major push towards a community-based annotation of both genomes and genes, as well as the annotation of gene-expression experiments. The availability of the GO terms already attached to those data, in combination with the anatomy ontologies, will therefore be of utmost importance. They will help the individual researcher to better understand the function of a gene, and the expected increased community feedback will certainly contribute to changing the evidence codes from those based on bioinformatic analysis to codes based on actual experiments.

Needs of vector biology

Will the GO and the additional anatomy ontologies dealing with individual disease vectors be sufficient for the annotation of genomes and experiments in the future? It is clear that this rhetorical question expects a negative answer, but to what extent? The simple annotation of genomes and genes might not require too much additional effort. There is a constantly increasing number of genomes being sequenced, mostly because of the new sequencing technology available [30]. This naturally leads to an improved knowledge on 'genes' in general, and the result is that the number of 'orphan' or 'hypothetical' genes (to which no GO terms can be assigned) in most organisms studied becomes smaller. Moreover, based on the usage of more-sophisticated bioinformatics tools, more genes can be tagged with GO terms. This is reflected in the GO by the listing of new additional evidence codes such as RCA (inferred by Reviewed Computational Analysis). One can thus expect individual researchers to fill in several blanks through their own work, and it seems *prima facie* that the area that remains to be completed with newly constructed ontologies would only be that of (insect or vector) physiology in the wide sense.

However, specific pathogen–host interactions are not covered by the GO and they will not be covered in the near future because this ontology is mainly dedicated to generic, rather than species-specific terms. Concepts such as 'midgut penetration by the ookinete' or 'probing' (in the sense of probing while feeding on a human), or 'outdoor resting' are not expected to be covered by the GO soon. Of course these

processes might subsequently be genetically and molecularly dissected into more precise components, to which a GO term could be assigned. Nevertheless, physiological functions such as those cited here are extremely important for disease transmission, and from the point of view of a vector biologist, these refer to processes that need to be annotated on a level higher than the gene. For this reason VectorBase has initiated the development of mosquito- and tick-specific ontologies that will cover physiology, using the already available anatomy ontologies as a starting block and in particular the aspects that are important for transmission.

Future perspectives

In addition to genome and gene annotation, in our opinion there is an urgent need for the development of ontologies that will be of particular importance to vector biology on a completely different level. This relates to the recent increasing efforts to fight vector-borne diseases using modern tools based on molecular biological strategies and information technology. It was briefly justified in this article why the generation of genome-scale datasets requires controlled vocabularies for a better interoperability of databases and, particularly, consistency in the usage of terms that have to be understood by scientists and computers. However, at the same time the non-genome related data-flow concerning the diseases transmitted by vectors is also increasing dramatically and efforts to improve epidemiological surveillance are increasing (see, for example, [31–33]). Therefore, in addition to the classical control measures, information systems need to be designed to manage these data and help obtain insights that will lead to the control of these diseases [34,35]. Ontologies will be important for data integration and data management (including database construction, complicated searches, interoperability). Areas that need to be covered include clinical aspects of the individual diseases, population biology of arthropod vectors, insecticide resistance and resistance of the pathogens to drugs, as well as the epidemiology of the diseases and further aspects, including pathology and mechanisms of disease. Efforts to construct vector disease data-management and decision-support systems based on such ontologies are already underway, and we are certain that in a couple of years a positive answer to the question asked in the title of this article will be based on practice and positive experience, rather than on future plans.

Acknowledgements

Work from the authors' laboratories related to the VectorBase project was funded by contract HHSN266200400039C from the National Institute of Allergy and Infectious Diseases and, in part, by the BioMalPar European Network of Excellence supported by a European grant (LSHP-CT-2004–503578) from the Priority 1 'Life Sciences, Genomics and Biotechnology for Health' in the 6th Framework Programme. The authors would like to thank all of their colleagues at VectorBase, particularly Rob Bruggner, Ryan Butler, Emmanuel Dialynas and Dan Sonenshine for a fruitful collaboration.

References

1 Wu, R. (1978) DNA sequence analysis. *Annu. Rev. Biochem.* 47, 607–634
2 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
3 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
4 Hamm, G.H. and Cameron, G.N. (1986) The EMBL data library. *Nucleic Acids Res.* 14, 5–9
5 Bilofsky, H.S. *et al.* (1986) The GenBank genetic sequence databank. *Nucleic Acids Res.* 14, 1–4
6 Drysdale, R. and Ashburner, M. (1994) FlyBase - the *Drosophila* genetic database. *Development* 120, 2077–2079
7 Crosby, M.A. *et al.* (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res.* 35, D486–D491
8 Bahl, A. *et al.* (2003) PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.* 31, 212–215
9 Stoeckert, C.J., Jr *et al.* (2006) PlasmoDB v5: new looks, new genomes. *Trends Parasitol.* 22, 543–546
10 Hubbard, T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.* 30, 38–41
11 Hubbard, T.J. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.* 35, D610–D617
12 Morel, C.M. *et al.* (2002) The mosquito genome – a breakthrough for public health. *Science* 298, 79
13 Holt, R.A. *et al.* (2002) The genome sequence of the Malaria Mosquito *Anopheles gambiae*. *Science* 298, 129–149
14 Topalis, P. *et al.* (2005) AnoBase: a genetic and biological database of Anophelines. *Insect Mol. Biol.* 14, 591–597
15 Lawson, D. *et al.* (2007) VectorBase: a home for invertebrate vectors of human pathogens. *Nucleic Acids Res.* 35, D503–D505
16 Ashburner, M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29
17 Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* 34, D322–D326
18 Camon, E. *et al.* (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in Swiss-Prot, TrEMBL and InterPro. *Genome Res.* 13, 662–672
19 Stevens, R. *et al.* (2000) Ontology-based knowledge representation for bioinformatics. *Brief. Bioinform.* 1, 398–414
20 Stenzhorn, H. *et al.* (2007) Towards a top-domain ontology for linking biomedical ontologies. *Medinfo* 12, 1225–1229
21 Smith, B. *et al.* (2005) Relations in biomedical ontologies. *Genome Biol.* 6, R46
22 Topalis, P. *et al.* (2008) Anatomical ontologies of mosquitoes and ticks, and their web browsers in VectorBase. *Insect Mol. Biol.* 17, 87–89
23 Harbach, R.E. and Knight, K.L. (1980) *Taxonomists' Glossary of Mosquito Anatomy,* Plexus Publishing Inc.
24 Sonenshine, D.E. (1991) *Biology of Ticks,* Oxford University Press
25 Haendel, M.A. *et al.* (2008) CARO – The Common Anatomy Reference Ontology. In *Anatomy Ontologies for Bioinformatics: Principles and Practice* (Burger, A., Davidson, D. and Baldock, R., eds), pp. 327–350, Springer
26 Day-Richter, J. *et al.* (2007) OBO-Edit, an ontology editor for biologists. *Bioinformatics* 23, 2198–2200
27 Nene, V. *et al.* (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316, 1718–1723
28 Mulder, N.J. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.* 35, D224–D228
29 Quevillon, E. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.* 33, W116–W120
30 Metzker, M.L. (2005) Emerging technologies in DNA sequencing. *Genome Res.* 15, 1767–1776
31 Kelly-Hope, L.A. *et al.* (2006) Negative spatial association between lymphatic filariasis and malaria in West Africa. *Trop. Med. Int. Health* 11, 129–135
32 Betanzos-Reyes, A.F. *et al.* (2007) Comparative analysis of two alternative models for epidemiological surveillance in the Mexican Malaria Control Program. *Health Policy* 80, 465–482
33 Chansang, C. and Kittayapong, P. (2007) Application of mosquito sampling count and geospatial methods to improve dengue vector surveillance. *Am. J. Trop. Med. Hyg.* 77, 897–902
34 Breman, J.G. *et al.* (2004) Conquering the intolerable burden of malaria: what's new, what's needed: a summary. *Am. J. Trop. Med. Hyg.* 71, 1–15
35 Hemingway, J. *et al.* (2006) The Innovative Vector Control Consortium: improved control of mosquito-borne diseases. *Trends Parasitol.* 22, 308–312

OPEN ACCESS Freely available online

PLoS NEGLECTED TROPICAL DISEASES

MIRO and IRbase: IT Tools for the Epidemiological Monitoring of Insecticide Resistance in Mosquito Disease Vectors

Emmanuel Dialynas[1], Pantelis Topalis[1], John Vontas[2,3,4], Christos Louis[1,4]*

1 Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas, Heraklion, Crete, Greece, **2** Laboratory of Pesticide Science, Agricultural University of Athens, Athens, Greece, **3** Vector Research, Liverpool School of Tropical Medicine, Liverpool, United Kingdom, **4** Department of Biology, University of Crete, Heraklion, Crete, Greece

Abstract

Background: Monitoring of insect vector populations with respect to their susceptibility to one or more insecticides is a crucial element of the strategies used for the control of arthropod-borne diseases. This management task can nowadays be achieved more efficiently when assisted by IT (Information Technology) tools, ranging from modern integrated databases to GIS (Geographic Information System). Here we describe an application ontology that we developed *de novo*, and a specially designed database that, based on this ontology, can be used for the purpose of controlling mosquitoes and, thus, the diseases that they transmit.

Methodology/Principal Findings: The ontology, named MIRO for Mosquito Insecticide Resistance Ontology, developed using the OBO-Edit software, describes all pertinent aspects of insecticide resistance, including specific methodology and mode of action. MIRO, then, forms the basis for the design and development of a dedicated database, IRbase, constructed using open source software, which can be used to retrieve data on mosquito populations in a temporally and spatially separate way, as well as to map the output using a Google Earth interface. The dependency of the database on the MIRO allows for a rational and efficient hierarchical search possibility.

Conclusions/Significance: The fact that the MIRO complies with the rules set forward by the OBO (Open Biomedical Ontologies) Foundry introduces cross-referencing with other biomedical ontologies and, thus, both MIRO and IRbase are suitable as parts of future comprehensive surveillance tools and decision support systems that will be used for the control of vector-borne diseases. MIRO is downloadable from and IRbase is accessible at VectorBase, the NIAID-sponsored open access database for arthropod vectors of disease.

Citation: Dialynas E, Topalis P, Vontas J, Louis C (2009) MIRO and IRbase: IT Tools for the Epidemiological Monitoring of Insecticide Resistance in Mosquito Disease Vectors. PLoS Negl Trop Dis 3(6): e465. doi:10.1371/journal.pntd.0000465

Editor: Pattamaporn Kittayapong, Mahidol University, Thailand

Received March 6, 2009; **Accepted** May 22, 2009; **Published** June 23, 2009

Funding: The work in C. Louis' laboratory was funded by contract HHSN266200400039C from the National Institute of Allergy and Infectious Diseases in the frame of the VectorBase project, and by the BioMalPar European Network of Excellence, which is supported by a European grant (LSHP-CT-2004-503578) from the Priority 1 "Life Sciences Genomics and Biotechnology for Health" in the 6th Framework Programme. JV was supported by the Hellenic Secretariat General for Research and Technology (HSGRT) and the Innovative Vector Control Consortium (IVCC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: louis@imbb.forth.gr

Introduction

Diseases transmitted by arthropod vectors and, in particular, mosquitoes pose an immense load on global health, with malaria alone being responsible for more than 46,000,000 DALYs (Disease-adjusted Life Years); pertinent calculations are based solely on official, yet largely incomplete statistical estimates [1], and the global burden of *falciparum* malaria is nowadays estimated by some to be lower than originally thought [2]. Nevertheless, given the fact that arthropod-borne diseases affect mostly the populations of tropical regions, these huge numbers directly imply that their control is a *conditio sine qua non* for the socio-economic development of many of the poor areas of the world. Control of disease, then, directly entails the control of the arthropod vector populations and, most prominently among them, mosquitoes. Of course, economic development itself is one of the key players in the control of vector-borne diseases, unfortunately leading to an argument of a spiral form [3]. However, since the original recognition of the causes of malaria and other tropical diseases, campaigns aiming at eradicating vector-borne diseases included environmental management [4], indoor residual spraying (IRS) with the widespread use of DDT (Dichloro-Diphenyl-Trichloro-ethane) or other insecticides [5,6], as well as the use of impregnated nets (Insecticide-Treated Nets, ITN [7]; and Long-Lasting Insecticide-treated Nets, LLIN [8]). These approaches, combined with extensive use of drugs, soon led to the disappearance of the disease from most non-tropical areas of the world and notably Europe [9].

In spite of the initial wide successes achieved in the temperate zones, eradication of vector-borne diseases proved to be elusive in

Author Summary

It is a historical fact that a successful campaign against vector populations is one of the prerequisites for effectively fighting and eventually eradicating arthropod-borne diseases, be that in an epidemic or, even more so, in endemic cases. Based mostly on the use of insecticides and environmental management, vector control is now increasingly hampered by the occurrence of insecticide resistance that manifests itself, and spreads rapidly, briefly after the introduction of a (novel) chemical substance. We make use here of a specially built ontology, MIRO, to drive a new database, IRbase, dedicated to storing data on the occurrence of insecticide resistance in mosquito populations worldwide. The ontological approach to the design of databases offers the great advantage that these can be searched in an efficient way. Moreover, it also provides for an increased interoperability of present and future epidemiological tools. IRbase is now being populated by both older data from the literature and data recently collected from field.

the tropics. Moreover, the failure of vaccine development for vector-borne diseases, with the exception of the relatively early production of a vaccine directed against yellow fever [10], complicated the strategies aimed at controlling these diseases. Perhaps, most prominent among several problems that were faced by the national and international public health agencies were the occurrences of resistance relating to both parasites becoming resistant to anti-parasitic drugs [11] and mosquitoes to insecticides [12]. The gradual development of insecticide resistance against all classes of insecticides used today soon after their introduction [13], which was exacerbated by the use of such chemicals in agriculture [14], is considered by some to be presently the most important impediment in the successful control of vector-borne diseases.

Resistance to one or more insecticides used in vector control can have a crucial impact on the management of arthropod vector-borne diseases. In the case of ITN and LLIN measures [15,16], monitoring of insecticide resistance needs to become a key component for the efficient usage of control strategies [17]. Although overall data on pesticide resistance have been collected over a long period of time [18–20], these often remain inaccessible to public health workers around the world for a variety of reasons. One of them is the lack of a central database tool that would gather, store and exploit such data. Although pertinent studies are often published in refereed journals, their accessibility is limited by the use of restrictions, such as expensive subscription, something that is of extreme importance to scientists from disease-endemic countries, namely the very ones who urgently need to access these data.

With this in mind we decided to develop IT tools that could offer solutions to some of the problems and most importantly to help monitor the occurrence of insecticide resistance in vector populations; we decided to first focus on mosquitoes as these represent the most important vectors of disease. Rather than only expanding the simple repository of insecticide resistance studies that we had previously developed [21], we decided to completely restructure the database and support it by a dedicated ontology (or controlled vocabulary). This type of tool, which among others helps standardize terminology in a computer-comprehensible form, has already proved its immense potential in cases such as, most prominently, the Gene Ontology (GO) project [22]. Both the ontology (hereafter called MIRO for Mosquito Insecticide Resistance Ontology) and the novel, enhanced database (called IRbase) are freely accessible to the research community through their incorporation in VectorBase [23,24].

Materials and Methods

A Dell PowerEdge 850, with a dual core Intel Pentium D CPU running at 3 GHz, 3 GB of RAM, and 150 GB of hard disk storage was used for the development of IRbase. The operating system used is CentOS 4.5 and the web service is handled by the Apache server. Both MySQL and PostgresQL database servers were used for data storage. Webpage scripts and command line scripts are written in PHP. For PHP development we have been using the Zend Development Environment (ZDE). The OBO-edit software package [25] was used for the development of the MIRO.

To display the locations of the collection sites the Google Maps API and maps are used. Geographic data are exchanged between the applications using the Keyhole Markup Language (KML), a data schema for annotating and visualizing two or three dimensional maps. All coordinates are based on the World Geodetic System (WGS) 84 projection standard.

Data are entered through the online AJAX web interface, which is ontology based. Alternatively, submitters may send in their data in Open Office (ods), Excel (xls), comma separated values (csv), or tab separated values (tsv) files, which are processed and imported into the database using PHP scripts.

The MIRO can be accessed and browsed at the URL http://www.vectorbase.org/Search/CVSearch/ and its latest version can be downloaded from http://anobase.vectorbase.org/miro/miro_release.obo; it is also available through the OBO-Foundry at http://obofoundry.org/cgi-bin/detail.cgi?mosquito_insecticide_resistance; the home page for the IRbase is at the URL http://anobase.vectorbase.org/ir/. Both MIRO and IRbase are freely accessible. To access all necessary files for a local usage of IRbase the authors should be contacted by e-mail (louis@imbb.forth.gr).

Results/Discussion

The MIRO ontology

For the construction of the MIRO we followed the rules established by the OBO Foundry [26] in order to establish maximum interoperability in the future. This implied the use, to some extent, of already established ontologies, rather than the *de novo* development of new ones, such as the geographical component (see below). This decision obviously restricted the usage of relations linking terms to those allowed by the OBO Foundry rules and thus only *is_a*, *part_of* and *agent_in* are used throughout [27]. We are convinced, though, that this choice increases cross-ontology coordination and makes the tools developed more amenable to integration in a suite of malaria decision tools that are being developed.

The next choice we were faced with was the one of whether this ontology should follow the ontological scaffold and the rules and conventions described for the Basic Formal Ontology [28]. This ontological arrangement is already used for a variety of biomedical ontologies, including anatomical ontologies of disease vectors, notably mosquitoes and ticks [29]. Although the obvious advantages of a BFO-based ontology such as, for example, the ease of expansion that is based on its modularity cannot be easily discarded, we decided to initially design the MIRO on a more "traditional" scheme that would make it easily recognizable by users who are not proficient in ontologies. The single reason for this is to be able to provide the insecticide resistance community with a module that can be easily incorporated into other IT tools currently being devised. Nevertheless we are in the process of

transporting the MIRO into a BFO-based format in order to be able to integrate that version in future constructs that would potentially require such a layout.

MIRO is based on five top-level classes that actually form independent sub-ontologies (see Figure 1); four of them, "*biological material*", "*insecticidal substance*", "*method*" and "*resistance*", were developed *de novo* by us explicitly for the MIRO. In Figure 1 (left part) the ontology's terms are shown in a depth of two levels with the exception of the fifth class, the "*gazetteer*". This class represents a full importation of the Gazetteer (GAZ), a controlled vocabulary following ontological rules that describes named geographical locations (http://darwin.nerc-oxford.ac.uk/gc_wiki/index.php/GAZ_Project). GAZ is a community-based project of the EnvO Consortium for describing instances of organism environments and biological samples, supporting consistent annotation of locations and environments. The Gazetteer describes places and place names and the relations between them. Here, GAZ is basically used to describe the locations of sampling. Although it is a fully integrated component of MIRO, due to its size GAZ is not incorporated as such in our ontology, but it is automatically loaded through the Internet every time that one opens the MIRO using the OBOedit software. At this moment the MIRO contains 4,291 terms excluding, of course, the GAZ component that contains more than one hundred and fifty thousand geographical names from all over the world; more than 99% of the MIRO terms have full definitions. It should be noted that terms are not fixed and more are being added as these become necessary.

Biological material. This class, the largest one in the MIRO with 3,790 terms, describes all parameters that define the mosquito populations investigated (Figure 2). Its two main nodes are self-explanatory, one defining details of the population under study, including the biological stages of the individual specimens collected and sampled, as well as the kind of populations studied (field or established laboratory stock), while the other eventually defining the species under study (Figure 2A). As mentioned above, we have restricted the taxa listed in the MIRO to mosquitoes as these represent the main vectors of disease (e.g. Dengue, filariasis, malaria, yellow fever, etc.). We will eventually restrict the species of mosquitoes listed in the ontology to those that have already been described as actual vectors, and expand the ontology to cover non-mosquito vector arthropods (e.g. ticks, sand flies, etc.). For the present compilation of the different mosquito taxa we used primarily the Systematic Catalog of the *Culicidae* found at the Walter Reed Biosystematics Unit (WRBU, http://www.mosquitocatalog.org/species/taxonomy.asp). All taxa were linked to their parents, *i.e.* to the respective subgenus and genus, by *is_a* relationships and all synonyms listed in the WRBU catalogue were also registered in the ontology. We have also gone beyond the WRBU catalogue by including in the MIRO cryptic species such as incipient species "*sensu* Coluzzi" or chromosomal and molecular forms [30]. This obviously means that at present the S and M forms of *An. gambiae* s.s., for which extensive studies are being conducted, can equally be found in the ontology, and a particular analysis can be annotated accordingly (Figure 2B). Should future entomological research make it necessary to include similar data for other insect species groups the ontology can naturally be expanded in this respect.

Insecticidal substance. Two "catalogues" are available for the construction of the sub-ontology defining insecticides. These are ChEBI an open ontology of Chemical Entities of Biological Importance [31] and the IRAC catalogue (http://www.irac-online.org/eClassification/) a structured vocabulary compiled by the Insecticide Resistance Action Committee (http://www.irac-online.org/). Upon our request, the ChEBI group included in their

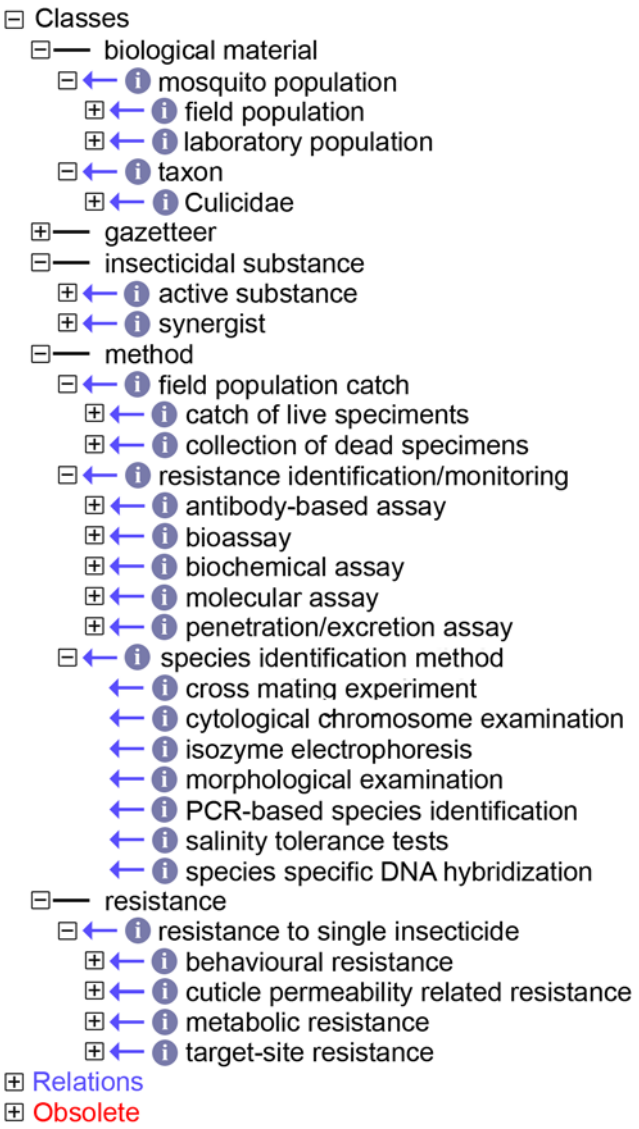


Figure 1. The upper levels of the MIRO. The figure shows the upper levels of the ontology. The small circles denote an "*is_a*" relation between the term and its parent, and small rectangles show the presence (plus) or absence (minus) of children for a given term. The children of the four "*biological material*", "*insecticidal substance*", "*method*" and "*resistance*" are shown in a depth of 2 levels. The "*gazetteer*" class has been loaded into MIRO (see Results and Discussion) and is therefore visible here.
doi:10.1371/journal.pntd.0000465.g001

ontology all insecticides listed by IRAC, and it now represents an optimal controlled vocabulary for those substances that could be used in the MIRO. Nevertheless, the IRAC eClassification list has the advantage of being immediately recognized and accepted by the IR (Insecticide Resistance) community as standard reference. Its structure is based on the mode of action of the individual insecticides but, to some extent, it is rather problematic on the level of an ontology. For example, it is based on a fairly rigid classification that, among others, leads to nameless or "non-existent" classes, or to classes that are not definable on either chemical or functional level (e.g. a class of compounds of "unknown mode of action" or a class of "miscellaneous non-specific inhibitors" [*sic*]. Nevertheless, given 1) the familiarity of the IR community with this classification and 2) the fact that the MIRO is meant to be an application ontology, we

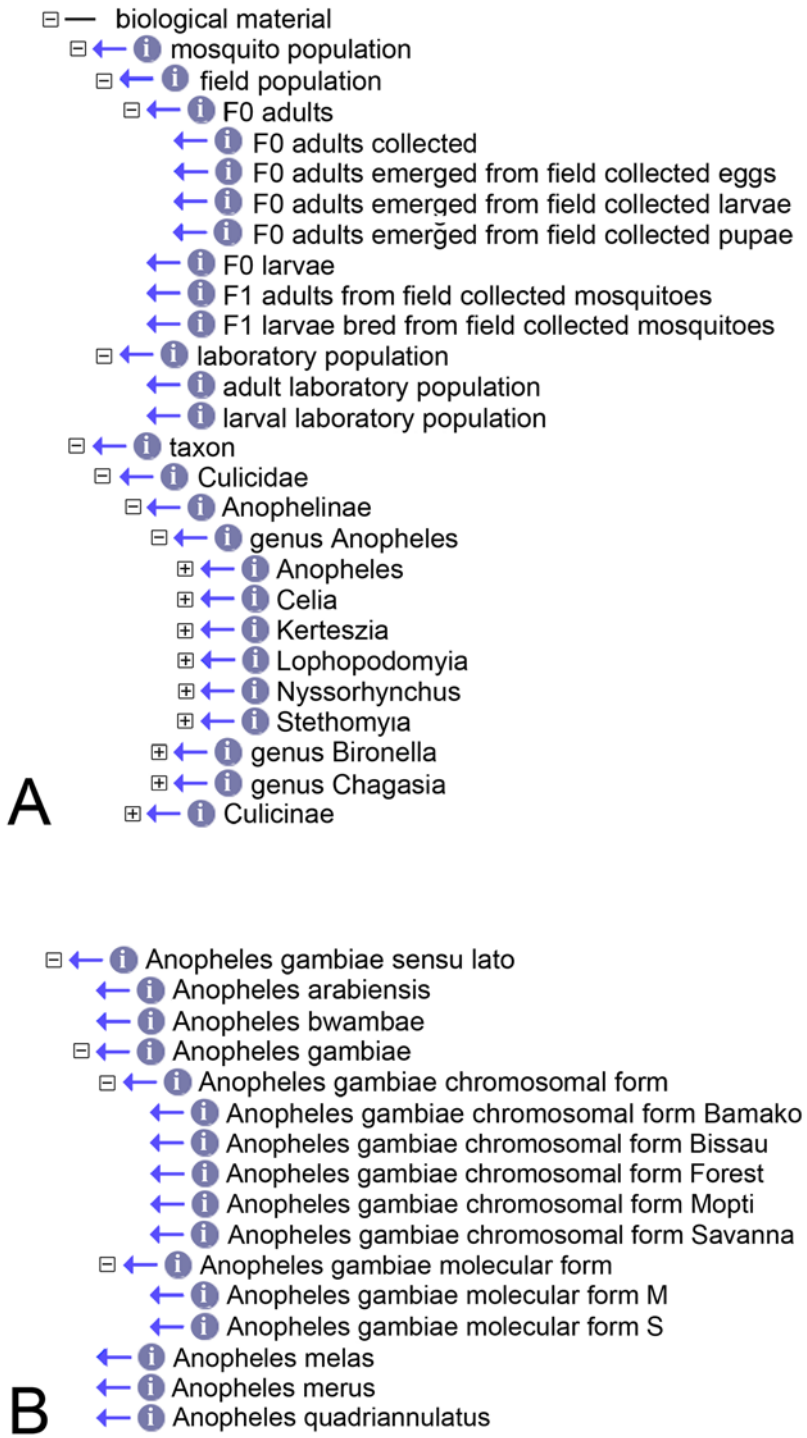


Figure 2. The ''*biological material*'' class. The figure shows, in a depth of 5 levels, the class ''biological material'' (A) and, within the ''*Cellia*'' subgenus, the *Anopheles gambiae* s.s.-related terms (B), that also include the chromosomal and molecular forms.
doi:10.1371/journal.pntd.0000465.g002

decided to use this, after some reorganization, as the primary scaffold for the construction of the ontology (Figure 3). All insecticides were, nevertheless, cross-referenced to ChEBI, as this ontology already represents the key ontology that links biology to chemical substances. The reason for cross-referencing, rather than using the ChEBI ID codes is to be found in the fact that in the MIRO all insecticides are defined, which is not the case for ChEBI.

A total of 22 different modes of action were retained for the classification of insecticides developed and included in the IRAC list effective December 2008. Finally we also included in the ontology a class containing synergists; two pertinent groups of synergists were listed in previous versions of the IRAC eClassification, but are no longer present. In spite of this, we incorporated them in the MIRO, given their potential significance in the actual usage of insecticides.

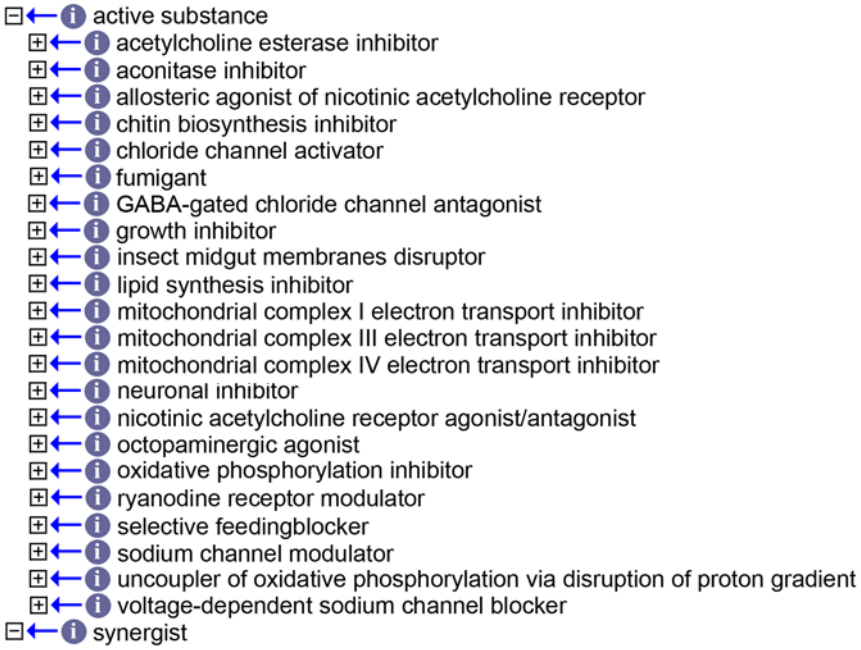


Figure 3. The groups of insecticides in MIRO. The figure shows the list of the groups of substances with differing modes of action containing the individual insecticides.
doi:10.1371/journal.pntd.0000465.g003

The "*Insecticidal substance*" class now contains 287 terms across all groups of substances.

Resistance. This class (73 terms), somewhat hindered by the jargon used by the IR community, lists all mechanisms known at this time. The four main categories used are behavioral and metabolic resistance, resistance due to changes in the permeability of the insect's cuticle and, finally, target-site resistance (Figure 4). Both "*behavioral resistance*" and "*cuticle permeability related resistance*" only list two self-explanatory children each: "*stimulus-dependent*" and "*stimulus-independent*" for the former, and "*enhanced excretion*" and "*reduced penetration*" for the latter. In contrast, the remaining two classes are more complex. The "*metabolic resistance*" class includes different facets of resistance connected to qualitative and quantitative changes of the activity of carboxyesterases (COE) and glutathione S-transferases (GST) and P450 monoxygenases. Furthermore a single child describes resistance due to modified midgut protease activity, *i.e.* processes related to the usage of biological insecticides such as the ones derived from *Bacillus thuringensis* or *B. sphaericus*. Finally target site resistance deals with known described mutations of specific genes.

Method. The final class was to some extent a problematic one. The reason for this was not the actual ontology construction but, rather, a problem of orthogonality. This class covers most, if not all methods that are directly used for the analysis of insecticide resistance in a mosquito population (a total of 137 terms). The methods vary from catch methods for field populations to molecular biological techniques (see Figure 1). While the former are straightforward and relatively easy to catalogue the latter pose certain dilemmas. These range from the terms used as such (e.g. "*bioassay*" or "*biochemical assay*" which may be too general) to the question of whether terms such as "*real-time PCR*" or "*RT-PCR*" that are outside the "narrow" field of insecticide resistance should be included. We decided eventually to include all techniques that are routinely used for the analysis of insecticide resistance. The choice was made based on the fact that the ontology OBI (personal communication, the OBI Consortium http://purl.obofoundry.org/obo/obi), which is currently being developed and which will describe life science and clinical investigations, is far from completion, and the MIRO would be missing a crucial

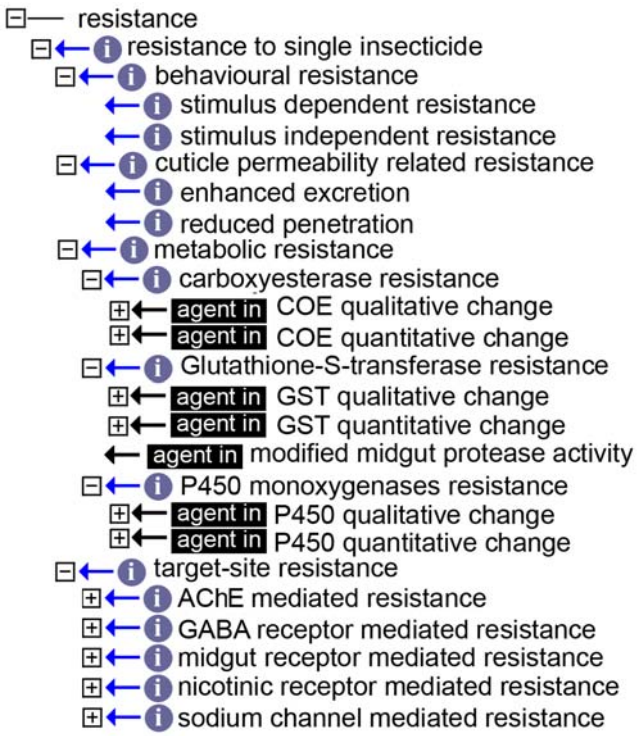


Figure 4. The "*resistance*" class. The "*resistance*" class has been opened to show the different contents. The black boxes denote an "*agent_in*" relation.
doi:10.1371/journal.pntd.0000465.g004

component if we were to exclude the relevant terms. Similar limitations existed for the techniques used for species identification and here we decided to keep the terms general without going into details. The species identification part is short and only describes the seven most common ways of identifying individual mosquito species. These include classical procedures (chromosomal banding patterns, cross mating experiments, morphology, and salinity tolerance tests) as well as biochemical (isozyme electrophoresis) and molecular (DNA probes, PCR). Of course, like is the case for all components of the MIRO, the ontology can be expanded or changed accordingly in the future if changes are deemed necessary.

The IRbase database

Based on feedback from the malaria entomology research community it was decided several years ago to include in AnoBase, the *Anopheles* database [21], a section on insecticide resistance; this tool was later transferred to and included in VectorBase [23,24] after this comprehensive genome database was established. The section consisted only of a series of manually-curated, already published studies; its role, therefore, was mostly to make data available to the community in a fashion that would be independent of the need for a library, rather than a use as an on-line epidemiological tool. The new IRbase in contrast is meant to serve as an expanding repository of associated data, which can be searched in a detailed fashion, thus providing immediately applicable information. Furthermore, IRbase now covers vectors of more diseases than the previous database that was only restricted to malaria. These are the reasons for designing a relational schema *de novo* (see Figure 5). It was our intention to design a schema that would easily enable both the addition of novel tables and the incorporation of IRbase into a larger and more complicated entity, which could be expanded later to encompass additional items linked to the control of vector-borne diseases.

The nine distinct tables can be distinguished in two major categories: While two of them (cv_term) handles all MIRO terms, including GAZ, and their relationships, the remaining are there to handle, mostly, ontology-independent items. These include, most prominently, description of the study in terms of details of the collection site, the mosquito population sampled (including collection dates, etc.) and the assay(s) performed. The "household" table is presently not in use by IRbase, but it has been included by request as it could be needed by decision support systems currently under development for Dengue and malaria [32]. The schema allows for a high degree of interoperability due to the enhanced usage of the ontology component, and it enhances the two distinctive features of IRbase, *i.e.* the two interactive components, *search* and *curator's tool*, both of which are accessible through a simplified web interface.

In addition to the completely new architecture of the database and to the fact that the software used is free and open source, IRbase has some key characteristic features: i) The data are stored in the database using MIRO terms wherever possible; ii) the Gaz geographic ontology is used for storing location data and the output can be viewed using maps; iii) extensive use of Ajax (Previously AJAX: Asynchonous JAvascript XML) is made in order to minimize network traffic and improve look and feel [33]. Moreover IRbase was built around basic entities:

1. "S*tudy data*" - a storage space for the data pertaining to an individual "study". The "study" could be an entire study,

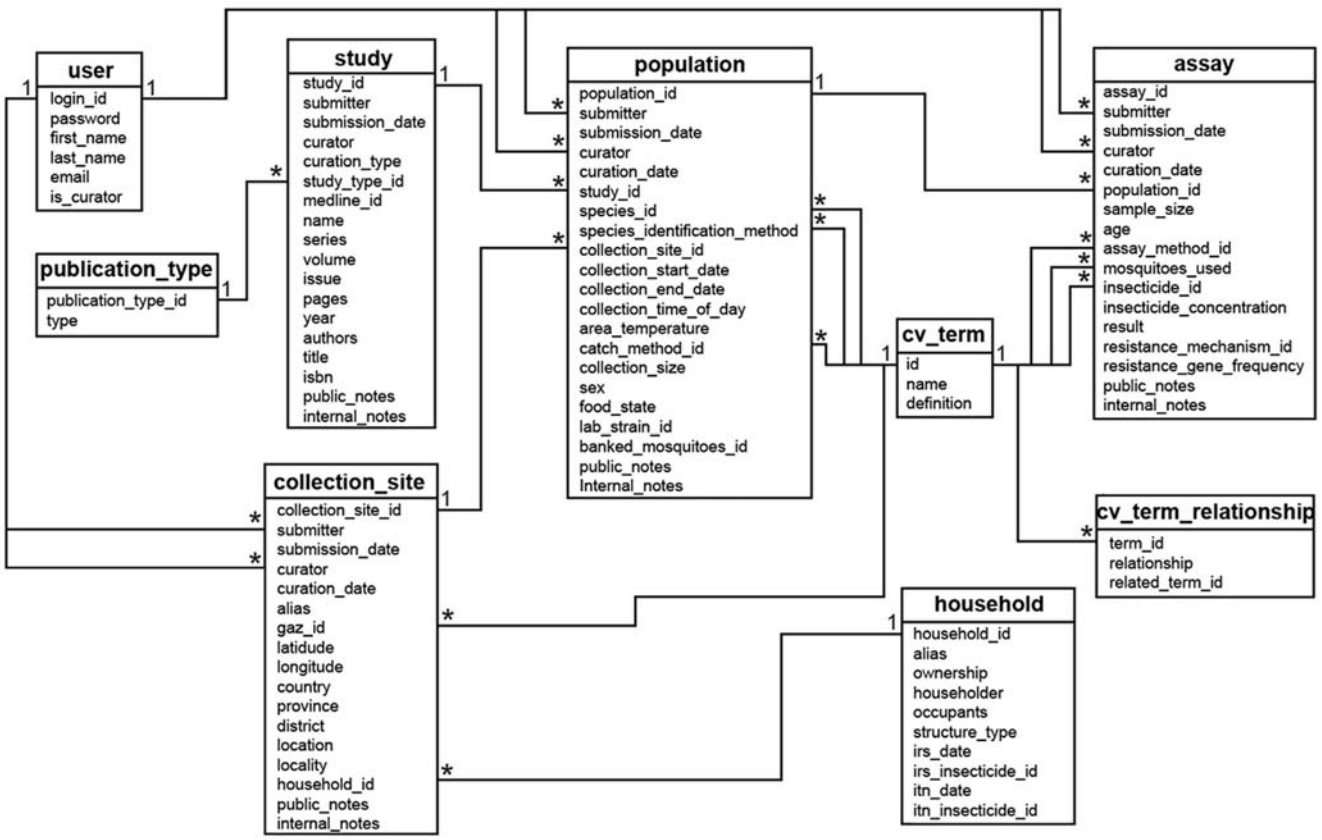


Figure 5. The IRbase schema. The figure shows the different tables that make up the schema. 1 and * denote a "one to many" relationship.
doi:10.1371/journal.pntd.0000465.g005

previously published or not, on an entire population or parts thereof, pertaining to one or more insecticides; the pertinent data would include the "owner" of the particular data, time it was carried out, the publication record when available, etc.

2. "*Collection site*" - common names of the collection site(s), their alias(es) and, most importantly, the geographic coordinates. Should these not be available through the submitter of the data, the IRbase curators will assign values based on available information and feedback. For those names that already exist in Gaz the Gaz ID is also stored.

The alias is an ID that the submitter can use for faster data entry: the collection site needs to be defined once and from thereon the alias can be used to identify that particular site. Collection sites that have no Gaz ID are exported and sent to the curators of that ontology for ID assignment.

3. "*Insect collections*" - this area holds information such as the species name, the collection date, the catch method, the sex, food state etc. of the specimens (field collected or lab bred) that were subsequently used to test resistance.

4. "*Assay data*" - The actual data expressing the findings and referring to the methods used, the conditions (insecticide concentration) and the results of an assay, etc.

The user interface. A brief manual is presented along with the search forms. There are two ways for entering search parameters into them. The first one is to use the drop-down menus and find the requested term by following the correct path. This obviously implies that the user is familiar with the MIRO or has a good knowledge for some of the properties of the requested term. For example, to find the insecticide "deltamethrin", the user must know that this insecticide belongs to the pyrethroid family and that this family of insecticides modulates sodium channels. Alternatively, auto-complete input boxes can be used. In these boxes the user needs to type two letters from the requested term and a list of all possible matching terms will appear. As more letters are typed the search is narrowed down to the decreasing number of options in the list. When the requested term appears on the list it can be clicked on and it will now appear in the input box. All terms are listed alphabetically in both the drop-down and the autocomplete menus.

Search criteria include species sites, year of collection, pertinent insecticide resistance mechanism, assay method, mosquitoes used, catch method and more. One criterion only may be used, or any combination of two or more of the above criteria. With the exception of the year of collection all remaining search criteria are ontology-based searches. As a result of this, the search algorithm implemented will also search for all the descendant terms of the term specified, and therefore searches can be narrowed in the process. Returned data are presented in descending chronological order, regardless of whether the collection year was set as a search criterion or not.

Users who want to utilize IRbase's data to run their own tests can set their export criteria and obtain the relevant data in a tab separated values (tsv) text file. This file can be opened in any spreadsheet application or be imported into a database.

Maps. In addition to the text-based interface to view data, IRbase also provides a map-based interface to access the same data (Figure 6). This interface utilizes Google Maps to visualize the data and is very rich in features such as grouping by color, zooming in and out, adding layers of related data, etc. By clicking at the collection sites marked on the map, a pop up balloon will appear with the same data, but also with a link to a detailed report (Figure 3). The map tool fully depends on the availability of geographical coordinates. As some of the older data are not linked to such information, this will have to be supplied manually by the IRbase curators before these studies can be incorporated. When the page is first loaded, a world map with all the collection sites spotted with small markers will appear. After leaving the mouse pointer on one of these markers, one line of text will appear displaying some of the information regarding the particular collection site such as species name, collection dates, and insecticide used. We stress here that the map section is continuously being improved in order to provide the users with a "friendlier" tool.

Data input. Data can be submitted to IRbase either online, via a web interface, or offline, using a spreadsheet template. These tools are available to the community upon request. Similarly a streamlined edition of the user and submitter/curator interface as well as the database can be loaded onto a laptop for offline data entry. This offers the advantage of a "limited" usage of IRbase even under conditions of limited access to the Internet (e.g. field trips). Again, users wanting to take advantage of this facility can contact VectorBase in order to obtain a user name and password.

Conclusion

We described here a set of IT tools to be used for the analysis of insecticide resistance in wild populations of insect disease vectors and in particular mosquitoes. The concept of intimately linking a dedicated database to a specific application ontology describing the field offers the advantage that the database can later be easily expanded to include additional items and offer further tools. This fact, in our case, can form the overall foundation or one of the pillars of a comprehensive tool, which could be used to globally monitor insecticide resistance; this would form the basis for a global decision support system for malaria and/or other vector-borne diseases.

A database on insecticide resistance, the Arthropod Pesticide Resistance Database (APRD), can already be found in the world wide web (http://www.pesticideresistance.org/). APRD covers a large variety of arthropods, but its philosophy is different from the one of IRbase. It provides reports of instances of occurrence of resistance, without any precision as to the exact location and the actual data. Although useful as a general indicator of resistance, especially in the domain of agriculture, the lack of geographic accuracy, combined with the lack of a map interface makes this database less suitable as a tool that could be used either by itself, or in combination to a modern, IT-based decision support system.

Such decision support systems are considered to be a prerequisite for the efficient control of insect vector populations. Many potential components of such systems have been described (see for example [32,34–35]), especially components that are based on GIS. Our tool has for the moment the capacity to depict data of insecticide resistance on a map provided the geographic coordinates have been incorporated in the data collection. Since many of the data that will populate IRbase are old, some of the coordinates will have to be input manually; once this has been the case, it will be possible to link all available information to maps based on, and retrieved from Google Earth.

The MIRO/IRbase set of tools is presently focused completely on insecticide resistance linked to mosquitoes of medical importance. The open source policy linked to the MIRO, an ontology that abides with the OBO Foundry rules, makes it easy to further develop these tools in order to later include data of agricultural interest as well, should an interested party turn up. In that sense one should also consider the fact that development of resistance detected in disease vectors can often be traced back to the often-improper use of insecticides in agriculture (see [36] for a discussion of that problem).

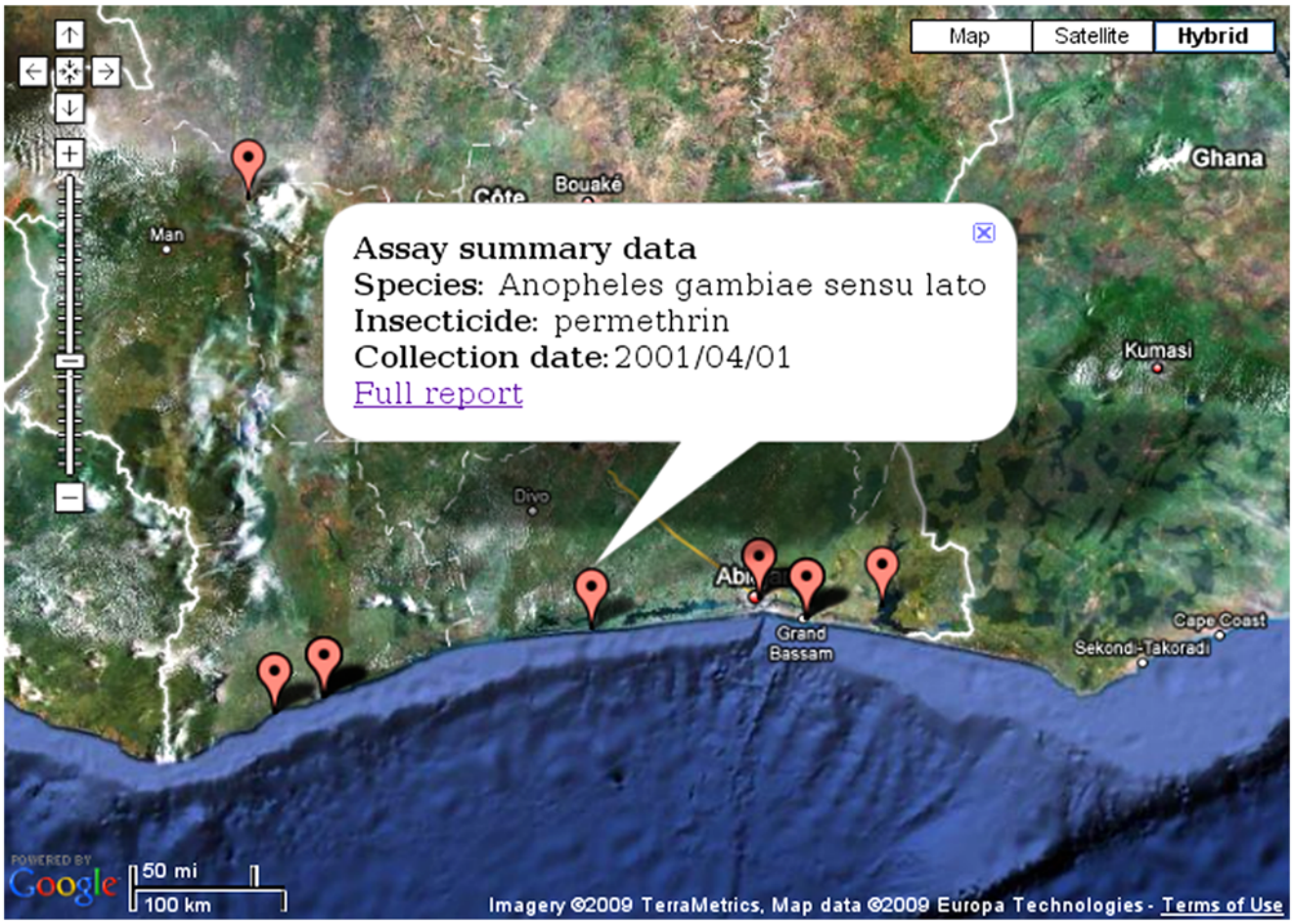


Figure 6. A map output providing summary information. The figure shows a screen shot of a search involving studies in Cote d'Ivoire. The pink "droplets" point to the sites of mosquito collections while the balloon that appears when clicking on any one of them includes summary data for the collection site. A link to the respective full report is also provided in the balloon.
doi:10.1371/journal.pntd.0000465.g006

We are currently in the process of populating IRbase with both data from the literature and data that are being collected from the field. This is done in collaboration with the international community in the frame of large consortia (e.g. African Network on Vector Research, Innovative Vector Control Consortium, WHO/Gates Foundation Vector Biology and Control Project, etc.), as well as on the basis of smaller individual research networks. We hope that, this way, IRbase will soon be established as the global repository for data insecticide resistance.

Acknowledgments

The authors would like to thank Drs. Michael Ashburner for critically evaluating the MIRO at its beginning stages, Kiril Degtyarenko and the ChEBI team for incorporating the full set of insecticides into the ChEBI ontology upon request, and Frank Collins for his encouragement and support in the frame of VectorBase. Moreover we are indebted to the international insecticide community for fruitful discussions, and in particular Drs. Magaran Bagayoko, Joseph Burhani, Maureen Coetzee, Marlize Coleman, Michael Coleman, Janet Hemingway, Louise Kelly-Hope, Lucien Manga and Hilary Ranson.

Author Contributions

Conceived and designed the experiments: CL. Performed the experiments: ED PT. Analyzed the data: ED PT CL. Contributed reagents/materials/analysis tools: ED JV. Wrote the paper: PT CL.

References

1. TDR/SWG/03 (2003) Report of the Scientific Working Group meeting on Malaria. Geneva: UNICEF-UNDP-World Bank-WHO Special programme for Research and Training in Tropical Diseases 2003.
2. Guerra CA, Gikandi PW, Tatem AJ, Noor AM, Smith DL, et al. (2008) The limits and intensity of Plasmodium falciparum transmission: implications for malaria control and elimination worldwide. PLoS Med 5: e38. doi:10.1371/journal.pmed.0050038.
3. Nájera JA (1994) The control of tropical diseases and socioeconomic development. Parassitologia 36: 17–33.
4. Coluzzi M (1994) Malaria and the Afrotropical ecosystems: impact of man-made environmental changes. Parassitologia 36: 223–227.
5. Davidson G (1951) Results of recent experiments on the use of DDT and BHC against adult mosquitos at Taveta Kenya. Bull World Health Organ 4: 329–332.
6. Wright JW, Fritz RF, Haworth J (1972) Changing concepts of vector control in malaria eradication. Annu Rev Entomol 17: 75–102.
7. Curtis CF (1991) Impregnated bed nets and curtains against malaria mosquitoes. In: Control of disease vectors in the community. Curtis CF, ed. London: Wolfe. pp 5–46.
8. Dabiré RK, Diabaté A, Baldet T, Paré-Toé L, Guiguemdé RT, et al. (2006) Personal protection of long lasting insecticide-treated nets in areas of *Anopheles gambiae* s.s. resistance to pyrethroids. Malar J 10: 5–12.

9. de Zulueta J (1973) Malaria eradication in Europe: the achievements and the difficulties ahead. J Trop Med Hyg 76: 279–82.
10. Barrett ADT (1997) Yellow fever vaccines. Biologicals 25: 17–25.
11. Renslo AR, McKerrow JH (2006) Drug discovery and development for neglected parasitic diseases. Nat Chem Biol 2: 701–710.
12. Brown AWA (1958) The insecticide resistance problem: a review of developments in 1956 and 1957. Bull World Health Organ 18: 309–321.
13. Hemingway J, Ranson H (2000) Insecticide resistance in insect vectors of human disease. Annu Rev Entomol 45: 371–391.
14. Roberts DR, Andre RG (1994) Insecticide resistance issues in vector-borne disease control. Am J Trop Med Hyg 50: 21–34.
15. Curtis CF, Maxwell CA, Magesa SM, Rwegoshora RT, Wilkes TJ (2006) Insecticide-treated bed-nets for malaria mosquito control. J Am Mosq Control Assoc 22: 501–506.
16. Hill J, Lines J, Rowland M (2006) Insecticide-treated nets. Adv Parasitol 61: 77–128.
17. Kulkarni MA, Malima R, Mosha FW, Msangi S, Mrema E, et al. (2007) Efficacy of pyrethroid-treated nets against malaria vectors and nuisance-biting mosquitoes in Tanzania in areas with long-term insecticide-treated net use. Trop Med Int Health 12: 1061–1073.
18. Chareonviriyahpap T, Aum-aung B, Ratanatham S (1999) Current insecticide resistance patterns in mosquito vectors in Thailand. Southeast Asian J Trop Med Public Health 30: 184–194.
19. Briët OJ, Galappaththy GN, Konradsen F, Amerasinghe PH, Amerasinghe FP (2005) Maps of the Sri Lanka malaria situation preceding the tsunami and key aspects to be considered in the emergency phase and beyond. Malar J 27: 4–8.
20. Jonsson NN, Hope M (2007) Progress in the epidemiology and diagnosis of amitraz resistance in the cattle tick *Boophilus microplus*. Vet Parasitol 146: 193–198.
21. Topalis P, Koutsos A, Dialynas E, Kiamos C, Hope LK, et al. (2005) AnoBase: a genetic and biological database of anophelines. Insect Mol Biol 14: 591–597.
22. The Gene Ontology Consortium (2008) The Gene Ontology project in 2008. Nucleic Acids Res 36: D440–D444.
23. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, et al. (2007) VectorBase: a home for invertebrate vectors of human pathogens. Nucleic Acids Res 35: D503–D505.
24. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, et al. (2009) VectorBase: a data resource for invertebrate vectorgenomics. Nucleic Acids Res 37: D583–D587.
25. Day-Richter J, Harris MA, Haendel M, Gene Ontology OBO-Edit Working Group, Lewis S (2007) OBO-Edit an ontology editor for biologists. Bioinformatics 23: 2198–2200.
26. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25: 1251–1255.
27. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, et al. (2005) Relations in biomedical ontologies. Genome Biol 6: R46.
28. Grenon P, Smith B, Goldberg L (2004) Biodynamic ontology: applying BFO in the biomedical domain. Stud Health Technol Inform 102: 20–38.
29. Topalis P, Tzavlaki C, Vestaki K, Dialynas E, Sonenshine DE, et al. (2008) Anatomical ontologies of mosquitoes and ticks, and their web browsers in VectorBase. Insect Mol Biol 17: 87–89.
30. della Torre A, Costantini C, Besansky NJ, Caccone A, Petrarca V, et al. (2002) Speciation within *Anopheles gambiae*–the glass is half full. Science 298: 115–117.
31. Degtyarenko K, de Matos P, Ennis M, Hastings J, Zbinden M, et al. (2008) ChEBI: a database and ontology for chemical entities of biological interest. Nucleic Acids Res 36: D344–D350.
32. Lozano-Fuentes S, Elizondo-Quiroga D, Farfan-Ale JA, Loroño-Pino MA, Garcia-Rejon J, et al. (2008) Use of Google Earth™ to strengthen public-health capacity and facilitate management of vector-borne diseases in resource-poor environments. Bull World Health Organ 86: 718–725.
33. Ullman C, Dykes L (2008) Beginning Ajax. Hoboken: Wrox.
34. Indaratna K, Hutubessy R, Chupraphawan S, Sukapurana C, Tao J, et al. (1998) Application of geographical information systems to co-analysis of disease and economic resources: dengue and malaria in Thailand. Southeast Asian J Trop Med Public Health 29: 669–684.
35. Hay SI, Snow RW (2006) The malaria Atlas Project: developing global maps of malaria risk. PLoS Med 3: e473. doi:10.1371/journal.pmed.0030473.
36. Lines JD (1988) Do agricultural insecticides select for insecticide resistance in mosquitoes? A look at the evidence. Parasitol Today 4: S17–S20.

Journal of Biomedical Informatics 44 (2011) 42–47

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A set of ontologies to drive tools for the control of vector-borne diseases

Pantelis Topalis [a], Emmanuel Dialynas [a], Elvira Mitraka [b], Elena Deligianni [a], Inga Siden-Kiamos [a], Christos Louis [a,b,*]

[a] Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas, 711 10 Heraklion, Crete, Greece
[b] Department of Biology, University of Crete, 711 10 Heraklion, Crete, Greece

ARTICLE INFO

Article history:
Received 20 October 2009
Available online 2 April 2010

Keywords:
Anatomy
Database
Decision support system
Insecticide resistance
Malaria
Mosquito
Tick
Transmission
Arthropod vector

ABSTRACT

We are developing a set of ontologies dealing with vector-borne diseases as well as the arthropod vectors that transmit them. After building ontologies for mosquito and tick anatomy we continued this project with an ontology of insecticide resistance followed by a series of ontologies that describe malaria as well as physiological processes of mosquitoes that are relevant to, and involved in, disease transmission. These will later be expanded to encompass other vector-borne diseases as well as non-mosquito vectors. The aim of the whole undertaking, which is worked out in the frame of the international IDO (Infectious Disease Ontology) project, is to provide the community with a set of ontological tools that can be used both in the development of specific databases and, most importantly, in the construction of decision support systems (DSS) to control these diseases.

1. The problem of vector-borne diseases

Epidemiologists have brought together in one "functional" group a series of diseases of different aetiology and pathogenesis that share one key component: their mode of transmission (see [1] and several chapters of [2] for specific questions addressing insect-borne diseases and their vectors). These diseases are transmitted by the bite of a specific arthropod vector, which is usually (but not exclusively) an insect. The pathogenic agent is usually passed with the saliva transferred during the bite to the potential patient. Two additional characteristics are shared by most vector-borne diseases, namely most people affected live in the tropical regions of the world and, connected to this, the diseases affect mostly populations that are also heavily affected by poverty. The pathogens responsible for these diseases are very diverse, ranging from protozoan parasites (e.g. *Plasmodium* spp. in malaria, *Leishmania* spp. in leishmaniosis) and bacteria (e.g. *Borrelia* spp. in Lyme disease), to worms (e.g. Nematodes in lymphatic filariasis and river blindness) and viruses (e.g. Dengue, Yellow fever). Similarly, the vectors that transmit them are also very diverse and range from mosquitoes (e.g. malaria and Dengue) and flies (e.g. Tsetse in African trypanosomiasis) to kissing bugs (Chagas' disease) and ticks (e.g. Lyme disease). This diversity is shown in Table 1, which lists several diseases along with the corresponding aetiologic agents and arthropod vectors. The great variation in the biology of both pathogens and vectors, and the ensuing differences in the illnesses caused, makes it impossible to address vector-borne diseases as a cohesive clinical entity. Importantly, these difficulties also affect significant aspects such as prevention, epidemiology, therapy, etc.

A common theme, which in a sense unites these diseases, is the fact that their transmission can be blocked if the agents that transmit them, i.e. the arthropod vectors, are removed from the pertinent chain of events [3]. Vector control has therefore historically become a *conditio sine qua non* for the control of these infections [4,5], and this fact has been exemplified by the elimination of malaria from most non-tropical areas of the globe [6]. While leading to about half a billion cases in the tropics and still being responsible for anything between one and three million deaths (mostly children in sub-Saharan Africa) every year, this killer illness has practically disappeared from Europe and North America through intense insecticidal measures aimed at eliminating the anopheline vectors [6]. It should be stressed here that, with the exception of the Yellow fever [7], no vaccine is currently available for any vector-borne disease as an alternative prevention strategy that would act on a different level than that of the actual vector. Prevention focused on the vector includes not only control of insect populations through environmental management or the use of chemicals, but also the protection of individuals through the use of clothing, repellents, nets and screens [8]. In addition, prevention

* Corresponding author at: Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas, N. Plastira 100, 700 13 Heraklion, Greece. Fax: +30 281 0391104.
E-mail address: louis@imbb.forth.gr (C. Louis).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.03.012

Table 1
Some important vector-borne diseases, their pathogens and their arthropod vectors; commonly used synonyms are also listed.

Disease	Pathogen	Vector(s)
(i) Bacterial diseases		
Louse-borne relapsing fever	*Borrelia recurrentis*	Louses
Lyme disease	*Borrelia burgdorferi*	Ticks
Plague	*Yersinia pestis*	Fleas
Tick-borne relapsing fever	var. spirochetes	Ticks
Tularemia	*Francisella tularensis*	Ticks, deer flies
(ii) Viral diseases		
Chikungunya fever	Chikungunya virus	*Aedes aegypti*, *Ae. albopictus*
Dengue fever	DENV	*Ae. aegypti*, *Ae. albopictus*
Eastern equine encephalitis	EEEV	*Aedes* spp., *Coquillettidia* spp. *Culex* spp.
Japanese encephalitis	JEV	*Culex tritaeniorhynchus*
La Crosse encephalitis	La Crosse virus	*Ae. triseriatus*
Saint Louis encephalitis	SLE	*Culex* spp.
West Nile encephalitis	WNV	*Culex* spp.
Western Equine Encephalitis	WEEV	Various mosquito species
Yellow fever	YFV	*Aedes* spp., *Haemagogous* spp.
(iii) Parasitic diseases		
African trypanosomiasis[1]	*Trypanosoma brucei*	*Glossina* spp.[2]
American trypanosomiasis[3]	*Trypanosoma cruzi*	*triatominae*[4]
Leishmaniasis	*Leishmania* spp.	*Lutzomyia* spp.[5], *Phlebotomus* spp.[5]
Lymphatic filariasis[6]	*Wuchereria bancrofti*, *Brugia malayi*	Various mosquito species
Malaria	*Plasmodium* spp.	*Anopheles* spp.
Onchocerciasis[7]	*Onchocerca volvulus*	*Simullium* spp.[8]

Commonly used synonyms: [1]Asian tiger mosquito, [2]sleeping sickness, [3]Tsetse, [4]Chagas' disease, [5]kissing bugs, [6]sand flies, [7]elephantiasis, [8]river blindness, and [9]black flies.

is complemented, in cases in which this is possible, by the use of drugs that block infection in its very initial stages [9].

Although greatly successful in the previous century, insect-control programmes are now immensely obstructed by a variety of factors. These range from community opposition to the widespread use of chemicals [10], to the development of resistance against these very chemicals by the insect vectors to be controlled [11]. Moreover, these problems are aggravated by several additional facts: resistance against drugs is also encountered in the pathogens [12]; vaccine development, if at all possible, is slow [13]; new drug development is not only slow but extremely expensive and the areas affected by the diseases in question are certainly not the ones that can easily spearhead such efforts due to the lack of economic and scientific resources in them [14]. All of the difficulties addressed above have led to a resurgence of vector-borne diseases, which now pose again a threat to more than just the tropical regions of the world [15]. It is therefore of utmost importance to develop innovative strategies for the control of vector-borne diseases. One novel approach is to use information technologies (IT) as a complement to the application of modern biochemical/biological techniques, often based on molecular biology, in the study of the biology of disease vectors. While the latter approaches make use of scientific research products such as whole genome sequences [16,17], transgenesis [18], and the use of other "intelligent" approaches [19], the former can introduce new specific tools, such as databases and DSS, that can be used for a more efficient, and often close-to-the-field management of pertinent disease data, including entomological data.

In this context, several years ago our group embarked on a long project that involves the development of ontologies dealing with vector-borne diseases and their vectors [20,21]. The obvious rationale behind this is the potential of such ontologies to unify the "language" spoken by vector biologists, epidemiologists and other specialists. It should be noted here that the usage of very specific terms and, even worse, jargon often makes it more difficult to obtain a wide understanding of certain terms. For example, the terms "refractory to" or "resistant to", combined with the words "malaria" or "*Plasmodium*" or "infection" are all synonyms. We obviously do not see the need for ontologies restricted only to the actual vectors of the vector-borne diseases but also expanding into the "area" of the diseases and, most importantly, the two have to be interoperable. The ultimate end, thus, is to build a comprehensive ontology for insect-borne diseases that may consist of sub-ontologies, each addressing a specific aspect of the whole. In the frame of the Infectious Disease Ontology project, IDO [22 and http://www.infectiousdiseaseontology.org/Home.html], we initiated this effort focusing on malaria, but we are already expanding this to encompass most other vector-borne diseases as well. The choice of developing the malaria ontology in close partnership with the IDO project was made because of its specific advantages. Having IDOMAL as an extension of such a reference ontology, as opposed to an autonomous approach, allows for a superior interoperability of all individual application ontologies in the context of the greater infectious disease domain. The ontologies that we are working on, some of which are already available and some under development, will be presented below in a summary form.

2. Ontologies and vector-borne diseases: an ephemeral account

There are several aspects of vector-borne diseases that are in need of ontological description; they range from those that deal with the diseases as such (e.g. pathogenesis, clinical aspects, therapy, etc.), to vector biology (physiological processes of the vectors) and to epidemiology and control in the widest sense of the terms (prevention, insect control, etc.). As stated earlier, these aspects are extremely diverse and complex, simply given the multitude of organisms involved (vectors and pathogens in addition to the human host) and the fact that we are often dealing with populations, rather than individuals (additional level of granularity!). The construction of a comprehensive ontology, thus, if at all feasible, must be addressed using a piecemeal approach. It is clear that certain fundamental decisions have to be taken at the initial phases, and an open-ended advance is, in our mind, a must. We therefore decided, early on, that the end product (i) would have to follow the rules set by the OBO Foundry [23] and, if no other reasons dictated a different decision, (ii) should be based on BFO, the basic formal ontology [24,25]. The reasoning behind the decision

was that if long-term interoperability of future databases and IT tools is to be achieved, these two choices are a prerequisite. We considered this choice, though, to be of more relevance to the final goal and we therefore decided to keep a certain degree of flexibility throughout the project until a unified vector-borne disease ontology is fully developed. One example for such a flexible approach is the fact that the ontology of insecticide resistance in mosquitoes, MIRO [26], which we constructed, does not comply with the BFO in its initial versions; rather, it is structured such that it can be adopted, without many problems, by the community that immediately needs to apply it in the field (see below). The MIRO forms the core of the related database on insecticide resistance (IRbase; http://anobase.vectorbase.org/ir/) that we also developed [26], and which was adopted for immediate use by the Regional Office for Africa of the World Health Organization (WHO-AFRO): all field studies that are run under direct or indirect support by WHO-AFRO are asked to submit their data to IRbase. Furthermore, it is planned to move the curation of both database and ontology to an African country with the support of WHO-AFRO. As we have not abandoned the goal of ultimately unifying all ontologies we currently construct, we are in the process of restructuring the MIRO along BFO standards, such that its contents can be later directly imported and incorporated into the comprehensive ontology on vector-borne diseases. At the same time, this will also add accuracy to the ontology (see below). A similar restructuring after the first version was made publicly available occurred with the TGMA, the ontology of the mosquito gross anatomy [20]; we decided for the reasons outlined above, that we should conform the TGMA to the CARO, the Common Anatomy Reference Ontology [27], which is BFO-based. The first version was, thus, retracted and a CARO-compliant TGMA version was then submitted to the OBO Foundry and is now available (http://www.obofoundry.org/cgi-bin/detail.cgi?id=mosquito_anatomy). In contrast to TGMA, TADS, the tick anatomy ontology that we constructed next [20, http://www.obofoundry.org/cgi-bin/detail.cgi?id=tick_anatomy], was directly built as an extension to CARO.

The MIRO has also been already submitted to, and is listed by the OBO Foundry (http://www.obofoundry.org/cgi-bin/detail.cgi?id=mosquito_insecticide_resistance). It consists of four sub-ontologies that cover all aspects of insecticide resistance of mosquito disease vectors, with a special emphasis on fieldwork and monitoring. Thus, although genetic mechanisms of resistance are covered, this is not done in detail, since many of those are processes already covered by the Gene Ontology [28,29]. Furthermore, the MIRO's fifth major component, a geographical one, uses *in toto* the controlled vocabulary Gazeteer (http://darwin.nerc-oxford.ac.uk/gc_wiki/index.php/GAZ_Project) in order to provide IRbase curators with records describing the areas in which data were collected. The MIRO is constantly being updated, upon request, by members of the international community that is involved in the study of insecticide resistance. To help cover the wishes of the pertinent community, we recruited the help of an expert on insecticide resistance who also co-authored the publication of the MIRO [26]. Moreover, all geographical locations reported to IRbase, which so far are spread over 5 continents, are annotated using the GAZ. Should a location not be listed, it is communicated to the curators of the GAZ ontology to be placed at the appropriate position. As mentioned earlier, the fact that MIRO is not BFO compliant renders it easier to be understood by non-experts, but at the same time it looses in accuracy. Table 2 illustrates this point. In MIRO, two steps are enough to define DEF (S,S,S-tributyl phosphorotrithioate), to some extent wrongly, as an insecticidal substance. In contrast, five steps are necessary in IDOMAL, but DEF is defined much more accurately here. As is the case with most, if not all biomedical ontologies, MIRO cannot be considered as complete. More insecticides are being developed, new modes of action are discovered and, unfortunately, and spread of insecticide resistance simply cannot be stopped. MIRO is therefore under steady curation and new versions are made public as soon as the community requires pertinent changes.

Table 2
The table shows the comparison between the non-BFO compliant MIRO and the corresponding term(s) after introduction into IDOMAL and adherence to BFO.

MIRO	IDOMAL
DEF *is_a* synergist	DEF *has_role* insecticide synergist
Synergist *is_a* insecticidal substance	DEF *is_a* chemical compound
	Chemical compound *is_a* abiotic object
	Abiotic object *is_a* object
	Insecticide synergist *is_a* role

Another, originally nameless, ontology that we build covers physiological processes of mosquitoes that are involved in disease transmission. Our original decision to make this an autonomous ontology was later modified, and we are presently in the process of fully incorporating it into IDOMAL (see below). The processes listed do not only address the actual disease transmission, i.e. the interplay between vectors and pathogens but, importantly, also the actual progression of events in the vector. We want to stress that the processes mentioned here are, in their vast majority, processes on the level of the organism and not cellular or sub-cellular ones, such as the ones covered by the GO [28,29]. Moreover, many of these processes are species-specific, and therefore also excluded from the GO, which is focused on processes of a general nature (but see below). Thus, (near) top level classes are, among others, behavior, sensory perception, processes of the immune system and nutrition, all physiological components that directly affect the transmission potential of disease vectors. As an example, when looking at the children terms of "behavior", one will find a line of terms leading through the adult feeding behavior, to entities such as the four phases of "interrupted feeding" (exploratory phase, imbibing phase, probing phase and withdrawal phase). The ontology also covers processes that are not directly "linked" to disease transmission and this, obviously, for reasons of completion. Because of the principle of orthogonality, as was the case with GAZ and MIRO, in all cases in which terms are already covered by established public ontologies we adhere to these, along with their descendants. For that, we search ontologies at the NCBO Bioportal (http://bioportal.bioontology.org/) and directly import relevant hits (IDs and definitions) into our ontologies. This is notably the case for the Biological Processes sub-ontology of the GO, as can be seen in Table 3 that lists a part containing metabolic

Table 3
The table shows a small part of the physiological processes of vectors, described in IDOMAL, that has extensive overlap with the GO-Biological Process sub-ontology (GO IDs are indicated in parentheses). All terms are connected with terms lying above and to the left of them with *is_a* relations. ("IDOMAL:XXXXXXX, no GO term in BP") refers to terms for which no corresponding term is found in the Biological Process sub-ontology. For three cases similar terms, indicated at the bottom part of the table, are found in the sub-ontology of molecular function.

Metabolic process (GO:0008152)
- Catabolic process (GO:0009056)
  - Carbohydrate catabolism (GO:0016052)
    - Glycolysis (GO:0006096)
    - Cleavage by carbohydrases (IDOMAL:0001299, no GO term in BP)
  - Lipid catabolic process (GO:0016042)
    - Fatty acid β-oxidation (GO:0006635)
    - Cleavage by esterases (IDOMAL:0001298, no GO term in BP)[a]
  - Protein catabolic process (GO:0030163)
    - Cleavage by peptidases (IDOMAL:0001300, no GO term in BP)[b]
      - Cleavage by serine proteases (IDOMAL:0001297, no GO term in BP)[c]
- Pigment metabolic process (GO:0042440)

GO terms describing molecular function.
[a] GO:0016788 (hydrolase activity, acting on ester bonds).
[b] GO:0008233 (peptidase activity).
[c] GO:0008236 (serine-type peptidase activity).

processes. In this example, eight of 12 terms do have corresponding terms in the GO, while for the remaining four, three have similar (but obviously not identical) terms in the molecular function sub-ontology of the GO. The decision to use terms (and their IDs) *verbatim* from ontologies that have previously found their way into the public domain is one that we strictly adhere to, as this provides one of the most crucial advantages linked to the usage of ontologies, namely the possibility of cross-talk between databases that share biological metadata. Initially we had decided against using the parent term ID, as we often did not want to import the whole tree associated to the terms. For example, in ChEBI, some insecticides are listed as acaricides, while we consider them to be bona fide insecticides. In the meanwhile we have modified this standard and decided, in most cases, to use the original IDs from OBO Foundry ontologies such as GO and GAZ. This "transcription" is now in progress and soon most such IDs will cease appearing as xrefs.

IDOMAL, is an ontology describing malaria; it is the ontology that we are in the process of populating with terms and this is the actual ontology that we decided to develop in the frame of IDO, and which we plan to expand in the near future in order to cover other vector-borne diseases as well. It is built based on BFO and the IDO reference ontology (http://www.infectiousdiseaseontology.org/IDO_files/IDO_10.08.07.obo.txt), and it is meant to cover malaria on all possible levels. More than 1800 terms currently exist in IDOMAL, even though it cannot be considered as complete. Table 4 shows semi-schematically, the contents of IDOMAL. These obviously include both the clinical aspects of the disease in the widest sense (i.e. including epidemiology, etc.) and the biology of the disease that describes processes and objects of no immediate clinical relevance. We consider as such items (e.g. proteins) involved in the penetration of both mosquito and human/vertebrate cells as well as their interacting partners in the *Plasmodium* parasites. Again, similarly to the case of the ontology of physiological processes, we have taken care to include, wherever possible, direct imports of pre-existing ontologies. This is again the case with terms already described by the GO, but an additional example is the *Plasmodium* parasite life cycle; all stages have cross-references to the, at the moment, inactive *Plasmodium* life cycle ontology.

We have now finished expanding the IDOMAL to cover the immunology of malaria. This now covers the immune responses and the immune state of the vertebrate hosts of the parasites and in particular humans, and it is planned to also include, in the future, the immune responses depicted by anopheline vectors when they are "infected" with *Plasmodium* parasites during a blood meal. While insect immunity's possible interaction with pathogens carried by the vector could be potentially used for intelligent schemes aiming at halting pathogen transmission [30], the human immune system could also be "recruited" in strategies aiming at stopping malaria [31]: it should be stressed again that no vaccines are available for malaria, and therefore any instrument that may be of help in developing them is of utmost importance.

Since both IDO and IDOMAL are still in development, even if at an advanced stage, our ontology may well have to be modified later to take care of discrepancies between the two ontologies, given the fact of their intimate relationship. Table 5 shows identical entities in the IDO and IDOMAL, which obviously share the same definition despite the fact that their names are slightly different; eventually, IDOMAL will switch to IDO's ID number, keeping the alternate name, where necessary, as a synonym. In contrast, Table 6 lists five examples of terms that, although very similar, have a clearly different meaning in the two ontologies; the IDOMAL terms, here, have a more specific meaning, thus at the end the terms will continue appearing with different ID numbers in IDO and IDOMAL.

Table 4
Semi-schematic listing of the contents of IDOMAL. The hierarchy of terms listed here does not correspond to what is to be found in IDOMAL, due to the BFO format followed. Not all classes are listed.

Biology of disease		
	Malaria immunology	
	Malaria pathogens	
	Parasite–vector interactions	
	Parasite–vertebrate interactions	
Clinical features		
	Malaria forms	
	Severe malaria	
	Cerebral malaria	
	Malaria in pregnancy	
	Malaria in children	
Diagnostic procedures		
Epidemiology		
Disease control		
	Malaria eradication	
	Vector control	
	Treatment	
		Chemoprophylaxis
		Chemotherapy
		Immunization
	Treatment of severe malaria	
Parasite biology		
	Plasmodium cycle	
	Plasmodium species	
	Drug resistance	
Vector biology		
	Anopheline species	
	Insecticide resistance	
	Mosquito immunology	
	Transmission-related physiology	
	Population biology	

3. Ontologies and vector-borne diseases: concluding remarks

The ontologies that we are constructing could be described, in a sense, as pure application ontologies that are meant to form the basis for specific tools such as specific databases or decision support systems for various diseases. The need for such tools became apparent immediately after the first working version of the MIRO and its sibling IRbase were made public. Not only did the international community, most prominently WHO-AFRO, immediately decide to adopt both tools, but also already within a few months after the initiation of data submission, there are about 1500 population records in the database. This is about 1400 more samples than what the previous insecticide resistance section in VectorBase carried, the only repository for data of this kind. In addition to databases that are driven by ontologies in an increasing fashion (see for example databases using the ontology-depending schema Chado [32], such as FlyBase [33,34] and VectorBase [35,36]), ontologies are ideal tools for the design and function of intelligent DSS. As a matter of fact, we are aware of at least two such IT tools being developed presently, the malaria decision support system MDSS (http://www.ivcc.com/projects/mdss.htm) and the Dengue decision support system (http://www.ivcc.com/projects/ddss.htm), that are driven in part, by ontologies developed or specifically adapted for that purpose. In cases such as vector-borne diseases, whose control is also hampered by weak infrastructure in endemic countries, these DSS can be used by medical workers and health agencies in remote areas, either for ongoing studies or, most crucially, in cases that need immediate attention [37,38] such as emerging epidemics.

One of the intricacies that we are already faced with is the planned expansion of the malaria-oriented ontologies, to cover many other vector-borne diseases. To understand the magnitude of the challenge one should think of the fact, as stated earlier, that

Table 5
Terms in IDO with their counterparts in the draft IDOMAL.

IDO: term, ID	IDOMAL: term, ID	Common definition
Host role, 408	Host, 0000055	A role borne by an organism by virtue of the fact it provides an environment supportive for the survival or reproduction of an entity of another type
Parasite role, 443	Parasite, 0000995	A symbiont role borne by an organism in virtue of the fact that it derives a growth, survival, or fitness advantage from symbiosis while the other symbiont's growth, survival, or fitness is reduced
Reservoir of infectious agent role, 424	Reservoir, 0000058	A role borne by a material entity by virtue of the fact that it is a habitat in which an infectious agent is persisting and multiplying and from which the infectious agent can be transmitted
Pathogen role, IDO:405	Pathogen, 0000063	A role borne by an object in virtue of the fact that it is sufficiently close to an organism towards which it has the pathogenic disposition to allow processes resulting in disorder to occur
Infectious disease, 436	Infectious disease, 000001051	A disease whose physical basis is an infection
Virulence, 466	Virulence, 0000004	A quality that inheres in an infectious agent and is the degree to which realizations of the infectious disease caused by the infectious agent become severe or fatal
Organism population, 509	Population, 0001254	An aggregate of organisms
Susceptibility, 467	Susceptibility 0001048	A quality that inheres in an entity and is the degree to which it can be harmed by another entity of a certain type
Immunization against infectious agent, 497	Immunization, 0001039	A process by which an organism acquires immunity to an infectious agent

Table 6
Terms in IDO with their counterparts in the draft IDOMAL.

IDO: term, ID	Definition	IDOMAL: term, ID	Definition
Infectious disease prevalence, 485	A quality that inheres in an organism population and is the number of realizations of an infectious disease of a certain type in the population at a specified time	Prevalence of malaria, 0000019	The number of malaria cases existing in a given population at any given time
Infectious disease incidence, 479	A quality that inheres in an organism population and is the number of realizations of an infectious disease of a certain type for which the infectious disease course begins during a specified period of time	Incidence of malaria, 0001243	Prevalence over a stated time period independent of whether the disease resulted from a new infection or not
Infectious disease epidemic, 502	A process in which there is a relatively significant increase in the infectious disease incidence of a certain type of infectious disease, relative to the endemic level of realizations of that infectious disease, in an organism population located in a geographically connected region	Epidemic malaria, 0000116	Spread of malaria across a population beyond what is characterized as endemic
Infectious disease course, 495	A disease course that is the realization of an infectious disease	Progression of malaria, 0000091	All clinical features of malaria from infection to cure or death
Herd immunity to infectious organism, 447	A collective resistance disposition that inheres in an infectious population in virtue of the fact that a sufficient number of members of the population have immunity to an infectious agent	Herd immunity, 00000352	Resistance of a group to a pathogen due to immunity of a large proportion of the group to that pathogen
Acquired immunity to infectious agent, 621	An immunity to infectious agent that inheres in an organism in virtue of lymphocytes and lymphocyte receptors that came into being as a result of a primary immune response	Acquired immunity to malaria, 0000543	Immunity to malaria gradually acquired by infection
Vaccination against infectious agent, 499	An active immunization process that begins with exposure of an organism to a vaccine and results in immunity against an infectious agent	Malaria vaccination, 0000021	The administration of antigenic material from malaria pathogen to produce immunity to malaria

vector-borne diseases represent major threats to public health in wide and ecologically diverse areas of the world, that they are caused by completely different pathogens and that completely different species of vectors transmit them. Thus, the challenge now is how to cover this broad spectrum of facts in a single ontology. There is naturally the possibility of cutting through the Gordian knot by devising separate ontologies for each disease. The counter-argument in this case would be that, brought to an extreme, each pathogen-related malaria form (i.e. tertian, malignant and benign, and quartan, which all have some distinct clinical features) should have its own ontology, similar to the different forms of lymphatic filariasis, which are caused by different species of nematodes but whose clinical aspect differ only slightly. In addition, similarities between these diseases and the agents that transmit them may be obscured if different ontologies were used, and this would certainly have a negative impact on their value in the long term. Therefore, we are still trying to solve the knot in a non-Alexandrian way. By attempting to merge the ontologies *in spe* into one, we can also actively support the rules of the OBO Foundry and provide an example of how the construction of a large and comprehensive ontology can, later on, provide advantages to its users.

Acknowledgments

The work was funded by contract HHSN266200400039C from the National Institute of Allergy and Infectious Diseases in the frame of the VectorBase project and by the BioMalPar and EViMalR European Networks of Excellence supported by European Grants (LSHP-CT-2004-503578 and HEALTH-F3-2009-242095) in the frame of the 6th and 7th Framework Programmes. The authors would like to thank numerous colleagues who helped at different stages of the work, and in particular Drs. John Vontas for his contribution to the MIRO, Frank Collins for his encouragement and support in the frame of VectorBase and Barry Smith and Lindsay Cowell for hosting us in the IDO community.

References

[1] Goddard J. Infectious diseases and arthropods. Totowa, NJ: Humana Press; 2000.
[2] Marquardt WC, Kondratieff BC. Biology of disease vectors. 2nd ed. Burlington, MA: Elsevier Academic Press; 2005.

[3] Hemingway J, Beaty BJ, et al. The innovative vector control consortium: improved control of mosquito-borne diseases. Trends Parasitol 2006;22:308–12.
[4] della Torre A, Arca B, et al. The role of research in molecular entomology in the fight against malaria vectors. Parassitologia 2008;50:137–40.
[5] Peter RJ, Van den Bossche P, et al. Tick, fly, and mosquito control – lessons from the past, solutions for the future. Vet Parasitol 2005;132:205–15.
[6] de Zulueta J. The end of malaria in Europe: an eradication of the disease by control measures. Parassitologia 1998;40:245–6.
[7] Roukens AH, Visser LG. Yellow fever vaccine: past, present and future. Expert Opin Biol Ther 2008;8:1787–95.
[8] Hill J, Lines J, Rowland M. Insecticide-treated nets. Adv Parasitol 2006;61:77–128.
[9] Greenwood BM. Control to elimination: implications for malaria research. Trends Parasitol 2008;24:449–54.
[10] Schapira A. DDT: a polluted debate in malaria control. Lancet 2006;368:2111–3.
[11] Hemingway J, Ranson H. Insecticide resistance in insect vectors of human disease. Annu Rev Entomol 2000;45:371–91.
[12] Laufer MK. Monitoring antimalarial drug efficacy: current challenges. Curr Infect Dis Rep 2009;11:59–65.
[13] Langhorne J, Ndungu FM, et al. Immunity to malaria: more questions than answers. Nat Immunol 2008;9:725–32.
[14] Craft JC. Challenges facing drug development for malaria. Curr Opin Microbiol 2008;11:428–33.
[15] Gubler DJ. Resurgen vector-borne diseases as a global health problem. Emerg Infect Dis 1998;4:442–50.
[16] Holt RA, Subramanian GM, et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 2002;298:129–49.
[17] Nene V, Wortman JR, et al. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 2007;316:1718–23.
[18] James AA. Preventing the spread of malaria and dengue fever using genetically modified mosquitoes. J Vis Exp 2007:231.
[19] Rasgon JL. Using predictive models to optimize Wolbachia-based strategies for vector-borne disease control. Adv Exp Med Biol 2008; 627:114–25.
[20] Topalis P, Tzavlaki C, et al. Anatomical ontologies of mosquitoes and ticks, and their web browsers in VectorBase. Insect Mol Biol 2008;17:87–9.
[21] Topalis P, Lawson D, Collins FH, Louis C. How can ontologies help vector biology? Trends Parasitol 2008;24:249–52.
[22] Cowell LG, Smith B. Infectious disease ontology. In: Sintchenko V, editor. Infectious disease informatics. New York: Springer; 2010. p. 373–96.
[23] Smith B, Ashburner M, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25:251–5.
[24] Simon J, Dos Santos M, Fielding J, Smith B. Formal ontology for natural language processing and the integration of biomedical databases. Int J Med Inform 2006;75:224–31.
[25] Grenon P, Smith B, Goldberg L. Biodynamic ontology: applying BFO in the biomedical domain. Stud Health Technol Inform 2004;102:20–38.
[26] Dialynas E, Topalis P, et al. MIRO and IRbase: IT tools for the epidemiological monitoring of insecticide resistance in mosquito disease vectors. PLOS Negl Trop Dis 2009;3(6):e465. doi:10.1371/journal.pntd.000046.
[27] Haendel MA, Neuhaus F, Osumi-Sutherland D, Mabee P, Mejino Jr JLV, Mugall CJ, et al. CARO – the common anatomy reference ontology. In: Burger A, Davidson D, Baldock R, editors. Anatomy ontologies for bioinformatics: principles and practice. New York: Springer; 2008. p. 327–50.
[28] Ashburner M, Lewis S. On ontologies for biologists: the Gene Ontology – untangling the web. Novartis Found Symp 2002;247:66–80. discussion 80–3, 84–90, 244–52.
[29] Harris MA, Clark J, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 2004;32:D258–61.
[30] Christophides GK, Vlachou D, Kafatos FC. Comparative and functional genomics of the innate immune system in the malaria vector *Anopheles gambiae*. Immunol Rev 2004;198:127–48.
[31] Pierce SK, Miller LH. World Malaria Day 2009: what malaria knows about the immune system that immunologists still do not. J Immunol 2009;182(9):5171–7.
[32] Mungall CJ, Emmert DB. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics 2007;23:i337–46.
[33] Gelbart WM, Crosby M, et al. FlyBase: a *Drosophila* database. The FlyBase consortium. Nucleic Acids Res 1997;25:63–6.
[34] Tweedie S, Ashburner M, et al. FlyBase: enhancing *Drosophila* Gene Ontology annotations. Nucleic Acids Res 2009;37:D555–9.
[35] Megy K, Hammond M, et al. Genomic resources for invertebrate vectors of human pathogens, and the role of VectorBase. Infect Genet Evol 2009;9(3):308–13.
[36] Lawson D, Arensburger P, et al. VectorBase: a data resource for invertebrate vector genomics. Nucleic Acids Res 2009;37:D583–7.
[37] Thomson MC, Connor SJ, et al. The ecology of malaria – as seen from Earth-observation satellites. Ann Trop Med Parasitol 1996;90:243–64.
[38] Coleman M, Sharp B, et al. Developing an evidence-based decision support system for rational insecticide choice in the control of African malaria vectors. J Med Entomol 2006;43:663–8.

MALARIA JOURNAL

RESEARCH Open Access

IDOMAL: an ontology for malaria

Pantelis Topalis[1], Elvira Mitraka[1,2], Ioana Bujila[3], Elena Deligianni[1], Emmanuel Dialynas[1], Inga Siden-Kiamos[1], Marita Troye-Blomberg[3], Christos Louis[1,2*]

Abstract

Background: Ontologies are rapidly becoming a necessity for the design of efficient information technology tools, especially databases, because they permit the organization of stored data using logical rules and defined terms that are understood by both humans and machines. This has as consequence both an enhanced usage and interoperability of databases and related resources. It is hoped that IDOMAL, the ontology of malaria will prove a valuable instrument when implemented in both malaria research and control measures.

Methods: The OBOEdit2 software was used for the construction of the ontology. IDOMAL is based on the Basic Formal Ontology (BFO) and follows the rules set by the OBO Foundry consortium.

Results: The first version of the malaria ontology covers both clinical and epidemiological aspects of the disease, as well as disease and vector biology. IDOMAL is meant to later become the nucleation site for a much larger ontology of vector borne diseases, which will itself be an extension of a large ontology of infectious diseases (IDO). The latter is currently being developed in the frame of a large international collaborative effort.

Conclusions: IDOMAL, already freely available in its first version, will form part of a suite of ontologies that will be used to drive IT tools and databases specifically constructed to help control malaria and, later, other vector-borne diseases. This suite already consists of the ontology described here as well as the one on insecticide resistance that has been available for some time. Additional components are being developed and introduced into IDOMAL.

Background

The failure of the campaign to eradicate malaria about 40 years ago led, among others, to a widespread notion that this disease can simply not be wiped out. This modified the goals of the majority of malaria workers worldwide towards achieving a mitigation of the problem, rather than seeking a final solution. On the other hand it is evident that campaigns based both on novel and traditional concepts, have been highly successful; the key example is the European paradigm of malaria eradication. Moreover, the advent of modern molecular biological techniques, today ranging into genomics and post-genomics, have also provided an impetus towards the development of original and groundbreaking approaches. For example, on the level of malaria entomology, an increased understanding of vector biology in areas such as genetics, molecular and population biology has formed the basis for the design of potential future anti-malarial strategies: these are to be based on the use of genetically modified mosquitoes in order to accomplish a (permanent?) break of transmission cycles.

The recent resurrection of the idea of malaria eradication attributed to Melinda and Bill Gates [1] and immediately adopted by many malariologists, even if only as a "distant dream" [see [2,3]], has moved many research efforts towards schemes aiming at this ultimate goal. The relative optimism with which such a possibility was met was based, among others, on a series of realities that differentiate the present situation from that of the second half of the previous century. These facts primarily include the increased knowledge on all aspects of the biology of the disease, and most importantly, the availability of tools that, fifty years ago, could only be found in the realm of science fiction. Modern information technology (IT) and logistics are good examples of this.

Bioinformatics, as a specialized and logical descendant of computer sciences and IT, evolved mainly due to the development of DNA sequencing and the need to access and understand those primary data. It received its first boost through automated sequencing and it has

* Correspondence: louis@imbb.forth.gr
[1]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas, 700 13 Heraklion, Crete, Greece
Full list of author information is available at the end of the article

progressed even more in order to be able to handle the immense accrual of information that keeps accumulating through genomics in the widest sense. In parallel to the actual sequence analysis, a major part of bioinformatics deals with the development and maintenance of databases in terms of, among others, the organization of their contents, their accessibility, and the cross-talk between them.

It was recently suggested to use ontologies as an efficient instrument to enhance the impact of IT tools in vector biology and malaria entomology [4,5]. This can be achieved by building databases and/or decision support systems driven by wide-ranging ontologies that follow common and established rules. In information science an ontology is a formal representation of the knowledge, which includes the definition of concepts within a given domain as well as the relations between these concepts. In a simplified example, a given biomedical ontology would provide the definition of the term "translation", list its synonym "protein synthesis", and also include its parent (e.g. biological process, metabolic process, gene expression, etc.) and child terms (e.g. initiation, elongation, termination, tRNA aminoacylation, etc.). All of these terms and the relations (in our examples, "*is_a*" and "*part_of*" relations) are well understood by humans but also, most importantly, all computers that have adopted the usage of a given ontology. Although an ontology is often confused with a controlled vocabulary, the latter does not usually use relations and, thus, looses power in terms of computer use: For example, a search of a database driven by the example ontology just mentioned would list, in searches using the string "translation" all items annotated with the term "elongation", since it would be known that the former is a parent term of the latter. It is apparent that if this kind of data exchange and comprehension by information systems can be achieved, a world-wide malaria eradication campaign would greatly benefit from the adoption of standardized ontologies, which would allow for an extensive data exchange across national boundaries and specific projects.

The power of such biomedical ontologies can be best exemplified through the immense success of the Gene Ontology (GO) [6], that has not only allowed improved annotation of experimental data but which, concomitantly, led to an easier comprehensive data mining as well as understanding of molecular biology. VectorBase [7], the database of genomic information on disease vectors, therefore recently incorporated a section on insecticide resistance (IRbase) that fully relies on a specially designed ontology called MIRO [8]. Here, the first version of an ontology for malaria called IDOMAL (Infectious Disease Ontology-MALaria) is described; it is made publicly available in order to seek feedback from the wide community of malariologists.

Methods

The OBOEdit2 software [9], which is freely available for downloading [10] was used for the construction of the IDOMAL. The malaria ontology is based on BFO, the Basic Formal Ontology [11,12] and it follows, in full, the rules set by the OBO Foundry consortium [13]. IDOMAL can be downloaded from Vectorbase [14], and it can be viewed and browsed on line at the NCBO bioportal [15].

Results and Discussion

IDOMAL: the format and the contents

A decision to build an ontology immediately raises some crucial questions that should be answered at the very beginning of the project. Perhaps the first one is the question concerning the primary reasoning that led to the initiation of a project: what is the real need for a given ontology? In the case of the IDOMAL it was clear that there is a vast wealth of knowledge available that could be put to use for the purpose of malaria control by malaria experts, database developers and technicians constructing decision support tools for the disease. Unfortunately, though, the data that range back several decades have been annotated using a multitude of different criteria. This makes it tedious to "unify" the information in order to exploit it to the maximum. We therefore decided to develop a global malaria ontology having in mind two pre-requisites a) the ontology will aim at maximum interoperability and b) it will be amenable to future expansion to encompass aspects that would not be part of it in the initial versions. For several reasons that will be laid out below, it was decided to construct the ontology in the frame of IDO, the Infectious Disease Ontology [16], a loose consortium of research groups aiming at developing ontologies for a variety of infectious diseases that include brucellosis, Dengue fever, infective endocarditis, influenza, tuberculosis and others. Within this consortium, it was decided to initiate the implementation of the project of vector-borne disease ontologies with malaria, indisputably the most important vector-borne disease for global health.

IDO will be a top-level ontology that will form the neutral core for all other sub-domain-specific ontologies to be developed as disease-specific extensions of the core. In this sense, IDO will function similarly to the CARO, the Common Anatomy Reference Ontology [17], that is the nucleus for many anatomical ontologies, and which also served as the basis for the two ontologies built by our group for the anatomy of arthropod disease vectors, TGMA for mosquitoes and TADS for ticks [18].

Similar to the CARO, it has been decided to base IDO and all its "components" on BFO [11,12]. Being domain neutral, BFO allows for a unified treatment of all biomedical items that are to be described in different ontologies. The BFO structure was therefore exploited to incorporate in IDOMAL parts of the previously developed insecticide resistance ontology MIRO [8] without having to go through its intricate restructuring (see below). Given the dependence of IDOMAL on BFO, its architecture and the details of its structure may not be immediately decipherable by non-experts. For example, not only are the distinctions between *process* and *fiat process part* (a processual entity that is part of a process but that does not have *bona fide* beginning and ending that correspond to real discontinuities [19]) or between *disposition*, *state* and *condition* hard to recognize for the uninitiated, but the overall "architecture" may seem complicated in spite of being ontologically correct. The example chloroquine, outlined in figure 1, illustrates this point. The term is found five times in the ontology, once as an *object*, three times as a child of the class *process* and once as a child of the class *role*. Of these five times, only one links chloroquine to the top level, *object*, using a complete *is_a* relation path (figure 1A). There are many similar cases in IDOMAL; a second example can be found within the class *process/process of malaria* where, at a high level, the term *ancillary treatment of malaria* can be identified, containing the terms relating to both severe and uncomplicated malaria. A sibling to this term, though, is *treatment of malaria*, which itself contains the mentioned term *ancillary treatment of malaria*, again with all of its children. Although this may seem illogical at first glance, this is not so: in the former case the indispensable *is_a* path is set up while in the latter the parthood relation is described. Figure 2 shows the contents of the *process* class illustrating the example mentioned while Table 1 shows a summary of the overall contents of the ontology, and the top-level classes that they have been ontologically attributed to. This table is obviously only a schematic, summary representation.

It should be stressed here that as the ontology is a specialized tool and not a simple "dictionary", the immediate advantages of the BFO representation, central for database development, not being immediately noticeable when browsing through the IDOMAL. At this time (version 1.2) it contains 2392 unique terms of which 2377 (> 99.3%) are defined. These terms are distributed in 12 upper level classes, all defined by BFO. Table 1 also shows the number of individual descendants for each of these classes in IDOMAL. The two classes dealing with processes, *process* and *fiat process part*, are made out of a total of 1441 descendants, whereby the latter class contains, almost exclusively, vector-related terms that are listed in Table 2. Another densely populated class is "object" (1148 terms); this is due, on one hand, to the inclusion of such terms as a comprehensive list of hosts, parasites and malaria vectors, all of them under "biotic", and on the other, to the inclusion of "chemical compounds" that includes extensive lists of anti-malarial drugs, insecticides and several proteins, again, from host, parasite and vector. A "similar" class is *object aggregate*, which, among others such as populations-related terms, also lists drug combinations and diagnostic tests. *quality* and *role* are two additional heavily populated classes (253 and 576 terms respectively), while the other classes are presently not very densely populated. Obviously, the total numbers of descendants of the top classes (3699) don't add up to the total number of terms listed in the ontology; the reason for this is that in addition to being connected to their parent directly through an *is_a* relation, several terms are also equally connected to other terms through relations of a different kind such as, for example, *part_of*, *realizes*, *preceded_by* and others, similar to the example illustrated previously in figure 1. Finally, no indications as to the contents for some of the upper classes were included in Table 1; the reason for this is the scarce population with terms. For example, "state" contains only two terms, oostasis and diapause, while "spatiotemporal region" lists the five developmental gates described for follicular development in mosquitoes.

IDOMAL: the disease-related terms

Terms pertinent to the malaria disease as such relate to several distinct aspects of malaria. These obviously include clinical manifestations, therapeutic approaches and epidemiology, but also terms that relate to *Plasmodium* parasites as aetiologic agents. As the aim of IDOMAL is not to build a general disease ontology, the contents focus on terms that are pertinent to malaria as such; nevertheless they are quite complete, aiming at annotating, when need is, all aspects of clinical malaria. Among others, the ontology lists the generic names of all currently available anti-malarial drugs (proprietary names are often listed as synonyms) and commonly used combination therapies; all available diagnostic procedures, including all available rapid diagnostic tests (RDT, as of 2008); therapeutic approaches, including ancillary treatment of malaria. It should be stressed at this point that terms that are already described, defined and given a separate ID number by a higher order ontology such as, for example, IDO or another generic and publically available (open) biomedical ontology, may in the future replace, in full or in part, some of the terms used in IDOMAL. Should this be the case, obviously, the current ID numbers will be kept as cross-references and terms with a slightly different wording

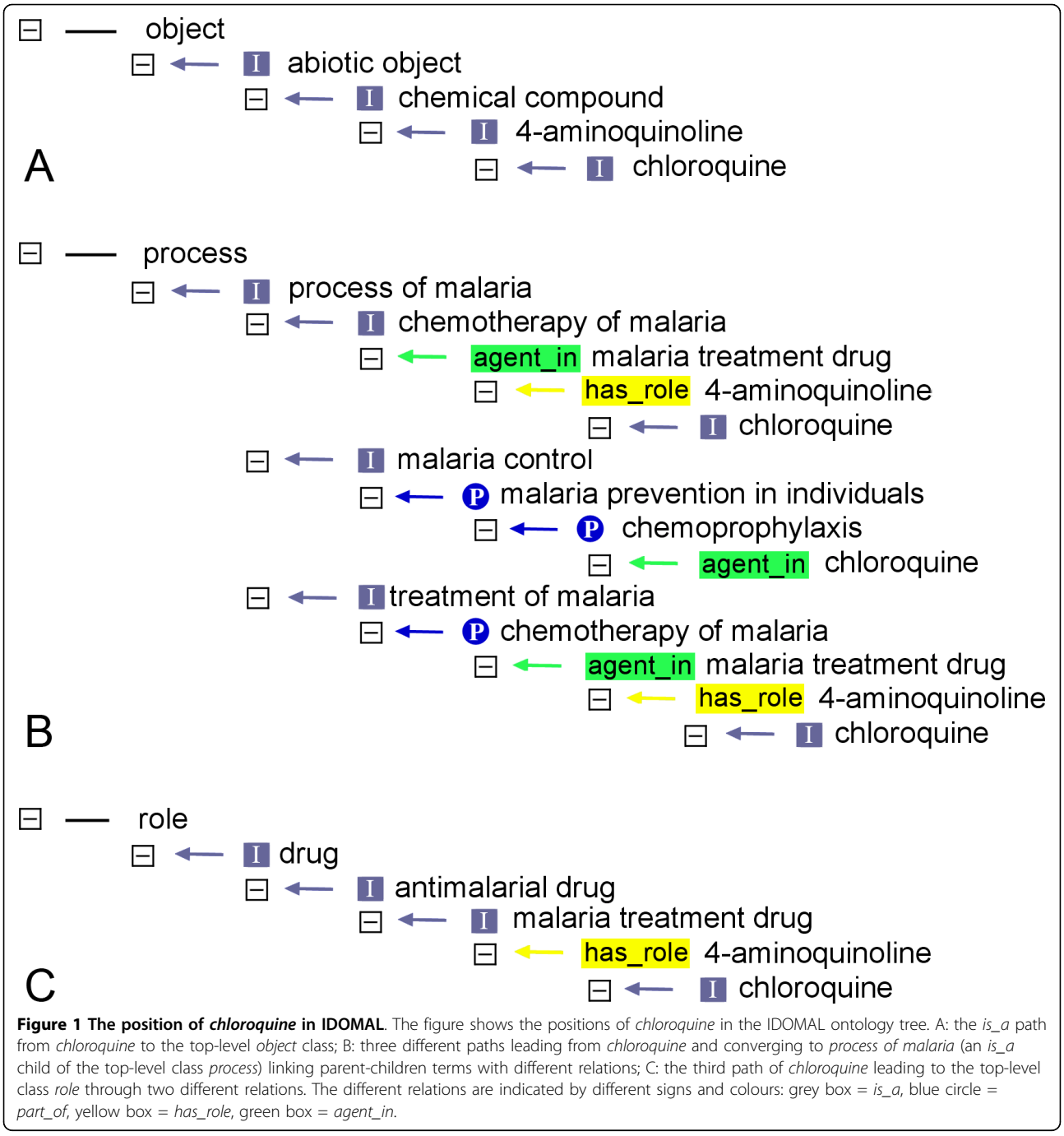


Figure 1 The position of *chloroquine* in IDOMAL. The figure shows the positions of *chloroquine* in the IDOMAL ontology tree. A: the *is_a* path from *chloroquine* to the top-level *object* class; B: three different paths leading from *chloroquine* and converging to *process of malaria* (an *is_a* child of the top-level class *process*) linking parent-children terms with different relations; C: the third path of *chloroquine* leading to the top-level class *role* through two different relations. The different relations are indicated by different signs and colours: grey box = *is_a*, blue circle = *part_of*, yellow box = *has_role*, green box = *agent_in*.

may also be kept as synonyms. A good example for this are the *Anopheles* breeding sites (ontologically: roles!), which have been described by, and are already listed in IDOMAL with the ID of ENVO, the Environment Ontology [20].

In addition to clinical aspects of malaria one additional feature that is also dealt with in IDOMAL is disease biology, including immunology. Here, we were faced with the choice of describing several terms in the ontology in detail or of handling them on a shallow level and relying on a future database for their potential detailed "description". The best cases in point for this are the proteins that have been described as being involved in different crucial host-vector interactions. One example can demonstrate the question faced, as well as the possibility to tackle its solution.

The thrombospondin-related anonymous protein (TRAP) from *Plasmodium* was first identified more than

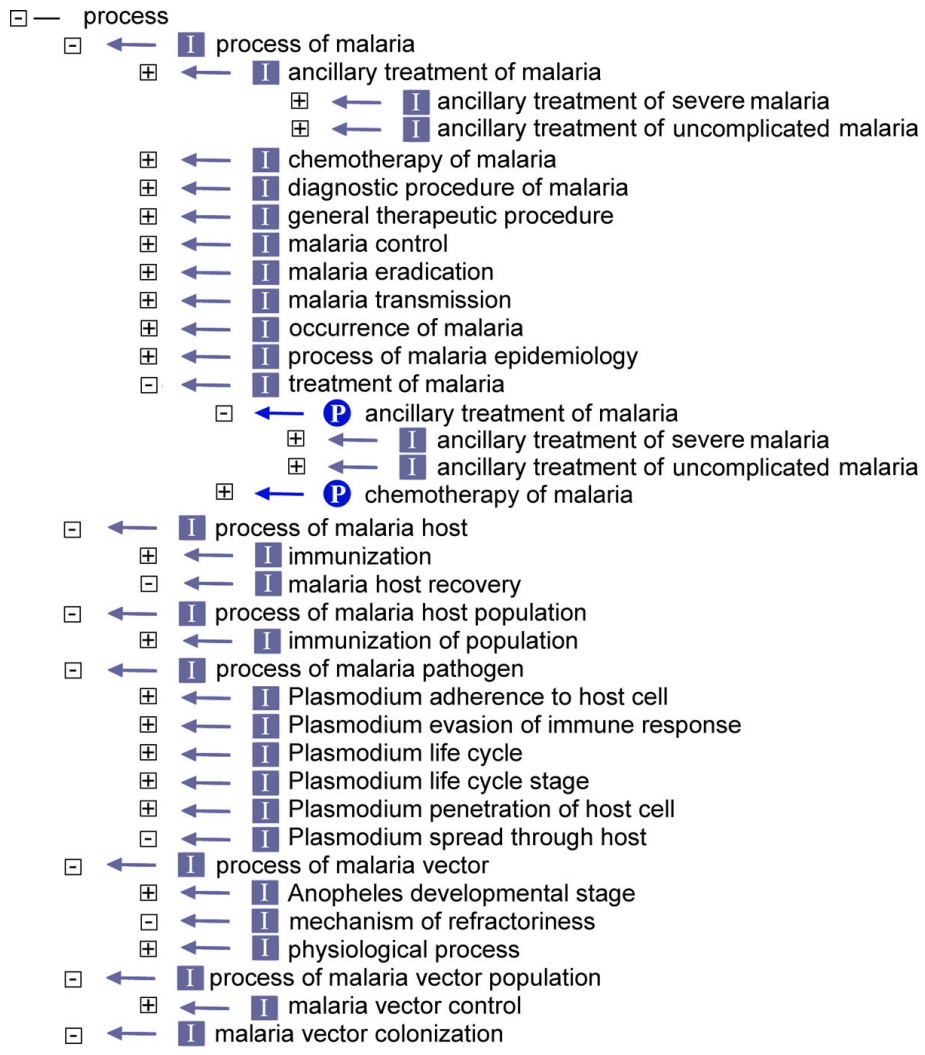


Figure 2 The class *process* in IDOMAL. The upper-most terms within the class *process* are indicated here. Most relations shown are *is_a* relations (grey box) with two exceptions in which *part_of* relations (blue circles) are indicated. The small boxes containing a plus sign signify that the term contains a number of children terms, while the boxes with a minus sign show a "terminal" term that has no children in the ontology.

20 years ago in the human parasite *P. falciparum* [21], and since then in several more *Plasmodium* species. The function of TRAP, a sporozoite transmembrane protein, is to interact with the substrate in the process of motility [22,23]. Moreover, it was later found that TRAP plays an active role in invasion of hepatocytes [24,25]. Should TRAP be included in the malaria ontology as a term? To begin with, there are several *Plasmodium* species for which there is complete lack of information on the respective protein and the gene that encodes it; these species include, unfortunately, even human parasites. Furthermore, specific information on already identified genes/proteins is often stored in databases such as PlasmoDB [26], and so far no need for an annotation in terms of the TRAP name has surfaced or, at least, no such need is described in any major publication. These facts, therefore, would imply that a protein such as TRAP should not be included in a malaria ontology. On the other hand, TRAP has been discussed as a potential vaccine [27,28]. Thus, in a database that deals with vaccines, it is possible that a generic TRAP term might be needed for potential annotation. Similar thoughts concerning other *Plasmodium* proteins which could potentially become pharmacological targets have led us to the inclusion in IDOMAL of TRAP and several more proteins that are potentially involved in vector-parasite and host-parasite interactions. Figure 3 shows the term *TRAP* in two clades of the ontology, a longer one describing the protein in its biological context using four different relations (A, top right) and a short one

Table 1 The upper classes of IDOMAL

Class	Number of terms	Contents summary
condition	45	clinical features of malaria host (e.g. symptoms and signs, etc.)
disposition	77	infectious disease (malaria - > transmission, progression - > clinical manifestation, etc.
fiat process part	121	mostly vector-related "processes"
object	1148	a) abiotic objects (chemical compounds, including insecticides, antimalarials), screening material, environmental/geographic features, etc.
	b) biotic objects (anatomical structures, host -, vector - and parasite species, etc.)	
object aggregate	89	populations (host, vector, parasite), protein complexes
process	1320	processes of malaria, host, vector, parasites, populations, combination therapy, diagnostic tests, etc.
process boundary	2	
quality	253	qualities of malaria, environment, host, vector, parasites and populations
role	576	roles of biological and chemical substances (e.g. drugs, enzymes, factors, etc.), parasites, breeding sites
spatiotemporal region	6	
temporal region	5	

The table lists the upper classes of the ontology and includes, only for the heavily populated ones, a summary of the main contents of the class. The numbers denote the number of terms found in each one of the classes.

showing the *is_a* path from *TRAP* to its uppermost parent class in five steps (B, bottom left). At this moment IDOMAL lists 86 *Plasmodium* proteins but, obviously, the number of such malaria-related molecules will certainly increase in the future as knowledge on the molecular biology of malaria increases and several more will have to be added to IDOMAL.

Similar considerations are valid for terms dealing with malaria immunology, in general, and malaria vaccines, in particular. The rapid progress achieved in these fields, combined with the complicated immunological aspects of the disease [29,30] are principally to blame for an initial relative scarcity of relevant information in IDOMAL. It is noted, though, that attention was focused on immunology-related terms that are "linked" to processes of malaria and not immunity in general, and certainly no description of the immune response in vectors is described yet. Table 3 lists all vertebrate host proteins that are currently listed in the ontology. Several more terms relating to host immunity can be found in both *process of malaria host* and *quality of host*. As stated above, these terms don't include important, yet malaria-unrelated entities.

Finally, the ontology evidently includes a series of terms that pertain to the parasite and its role as a pathogen. These terms deal with the biology of *Plasmodium* (including the aspects just mentioned above), and a brief section that is also in need of expansion deals with the resistance of the parasite against several anti-malarial drugs.

IDOMAL: the vector-related contents

Although it sounds relatively easy to determine what should be included in a disease ontology, the fact that malaria is a three-organism infectious disease complicates matters to some extent. Of course, it is expected that a malaria ontology will include clinical and epidemiological concepts, and naturally all aspects of the biology of the disease are also assumed to form part of IDOMAL. But should vector biology be included or should it form an independent ontology? And if the first part of the question is answered in a positive way, to what extent should vector-related terms be included? A decision was reached to include in IDOMAL all aspects of vector biology that are crucial to malaria transmission and epidemiology. Thus, two such major components were included, insecticide resistance (IR), which is already covered by a specific ontology, MIRO [8], as well as terms pertaining to mosquito physiology. In the case of IR, clearly there is no way that all of its aspects should form an integral part of IDOMAL and it was decided to first importing from MIRO only the mechanisms of resistance as well as the actual insecticides. Therefore, terms relating to pertinent methodology and to populations were omitted from IDOMAL. It is planned, though, to later import MIRO entirely into

Table 2 Physiological processes and "fiat process parts" of malaria vector listed in IDOMAL

process		*fiat process part*	
behavioural process	189	cell-to-cell communication	0
chorion formation	2	descent to the body surface and alighting	4
circulation	0	descent to water surface	0
developmental process	30	development of competence	0
distension of midgut	6	digestion of food	27
egg laying	1	equilibrium during flight	0
endocrine system process	1	exploration and examination of body surface	33
excretion	15	flight orientation	0
fertilization	0	food ingestion	6
formation of ovarian follicles	0	formation of assembly	9
formation of peritrophic matrix	2	gliding	0
growth	4	hovering	0
immune system process	17	internalization of vitellogenin	1
muscular system process	25	long-range approach	30
nervous system process	5	organelle synthesis in midgut cells	5
nutritional process	1	ovarian cycle	27
previtellogenic development	2	ovarian developmental stages (Christophers)	10
regulation of biological process	1	ovarian developmental stages (Troy et al.)	9
release of 20-hydroxyecdysone	0	oviposition	0
reproduction	90	persistent locomotion	0
respiration	5	process of oogenesis	40
response to stimulus	87	process of ovulation	1
rRNA synthesis in oocyte and nurse cells	0	production of digestive enzymes	4
saliva secretion	0	senses and flight response during mating	29
secretion of peritrophic matrix in larvae	0	short-range approach to the host	5
sensory perception	21	skin-hopping	0
stimulation of vitellogenin synthesis	0		
termination stage	0		
untrastructural change in the trophocyte	6		
vector metabolic process	32		
vitellogenesis	10		
vitellogenic stage	1		
vitellogenin synthesis	2		

The table lists, alphabetically, physiological processes and fiat process parts of malaria vectors that are currently listed in IDOMAL. The numbers refer to the numbers of individual child terms of a given term. When a zero (0) is indicated, the term in the table has no children listed in the ontology.

IDOMAL, when the former ontology has been re-organized according to the BFO format. It was also chosen to omit, in the first version, terms relating to mosquito immunity [31,32] although, again, these will be included those in a future release. For the time being, for both imported classes of terms the original ID numbers that have been assigned through their inclusion in MIRO were kept, this way allowing for an unambiguous identification of the various items and avoiding later confusion. In other words, the use of both IDOMAL and MIRO by an IT tool or a database to be developed in the future would not be faced with problems of disambiguation of the terms. Similar to what was done for terms imported from MIRO, in all cases in which a term was imported from an existing ontology (e.g. physiological process are covered, in part, by the GO [33]) the original IDs were kept (see below).

A series of terms dealing with vector physiology were incorporated in IDOMAL, which in their majority concerned processes in mosquitoes that are related to transmission, directly or indirectly. Thus, larval life is only poorly addressed; in contrast, behavioural parameters such as host seeking or blood meal-related processed are described in more detail. More than 600 fully defined terms make up this part of the ontology. Table 2 lists the categories of such terms that can be found in the ontology, i.e. the upper levels of the corresponding section of IDOMAL as well as the number of terms in each one of the classes. As stated earlier, it should be noted that the number of terms indicated does not

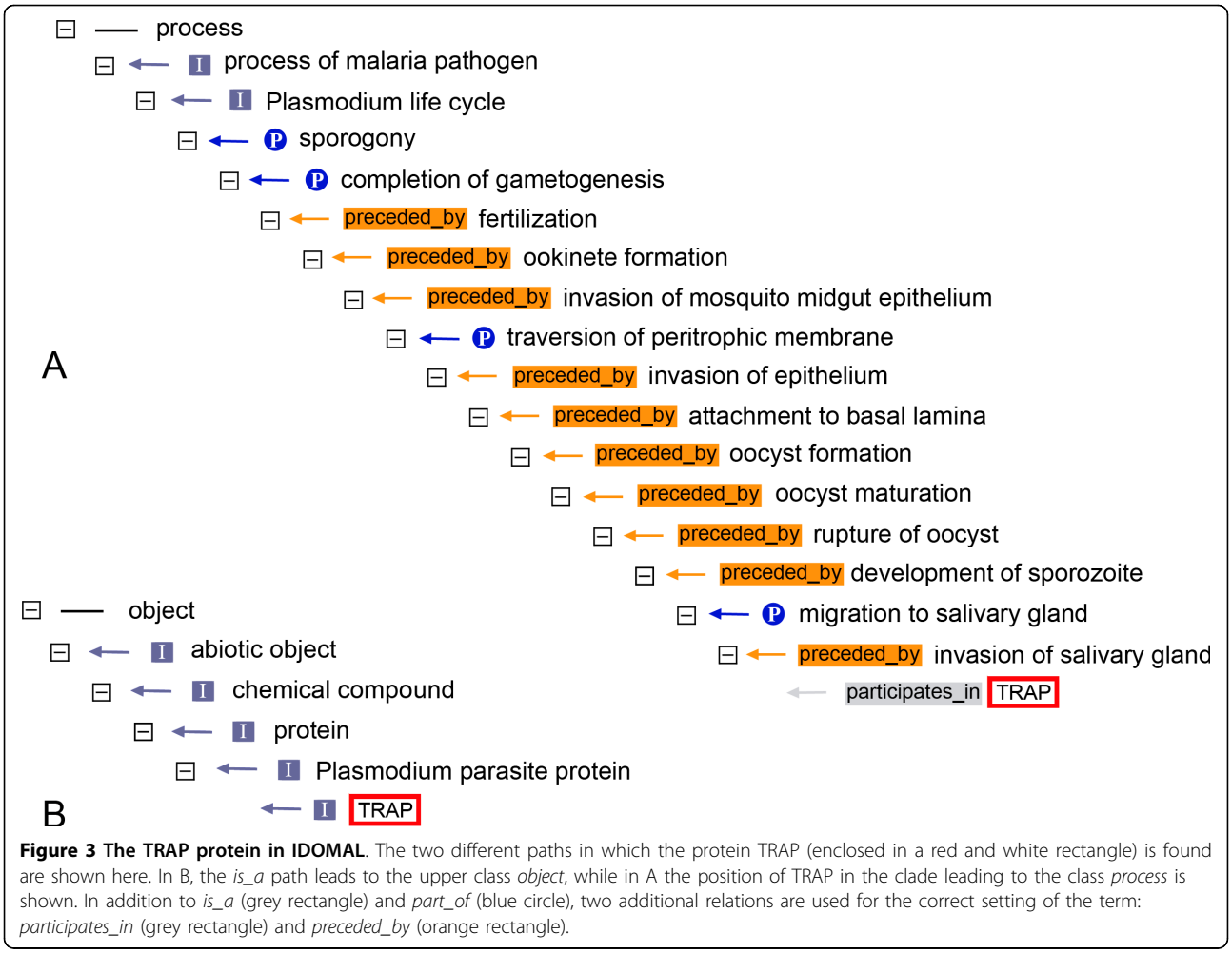


Figure 3 The TRAP protein in IDOMAL. The two different paths in which the protein TRAP (enclosed in a red and white rectangle) is found are shown here. In B, the *is_a* path leads to the upper class *object*, while in A the position of TRAP in the clade leading to the class *process* is shown. In addition to *is_a* (grey rectangle) and *part_of* (blue circle), two additional relations are used for the correct setting of the term: *participates_in* (grey rectangle) and *preceded_by* (orange rectangle).

reflect unique terms, for the additional reason that certain terms that can be found as parts of different processes. Importantly, some of the processes described in the ontology don't refer to physiological processes of the vectors in a strict sense but, rather, they relate to the interactions between the vector and the vertebrate host of *Plasmodium* as well as the vector and *Plasmodium* itself. Furthermore, some of the processes (and "fiat process parts") described in IDOMAL can also be found in the GO listed as biological processes. In all cases in which a 1:1 tautology exists, the GO ID has been used to identify the terms in the malaria ontology. One needs to differentiate, though, between processes and functions, as does the GO in its division into three sub-ontologies. Therefore, IDOMAL lists a process called "cleavage by peptidase", while the GO includes the molecular function "peptidase activity". In this case, and some other similar ones, the GO term is cross-referenced but not directly imported.

As is true for the remaining IDOMAL, the terms addressing the mosquito physiological processes are certainly not exhausted, and more terms can (and will) be added in the future. This will certainly be the case when the ontology is expanded (or, potentially, entirely reorganized) to include other vector-borne diseases; virus-host interactions are, here, the best example.

Finally, a series of terms relating to vector control as such are also included in IDOMAL. It is expected that this kind of terms would be of importance given the stated possibility of malaria eradication efforts. It is worth mentioning here that throughout IDOMAL, as is the case for the malaria vertebrate host, terms that can be unambiguously linked to either vector or vector population are listed separately.

Conclusions

The aim was to produce a tool that will be useful to the malaria community working towards effectively reducing the global malaria burden. Ontologies are such tools, as they provide the community with a common language that is equally well understood by computers and dedicated software. Thus, if/when widely accepted,

Table 3 Malaria-related vertebrate host proteins listed in IDOMAL

C3b			
CD36			
complement receptor 1			
defensin			
granzyme B			
human actin			
human ankyrin			
human band 3 protein			
human band 4.1 protein			
human Duffy blood group antigen			
human glucose-6-phosphate dehydrogenase			
human glycophorin A			
human glycophorin C			
human haemoglobin			
	variant haemoglobin		
		haemoglobin C	
		haemoglobin E	
		haemoglobin S	
		thalassaemia-related haemoglobin	
			alpha thalassaemia-related haemoglobin
			beta thalassaemia-related haemoglobin
	wild type haemoglobin		
human spectrin			
immunoglobulin			
	immunoglobulin E		
	immunoglobulin G		
		immunoglobulin G1	
		immunoglobulin G3	
	immunoglobulin M		
interferon gamma			
interleukin 10			
interleukin 12			
interleukin 13			
interleukin 2			
interleukin 4			
lysozyme			
perforin			
toll like receptor 2			
toll like receptor 9			
tumor necrosis factor-alpha			

The table lists, alphabetically, all malaria-related vertebrate host proteins that are currently listed in IDOMAL. Proteins that are found tab-shifted rightwards in any line of the table are *is_a* children of the respective higher order term.

ontologies provide the means to expand the information through interoperability and mutual understanding of database annotations. This possibility clearly enhances the usefulness of databases: rather than simple repositories, they advance to the level of complex tools. In spite of the fact that more than two thousand terms are included in IDOMAL, the fact that this first, working version of the malaria ontology is far from being complete has, indeed, to be emphasized. This is obviously the case with any ontology that expands and changes to satisfy advances such as scientific findings and novel ideas in any given field or domain. Moreover, mistakes and omissions are always part of such an effort, and the malaria community is invited, and urged, to provide constructive feedback. It may be a fact that in its present form, the ontology may be leaning slightly towards vectors than towards the other two key players of malaria, the vertebrate host and the parasite. An ontology is bound to constantly expand as new terms appear. Moreover, both for any expansion as well as for the

optimal description of existing terms, an input from the community is a *conditio sine qua non*.

This ontology is freely available to everyone wishing to use it. The only condition linked to its usage is that, following the rules established by the OBO Foundry, if this ontology is to be changed in any sense by a user for any purpose, the name IDOMAL can no longer be used. We hope that in the near future we will be able to provide the users from the malaria community with a much better product that will greatly rely on their own criticism.

List of Abbreviations used

BFO: Basic Formal Ontology; IDO: Infectious Disease Ontology; IT: Information tool; GO: gene Ontology; MIRO: Mosquito Insecticide Resistance Ontology; URL: Uniform Resource Locator; OBO: Open Biomedical Ontologies; NCBO: National Center for Biomedical Ontologies; CARO: Common Anatomy Reference Ontology; TGMA: Mosquito Gross Anatomy Ontology; TADS: Tick Gross Anatomy Ontology; ID: Identification Number; TRAP: thrombospondin-related anonymous protein; IR: Insecticide Resistance.

Acknowledgements

The authors would like to thank Drs. Lindsay G. Cowell and Barry Smith for bringing to life the IDO project, accepting us as their partners and discussing aspects of IDO and the individual components; Dr. Alan Ruttenberg for critically evaluating IDOMAL at its beginning stages; Dr. Frank Collins for his encouragement and support in the frame of VectorBase. The work was supported by the NIAID (contracts HHSN266200400039C and HHSN272200900039C to the core VectorBase project) and, in part, by the BioMalPar and the EVIMALAR networks of excellence (contracts LSHP-CT-2004-503578 and HEALTH-F3-2009-242095).

Author details

[1]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology-Hellas, 700 13 Heraklion, Crete, Greece. [2]Department of Biology, University of Crete, 711 10 Heraklion, Crete, Greece. [3]Department of Immunology, Stockholm University, SE-106 91 Stockholm, Sweden.

Authors' contributions

PT researched and constructed a major part of the ontology, reviewed the physiological processes of the vectors, reviewed the entire ontology and discussed open questions with representatives of the community. EM researched and constructed the part of the ontology dealing with the physiological processes of the vectors. IB researched and constructed the part of the ontology dealing with malaria immunology. EID researched and constructed the part of the ontology dealing with the vector-parasite and host-parasite interactions. EmD, ISK and MTB reviewed the ontology. CL conceived the project, obtained funding, constructed the medical part of the ontology, supervised the study and wrote the paper. All authors read, edited and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 21 April 2010 Accepted: 10 August 2010
Published: 10 August 2010

References

1. Roberts L, Enserink M: **Malaria. Did they really say ... eradication?** *Science* 2007, **318**:1544-1545.
2. Mendis K, Rietveld A, Warsame M, Bosman A, Greenwood B, Wernsdorfer WH: **From malaria control to eradication: The WHO perspective.** *Trop Med Int Health* 2009, **14**:802-809.
3. Greenwood B: **Can malaria be eliminated?** *Trans R Soc Trop Med Hyg* 2009, **103(Suppl 1)**:S2-5.
4. Topalis P, Lawson D, Collins FH, Louis C: **How can ontologies help vector biology?** *Trends Parasitol* 2008, **24**:249-252.
5. Topalis P, Dialynas E, Mitraka E, Deliyanni E, Siden-Kiamos I, Louis C: **A set of ontologies to drive tools for the control of vector-borne diseases.** *J Biomed Inform* 2010.
6. Gene Ontology Consortium: **The Gene Ontology project in 2008.** *Nucleic Acids Res* 2008, **36**:D440-D444.
7. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, Campbell KS, Christophides GK, Dialynas E, Hammond M, Hill CA, Konopinski N, Lobo NF, MacCallum RM, Madey G, Megy K, Meyer J, Redmond S, Severson DW, Stinson EO, Topalis P, Birney E, Gelbart WM, Kafatos FC, Louis C, Collins FH: **VectorBase: a data resource for invertebrate vector genomics.** *Nucl Acids Res* 2009, **37**:D583-587.
8. Dialynas E, Topalis P, Vontas J, Louis C: **MIRO and IRbase: IT Tools for the Epidemiological Monitoring of Insecticide Resistance in Mosquito Disease Vectors.** *PLoS Negl Trop Dis* 2009, **3**:e465.
9. Day-Richter J, Harris MA, Haendel M, Gene Ontology OBO-Edit Working Group, Lewis S: **OBO-Edit–an ontology editor for biologists.** *Bioinformatics* 2007, **23**:2198-2200.
10. **Browse Gene Ontology Files on SourceForge.net.** [http://sourceforge.net/projects/geneontology/files/OBO-Edit%202%20%5Bcurrent%20release%5D/oboedit2.0/.].
11. Simon J, Dos Santos M, Fielding J, Smith B: **Formal ontology for natural language processing and the integration of biomedical databases.** *Int J Med Inform* 2006, **75**:224-231.
12. Grenon P, Smith B, Goldberg L: **Biodynamic ontology: applying BFO in the biomedical domain.** *Stud Health Technol Inform* 2004, **102**:20-38.
13. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, OBI Consortium, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S: **The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.** *Nat Biotechnol* 2007, **25**:1251-1255.
14. **IDOMAL.** [http://anobase.vectorbase.org/idomal/IDOMAL.obo].
15. **NCBO BioPortal: Malaria Ontology.** [http://bioportal.bioontology.org/visualize/40463].
16. **Infectious Disease Ontology.** [http://www.infectiousdiseaseontology.org/Home.html].
17. Haendel MA, Neuhaus F, Osumi-Sutherland O, Mabee PM, Mejino JLV Jr, Mungall CJ, Smith B: **CARO – The Common Anatomy Reference Ontology.** *Anatomy Ontologies for Bioinformatics* London: SpringerBurger A, Davidson D, Baldock R 2007, 327-349.
18. Topalis P, Tzavlaki C, Vestaki K, Dialynas E, Sonenshine DE, Butler R, Bruggner RV, Stinson EO, Collins FH, Louis C: **Anatomical ontologies of mosquitoes and ticks, and their web browsers in VectorBase.** *Insect Mol Biol* 2008, **17**:87-89.
19. Smith B, Grenon P: **The cornucopia of formal-ontological relations.** *Dialectica* 2004, **58**:279-296.
20. **The Environment Ontology (EnvO)-Linking Environmental Data.** [http://www.environmentontology.org/].
21. Robson KJ, Hall JR, Jennings MW, Harris TJ, Marsh K, Newbold CI, Tate VE, Weatherall DJ: **A highly conserved amino-acid sequence in thrombospondin, properdin and in proteins from sporozoites and blood stages of a human malaria parasite.** *Nature* 1988, **335**:79-82.
22. Spaccapelo R, Naitza S, Robson KJ, Crisanti A: **Thrombospondin-related adhesive protein (TRAP) of Plasmodium berghei and parasite motility.** *Lancet* 1997, **350**:335.
23. Sultan AA, Thathy V, Frevert U, Robson KJ, Crisanti A, Nussenzweig V, Nussenzweig RS, Ménard R: **TRAP is necessary for gliding motility and infectivity of plasmodium sporozoites.** *Cell* 1997, **90**:511-522.
24. Akhouri RR, Sharma A, Malhotra P, Sharma A: **Role of Plasmodium falciparum thrombospondin-related anonymous protein in host-cell interactions.** *Malar J* 2008, **7**:63.
25. Morahan BJ, Wang L, Coppel RL: **No TRAP, no invasion.** *Trends Parasitol* 2009, **25**:77-84.
26. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS, Heiges M, Innamorato F, Iodice J, Kissinger JC,

Kraemer E, Li W, Miller JA, Nayak V, Pennington C, Pinney DF, Roos DS, Ross C, Stoeckert CJ Jr, Treatman C, Wang H: **PlasmoDB: a functional genomic database for malaria parasites.** *Nucleic Acids Res* 2009, **37**: D539-543.

27. Dolo A, Modiano D, Doumbo O, Bosman A, Sidibé T, Keita MM, Naitza S, Robson KJ, Crisanti A: **Thrombospondin related adhesive protein (TRAP), a potential malaria vaccine candidate.** *Parassitologia* 1999, **41**:425-428.
28. Epstein JE, Giersing B, Mullen G, Moorthy V, Richie TL: **Malaria vaccines: are we getting closer?** *Curr Opin Mol Ther* 2007, **9**:12-24.
29. Artavanis-Tsakonas K, Tongren JE, Riley EM: **The war between the malaria parasite and the immune system: immunity, immunoregulation and immunopathology.** *Clin Exp Immunol* 2003, **133**:145-152.
30. Hviid L: **Naturally acquired immunity to Plasmodium falciparum malaria in Africa.** *Acta Trop* 2005, **95**:270-275.
31. Dimopoulos G: **Insect immunity and its implication in mosquito-malaria interactions.** *Cell Microbiol* 2003, **5**:3-14.
32. Alphey L: **Natural and engineered mosquito immunity.** *J Biol* 2009, **8**:40.
33. Gene Ontology Consortium: **The Gene Ontology in 2010: extensions and refinements.** *Nucleic Acids Res* 2010, **38**:D331-335.

doi:10.1186/1475-2875-9-230
Cite this article as: Topalis *et al.*: IDOMAL: an ontology for malaria. *Malaria Journal* 2010 **9**:230.