

ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Ηλεκτρονικών Μηχανικών & Μηχανικών Υπολογιστών

**Μετατροπή, Ενοποίηση και Οπτικοποίηση
Μεταπληροφορίας
για Γονιδιακά Ρυθμιστικά Δίκτυα**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ιωάννη Μαράκη

Εξεταστική Επιτροπή

Επ. Καθ. Μιχαήλ Γ. Λαγουδάκης (επιβλέπων)

Αν. Καθ. Ευριπίδης Πετράκης

Καθ. Μιχάλης Ζερβάκης

Χανιά 2009

Περίληψη

Ένα από τα πιο σημαντικά προβλήματα στον τομέα της βιοπληροφορικής είναι η μεγάλη ετερογένεια των δεδομένων που μοντελοποιούν γενομικές οντότητες, καθώς προέρχονται από διαφορετικές πηγές, ακολουθούν διαφορετικές μοντελοποιήσεις και προορίζονται για διαφορετικού είδους επεξεργασία. Ένα τέτοιο παράδειγμα είναι τα Γενομικά Ρυθμιστικά Δίκτυα (ΓΡΔ), τα οποία περιγράφουν τις αλληλεπιδράσεις των γονιδίων κατά τη διάρκεια βιολογικών διεργασιών, και τα προγράμματα οπτικοποίησης βιολογικών (και όχι μόνο) γράφων. Στην παρούσα εργασία υλοποιούμε μία μέθοδο μετατροπής των ΓΡΔ από το πρότυπο KGML, όπως παρέχονται από το ίδρυμα KEGG (Kyoto Encyclopedia of Genes and Genomes), στο πρότυπο XGMML, το οποίο είναι ένα καθιερωμένο πρότυπο περιγραφής γράφων και της μεταπληροφορίας τους, ώστε να είναι δυνατή η οπτικοποίησή τους σε εφαρμογές ευρείας χρήσης, όπως το Cytoscape. Επίσης, υλοποιούμε μία μέθοδο συνένωσης και σύμπτυξης ΓΡΔ που οδηγεί σε καλύτερη οπτικοποίηση πολύπλοκων αλληλεπιδράσεων ακόμη και μεταξύ γονιδίων σε διαφορετικά ΓΡΔ. Τέλος, παρέχουμε ένα γενικότερο πλαίσιο διαχείρισης ΓΡΔ στο Cytoscape για εφαρμογή γραφοθεωρητικών μεθόδων, όπως η εύρεση ελάχιστου μονοπατιού, και διασύνδεση των κόμβων των ΓΡΔ με δικτυακές βάσεις γενομικής πληροφορίας.

Περιεχόμενα

1. Εισαγωγή.....	7
1.1 Περιγραφή της εργασίας – Σκοποί.....	7
1.2 Δομή της εργασίας.....	8
2. Επεξεργασία και μοντελοποίηση των Γενομικών Ρυθμιστικών Δικτύων.....	9
2.1 Βασικές έννοιες βιολογίας.....	9
2.2 Δίκτυα γονιδιακών αλληλεπιδράσεων.....	10
2.3 KEGG και KGML.....	12
2.3.1 Ασυμβατότητες μεταξύ KGML και HTML.....	17
2.4 Επεξεργασία KGML αρχείων.....	18
2.4.1 Parsing των KGML αρχείων.....	19
2.4.2 Αποθήκευση των ΓΡΔ σε Δομές Δεδομένων.....	20
2.4.3 Ανάλυση σε κλάσεις.....	21
2.5 Συνένωση ΓΡΔ.....	22
2.5.1 Βιολογική σημασία της συνένωσης.....	22
2.5.2 Συνένωση δύο ΓΡΔ.....	23
2.5.3 Παράδειγμα συνένωσης ΓΡΔ.....	26
2.5.4 Παρατηρήσεις σχετικά με τη συνένωση ΓΡΔ.....	28
3. Οπτικοποίηση των Γενομικών Ρυθμιστικών Δικτύων.....	31
3.1 Cytoscape.....	31
3.2 XGMML.....	33
3.2.1 Μετατροπή από KGML σε XGMML.....	34
4. Εφαρμογές και παραδείγματα.....	37
4.1 Παράδειγμα μετατροπής ΓΡΔ.....	37
4.2 Χειρισμός γράφου από το Cytoscape.....	41
4.3 Συνένωση δύο ή περισσότερων ΓΡΔ.....	42
4.4 Εύρεση ελάχιστου μονοπατιού.....	45
5. Επίλογος.....	47
Αναφορές.....	49
Βιβλιογραφία.....	49

1. Εισαγωγή

Το τελευταίο διάστημα, η επιστημονική περιοχή της βιολογίας, της επιστήμης των υπολογιστών, καθώς και των μαθηματικών έχουν συγκλίνει προς ένα κοινό θεματικό πλαίσιο, αυτό της βιοπληροφορικής. Η βιοπληροφορική μπορεί να οριστεί ως ο τομέας της θετικής επιστήμης ο οποίος μελετάει τη συμπεριφορά βασικών μονάδων της βιολογικής λειτουργίας μέσω υπολογιστικών μεθόδων. Στη κορυφή του ενδιαφέροντος της ερευνητικής κοινότητας σήμερα βρίσκονται τα Γονιδιακά Ρυθμιστικά Δίκτυα (Gene Regulatory Networks) ή αλλιώς ΓΡΔ [1]. Τα ΓΡΔ είναι γράφοι οι οποίοι περιγράφουν τις αλληλεπιδράσεις των γονιδίων κάτω από πολύ συγκεκριμένες συνθήκες. Κορυφαίοι ερευνητικοί οργανισμοί και ιδρύματα ασχολούνται όχι μόνο με την σύνθεση των ΓΡΔ, αλλά και με τη μετέπειτα επεξεργασία, διαχείριση και μοντελοποίησή τους. Ένας από αυτούς είναι ο KEGG (Kyoto Encyclopedia of Genes and Genomes) [2]. Το KEGG δεν παρέχει μόνο μια μεγάλη συλλογή από ΓΡΔ, αλλά παρέχει και ένα μοντέλο ανταλλαγής δικτύων βασισμένο στο γενικότερο μοντέλο ανταλλαγής δεδομένων XML. Το μοντέλο αυτό περιέχει με απλό τρόπο τις δικτυακές δομές (κόμβοι, ακμές) του δικτύου, καθώς και τη μετα-πληροφορία τόσο των κόμβων (γονιδίων) όσο και των ακμών (συσχετίσεις). Το KEGG επίσης παρέχει μέσα από την ιστοσελίδα του τη δυνατότητα για οπτικοποίηση του γράφου. Το πρόβλημα είναι ότι αυτή η οπτικοποίηση βασίζεται σε στατικές HTML σελίδες, με συνέπεια να μην είναι δυνατή η δυναμική διαχείριση των ΓΡΔ ως γράφων. Ως δυναμική διαχείριση εννοούμε την οπτικοποίηση του γράφου με διάφορους τρόπους, την εφαρμογή γραφοθεωρητικών μεθόδων και τη διασύνδεση με πηγές βιολογικών δεδομένων.

1.1 Περιγραφή της εργασίας – Σκοποί

Στη παρούσα εργασία παρέχουμε έναν αλγόριθμο επεξεργασίας μέσω parsing της δομής και της συνολικής πληροφορίας που υπάρχουν στα ΓΡΔ, όπως παρέχεται από το KEGG. Η πληροφορία αυτή αποθηκεύεται σε XGML [3] το οποίο είναι ένα διεθνώς αναγνωρισμένο πρότυπο για την αποθήκευση και διαχείριση γράφων. Ο τρόπος επεξεργασίας είναι τέτοιος ώστε να παρέχεται ένα προγραμματιστικό εργαλείο περαιτέρω επεξεργασίας των γράφων. Ως παράδειγμα χρήσης υλοποιήσαμε μία μέθοδο συνένωσης γράφων, όπου διαφορετικοί γράφοι που περιέχουν κάποιους

κοινούς κόμβους συνενώνονται σε έναν γράφο. Στη συνέχεια επεξεργαζόμαστε περαιτέρω τους γράφους με τη βοήθεια της ειδικής εφαρμογής Cytoscape [4], η οποία είναι ειδικό, ελεύθερα διαθέσιμο λογισμικό για την οπτικοποίηση και επεξεργασία γράφων κυρίως βιολογικού ενδιαφέροντος. Με το Cytoscape, εκτός από την οπτικοποίηση μεμονωμένων ή συνενωμένων γράφων, μπορέσαμε να διασυνδέσουμε την πληροφορία που περιέχουν οι κόμβοι με γνωστές δικτυακές πηγές βιολογικού περιεχομένου και να εφαρμόσουμε γραφοθεωρητικές μεθόδους, όπως είναι η εύρεση ελάχιστου μονοπατιού.

1.2 Δομή της εργασίας

Η παρούσα διπλωματική εργασία είναι δομημένη σε κεφάλαια ως εξής:

Στο δεύτερο κεφάλαιο παρουσιάζουμε μία εισαγωγή στα Γενομικά Ρυθμιστικά Δίκτυα, στο περιεχόμενο και τη σημασία τους. Αναλύουμε τον τρόπο που τα χειριζόμαστε προγραμματιστικά και με ποιο τρόπο επιτυγχάνουμε την συνένωσή τους.

Στο τρίτο κεφάλαιο παρουσιάζουμε με πιο τρόπο μπορούμε να οπτικοποιήσουμε μέσω ολοκληρωμένων εφαρμογών τα ΓΡΔ, πως συνδέονται οι κόμβοι με άλλες πηγές βιολογικής πληροφορίας και πως μπορούμε να εφαρμόσουμε γραφο-θεωρητικούς αλγόριθμους που έχουν αναπτυχθεί από τρίτους για την περαιτέρω ανάλυσή τους.

Στο τέταρτο κεφάλαιο παρουσιάζουμε με λεπτομέρεια την υλοποίηση μας και παραδείγματα χρήσης.

Τέλος, στο πέμπτο κεφάλαιο κάνουμε μία συνοπτική παρουσίαση των συμπερασμάτων που εξήχθησαν από την εργασία.

2. Επεξεργασία και μοντελοποίηση των Γενομικών

Ρυθμιστικών Δικτύων

Στην αρχή του 21^{ου} αιώνα διάφορα σημαντικά γεγονότα καθόρισαν τη δημιουργία της σύγχρονης επιστημονικής περιοχής της βιοπληροφορικής. Ένα από αυτά είναι η αποκωδικοποίηση του ανθρώπινου γονιδιώματος τον Απρίλιο του 2003, η οποία επέτρεψε τη συστηματική μελέτη όλης της ακολουθίας του ανθρώπινου γενομικού υλικού. Ένα άλλο γεγονός είναι η εφεύρεση των γονιδιακών μικροσυστοιχιών (microarrays) το 1996 το οποίο επέτρεψε τη συστηματική και ταυτόχρονη μέτρηση της έκφρασης πολλών χιλιάδων γονιδίων κάτω από τις ίδιες πειραματικές συνθήκες.

2.1 Βασικές έννοιες βιολογίας

Οι βασικές μονάδες που μελετάει η βιολογία και συνεπώς και η βιοπληροφορική καλύπτουν ένα ευρύ φάσμα από τις πιο βασικές στις πιο σύνθετες μονάδες. Ως πιο βασική μονάδα βιολογικής λειτουργίας σήμερα ορίζεται το νουκλεοτίδιο το οποίο είναι απλή χημική οργανική ένωση. Η δομή και οι διαδικασίες δημιουργίας των νουκλεοτιδίων ξεφεύγουν από τη σφαίρα της βιολογίας και φτάνουν στις περιοχές της βιοχημείας και της μοριακής φυσικής. Υπάρχουν 4 διαφορετικά είδη νουκλεοτιδίων τα οποία συνενώνονται δημιουργώντας μία ελικοειδή δομή γνωστή και ως DNA. Το DNA μπορεί να χαρακτηριστεί και ως ένα σειριακό πληροφοριακό αποθηκευτικό μέσο. Τα νουκλεοτίδια που περιέχονται στο DNA μέσω μιας βιολογικής διαδικασίας που ονομάζεται «μεταγραφή» (transcription) μπορούν να καθοδηγήσουν το υπόλοιπο κύτταρο στη δημιουργία μιας συγκεκριμένης πρωτεΐνης. Οι πρωτεΐνες είναι τα βασικά δομικά χαρακτηριστικά των έμβιων όντων. Παραδείγματα πρωτεϊνών είναι το σμάλτο των δοντιών, η ινσουλίνη του αίματος και το κολλαγόνο του δέρματος. Έρευνες έχουν δείξει ότι ένα πολύ μικρό μέρος της ακολουθίας του DNA μεταγράφεται ώστε να καθοδηγήσει την έκφραση κάποιας πρωτεΐνης. Συγκεκριμένα, μόνο το 2% της πληροφορίας που περιέχεται θεωρείται ότι έχει κάποιο σημαντικό ενεργό ρόλο, ενώ το υπόλοιπο 98% δεν έχει κάποιο προφανή ρόλο. Για τα νουκλεοτίδια, τα οποία περιέχουν πληροφορία για τη δημιουργία κάποιας πρωτεΐνης, έχει βρεθεί ότι έχουν μία πολύ συγκεκριμένη δομή. Αυτή η δομή ονομάζεται σήμερα γονίδιο. Αν και δεν υπάρχει ακόμα σταθερός και ευρέως αποδεκτός ορισμός του

γονιδίου, σήμερα τείνει να εννοεί την αφαιρετική δομή που περιέχει τα νουκλεοτίδια του DNA τα οποία περιέχουν την πληροφορία για τη δημιουργία είτε μίας είτε περισσότερων πρωτεϊνών που ως ομάδα κάνουν μία διακριτή βιολογική λειτουργία. Τα γονίδια αποτελούν σήμερα τον ακρογωνιαίο λίθο της σημερινής βιοτεχνολογίας καθώς διάφορες παθολογικές καταστάσεις έχουν άμεσα συνδεθεί με την παρουσία, απουσία ή τη διαφορετική δομή των γονιδίων από άνθρωπο σε άνθρωπο.

2.2 Δίκτυα γονιδιακών αλληλεπιδράσεων

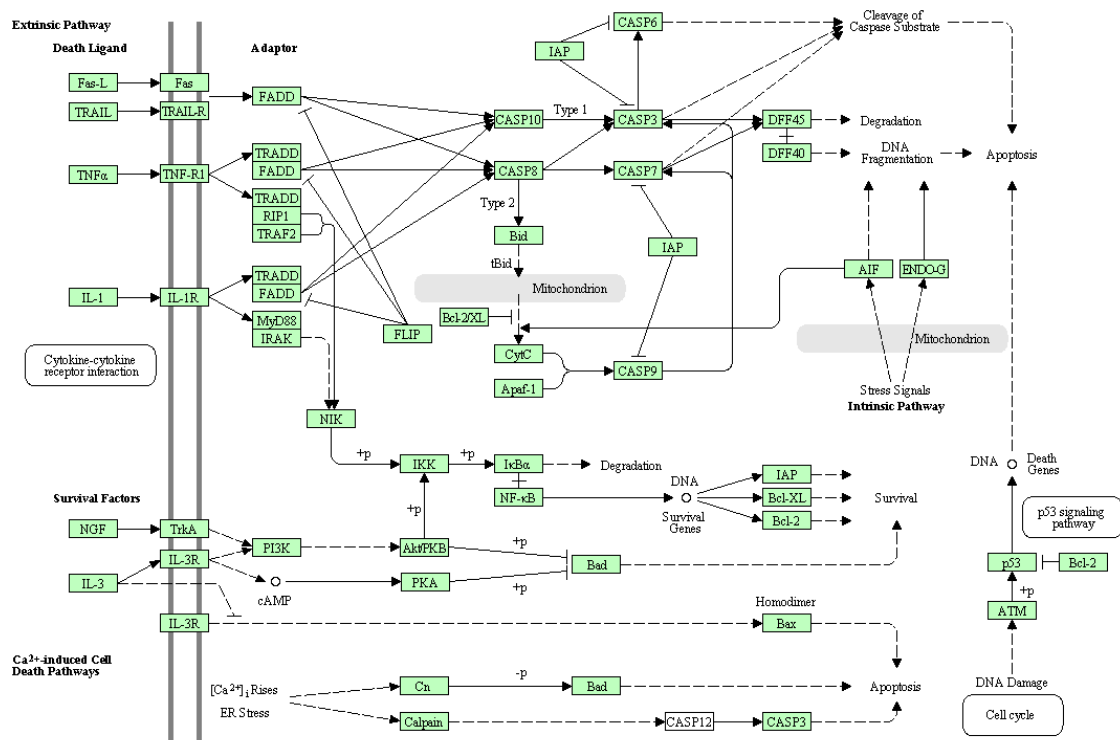
Η συνεχώς αυξανόμενη γνώση για τα γονίδια σήμερα μας έχει δώσει κάποιες επιπρόσθετες πληροφορίες, όπως το γεγονός ότι αλληλεπιδρούν μεταξύ τους με πολύπλοκο και χαοτικό τρόπο ώστε να δημιουργηθεί η πληθώρα των πρωτεϊνών και των βιολογικών συμπεριφορών που αντιλαμβανόμαστε στα έμβια όντα. Η αλληλεπίδραση αυτή μπορεί να χωριστεί σε πολλές κατηγορίες αλλά δύο είναι οι πιο σημαντικές. Η πρώτη είναι η ενεργοποίηση (activation), όπου ένα γονίδιο ενεργοποιεί κάποιο άλλο, και το άλλο η αποτροπή (inhibition), όπου ένα γονίδιο σταματάει την ενεργοποίηση κάποιου άλλου. Με τον όρο ενεργοποίηση εννοούμε τη μεταγραφική δραστηριότητα ενός γονιδίου, δηλαδή τη διαδικασία μετάφρασης της πληροφορίας που διαθέτει σε κάποια πρωτεΐνη. Είναι χαρακτηριστικό ότι υπάρχουν γονίδια των οποίων η πρωτεΐνη που κωδικοποιούν δεν έχει κάποιο βιολογικό ρόλο πέρα από την ενεργοποίηση ή απενεργοποίηση άλλων γονιδίων. Τα γονίδια αυτά ονομάζονται μεταγραφικοί παράγοντες (transcription factors).

Η μελέτη των γονιδιακών αλληλεπιδράσεων είναι ένας πολύ δύσκολος τομέας ο οποίος περιέχει πολλές προκλήσεις. Η παραδοσιακή μελέτη των γονιδιακών αλληλεπιδράσεων μέχρι σήμερα εστιαζόταν είτε σε μία μόνο αλληλεπίδραση μεταξύ δύο γονιδίων είτε στις συνολικές αλληλεπιδράσεις ενός γονιδίου. Αυτό γινόταν γιατί οι υπολογιστικές και εργαστηριακές συνθήκες δεν επέτρεπαν τη μαζική μελέτη περισσότερων από ένα περιορισμένο αριθμό γονιδίων. Γρήγορα έγινε αντιληπτό ότι αυτός ο τρόπος μελέτης επέτρεπε ένα περιορισμένο αλλά και περιοριστικό τρόπο θέασης της συνολικής γονιδιακής συμπεριφοράς. Η ανάγκη για τη μελέτη των αλληλεπιδράσεων μεγάλου πλήθους γονιδίων έφερε στην επιφάνεια μεθόδους που επέτρεπαν τη εξαγωγή γράφων γονιδιακών αλληλεπιδράσεων. Η γενική εικόνα της λειτουργίας αυτών των μεθόδων είναι ότι αν ένα γονίδιο A αλληλεπιδρούσε με ένα

άλλο γονίδιο B κάτω από τις ίδιες εργαστηριακές συνθήκες, ενώ παράλληλα το γονίδιο B αλληλεπιδρούσε με το γονίδιο Γ υπό τις ίδιες συνθήκες, τότε μπορούσε να υποτεθεί ότι υπάρχει ένας υφέρπων βιολογικός μηχανισμός ο οποίος άγεται από το γράφο γονιδιακών αλληλεπιδράσεων: $A \rightarrow B \rightarrow \Gamma$. Καθώς ο παραδοσιακός τρόπος μελέτης των γονιδίων μέχρι και τη δεκαετία του '90 ήταν «ένα γονίδιο ανά εργαστήριο», ήταν πολύ δύσκολο να βρεθούν γονιδιακές αλληλεπιδράσεις με κοινά γονίδια κάτω από τις ίδιες εργαστηριακές συνθήκες. Γι' αυτό και τα πρώτα δίκτυα αλληλεπιδράσεων ήταν περιορισμένα και σε μεγάλο βαθμό περιείχαν ελλιπή και λανθασμένη πληροφορία. Ο τρόπος σύνθεσης των δικτύων βασιζόταν (και σε μεγάλο βαθμό ακόμα και σήμερα βασίζεται) στη βιβλιογραφική έρευνα. Όπως αναφέρει και η Wikipedia:

GRNs (Gene Regulatory Networks) are inferred from the biological literature on a given system and represent a distillation of the collective knowledge about a set of related biochemical reactions.

Παρόλες τις δύσκολες συνθήκες σύνθεσης των δικτύων γονιδιακών αλληλεπιδράσεων σήμερα έχουμε φτάσει σε ένα επίπεδο όπου υπάρχουν δίκτυα για όλες σχεδόν τις βιολογικές λειτουργίες, όπως ο μεταβολισμός, η κυτταρική επεξεργασία της γενετικής και περιβαλλοντικής πληροφορίας, οι κυτταρικές λειτουργίες (όπως ο κυτταρικός κύκλος, η διαίρεση και η απόπτωση), οι παθολογικές καταστάσεις καθώς επίσης και η επίδραση συγκεκριμένων ουσιών και φαρμάκων. Πολλοί διεθνείς οργανισμοί βιοτεχνολογίας παρέχουν δεδομένα με δίκτυα αλληλεπιδράσεων ελεύθερα προς μελέτη. Παράλληλα έχει αναπτυχθεί και ο κλάδος της πληροφορικής για την αποθήκευση, διαχείριση και μοντελοποίηση της πληροφορίας που περιέχουν. Στην Εικόνα 2.1 υπάρχει ένα παράδειγμα ενός ΓΡΔ όπως βρίσκεται στη σύγχρονη μορφή.



Εικόνα 2.1 Παράδειγμα του ΓΡΔ της απόπτωσης από το KEGG

2.3 KEGG και KGML

Ένας από τους πιο βασικούς διεθνείς οργανισμούς που παρέχει πληροφορίες σχετικά με ΓΡΔ είναι το KEGG (Kyoto Encyclopedia for Genes and Genomes). Το KEGG επίσης παρέχει πληροφορίες για γονίδια, πρωτεΐνες, χημικές ενώσεις και ακολουθίες νουκλεοτιδίων, ενώ διατηρεί ένα συγκεκριμένο κωδικό για κάθε οντότητα ώστε να είναι πιο εύκολη η αναφορά σε αυτή. Πολλοί άλλοι οργανισμοί είτε ακολουθούν την κωδικοποίηση του KEGG είτε παρέχουν διασύνδεση της δικής τους κωδικοποίησης με τη κωδικοποίηση του KEGG. Για τη διάθεση των ΓΡΔ το KEGG διαθέτει ένα ειδικό πρότυπο που βασίζεται στο γενικότερο πρότυπο ανταλλαγής αρχείων XML. Το πρότυπο αυτό ονομάζεται KGML (KEGG Markup Language). Όπως όλα τα XML αρχεία έτσι και το KGML παρέχει μία δενδρική δομή παράθεσης της πληροφορίας. Βασικά στοιχεία είναι τα elements τα οποία είναι tags της μορφής:

<elementName1>

... Άλλα elements

</elementName1>

Κάθε element μπορεί να έχει επιπλέον πληροφορία μέσω attributes. Τα attributes παρατίθενται μαζί με τα elements ενώ κάθε attribute έχει ένα value. Η μορφή των attributes και των values είναι:

```
<elementName1 attributeName1="value1">
```

```
...Άλλα elements
```

```
</elementName1>
```

Στο πρότυπο KGML τα elements, τα attributes καθώς και η σχετική τους θέση καθορίζεται από ένα άλλο αρχείο σύμφωνα με το πρότυπο Document Type Definition (DTD). Η τρέχουσα έκδοση του DTD για τα αρχεία KGML είναι η 0.6.1¹. Στη συνέχεια περιγράφουμε αναλυτικά τη δενδρική δομή του KGML στην παρούσα έκδοση:

Element pathway: Είναι η ρίζα του δένδρου. Περιέχει όλες τις πληροφορίες χωρισμένες σε δύο μέρη. Το πρώτο περιέχει πληροφορίες για τους κόμβους και το δεύτερο για τις ακμές. Το element pathway έχει τα παρακάτω attributes:

Attribute name: Περιέχει το όνομα του ΓΡΔ σύμφωνα με τη κωδικοποίηση του KEGG παράδειγμα ονόματος είναι το: “path:hsa04010”. Αυτό το όνομα σημαίνει ότι αναφέρεται σε κάποιο path (το KEGG αναφέρεται στα ΓΡΔ ως pathways και σε συντομογραφία ως paths), το οποίο αναφέρεται στον οργανισμό hsa (συντομογραφία του Homo Sapiens) και ο κωδικός του είναι 04010.

Attribute org: Περιέχει τον οργανισμό στον οποίο ανήκει το ΓΡΔ. Παράδειγμα είναι το “hsa” (Homo Sapiens).

Attribute number: Περιέχει τον KEGG κωδικό για το συγκεκριμένο ΓΡΔ. Παράδειγμα είναι το “04010”.

Attribute title: Ο περιγραφικός τίτλος του ΓΡΔ. Παράδειγμα μπορεί να είναι το “MAPK signaling pathway”.

Attribute image: Σύνδεσμος προς ένα αρχείο που περιέχει μία γραφική αναπαράσταση του ΓΡΔ σε GIF μορφή.

Attribute link: Σύνδεσμος προς την HTML αναπαράσταση του ΓΡΔ.

Element entry: Το element αυτό είναι πάντα «παιδί» του element pathway.

Περιέχει πληροφορίες σχετικά με έναν κόμβο του ΓΡΔ. Τις περισσότερες (άλλα όχι όλες) αυτός ο κόμβος είναι γονίδιο.

¹ Η πλήρης περιγραφή του DTD βρίσκεται στη διεύθυνση:
http://www.genome.jp/kegg/xml/KGML_v0.6.1_.dtd

Attribute id: Περιέχει ένα μοναδικό ακέραιο αριθμό για τον συγκεκριμένο κόμβο. Ο αριθμός αυτός ξεκινάει πάντα από 1. Η πληροφορία αυτή είναι πολύ σημαντική γιατί όταν περιγραφούν οι ακμές του ΓΡΔ τότε κάθε ακμή θα ενώνει δύο κόμβους οι οποίοι θα αναφέρονται με βάση αυτούς τους κωδικούς.

Attribute name: Περιέχει το όνομα του κόμβου σύμφωνα με την κωδικοποίηση του KEGG. Παράδειγμα ονόματος είναι το “hsa:5495”. Το όνομα αυτό αναφέρει ότι ο κόμβος αυτός ανήκει στο είδος Homo Sapiens και ο κωδικός του είναι “5495”. Πρέπει να τονίσουμε ότι είναι πολύ κοινό να περιέχονται παραπάνω από ένα ονόματα αντικειμένων (π.χ. γονιδίων). Αυτό γίνεται στη περίπτωση που ο κόμβος αποτελεί μία ομαδοποίηση (group) πολλών γονιδίων. Σε αυτή τη περίπτωση τα ονόματα χωρίζονται μεταξύ τους με ένα κενό.

Attribute type: Περιέχει το είδος του κόμβου. Δυνατές τιμές είναι: “gene” (σημαίνει ότι ο κόμβος είναι ένα γονίδιο), “group” (σημαίνει ότι ο κόμβος είναι ένα σύνολο από γονίδια), “compound” (σημαίνει ότι ο κόμβος είναι μία χημική ένωση, ή κάποιο ένζυμο, το οποίο αλληλεπιδράει με κάποιο γονίδιο) και “map” (σημαίνει ότι ο κόμβος είναι κάποιο άλλο ΓΡΔ. Γενικότερα είναι δυνατό κάποιο ΓΡΔ να αλληλεπιδράει με κάποιο άλλο ΓΡΔ μέσω ενός γονιδίου. Για παράδειγμα το ΓΡΔ που αναφέρεται στη δημιουργία καρκινικών όγκων αλληλεπιδράει με το ΓΡΔ το οποίο αναφέρεται στη διαίρεση των κυττάρων).

Attribute link: Περιέχει διεύθυνση προς μία ιστοσελίδα που περιέχει επιπλέον πληροφορίες σχετικά με αυτόν τον κόμβο. Για παράδειγμα περιέχει ονοματολογίες του κόμβου σύμφωνα με άλλους οργανισμούς, τη γραφική του αναπαράσταση και άλλα.

Element graphics: Το element αυτό είναι παιδί του element entry και περιέχει πληροφορίες σχετικά με τη γραφική αναπαράσταση αυτού του κόμβου.

Attribute name: Το όνομα το οποίο θα αναγράφεται πάνω στον κόμβο κατά τη γραφική αναπαράστασή του. Παράδειγμα είναι το “MAP4K4”.

Attribute fgcolor: Το χρώμα της γραφικής αναπαράστασης του κόμβου.

Attribute bgcolor: Το χρώμα του φόντου στη γραφική αναπαράσταση του κόμβου.

Attribute type: Το σχήμα της γραφικής αναπαράστασης του κόμβου. Για τα γονίδια και τις ομάδες γονιδίων χρησιμοποιείται ορθογώνιο, για τις χημικές ενώσεις χρησιμοποιείται κύκλος, ενώ για τα ΓΡΔ χρησιμοποιείται τετράγωνο με στρογγυλοποιημένες άκρες.

Attribute x: Η οριζόντια θέση στον δισδιάστατο χώρο κατά τη γραφική αναπαράσταση.

Attribute y: Η κάθετη θέση στον δισδιάστατο χώρο κατά τη γραφική αναπαράσταση.

Attribute width: Το πλάτος που θα διατεθεί για τη γραφική αναπαράσταση του κόμβου.

Attribute height: Το ύψος που θα διατεθεί για τη γραφική αναπαράσταση του κόμβου.

Είναι πολύ σημαντικό να τονίσουμε ότι τα attributes του element graphics μας δίνουν πολύτιμες πληροφορίες που μας βοηθάνε να δώσουμε μια αρχική γραφική αναπαράσταση του ΓΡΔ παρόμοια με αυτή που δίνει η στατική HTML σελίδα του KEGG. Με αυτόν τον τρόπο ο χρήστης ο οποίος είναι εξοικειωμένος με την θέση και διάσταση των στοιχείων του γράφου στο KEGG να μπορεί εύκολα να τα εντοπίζει στη δική μας οπτικοποίηση.

Element relation: Το element αυτό είναι «παιδί» του element pathway και περιγράφει μια συσχέτιση μεταξύ δύο κόμβων του δικτύου. Περιέχει δηλαδή τη πληροφορία για τις ακμές του δικτύου.

Attribute entry1: Περιέχει τον ακέραιο κωδικό του κόμβου από τον οποίο ξεκινάει η σύνδεση.

Attribute entry2: Περιέχει τον ακέραιο κωδικό του κόμβου στον οποίο καταλήγει η σύνδεση.

Είναι σημαντικό να τονίσουμε ότι οι συνδέσεις στους γράφους είναι κατευθυνόμενες. Αυτό γίνεται γιατί η βιολογική σημασία των συνδέσεων είναι αιτιοκρατική. Δηλαδή το γονίδιο A δημιουργεί κάποια συνέπεια στο γονίδιο B. Συνεπώς έχει μεγάλη σημασία η σειρά των κωδικών που μπαίνει στα attributes entry1 και entry2.

Attribute type: Περιέχει τον τύπο της σύνδεσης. Η τιμή είναι πάντα “PPrel” οπότε δεν περιέχει καμία σημαντική πληροφορία.

Element subtype: Το element αυτό είναι πάντα «παιδί» του element relation. Περιέχει επιπλέον πληροφορίες σχετικά με το είδος της διασύνδεσης μεταξύ δύο κόμβων του δικτύου. Μπορεί να έχουμε παραπάνω από ένα subtype element κάτω από το ίδιο relation element. Αυτό σημαίνει ότι δύο κόμβοι μπορεί να ενώνονται με παραπάνω από έναν τρόπους.

Attribute name: Περιέχει το όνομα της συσχέτισης μεταξύ των δύο κόμβων. Δυνατές τιμές είναι:

- Activation. Είναι η πιο σημαντική και πιο συχνή σχέση στα ΓΡΔ. Το νόημα του activation είναι ότι ένα γονίδιο ενεργοποιεί κάποιο άλλο.
- Inhibition. Σημαίνει ότι ένα γονίδιο απενεργοποιεί κάποιο άλλο.
- Expression. Σημαίνει ότι ένα γονίδιο ενεργοποιεί κάποιο άλλο μέσω ενός μηχανισμού ο οποίος παρεμβάλλεται. Πρόκειται δηλαδή για έμμεση ενεργοποίηση.
- Indirect. Σημαίνει ότι υπάρχουν ισχυρές ενδείξεις ότι ένα γονίδιο ενεργοποιεί κάποιο άλλο, αλλά δεν έχει τεκμηριωθεί ακόμα αυτή η γνώση.
- Ubiquitination. Είναι μια ειδική κατηγορία γονιδιακής αλληλεπίδρασης της οποίας η συμπεριφορά είναι παρόμοια με το inhibition.
- Phosphorylation. Είναι επίσης μία ειδική κατηγορία γονιδιακής αλληλεπίδρασης. Μέσω της φωσφορυλίωσης μπορεί να έχουμε είτε activation είτε inhibition. Στο KGML αρχείο διασαφηνίζεται αυτό μέσω ενός δευτέρου subtype tag το οποίο δηλώνει για ποιο από τα δύο πρόκειται.
- Dephosphorylation. Είναι το αντίστροφο της σχέσης φωσφορυλίωσης. Επίσης μπορεί να πρόκειται είτε για activation είτε για inhibition και δηλώνεται ρητά στο KGML αρχείο.

- Association. Σημαίνει ότι ένα γονίδιο μπορεί να εκφραστεί αν και μόνο αν ένα άλλο γονίδιο είναι επίσης εκφρασμένο.
- Disassociation. Σημαίνει ότι ένα γονίδιο μπορεί να εκφραστεί αν και μόνο αν ένα άλλο γονίδιο δεν είναι εκφρασμένο.

Attribute value: Περιέχει μία αναπαράσταση με ASCII χαρακτήρες της συσχέτισης. Η πληροφορία αυτή δεν είναι σημαντική γιατί το είδος της συσχέτισης υπάρχει στο attribute name.

Ένα παράδειγμα ΓΡΔ σε κωδικοποίηση KGML φαίνεται στην Εικόνα 2.2.

```
<?xml version="1.0"?>
<!DOCTYPE pathway SYSTEM "http://www.genome.jp/kegg/xml/KGML_v0.6.1_.dtd">
<!-- Creation date: Aug 30 2008 00:24:20 +0900 (JST) -->
<pathway name="path:hsa04210" org="hsa" number="04210"
  title="Apoptosis"
  image="http://www.genome.jp/kegg/pathway/hsa/hsa04210.gif"
  link="http://www.genome.jp/dbget-bin/show_pathway?hsa04210">
  <entry id="1" name="path:hsa04115" type="map"
    link="http://www.genome.jp/kegg/pathway/hsa/hsa04115.html">
    <graphics name="p53 signaling pathway" fgcolor="#000000" bgcolor="#FFFFFF"
      type="roundrectangle" x="1048" y="547" width="95" height="39"/>
  </entry>
  <entry id="2" name="path:hsa04060" type="map"
    link="http://www.genome.jp/kegg/pathway/hsa/hsa04060.html">
    <graphics name="Cytokine-cytokine receptor interaction" fgcolor="#000000" bgcolor="#FFFFFF"
      type="roundrectangle" x="111" y="402" width="124" height="39"/>
  </entry>
  <entry id="3" name="hsa:11261 hsa:5530 hsa:5532 hsa:5533 hsa:5534 hsa:5535" type="gene"
    link="http://www.genome.jp/dbget-bin/www_bget?hsa:11261+5530+5532+5533+5534+5535">
    <graphics name="CHP..." fgcolor="#000000" bgcolor="#808080"
      type="rectangle" x="430" y="708" width="46" height="17"/>
  </entry>
  ...
  <relation entry1="54" entry2="65" type="PPrel">
    <subtype name="activation" value="-->"/>
  </relation>
  <relation entry1="54" entry2="66" type="PPrel">
    <subtype name="activation" value="-->"/>
  </relation>
  <relation entry1="53" entry2="67" type="PPrel">
    <subtype name="activation" value="-->"/>
  </relation>
  ...
</pathway>
```

Εικόνα 2.2. ΓΡΔ σε κωδικοποίηση KGML

2.3.1 Ασυμβατότητες μεταξύ KGML και HTML

Το KEGG παρέχει ΓΡΔ σε δύο διαφορετικές μορφές. Η μία είναι μέσω του KGML και η άλλη είναι μέσω των στατικών HTML σελίδων. Θεωρητικά θα έπρεπε αυτή η πληροφορία σε αυτές τις δύο μορφές να συνέπιπτε. Παρατηρούμε όμως ότι υπάρχουν ασυνέπειες με αποτέλεσμα η γραφική αναπαράσταση της HTML να μην συμφωνεί πάντα με την KGML έκδοση. Από τη πλευρά του KEGG θεωρείται ότι αυτό είναι ένα προσωρινό πρόβλημα, μπορεί όμως να δημιουργήσει προβλήματα όταν κάνουμε έλεγχο της σωστής οπτικοποίησης ενός KGML αρχείου χρησιμοποιώντας ως σημείο

αναφοράς την αναπαράσταση της HTML σελίδας. Οι ασυμφωνίες που συνήθως συναντάμε είναι:

- Τα ονόματα των γονιδίων δεν είναι πάντα τα ίδια. Κάποιο γονίδιο μπορεί να αναφερθεί με διαφορετικό όνομα.
- Δεν υπάρχουν όλα τα γονίδια της HTML έκδοσης στην KGML έκδοση. Συνεπώς λείπουν και οι συσχετίσεις τους.
- Υπάρχουν οντότητες στην HTML σελίδα οι οποίες ίσως λόγω της πολυπλοκότητάς τους δεν αναπαρίστανται στο KGML αρχείο. Αυτό δίνει την εντύπωση ότι το HTML αρχείο περιέχει πολύ πιο πλούσια πληροφορία.

Γενικότερα υπάρχει η εντύπωση ότι το KGML αρχείο είναι πιο φτωχό και πιο ανενήμερο από τη HTML σελίδα. Θεωρούμε όμως ότι είναι θέμα χρόνου να γίνει σωστή συντήρηση και μοντελοποίηση των αρχείων ώστε να περιέχουν τα ίδια δεδομένα.

Μία εύλογη απορία είναι γιατί δεν γίνεται απευθείας parsing του HTML αρχείου από τη στιγμή που φαίνεται ότι παρέχει πιο πλήρη πληροφορία. Η HTML σελίδα είναι στη πραγματικότητα ένα αρχείο εικόνας σε format png, στο οποίο έχουν προστεθεί σύνδεσμοι πάνω στους κόμβους που οδηγούν στη μετα-πληροφορία που διαθέτει το KEGG. Δηλαδή η HTML σελίδα δεν παρέχει καμία σημασιολογία για τους κόμβους και τις ακμές του δικτύου απλώς εμφανίζει μία γραφική αναπαράστασή τους.

2.4 Επεξεργασία KGML αρχείων

Γνωρίζοντας τη δομή και τις ιδιαιτερότητες των KGML αρχείων μπορούμε να τα επεξεργαστούμε με μία γλώσσα προγραμματισμού. Σκοπός αυτής της επεξεργασίας είναι να αποθηκεύσουμε σε ειδικές δομές δεδομένων όλη την πληροφορία που περιέχουν ώστε να μπορέσουμε να τη διαχειριστούμε δυναμικά μέσα από απλές προγραμματιστικές διαδικασίες. Προτεραιότητα δόθηκε ώστε η εφαρμογή μας να είναι βοήθημα προς τρίτους που επιθυμούν να επεξεργαστούν KGML αρχεία χωρίς να γνωρίζουν την ιδιαίτερη δομή τους. Ως γλώσσα προγραμματισμού επιλέχθηκε η Python η οποία έχει το πλεονέκτημα ότι είναι μία υψηλού επιπέδου γλώσσα η οποία δίνει μεγάλη έμφαση στην «αναγνωσιμότητα» του κώδικα που γράφεται, ενώ είναι σχετικά εύκολη στην εκμάθηση. Η σύνταξη της Python είναι τέτοια ώστε να ωθεί τους χρήστες οι οποίοι γράφουν τον ίδιο αλγόριθμο να γράψουν περίπου τον ίδιο κώδικα. Άλλα πλεονεκτήματα της Python είναι η μεγάλη συλλογή από βιβλιοθήκες

που διαθέτει για κάθε ειδική περίπτωση, ενώ ως “scripting” γλώσσα δεν απαιτεί την εγκατάσταση κάποιου συγκεκριμένου περιβάλλοντος και μπορεί να τρέξει σε όλα τα δημοφιλή λειτουργικά συστήματα. Τέλος, παρέχει δυναμικούς τύπους δεδομένων, όπως το λεξικό και οι λίστες. Για παράδειγμα, με το λεξικό μπορούμε να έχουμε μια δομή που αποτελείται από δυνάδες <κλειδί, περιεχόμενο>. Το κλειδί και το περιεχόμενο μπορεί να είναι οποιουδήποτε τύπου. Αυτό επιτρέπει τη γρήγορη κατηγοριοποίηση ενός οποιουδήποτε τύπου δεδομένου σε μορφή πίνακα. Δηλαδή είναι επιτρεπτές οι εκφράσεις $A[3] = 7$, $A['foo'] = d$, $A[3.2] = A$.

2.4.1 Parsing² των KGML αρχείων

Το πρώτο βήμα για την επεξεργασία των KGML αρχείων είναι το parsing. Ως XML αρχεία μπορούν να γίνουν εύκολα parse από μία από τις πολλές βιβλιοθήκες που υπάρχουν για αυτό τον σκοπό. Η βιβλιοθήκη που επιλέξαμε ήταν η expat της Python. Η expat είναι μία μη-επικυρωτική (non-validating) ελαφριά βιβλιοθήκη για XML parsing. Μη-επικυρωτική σημαίνει ότι δεν ελέγχει αν η δομή του XML αρχείου είναι σύμφωνη με κάποιο DTD, αν και ελέγχει αν το αρχείο έχει τη γενικότερη δομή ενός XML εγγράφου. Ένα αντικείμενο τύπου expat μπορεί να έχει ως μέλη συναρτήσεις του χρήστη οι οποίες καλούνται όταν συμβεί κάποιο «XML γεγονός», δηλαδή όταν ανιχνευθεί η αρχή κάποιου element ή το τέλος κάποιου element ή η παρουσία κάποιου attribute. Συγκεκριμένα, όταν ανιχνευθεί το element για παράδειγμα <entry> τότε καλείται η διαδικασία StartElementHandler την οποία έχουμε υλοποιήσει εμείς

με όρισμα τη λέξη “entry”. Ομοίως γίνεται και για τα υπόλοιπα «XML γεγονότα». Με αυτό τον τρόπο επιτυγχάνεται μία οντοκεντρική αντιμετώπιση του προβλήματος του parsing του XML εγγράφου. Η expat είναι μία κλάση η οποία παρέχει εικονικές συναρτήσεις για την διευθέτηση των XML γεγονότων. Δημιουργώντας μία κλάση παράγωγο (overload) που κληρονομεί την expat και υλοποιώντας τις διαδικασίες που χειρίζονται τα XML γεγονότα έχουμε μία «καθαρή» από πλευράς υλοποίησης προγραμματιστική λύση για το χειρισμό των αρχείων που περιγράφουν τα ΓΡΔ. Συνεπώς, δεν χρειάζεται να ασχοληθούμε με τη δομή, τη σύνταξη και την κωδικοποίηση (encoding) των XML εγγράφων.

² Σύμφωνα με το λεξικό δεν υπάρχει λέξη προς λέξη μετάφραση της λέξης Parsing στα Ελληνικά.

2.4.2 Αποθήκευση των ΓΡΔ σε Δομές Δεδομένων

Για την αποθήκευση των δεδομένων που περιγράφουν τους κόμβους και τις ακμές του γράφου, έχουμε αντίστοιχα δύο δομές δεδομένων τύπου λεξικό. Για την αποθήκευση ενός στοιχείου Σ σε μία δομή τύπου λεξικό χρειαζόμαστε ένα κλειδί K . Αν το όνομα της δομής είναι Δ , τότε η αποθήκευση αναπαρίσταται ως: $\Delta[K] = \Sigma$. Τα στοιχεία K και Σ μπορεί να είναι οποιουδήποτε τύπου (ακόμα και κάποιο άλλο λεξικό). Τέτοιος χειρισμός επιτρέπεται από τη γλώσσα προγραμματισμού Python.

Το λεξικό που χρησιμοποιούμε για την αποθήκευση των κόμβων ονομάζεται “entries”. Το κλειδί K του λεξικού αυτού είναι το “Attribute id” του εκάστοτε κόμβου όπως είναι αποθηκευμένο στο XML αρχείο. Το στοιχείο Σ είναι μία δομή επίσης τύπου λεξικού με το όνομα `currentEntry`. Το `currentEntry` περιέχει τις πληροφορίες του κόμβου που έχει περάσει από parsing πιο πρόσφατα. Τα κλειδιά του λεξικού `currentEntry` είναι τα ονόματα όλων των διαφορετικών XML arguments που περιγράφουν έναν κόμβο (`link`, `name`, `type`, `id`, `graphics.fgcolor`, `graphics.x`, ...) και τα στοιχεία του είναι οι τιμές αυτών των arguments. Ένα παράδειγμα χρήσης αυτής της δομής είναι:

```
currentEntry['name'] = 'hsa:598'
```

Όταν ανιχνευτεί το XML element `</entry>` το οποίο υποδηλώνει το τέλος της περιγραφής κάποιου κόμβου αποθηκεύουμε το λεξικό `currentEntry` ως στοιχείο στο λεξικό `entries` με κλειδί το “Attribute id” του κόμβου. Αυτή η πράξη υποδηλώνεται ως:

```
entries[currentEntry['id']] = currentEntry
```

Παρόμοιος μηχανισμός χρησιμοποιείται για την αποθήκευση των δεδομένων σχετικά με τις ακμές του γράφου. Η σημαντική διαφορά είναι ότι στα XML έγγραφα δεν υπάρχει μοναδικό `id` για κάθε ακμή. Γι’ αυτό χρησιμοποιήσαμε έναν μετρητή `relationCounter` ο οποίος μετράει το πλήθος των ακμών που έχουν περάσει από parsing σε κάθε δεδομένη στιγμή. Το λεξικό στο οποίο αποθηκεύονται τα δεδομένα των ακμών ονομάζεται `relations` και τα κλειδιά του λεξικού αυτού είναι το `relationCounter`. Τα στοιχεία του `relations` είναι μία άλλη δομή επίσης τύπου λεξικό η οποία ονομάζεται `currentRelation` και περιέχει τα δεδομένα της ακμής η οποία περνάει από parsing μία δεδομένη στιγμή. Τα κλειδιά του λεξικού `currentRelation` είναι τα ονόματα των XML arguments που υπάρχουν στο XML element `<relation>`

και έχουν ήδη περιγραφεί. Παράδειγμα αποτελεί το argument “entry1” το οποίο περιέχει το Attribute id του κόμβου από τον οποίο προέρχεται η ακμή και το argument “entry2” το οποίο περιέχει το Attribute id του κόμβου στον οποίο καταλήγει η τρέχουσα ακμή. Δηλαδή η έκφραση: `currentRelation['entry1'] = 5` και `currentRelation['entry2'] = 7` υποδηλώνει ότι η τρέχουσα ακμή ενώνει τον κόμβο με id=5 με τον κόμβο με id=7. Το XML element το οποίο υποδηλώνει το τέλος της περιγραφής κάποιας ακμής είναι το “</relation>”. Όταν ανιχνευτεί αυτό το element τότε αυξάνουμε τον μετρητή relationCounter κατά ένα, και αποθηκεύουμε το currentRelation ως στοιχείο στο λεξικό relations. Αυτό γίνεται με την παρακάτω έκφραση:

```
relations[relationCounter] = currentRelation
```

Στο τέλος αυτής της διαδικασίας έχουμε τα δύο λεξικά entries και relations τα οποία περιέχουν όλη την πληροφορία που υπάρχει στο KGML έγγραφο σε μορφή δομής δεδομένων της γλώσσας Python. Είναι σημαντικό να τονίσουμε ότι οι δομές δεδομένων που χρησιμοποιήθηκαν είναι απλές και εύκολα διαχειρίσιμες. Με αυτόν τον τρόπο παρέχουμε τα πρώτα προγραμματιστικά εργαλεία για την μετέπειτα επεξεργασία των ΓΡΔ. Ένας από τους κύριους σκοπούς αυτού του κομματιού της εργασίας ήταν το παραγόμενο προγραμματιστικό πλαίσιο να είναι εύχρηστο και κατανοητό από τρίτους που επιθυμούν να εφαρμόσουν επιπλέον μεθόδους ανάλυσης σε ΓΡΔ.

2.4.3 Ανάλυση σε κλάσεις

Για να μπορούμε να χειριστούμε με αφαιρετικό τρόπο τις δομές που περιέχουν τα δεδομένα των ΓΡΔ, σχεδιάσαμε την κλάση myKGMLParser. Εκτός από τις διαδικασίες για parsing XML εγγράφων, η κλάση αυτή περιέχει και τις δομές δεδομένων που περιγράφηκαν προηγουμένως. Συνοπτικά, η κλάση myKGMLParser έχει τη παρακάτω μορφή σε UML:

myKGMLParser	
start_element (elementName : String , arguments : List) Εκτελείται όταν ανιχνεύεται κάποιο XML element της μορφής: <code><elementName arg1="argVal1" arg2="argVal2"></code>	Αυτές οι διαδικασίες καλούνται αυτόματα από την κλάση expat κατά το parsing του XML εγγράφου
end_element (elementName : String) Εκτελείται όταν ανιχνεύεται το τέλος κάποιου XML element: <code></elementName></code>	
entries : Dictionary Περιέχει όλες τις πληροφορίες για τους κόμβους. Το κλειδί του λεξικού είναι το id του κάθε κόμβου	Σε αυτά τα πεδία περιέχονται όλες οι πληροφορίες του XML εγγράφου και μέσω αυτών είναι δυνατή η προσπέλασή τους.
relations : Dictionary Περιέχει όλες τις πληροφορίες για τις ακμές. Το κλειδί του λεξικού είναι ο αύξων αριθμός της ακμής.	
relationCounter : integer Περιέχει το πλήθος των ακμών.	

Ένα πρόγραμμα μπορεί να έχει πολλά αντικείμενα τύπου myKGMLParser κάθε ένα από τα οποία μοντελοποιεί και διαφορετικό XML αρχείο.

2.5 Συνένωση ΓΡΔ

Σ' αυτή την ενότητα, παρέχουμε ως ένα παράδειγμα χρήσης των λειτουργιών που περιγράφηκαν νωρίτερα, τη διαδικασία της συνένωσης δύο διαφορετικών ΓΡΔ.

2.5.1 Βιολογική σημασία της συνένωσης

Όπως έχουμε παρουσιάσει η γνώση μας για τα ΓΡΔ είναι ατελής και συνεχώς εμπλουτίζεται και βελτιώνεται. Επίσης, οι διαδικτυακές πηγές που παρέχουν

πληροφορίες σχετικά με ΓΡΔ είναι συχνά προσανατολισμένες σε μία συγκεκριμένη βιολογική λειτουργία. Στο δικτυακό τόπο KEGG³ τα διαθέσιμα ΓΡΔ είναι χωρισμένα σε πολλές κατηγορίες και υποκατηγορίες. Για παράδειγμα, υπάρχει το ΓΡΔ της κατηγορίας Cellular Processes της υποκατηγορίας Endocrine system με το όνομα «GnRH signaling pathway». Αυτή η κατηγοριοποίηση είναι πολύ χρήσιμη για τους ερευνητές που μελετάνε τη βιολογική λειτουργία ενός συγκεκριμένου κομματιού του κυττάρου ή μίας συγκεκριμένης λειτουργίας του οργανισμού. Παρόλα αυτά στη σύγχρονη βιοπληροφορική έρευνα θέλουμε να έχουμε μία συστηματική και όσο το δυνατόν ολιστική εικόνα των αλληλεπιδράσεων των γονιδίων γι' αυτό κρίνεται απαραίτητη η δημιουργία μίας διαδικασίας συνένωσης των υπαρχόντων ΓΡΔ. Άλλωστε τα υπάρχοντα ΓΡΔ είναι στην ουσία μία συνένωση μικρότερης κλίμακας αλληλεπιδράσεων οι οποίες ειδοθήκαν υπό την ίδια ερευνητική ματιά και συγκρότησαν πιο πολύπλοκα δίκτυα. Παρομοίως, η ανάγκη για συστημική θεώρηση των οργανισμών σήμερα μας επιβάλλει να δούμε τα ΓΡΔ σε ένα πιο αφαιρετικό πλαίσιο όπου έχουν συνενωθεί σε πιο πολύπλοκες δομές. Επίσης, αναμένουμε οι δομές και οι ιδιαιτερότητες οι οποίες θα φανούν σε πιο πολύπλοκες γονιδιακές αλληλεπιδράσεις να δώσουν περισσότερα στοιχεία τόσο για τη γονιδιακή λειτουργία και έκφραση όσο και για τη παθολογία συγκεκριμένων ασθενειών.

2.5.2 Συνένωση δύο ΓΡΔ

Για τη συνένωση δύο ΓΡΔ καταρχήν εκτελούμε parsing και δημιουργούμε δύο αντικείμενα της κλάσης myKGMLParser, έστω a και b. Επίσης δημιουργούμε το άδειο αντικείμενο c τύπου myKGMLParser στο οποίο θα αποθηκεύσουμε τη πληροφορία του συνενωμένου ΓΡΔ. Στη συνέχεια δημιουργούμε δύο σύνολα. Το namesA περιέχει όλα τα διαφορετικά ονόματα του a και το namesB περιέχει όλα τα διαφορετικά ονόματα του b. Στο επόμενο στάδιο υπολογίζουμε την τομή των δύο συνόλων, έστω namesCommon:

$$\text{namesCommon} = \text{namesA} \cap \text{namesB}$$

Με αυτόν τον τρόπο έχουμε τα ονόματα των γονιδίων τα οποία είναι κοινά και στα δύο ΓΡΔ. Ο σκοπός που το κάνουμε αυτό είναι ώστε να ξέρουμε ποια γονίδια θα μοντελοποιηθούν ως κοινά στο νέο ΓΡΔ που θα προκύψει. Στο αντικείμενο c αντιγράφουμε όλους τους κόμβους του a. Επίσης πρέπει να αντιγράψουμε και όλους

³ <http://www.genome.jp/kegg/pathway.html>

τους κόμβους του b με τη διαφορά όμως ότι δεν πρέπει οι κόμβοι του b να έχουν το ίδιο id με τους κόμβους του a. Τα ids ενός KGML αρχείου είναι ένας αριθμός από το 1 μέχρι το πλήθος των κόμβων του αρχείου. Οπότε πριν αντιγράψουμε τα entries του b στο c τους προσανυξάνουμε τα ids κατά έναν αριθμό ίσο με το πλήθος των κόμβων του a. Με αυτό τον τρόπο αν οι κόμβοι του a έχουν ids από το 1 μέχρι το 10 και οι κόμβοι του b έχουν ids από το 1 μέχρι το 15 τότε οι κόμβοι του c θα έχουν ids από το 1 μέχρι το 25 (οι κόμβοι από το 1 μέχρι το 10 θα είναι του a και οι κόμβοι από το 11 μέχρι 25 θα είναι του b). Φυσικά προσέχουμε ώστε να μην αλλάξουμε id στους κόμβους του b που είναι κοινοί με το a.

Τα ids δεν είναι τα μόνα στοιχεία που αλλάζουν στο b. Όπως έχουμε περιγράψει, υπάρχει το element graphics το οποίο το οποίο έχει το argument "x". Αυτό βρίσκεται στην παρακάτω θέση του XML:

```
<pathway>
  <entry>
    <graphics x="...">
```

Η τιμή του argument x περιέχει τη θέση στον οριζόντιο άξονα x όπου πρέπει να απεικονιστεί το συγκεκριμένο στοιχείο σε περίπτωση οπτικοποίησης του ΓΡΔ. Αν κρατήσουμε ανέπαφη αυτή την τιμή τότε η οπτικοποίηση του συνενωμένου ΓΡΔ θα περιέχει τα δύο συνιστώμενα ΓΡΔ το «ένα πάνω στο άλλο» κάτι τι οποίο είναι μη πρακτικό. Για αυτό το λόγο αυξάνουμε τη τιμή των x των entries του αντικειμένου b που δεν έχουν το ίδιο όνομα με τα entries του a κατά 1000 pixels. Με αυτόν τον τρόπο επιτυγχάνουμε να απεικονίσουμε τα μη κοινά στοιχεία του b πιο δεξιά κάνοντας εφικτή την πρακτική οπτικοποίηση του συνενωμένου ΓΡΔ. Η τιμή 1000 επιλέχθηκε μετά από πολλές δοκιμές και έδειξε ότι είναι η κατάλληλη για άρτιο οπτικό αποτέλεσμα. Φυσικά μπορεί κάποιος να μεταβάλει την τιμή αυτή ή να μην αλλάξει τα x του b αλλά τα y, έτσι ώστε το b να φαίνεται είτε πιο πάνω είτε πιο κάτω από το a. Στο κεφάλαιο 3 υπάρχουν πολλά παραδείγματα οπτικοποίησης συνενωμένων ΓΡΔ.

Το επόμενο βήμα είναι να υπολογίσουμε τα relations του συνενωμένου ΓΡΔ. Αρχικά τα relations του a αντιγράφονται στα relations του c. Στη συνέχεια προσθέτουμε όλα τα relations του b στα relations του c.

Καταρχήν αλλάζουμε το κλειδί των relations του b (το οποίο είναι ένας αριθμός από το 1 μέχρι το πλήθος των ακμών του b) προσθέτοντας το πλήθος των ακμών του a. Ο λόγος που το κάνουμε αυτό είναι για να εξασφαλίσουμε ότι όπως και στα a, b έτσι και στο c τα relations θα έχουν μοναδικό κωδικό έναν αύξοντα αριθμό από το 1 μέχρι το πλήθος των ακμών στον συνενωμένο γράφο. Δηλαδή, αν τα κλειδιά των relations του a είναι από το 1 μέχρι 20 και τα κλειδιά των relations του b είναι από το 1 μέχρι το 25 τότε αλλάζουμε τα κλειδιά των relations του b ώστε να είναι από το $20+1$ μέχρι το $20+25$ δηλαδή από το 21 μέχρι το 45. Στη συνέχεια προσθέτουμε τα relations του a και του b στο c. Με αυτόν τον τρόπο το c θα έχει όλα τα relations και του a και του b στη σειρά.

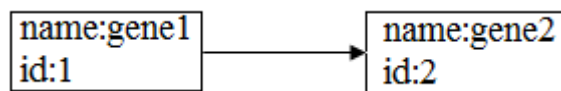
Όπως έχουμε πει, κάθε relation έχει ένα entry1 και ένα entry2. Τα πεδία αυτά προσδιορίζουν το id του κόμβου από όπου ξεκινάει και καταλήγει η ακμή αντίστοιχα. Αφού κατά τη συνένωση, όλα τα ids των κόμβων του b έχουν αλλάξει, πρέπει να αλλάξουν προφανώς και τα entry1 και entry2 όλων των relations του b έτσι ώστε να δείχνουν στους σωστούς κόμβους. Για να γίνει αυτό εξετάζουμε όλα τα relations του b. Αν το entry1 υπάρχει στο namesCommon τότε ο κόμβος από όπου ξεκινάει αυτή η ακμή είναι ένα κοινό γονίδιο. Άρα το entry1 πρέπει να έχει κοινό id, άρα πρέπει να αλλάξει. Για να αλλάξει ψάχνουμε στα ονόματα των κόμβων του a να βρούμε ένα γονίδιο το οποίο να έχει το ίδιο όνομα με το entry1. Αφού το βρούμε παίρνουμε το id του και το βάζουμε στο entry1 του c. Αν το entry1 δεν υπάρχει στο namesCommon τότε το entry1 πάλι αλλάζει αυτή τη φορά όμως στην τιμή που είχε πριν συν το πλήθος των κόμβων του a. Ακριβώς οι ίδιοι έλεγχοι γίνονται για το entry2.

2.5.3 Παράδειγμα συνένωσης ΓΡΔ

Για την καλύτερη κατανόηση της μεθόδου παραθέτουμε ένα παράδειγμα. Έστω το ΓΡΔ α:

```
<pathway>
<entry name="gene1" id="1"></entry>
<entry name="gene2" id="2"></entry>
<relation entry1="1" entry2="2"></relation>
</pathway>
```

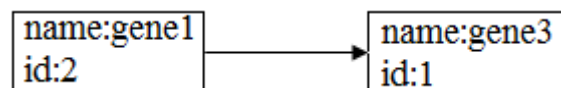
Το οποίο αναπαρίσταται γραφικά από τον εξής γράφο:



Έστω επίσης το ΓΡΔ β:

```
<pathway>
<entry name="gene1" id="2"></entry>
<entry name="gene3" id="1"></entry>
<relation entry1="2" entry2="1"></relation>
</pathway>
```

Το οποίο αναπαρίσταται γραφικά από τον εξής γράφο:



Αρχικά το συνενωμένο ΓΡΔ c θα είναι μία αντιγραφή του a. Στη συνέχεια αλλάζουμε τα ids του b ώστε να προσθέσουμε τη τιμή 2 (που είναι το πλήθος των κόμβων του a) και τα προσθέτουμε στο c. Οπότε το c γίνεται:

```
<pathway>
<entry name="gene1" id="1"></entry>
<entry name="gene2" id="2"></entry>
<relation entry1="1" entry2="2"></relation>

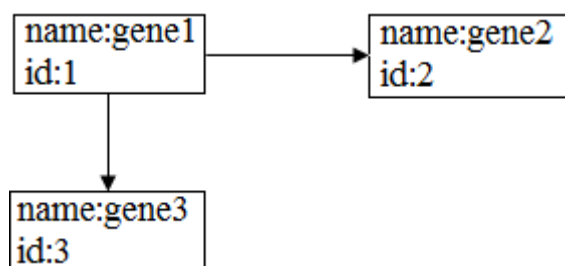
<entry name="gene1" id="4"></entry>
<entry name="gene3" id="3"></entry>
<relation entry1="4" entry2="3"></relation>
</pathway>
```

Στη συνέχεια παίρνουμε όλα τα relations του b που προσθέσαμε και κοιτάμε αν πρέπει να αλλάξει κάποιο entry1 ή entry2 λόγω κοινού ονόματος. Έτσι το entry1 είναι το 4 του οποίου το όνομα είναι gene1 το οποίο έχει το ίδιο όνομα με τον κόμβο του a που έχει id=1. Οπότε αλλάζει η τιμή του entry1 του συγκεκριμένου relation από 4 σε 1. Το entry2 έχει ήδη αλλάξει, δεν είναι κοινό όνομα οπότε το αφήνουμε όπως είναι. Το τελικό ΓΡΔ είναι το:

```
<pathway>
<entry name="gene1" id="1"></entry>
<entry name="gene2" id="2"></entry>
<relation entry1="1" entry2="2"></relation>

<entry name="gene1" id="1"></entry>
<entry name="gene3" id="3"></entry>
<relation entry1="1" entry2="3"></relation>
</pathway>
```

Του οποίου η γραφική παράσταση έχει την αναμενόμενη μορφή:



2.5.4 Παρατηρήσεις σχετικά με τη συνένωση ΓΡΔ

Ο συνενωμένος γράφος που δημιουργεί ο αλγόριθμος που παρουσιάσαμε δεν είναι εγγυημένο ότι θα είναι ιδανικά σωστός. Αυτό οφείλεται σε μία αδυναμία των KGML δεδομένων. Συγκεκριμένα, σε ένα KGML αρχείο μπορεί να υπάρχουν δύο κόμβοι με το ίδιο όνομα (π.χ. gene1) και διαφορετικά id (π.χ. 1 και 2). Αν αυτό το αρχείο θέλουμε να το συνενώσουμε με ένα άλλο KGML γράφο ο οποίος επίσης περιέχει γονίδιο με τον όνομα gene1, τότε δεν υπάρχει κάποιος τρόπος να βρούμε ποιο id είναι το σωστό (το 1 ή το 2). Με αυτόν τον τρόπο μπορεί ο συνενωμένος γράφος να μην περιέχει σωστή πληροφορία καθώς προκύπτουν δύο διαφορετικές οπτικοποιήσεις (visualizations) του συνενωμένου γράφου. Αν και δεν μπορούμε με τα υπάρχοντα δεδομένα να βρούμε ποιος είναι ο σωστός, ένας ειδικός μπορεί συνήθως να συμπεράνει ότι οι νέες συνδέσεις που εμφανίζει να έχει το gene1 στον συνενωμένο γράφο «φαίνεται» να είναι σωστές ή όχι. Μία πιθανή αντιμετώπιση του προβλήματος είναι η προεπισκόπηση των αρχείων KGML από κάποιον ειδικό ο οποίος θα δώσει διαφορετικά ονόματα στα γονίδια που έχουν κοινό όνομα.

Η συνένωση που παρουσιάσαμε μπορεί να ειπωθεί και ως μία αλγεβρική πράξη όπου όμως $a+b \neq b+a$. Αυτό συμβαίνει φυσικά λόγω των κοινών γονιδίων. Μπορούμε προφανώς να συνενώσουμε παραπάνω από δύο αρχεία ή και όλα τα ΓΡΔ τα οποία έχουμε για μία συγκεκριμένη περιοχή της βιολογίας. Για παράδειγμα, το KEGG παρέχει 10 ΓΡΔ σχετικά με το ανοσοποιητικό σύστημα των οποίων η συνένωση ενδεχομένως να παρέχει μία ολοκληρωμένη εικόνα όλης της ανοσοποιητικής λειτουργίας.

Η πολυπλοκότητα της μεθόδου εξαρτάται από δύο παραμέτρους. Ο πρώτος είναι το πλήθος των ακμών των ΓΡΔ. Ένα αρκετά πολύπλοκο ΓΡΔ δύσκολα φτάνει τις 300

συνδέσεις. Ο δεύτερος παράγοντας είναι η πολυπλοκότητα των διαδικασιών αναζήτησης που παρέχει η Python στη δομή δεδομένων τύπου λεξικό. Δεδομένης πάντως της βελτιστοποίησης που παρέχει η Python καθώς και του μικρού σχετικά πλήθους των ακμών των ΓΡΔ, η διαδικασία της συνένωσης διαρκεί το πολύ μερικά δευτερόλεπτα ακόμα και στα πιο πολύπλοκα δίκτυα.

3. Οπτικοποίηση των Γενομικών Ρυθμιστικών Δικτύων

Η μοντελοποίηση και η οντολογική κατηγοριοποίηση των στοιχείων ενός ΓΡΔ μέσω του προτύπου KGML ήταν ένα μέρος της εργασίας το οποίο αναλύθηκε στο προηγούμενο κεφάλαιο. Σε αυτό το κεφάλαιο θα αναλύσουμε τους τρόπους με τους οποίους η μοντελοποιημένη μέσω KGML πληροφορία που περιέχουν τα ΓΡΔ μπορεί να οπτικοποιηθεί. Γενικότερα, τα ΓΡΔ μπορούν να ιδωθούν ως μία προσπάθεια μοντελοποίησης των γενομικών αλληλεξαρτήσεων ώστε να είναι δυνατή η ολική αναπαράστασή τους. Μπορούμε να έχουμε δύο ειδών αναπαραστάσεις. Η πρώτη μέσω κάποιου αναγνωρισμένου προτύπου όπως είναι το KGML, και η δεύτερη μέσω της οπτικής αναπαράστασης. Μέσω της οπτικής αναπαράστασης περιμένουμε να βγάλουμε χρήσιμα συμπεράσματα σχετικά με τις εσωτερικές κανονιστικές δομές (regularities) που αναπτύσσονται μέσω της γονιδιακής έκφρασης. Ο τρόπος μέσω τον οποίων κατανέμονται στο χώρο τα γονίδια μέσω της οπτικοποίησης παρέχει τη δυνατότητα στον ειδικό να αναγνωρίσει γραφικά πρότυπα όπως είναι οι κλίκες, τα μονοπάτια, τα υπο-δίκτυα, τον εντοπισμό κεντρικών (hubs) και απομακρυσμένων (outliers) γονιδίων. Η βιολογική σημασία του εντοπισμού τέτοιων διατάξεων είναι πολύ μεγάλη τόσο για την ερμηνεία της βιολογικής λειτουργίας όσο και για τον εντοπισμό παθολογικών καταστάσεων. Επιπλέον, η οπτικοποίηση πρέπει να γίνεται με έναν ολοκληρωμένο τρόπο από ειδικό λογισμικό το οποίο να επιτρέπει την εφαρμογή αλγορίθμων γράφων, το μετασχηματισμό των γράφων σε διαφορετικές μορφές, την υποστήριξη μεταπληροφορίας στους κόμβους και στις ακμές και στη διασύνδεση με άλλες διαδικτυακές πηγές βιολογικής πληροφορίας.

3.1 Cytoscape

Το Cytoscape είναι ένα λογισμικό βιοπληροφορικής ανοικτού κώδικα για την οπτικοποίηση των δικτύων μοριακών αλληλεπιδράσεων. Η πληροφορία που οπτικοποιείται μπορεί να ολοκληρωθεί μέσω της προσθήκης επιπλέον δεδομένων, όπως είναι η γονιδιακή έκφραση και δεδομένα σχετικά με τη βιολογική κατάσταση των κόμβων και των ακμών. Επιπλέον λειτουργίες μπορούν να προστεθούν μέσω

plugins. Υπάρχουν plugins για την ανάλυση των δικτύων και του μοριακού προφίλ των κόμβων, για την εφαρμογή νέων γραφικών διατάξεων (layouts), για την υποστήριξη νέων τύπων αρχείων και για τη σύνδεση με εξωτερικές βάσεις δεδομένων. Μπορεί να γίνει ανάπτυξη νέων plugins μέσω προγραμματιστικού περιβάλλοντος που παρέχει το ίδιο το Cytoscape σε γλώσσα προγραμματισμού Java. Γενικότερα, έχει δημιουργηθεί μία ζωντανή κοινότητα από προγραμματιστές οι οποίοι αναπτύσσουν plugins για το Cytoscape υπό την έντονη υποστήριξη των δημιουργών του. Ένα από τα πλεονεκτήματα του είναι ότι αν και αρχικά σχεδιάστηκε για οπτικοποίηση βιολογικών γράφων, δεν περιορίζεται μόνο σε αυτούς. Μπορεί να γίνει οπτικοποίηση και χειρισμός γράφων οποιασδήποτε επιστημονικής περιοχής.

Αν και αρχικά δημιουργήθηκε από το Institute of Systems Biology του Seattle το 2002, τώρα συνεχίζεται η ανάπτυξή του από μία διεθνή ομάδα προγραμματιστών σύμφωνα με τις αρχές του ανοικτού κώδικα. Η τρέχουσα έκδοσή του είναι η 2.6.3 η οποία κυκλοφόρησε τον Ιούνιο του 2009, ενώ αναμένεται η έκδοση 3.0 η οποία θα παρέχει επιπλέον δυνατότητες καθώς και βελτιώσεις στη σταθερότητα.

Η επιλογή του Cytoscape για την οπτικοποίηση των ΓΡΔ έγινε επειδή παρέχει ορισμένες πολύ σημαντικές και χρήσιμες λειτουργίες. Αυτές οι λειτουργίες είναι:

- Εισαγωγή και εξαγωγή γράφων οι οποίοι περιγράφονται μέσω δημοφιλών τύπων αρχείων όπως είναι το GML (Graphics Modeling Language) και το XGMML (eXtensible Graph Markup Language).
- Χειρισμός πολλαπλών και αυθαίρετης μορφής μεταπληροφορίας σχετικά με τους κόμβους και τις ακμές. Αυτή η δυνατότητα κρίνεται πολύ σημαντική για την παρούσα εργασία καθώς το KGML format όπως το έχουμε παρουσιάσει περιέχει επιπλέον πληροφορίες σχετικά με την ονοματολογία των γονιδίων, τους διαδικτυακούς συνδέσμους σε γενομικές πηγές πληροφορίας, όπως είναι το KEGG, και πληροφορίες για τη γραφική αναπαράσταση των κόμβων, όπως για παράδειγμα το σχήμα του κόμβου (τετράγωνο, κύκλος, κτλ) και το μέγεθός του. Εκτός από αυτό είναι δυνατό το φιλτράρισμα της μεταπληροφορίας, δίνοντας μας τη δυνατότητα να εντοπίζουμε και να παρουσιάζουμε κόμβους οι οποίοι ικανοποιούν συγκεκριμένα κριτήρια. Για παράδειγμα, μπορούμε να εντοπίζουμε μόνο τους κόμβους οι οποίοι

περιέχουν γονίδια που βρίσκονται μόνο στο ανθρώπινο γονιδίωμα ή τους κόμβους οι οποίοι περιέχουν γονίδια τα οποία είναι γνωστό ότι ευθύνονται για συγκεκριμένες ασθένειες.

- Διασύνδεση με εξωτερικές πηγές γονιδιακής πληροφορίας. Ήδη είδαμε ότι το KGML περιέχει συνδέσμους στο KEGG. Οι σύνδεσμοι αυτοί διαχειρίζονται ως μετα-πληροφορία από το Cytoscape. Για να μεταφερθούμε σε μία από αυτές τις πηγές αρκεί να κάνουμε κλικ στον αντίστοιχο κόμβο.
- Για κάθε κόμβο μπορούμε να εισάγουμε επιπλέον πληροφορίες οι οποίες δεν βρίσκονται στο αρχείο το οποίο περιγράφει το ΓΡΔ. Το Cytoscape μπορεί να εντοπίσει αντιστοιχίες μεταξύ ονοματολογιών. Οι ονοματολογίες τις οποίες μπορεί να εντοπίσει είναι εκτός από το KEGG και το GO⁴ (Gene Ontology).
- Επιτρέπει την αποθήκευση ολόκληρου του περιβάλλοντος εργασίας σε ένα αρχείο ώστε να είναι δυνατή η επαναφόρτωσή του και ο διαμοιρασμός του με άλλες ομάδες εργασίας.

Από την πλευρά της οπτικοποίησης το Cytoscape επιτρέπει την εφαρμογή πολλών οπτικών στυλ καθώς και την αντιστοιχία μεταξύ κάποιου οπτικού χαρακτηριστικού με κάποια από τις μεταπληροφορίες. Δηλαδή, είναι δυνατόν να εμφανίζουμε το μέγεθος κάποιου κόμβου ανάλογα με την αριθμητική τιμή κάποιας από τις μεταπληροφορίες. Γενικότερα, επιτρέπει με ευκολία τον οπτικό χειρισμό (zoom in, zoom out, navigation) πολύ μεγάλων γράφων (πάνω από 100.000 κόμβους).

3.2 XGMML

Το XGMML (the Extensible Graph Markup and Modeling Language) είναι ένα πρότυπο σε μορφή XML το οποίο βασίζεται στο GML (Graph Modeling Language) που είναι ένα καθολικό πρότυπο για περιγραφή γράφων. Πρακτικά, αν και το GML δεν σχετίζεται με το XML ούτε με τη SGML, το XGMML είναι ένα XML πρότυπο το οποίο έχει σχεδιαστεί ώστε να παρέχει ένα προς ένα αντιστοιχία με τους κόμβους του GML. Στην ουσία δηλαδή το XGMML είναι μία XML υλοποίηση του GML έτσι ώστε να καταστεί εύκολη η μετάβαση από το ένα format στο άλλο. Το πιο σημαντικό πρόγραμμα το οποίο παρέχει υποστήριξη στο XGMML είναι το Cytoscape. Εκτός από αυτό, υποστήριξη παρέχουν και άλλες γνωστές εφαρμογές χειρισμού γράφων

⁴ <http://www.geneontology.org/>

όπως είναι το BIOMAX και το JNets. Περισσότερες πληροφορίες για το XGMML υπάρχουν στο <http://www.cs.rpi.edu/~puninj/XGMML/>. Στην Εικόνα 3.1 φαίνεται ένα παράδειγμα ενός αρχείου σε XGMML.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<graph label="Network" xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:cy="http://www.cytoscape.org" xmlns="http://www.cs.rpi.edu/XGMML" directed="1">
  <att name="documentVersion" value="1.1"/>
  <att name="networkMetadata">
    <rdf:RDF>
      <rdf:Description rdf:about="http://www.cytoscape.org/">
        <dc:type>Protein-Protein Interaction</dc:type>
        <dc:description>N/A</dc:description>
        <dc:identifier>N/A</dc:identifier>
        <dc:date>2009-01-15 18:38:04</dc:date>
        <dc:title>Network</dc:title>
        <dc:source>http://www.cytoscape.org/</dc:source>
        <dc:format>Cytoscape-XGMML</dc:format>
      </rdf:Description>
    </rdf:RDF>
  </att>
  <att type="string" name="backgroundColor" value="#ccccff"/>
  <att type="real" name="GRAPH_VIEW_ZOOM" value="1.0"/>
  <att type="real" name="GRAPH_VIEW_CENTER_X" value="0.0"/>
  <att type="real" name="GRAPH_VIEW_CENTER_Y" value="0.0"/>
  <node id="56" label="FAS" name="hsa:355">
    <att type="string" name="canonicalName" value="FAS"/>
    <att type="string" name="NODE_TYPE" value="DefaultNode"/>
    <att type="string" name="KGML_CODE" value="hsa:355"/>
    <att type="string" name="KGML_SHAPE" value="rectangle"/>
    <att type="string" name="KGML_LINK" value="http://www.genome.jp/dbget-bin/www.bget?hsa:355"/>
    <graphics type="ELLIPSE" h="40.0" w="40.0" x="221" y="166" fill="#ff9999" width="1" outline="#666666" cy:nodeTransparency="1.0" cy:nodeLabelFont="SansSerif.bold-0-12" cy:borderLineType="solid"/>
  </node>
  <edge label="59-53" source="59" target="53">
    <att type="string" name="canonicalName" value="59-53"/>
    <att type="string" name="EDGE_TYPE" value="DefaultEdge"/>
    <graphics width="1" fill="#0000ff" cy:sourceArrow="0" cy:targetArrow="3" cy:sourceArrowColor="#000000" cy:targetArrowColor="#000000" cy:edgeLabelFont="Default-0-10" cy:edgeLineType="SOLID" cy:curved="STRAIGHT_LINES"/>
  </edge>
</graph>
```

Εικόνα 3.1. Παράδειγμα XGMML αρχείου

3.2.1 Μετατροπή από KGML σε XGMML

Όπως έχουμε παρουσιάσει η οπτικοποίηση των ΓΡΔ είναι ένα κρίσιμο και σημαντικό ζητούμενο στην περιοχή της βιοπληροφορικής. Επίσης μία από τις πιο σημαντικές εφαρμογές για τον χειρισμό γράφων είναι το Cytoscape το οποίο κυρίως χειρίζεται γράφους σε μορφή XGMML. Οπότε για την οπτικοποίηση των ΓΡΔ είναι απαραίτητο να αναπτύξουμε μία εφαρμογή η οποία θα μετατρέπει από KGML σε XGMML.

Η εφαρμογή που αναπτύξαμε κάνει χρήση των μεθόδων μοντελοποίησης και επεξεργασίας των KGML αρχείων οι οποίες παρουσιάστηκαν στο κεφάλαιο 2. Αρχικά τυπώνουμε την επικεφαλίδα του αρχείου η οποία είναι η εξής:

```

<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<graph label="Network" xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:cy="http://www.cytoscape.org"
xmlns="http://www.cs.rpi.edu/XGML" directed="1">
  <att name="documentVersion" value="1.1"/>
  <att name="networkMetadata">
    <rdf:RDF>
      <rdf:Description rdf:about="http://www.cytoscape.org/">
        <dc:type>Protein-Protein Interaction</dc:type>
        <dc:description>N/A</dc:description>
        <dc:identifier>N/A</dc:identifier>
        <dc:date>2009-01-15 18:38:04</dc:date>
        <dc:title>Network</dc:title>
        <dc:source>http://www.cytoscape.org/</dc:source>
        <dc:format>Cytoscape-XGML</dc:format>
      </rdf:Description>
    </rdf:RDF>
  </att>
  <att type="string" name="backgroundColor" value="#ccccff"/>
  <att type="real" name="GRAPH_VIEW_ZOOM" value="1.0"/>
  <att type="real" name="GRAPH_VIEW_CENTER_X" value="0.0"/>
  <att type="real" name="GRAPH_VIEW_CENTER_Y" value="0.0"/>

```

Σε αυτή την επικεφαλίδα δηλώνουμε διάφορες παραμέτρους του γράφου όπως την ημερομηνία δημιουργίας του, το όνομα του γράφου και το όνομα της εφαρμογής η οποία το δημιούργησε. Επίσης, δηλώνουμε το κεντρικό σημείο αναφοράς στη δισδιάστατη επιφάνεια το οποίο είναι το σημείο [0,0].

Στη συνέχεια εκτελούνται τα παρακάτω βήματα:

1. Για όλους τους κόμβους τυπώνουμε το id του κόμβου, το όνομά του και το KGML όνομά του (σε διπλά άγκιστρα έχουμε βάλει τις απαραίτητες μεταβλητές):

```
<node id='{{id}}' label='{{name}}' name='{{kgml_name}}' >
```

2. Προσθήκη μεταπληροφορίας των κόμβων σχετικά με το σχήμα του κόμβου

```
<att type="string" name="KGML_SHAPE" value= {{shape}}>
```

Αν ο κόμβος είναι ένα γονίδιο τότε το σχήμα είναι ένα τετράγωνο, ενώ αν είναι ένα άλλο ΓΡΔ τότε είναι ένα τετράγωνο με στρογγυλοποιημένες άκρες.

3. Προσθήκη μεταπληροφορίας των κόμβων σχετικά με το σύνδεσμο στη βάση δεδομένων του KEGG:

```
<att type="string" name="KGML_LINK" value={{KGML_LINK}}>
```

4. Προσθήκη μεταπληροφορίας των κόμβων σχετικά με τη x,y θέση των γονιδίων στο ΓΡΔ:

```
<graphics type="ELLIPSE" h="40.0" w="40.0" x={{x_pos}} y={{y_pos}}>
```

5. Για όλες τις ακμές τυπώνουμε το id του κόμβου προέλευσης και του κόμβου προορισμού τους καθώς και ένα μοναδικό label για κάθε ακμή.

```
<edge label={{id1}}-{{id2}} source={{id1}} target={{id2}} >
```

6. Για κάθε ακμή το βέλος έχει διαφορετικό σχήμα ανάλογα με το είδος της ακμής:

```
<graphics width="1" fill="#0000ff" cy:sourceArrow="0"  
cy:targetArrow={{arrow_type}}>
```

Στη συνέχεια περιγράφουμε το νόημα των διαφορετικών ειδών ακμών:

- Activation. Μορφή βέλους: $A \rightarrow B$. Δηλώνει ότι το γονίδιο B ενεργοποιείται από το γονίδιο A.
- Inhibition. Μορφή βέλους: $A \dashv B$. Δηλώνει ότι το γονίδιο B απενεργοποιείται από το γονίδιο A.
- Association. Μορφή βέλους: $A \dashv\!\!\!\rightarrow B$. Δηλώνει ότι το γονίδιο A ενεργοποιείται μόνο αν είναι ενεργοποιημένο το γονίδιο B και απενεργοποιείται μόνο αν το γονίδιο B είναι απενεργοποιημένο.
- Dissociation. Μορφή βέλους: $A \dashv\!\!\!\leftarrow B$. Δηλώνει ότι το γονίδιο A ενεργοποιείται μόνο αν το γονίδιο B είναι απενεργοποιημένο και απενεργοποιείται μόνο αν το γονίδιο B είναι ενεργοποιημένο.

Στο τέλος αυτής της διαδικασίας έχουμε έναν γράφο σε μορφή XGMML ο οποίος περιέχει τα ίδια περίπου⁵ οπτικά χαρακτηριστικά με τον αρχικό γράφο KGML ο οποίος κωδικοποιεί κάποιο ΓΡΔ.

⁵ Βλ. Παράγραφο 2.3.1

4. Εφαρμογές και παραδείγματα

Σε αυτό το κεφάλαιο παρουσιάζουμε λεπτομέρειες της εφαρμογής η οποία υλοποιεί τις μεθόδους που έχουμε παρουσιάσει. Η παρουσίαση γίνεται με πραγματικά παραδείγματα σε υπάρχοντα ΓΡΔ. Όπως έχουμε παρουσιάσει η εφαρμογή έχει γραφτεί στη γλώσσα προγραμματισμού Python. Το αρχείο το οποίο περιέχει τον κώδικα ονομάζεται KGML2XGMMML.py και η εκτέλεσή του γίνεται απευθείας μέσω του μεταφραστή Python. Η εκτέλεσή του χωρίς παραμέτρους δίνει μία περιγραφή των δυνατών παραμέτρων:

```
C:\>python KGML2XGMMML.py
Converting/Combining tool for KGML files to XGMMML format
(c) Ioannis Marakis. Technical University of Crete,
Electronic & Computer Engineering
Usage:
KGML2XGMMML -convert <KGML File(s)>
KGML2XGMMML -combine <KGML Files>
```

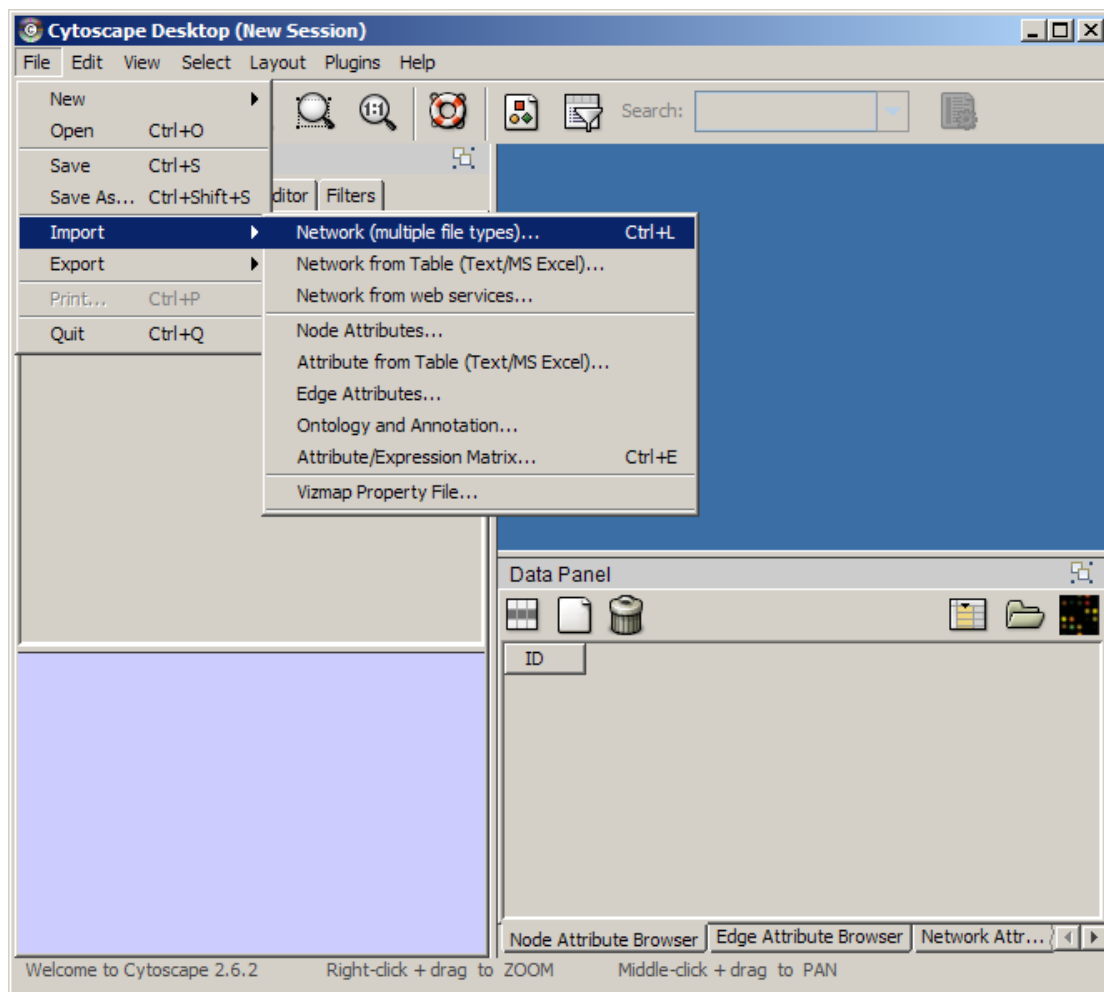
Ακολουθούν διάφορα παραδείγματα χρήσης μαζί με τα αποτελέσματα όπως οπτικοποιούνται από το Cytoscape.

4.1 Παράδειγμα μετατροπής ΓΡΔ

Έστω αρχικά ότι έχουμε το ΓΡΔ το οποίο περιγράφει τη βιολογική λειτουργία του φυσιολογικού κυτταρικού θανάτου γνωστού και ως απόπτωση. Αυτό το ΓΡΔ υπάρχει σε KGML μορφή με το όνομα hsa04210.xml . Για να το μετατρέψουμε σε XGMMML εκτελούμε την εντολή:

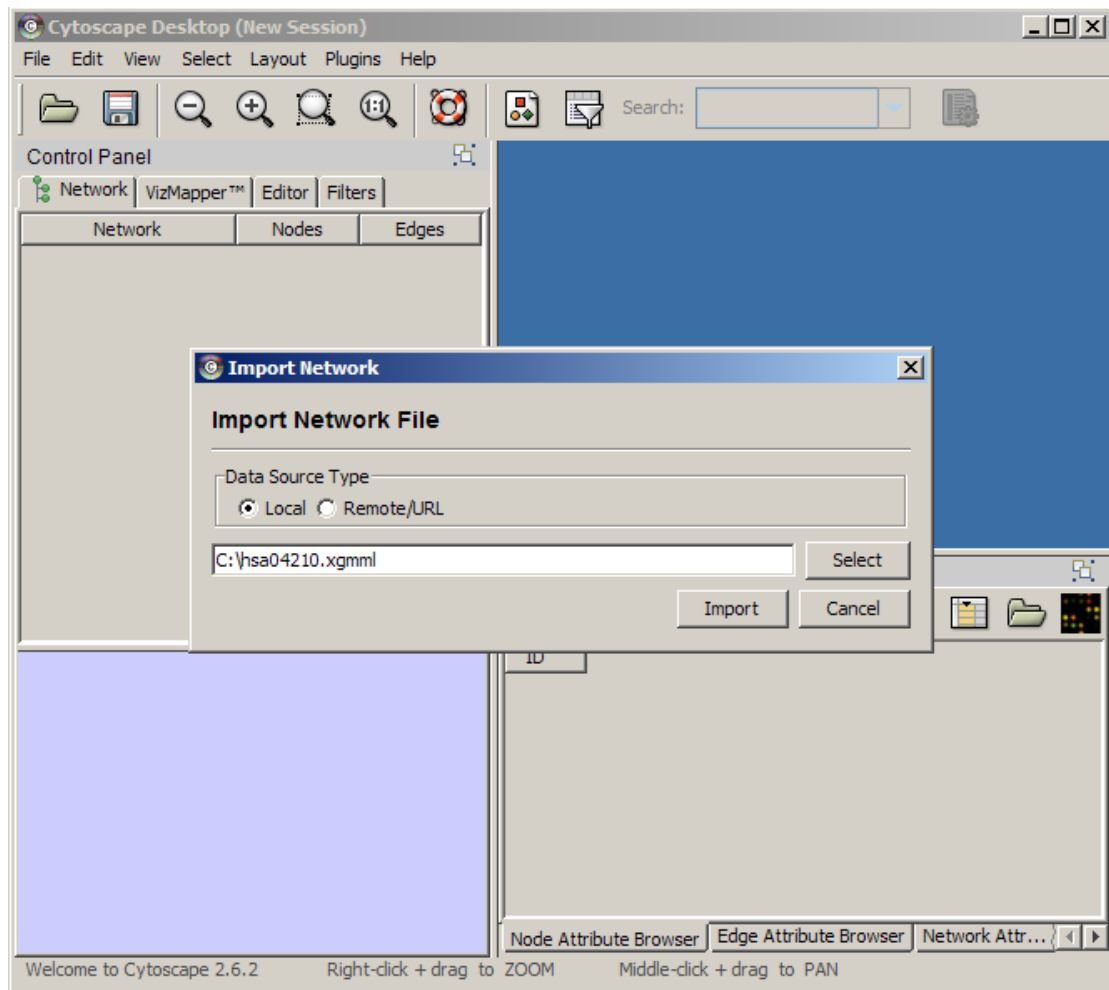
```
C:\>python KGML2XGMMML.py -convert hsa04210.xml
```

Η εντολή αυτή παράγει το αρχείο hsa04210.xgmmml το οποίο μπορούμε να φορτώσουμε από το Cytoscape μέσω File → Import → Network (multiple file types):



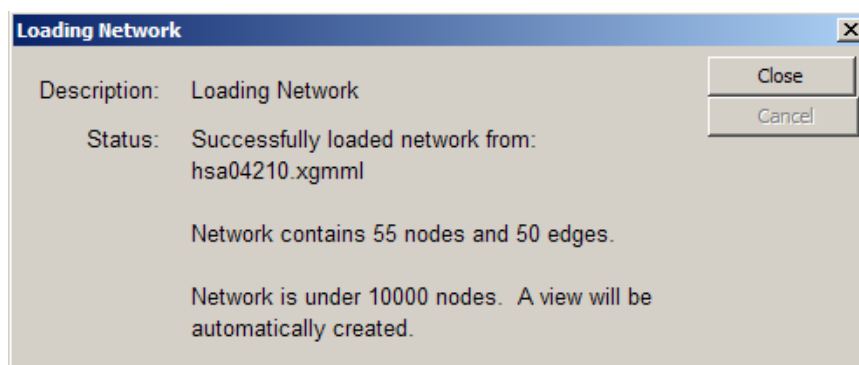
Εικόνα 4.1. Εισαγωγή XGMML αρχείου από το Cytoscape

Στη συνέχεια επιλέγουμε το αρχείο hsa04210.xgmml :



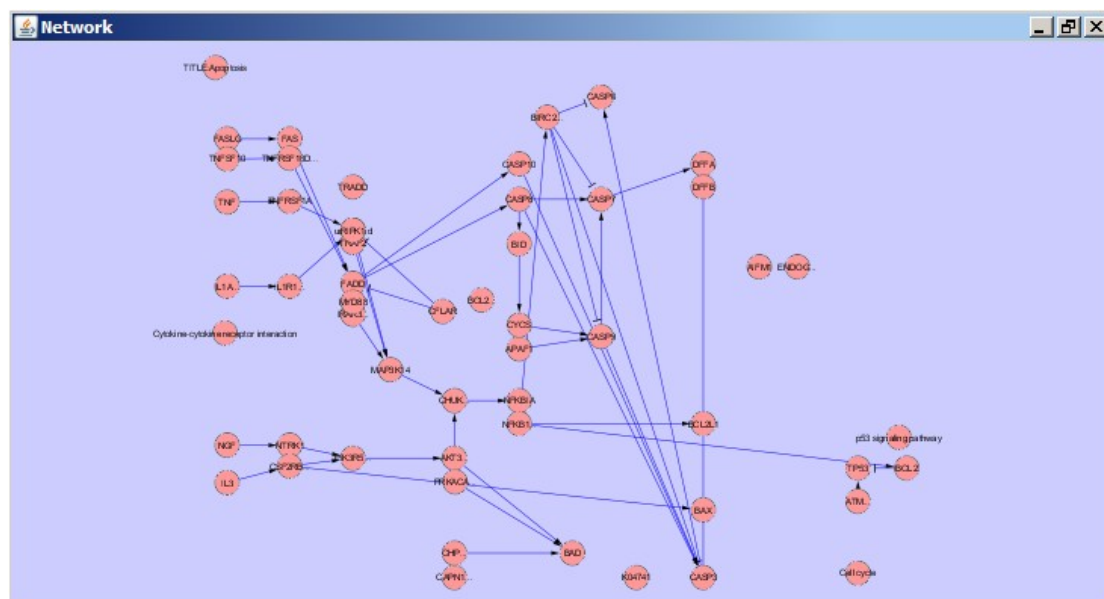
Εικόνα 4.2. Επιλογή του XGMML αρχείου

Το αρχείο φορτώνεται από το Cytoscape και διάφορες πληροφορίες σχετικά με το πλήθος των κόμβων και των ακμών του ΓΡΔ εμφανίζονται:



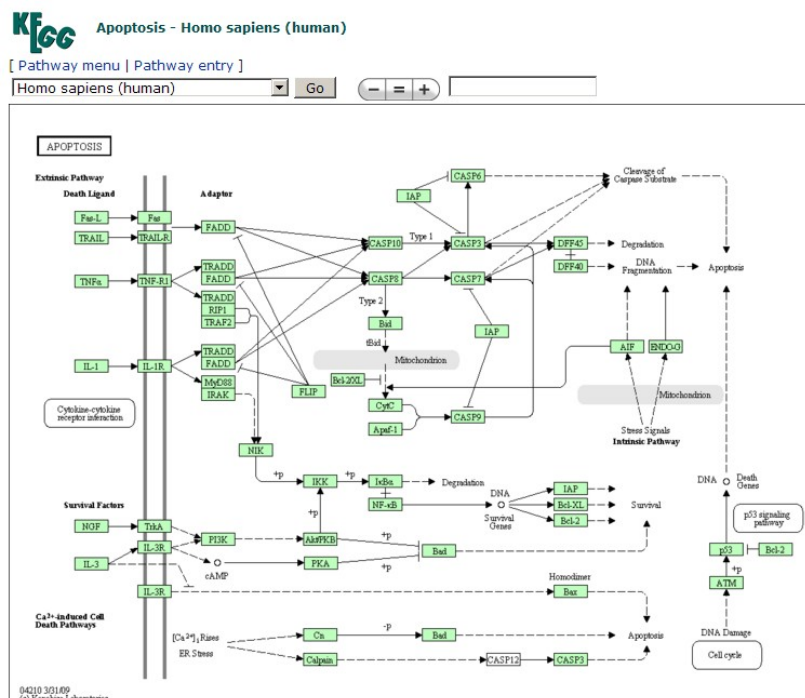
Εικόνα 4.3. Πληροφορίες για το πλήθος κόμβων / ακμών του γράφου

Η γραφική αναπαράσταση του ΓΡΔ της απόπτωσης φαίνεται στο Cytoscape ως εξής:



Εικόνα 4.4. Εμφάνιση του ΓΡΔ της απόπτωσης

Είναι σημαντικό να αντιπαραβάλουμε τη γραφική αναπαράσταση του ίδιου ΓΡΔ από το KEGG:



Εικόνα 4.5. Το ΓΡΔ της απόπτωσης από το KEGG

Παρατηρούμε ότι υπάρχει μεγάλη ποιοτική διαφορά μεταξύ των δύο αναπαραστάσεων αλλά όχι ποσοτική, καθώς όλοι οι κόμβοι και σχεδόν όλες οι ακμές υπάρχουν. Στη παράγραφο 2.3.1 αναλύσαμε το συγκεκριμένο ζήτημα. Το σημαντικό είναι ότι μέσω του Cytoscape μπορούμε να χειριστούμε δυναμικά τον γράφο κάτι που δεν μπορούσαμε να κάνουμε από την HTML παρουσίαση που διαθέτει το KEGG. Είναι επίσης δυνατόν να γίνει μετατροπή περισσοτέρων από ένα KGML αρχείων. Αρκεί να δοθούν παραπάνω από ένα όνομα αρχείου κατά την εκτέλεση της εντολής. Π.χ:

```
C:\>python KGML2XGMMML.py -convert hsa04210.xml hsa04110.xml
```

Σε περιβάλλον UNIX μπορούν να χρησιμοποιηθούν και wildcards:

```
#>python KGML2XGMMML.py -convert *.xml
```

4.2 Χειρισμός γράφου από το Cytoscape

Μέσω του Cytoscape μπορούμε να διερευνήσουμε τον γράφο εφαρμόζοντας απλές πράξεις όπως zoom-in, zoom-out. Επίσης, μπορούμε να χειριστούμε τη μεταπληροφορία η οποία εμφανίζεται με τη παρακάτω μορφή:

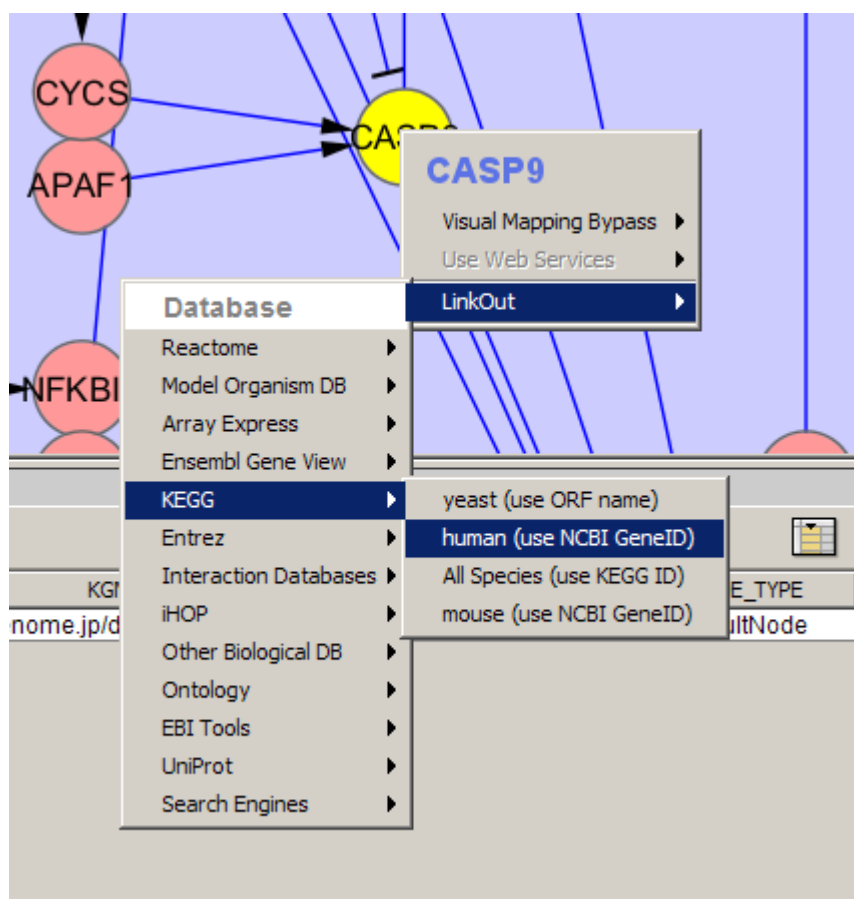
ID	KGML_CODE	KGML_LINK	
TNFRSF10D...	hsa:8793 hsa:87...	http://www.genome.jp/dbget-bin/www_bget?hsa+8...	r
FASLG	hsa:356	http://www.genome.jp/dbget-bin/www_bget?hsa+3...	r
NTRK1	hsa:4914	http://www.genome.jp/dbget-bin/www_bget?hsa+4...	r
TRAF2	hsa:7186	http://www.genome.jp/dbget-bin/www_bget?hsa+7...	r
AIFM1	hsa:9131	http://www.genome.jp/dbget-bin/www_bget?hsa+9...	r
NFKB1...	hsa:4790 hsa:59...	http://www.genome.jp/dbget-bin/www_bget?hsa+4...	r
CASP3	hsa:836	http://www.genome.jp/dbget-bin/www_bget?hsa+8...	r
PIK3R5...	hsa:23533 hsa:5...	http://www.genome.jp/dbget-bin/www_bget?hsa+2...	r
◀			
Node Attribute Browser Edge Attribute Browser Network Attribute Browser			

Εικόνα 4.6. Η μετα-πληροφορία των κόμβων (σε άλλο tab υπάρχουν και για τις ακμές)

Μέσω αυτού του πίνακα μπορούμε να δούμε τη μεταπληροφορία των ακμών και των κόμβων που έχουμε επιλέξει από το ΓΡΔ.

Μία άλλη λειτουργία η οποία παρέχεται όπως έχουμε παρουσιάσει είναι η διασύνδεση με εξωτερικές πηγές βιολογικών δεδομένων. Κάνοντας δεξί κλικ σε έναν κόμβο εμφανίζεται μία λίστα με όλες τις διαθέσιμες πηγές δεδομένων. Επιλέγοντας μία από αυτές ανοίγει μία σελίδα του διαδικτύου με την κατάλληλη εγγραφή σχετικά

με τον επιλεγθέντα κόμβο. Παράλληλα αν κάποια από τις μεταπληροφορίες περιέχει συνδέσμους στο διαδίκτυο, επιλέγοντάς τον μεταφερόμαστε επίσης στον αντίστοιχο δικτυακό τόπο. Στην Εικόνα 4.7 βλέπουμε τις επιλογές που εμφανίζονται για το γονίδιο CASP9.



Εικόνα 4.7. Σύνδεση με online πηγή γενομικών δεδομένων.

4.3 Συνένωση δύο ή περισσότερων ΓΡΔ

Όπως έχουμε δει μπορούμε να συνενώσουμε δύο ή περισσότερα ΓΡΔ σε ένα αρχείο XGMML και μετά να το οπτικοποιήσουμε μέσω του Cytoscape. Η συνένωση γίνεται ως εξής:

```
C:\>python KGML2XGMML.py -combine hsa04210.xml hsa04110.xml
```

Στο παραπάνω παράδειγμα συνενώσαμε τα δύο αρχεία τα οποία περιέχουν τα ΓΡΔ τη απόπτωσης και του κυτταρικού κύκλου αντίστοιχα. Κατά την εκτέλεση του προγράμματος εμφανίζεται στην οθόνη μία λίστα με όλα τα κοινά γονίδια μεταξύ των

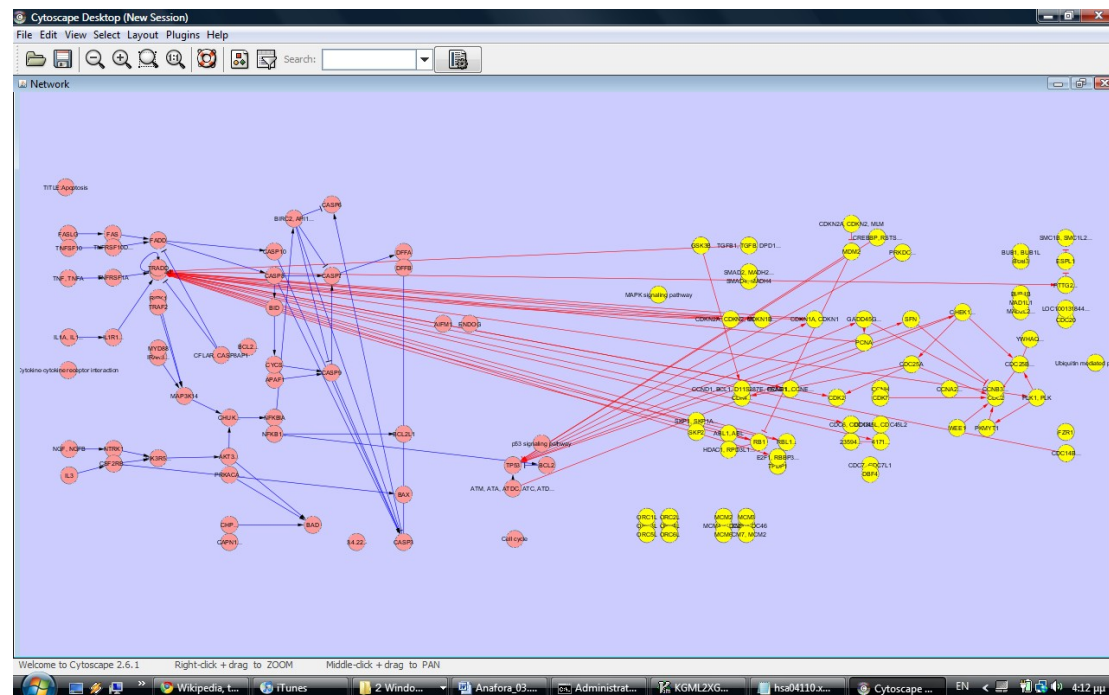
δύο ΓΡΔ με τη μορφή της δομής δεδομένων set που χειρίζεται τα σύνολα σε μορφή Python:

Common Genes:

```
set([u'path:hsa04210', u'path:hsa04110', u'undefined', u'hsa:7157'])
```

Ο συνενωμένος γράφος τυπώνεται στο αρχείο: hsa04210_hsa04110.xgml

Μπορούμε να δώσουμε το αρχείο αυτό στο Cytoscape για οπτικοποίηση:



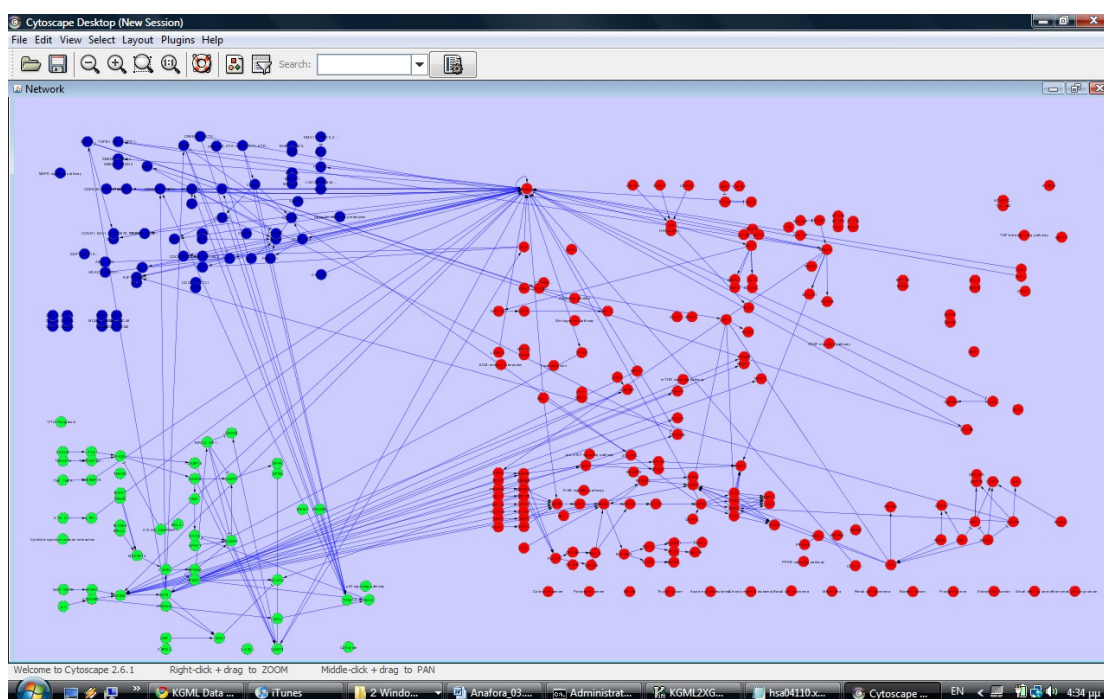
Εικόνα 4.8. Αναπαράσταση του συνενωμένου γράφου της απόπτωσης (κόκκινο) και του κυτταρικού κύκλου (κίτρινο)

Στην αριστερή πλευρά με κόκκινο χρώμα στους κόμβους φαίνεται το πρώτο ΓΡΔ (hsa04210.xml, απόπτωση), ενώ στα δεξιά με κίτρινο χρώμα στους κόμβους βρίσκεται το δεύτερο ΓΡΔ (hsa04110.xml, κυτταρικός κύκλος). Αυτό το γράφημα περιέχει πολύ σημαντική και πρωτότυπη πληροφορία. Με τη συνένωση μπορούν να φανούν τα γονίδια που έχουν κυρίαρχο ρόλο σε παραπάνω από μία βιολογική λειτουργία, όπως στο συγκεκριμένο παράδειγμα το γονίδιο TP53 το οποίο είναι γνωστό ως καρκινικός καταστολέας (tumor suppressor).

Ως συνέπεια της συνένωσης είναι τα μη ταυτοποιημένα γονίδια τα οποία υπάρχουν στα KGML αρχεία ως “undefined” να εμφανίζονται μία φορά στον γράφο και να

έχουν υψηλό αριθμό συνδέσεων. Αναμένουμε σε μελλοντικές εκδόσεις των ΓΡΔ από τον KEGG να υπάρχουν λιγότερα μη ταυτοποιημένα γονίδια.

Επίσης είναι δυνατόν να συνενώσουμε παραπάνω από 2 ΓΡΔ. Απλώς γράφουμε τη λίστα των ΓΡΔ που θέλουμε να συνενώσουμε κατά την εκτέλεση του προγράμματος ή χρησιμοποιούμε wildcards αν είμαστε σε περιβάλλον UNIX. Για παράδειγμα μπορούμε να συνενώσουμε την απόπτωση, τον κυτταρικό κύκλο και το γενικότερο ΓΡΔ για την καρκινική εκδήλωση:



Εικόνα 4.9. Συνενωμένου γράφος της απόπτωσης (πράσινο), κυτταρικού κύκλου (μπλε) και της καρκινικής εκδήλωσης (πράσινο)

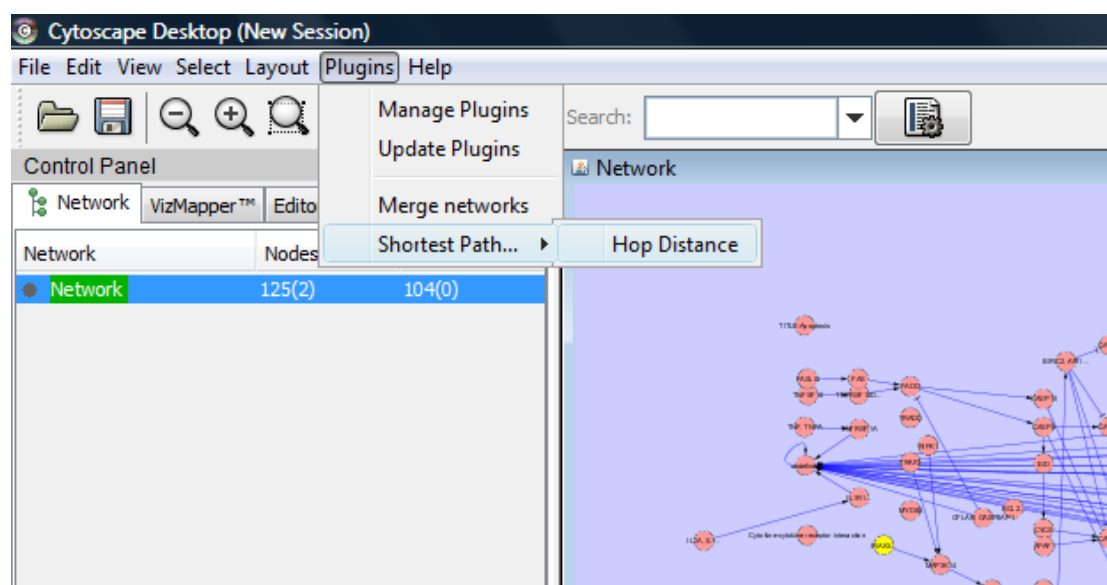
Σε αυτό το παράδειγμα υπάρχουν συνενωμένα η απόπτωση (πράσινο), ο κυτταρικός κύκλος (μπλε) και η καρκινική εκδήλωση (Pathways in cancer hsa05200.xml). Ο χρωματισμός είναι αυθαίρετος και έγινε χρησιμοποιώντας τα εργαλεία της διεπαφής του Cytoscape.

Σε περιβάλλον UNIX μπορούμε να συνενώσουμε ΓΡΔ με πιο ευέλικτο τρόπο χρησιμοποιώντας wildcards. Π.χ για να συνενώσουμε όλα τα ΓΡΔ που το όνομα του αρχείου τους έχει τη δομή hsa045XX.xml (είναι όλα τα ΓΡΔ που αφορούν την κυτταρική επικοινωνία) μπορούμε να χρησιμοποιήσουμε την εντολή:

```
python KGML2XGMML.py -convert hsa045???.xml
```

4.4 Εύρεση ελάχιστου μονοπατιού

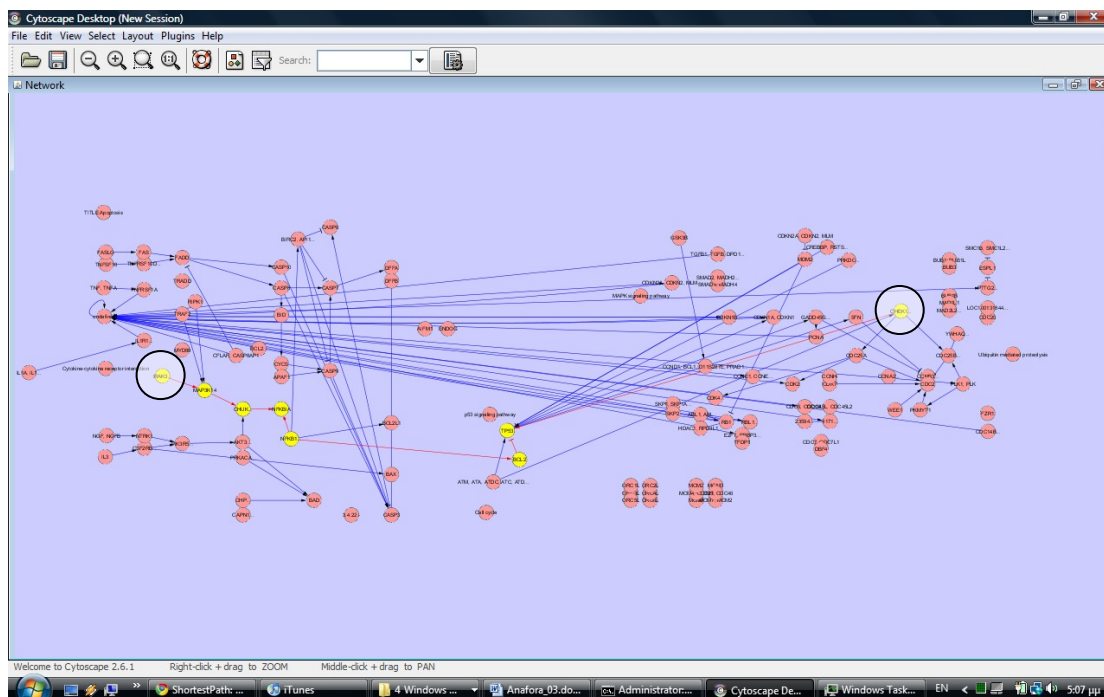
Η εύρεση ελάχιστου μονοπατιού είναι μία από τις πολλές λειτουργίες που παρέχει το Cytoscape σχετικά με χειρισμό γράφων. Η λειτουργία αυτή είναι διαθέσιμη μέσω plugin. Γενικότερα υπάρχει μία μεγάλη συλλογή από plugins διαθέσιμα στο διαδίκτυο για το Cytoscape για εξειδικευμένες ή μη λειτουργίες. Το plugin για την εύρεση του ελάχιστου μονοπατιού ονομάζεται ShortestPath⁶. Εγκαθίσταται με απλή αντιγραφή του αρχείου ShortestPath.jar στον κατάλογο plugins του Cytoscape. Αφού το εγκαταστήσουμε ο υπολογισμός γίνεται με την επιλογή των δύο κόμβων που θέλουμε να βρούμε την ελάχιστη απόσταση μεταξύ τους. Στη συνέχεια τρέχουμε το plugin (Plugins → Shortest Paths) και μας εμφανίζει το μήκος του ελάχιστου μονοπατιού.



Εικόνα 4.10. Επιλογή του plug in για την εύρεση του ελάχιστου μονοπατιού.

⁶ <http://www.rbvi.ucsf.edu/Research/cytoscape/shortestPath/index.html>

Οι κόμβοι που ανήκουν στο ελάχιστο μονοπάτι επιλέγονται αυτόματα:



Εικόνα 4.11. Το ελάχιστο μονοπάτι μεταξύ των κόμβων που βρίσκονται σε κύκλο.

Στο παραπάνω σχήμα βλέπουμε το ελάχιστο μονοπάτι μεταξύ των κόμβων που βρίσκονται στον κύκλο στον συνενωμένο γράφο της απόπτωσης και του κυτταρικού κύκλου. Παρόμοιες λειτουργίες μπορούν να γίνουν με την εγκατάσταση άλλων plugins.

5. Επίλογος

Τα Γενομικά Ρυθμιστικά Δίκτυα (ΓΡΔ) αποτελούν μία πολύτιμη πηγή πληροφορίας στη περιοχή της βιοπληροφορικής καθώς μοντελοποιούν τις αλληλοσυσχετίσεις των γονιδίων κατά τη διάρκεια των βιολογικών διεργασιών. Παρόλο που υπάρχει πληθώρα πηγών για ΓΡΔ δεν υπήρχε κάποιος ενοποιημένος τρόπος μοντελοποίησης και επεξεργασίας τους. Στην παρούσα εργασία παρουσιάσαμε μία ολοκληρωμένη λύση για τη διαχείρισή τους. Μέσω της υλοποίησης μας παρέχουμε πρωταρχικά έναν υπολογιστικό τρόπο χειρισμού των κόμβων και των ακμών των δικτύων με έναν τρόπο ώστε να μπορεί να χρησιμοποιηθεί από τους ερευνητές που δραστηριοποιούνται στη συγκεκριμένη περιοχή.

Αρχικά μέσω της γλώσσας προγραμματισμού Python παρέχουμε έναν τρόπο εσωτερικής αναπαράστασης των KGML αρχείων όπως αυτά παρέχονται από τον δημοφιλή οργανισμό KEGG. Στη συνέχεια υλοποιήσαμε μία μέθοδο συνένωσης πολλών ΓΡΔ. Με αυτόν τον τρόπο έγινε εφικτή η συνολική και συστημική αντιμετώπιση των ΓΡΔ, όχι ως μεμονωμένα δίκτυα, αλλά ως κομμάτια ενός πολύπλοκου μηχανισμού ο οποίος διέπει τη βιολογική λειτουργία και μας δίνει ενδείξεις για την φυσιολογική και παθολογική κυτταρική συμπεριφορά.

Για την επισκόπηση των βιολογικών λειτουργιών όπως αυτές μοντελοποιούνται από μεμονωμένα ή συνενωμένα ΓΡΔ παρείχαμε επίσης διαδικασίες μετατροπής των γράφων στο δημοφιλές πρότυπο γράφων XGMML. Με αυτόν τον τρόπο γίνεται εφικτή η οπτικοποίηση των ΓΡΔ μέσω ειδικών εφαρμογών όπως είναι το Cytoscape και ο δυναμικός χειρισμός τους. Τέλος, παρείχαμε οδηγίες για την περαιτέρω επεξεργασία των γράφων μέσω εφαρμογής γνωστών γραφο-θεωρητικών αλγορίθμων, όπως είναι η εύρεση ελάχιστου μονοπατιού.

Σε ένα μελλοντικό επίπεδο η παρούσα εργασία θα μπορούσε να επεκταθεί σε αρκετούς τομείς. Καταρχήν μπορούν να συμπεριληφθούν επιπλέον διεθνείς οργανισμοί που παρέχουν ΓΡΔ, ώστε να υπάρχει μία από κοινού μοντελοποίησή τους. Έτσι θα μπορούσαμε να διακρίνουμε τις διαφορές και τις ομοιότητες τους.

Περιληπτικά αναφέρουμε τη BioCarta⁷ το reActome⁸ και το BioPax⁹. Επίσης, εκτός από τη συνένωση θα μπορούσαμε να παρέχουμε επιπλέον αλγόριθμους χειρισμού πολλαπλών γράφων όπως είναι η τομή και η διαφορά. Το πλαίσιο που παρέχουμε επιτρέπει γενικότερα την εφαρμογή αυθαίρετων αλγορίθμων οπότε μπορούν να δοθούν αλγόριθμοι περεταίρω επεξεργασίας όπως είναι η εύρεση κύκλων και η εύρεση κεντρικών κόμβων (hubs). Τέλος το Cytoscape παρέχει επιπλέον επιλογές για την οπτική παρουσίαση των γράφων, οπότε μία πιθανή επέκταση είναι να χρησιμοποιήσουμε τις επιλογές αυτές για να παρουσιάσουμε οπτικά κάποιες από τις μετα-πληροφορίες (π.χ. εναλλακτικές ονομασίες γονιδίων, γονιδιακή έκφραση) ή τις ιδιότητες των κόμβων (π.χ. αν είναι μεταγραφικός παράγοντας ή γονίδιο με γνωστή παθολογία).

⁷ <http://www.biocarta.com/genes/index.asp>

⁸ <http://www.reactome.org/>

⁹ <http://www.biopax.org/>

Αναφορές

- [1] Thomas Schlitt and Alvis Brazma. Current approaches to gene regulatory network modeling. BMC Bioinformatics 2007, 8(Suppl 6):S9
- [2] Kanehisa M, Goto S: KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 2000, 28:27-30
- [3] XGMML (eXtensible Graph Markup and Modeling Language).
<http://www.cs.rpi.edu/~puninj/XGMML/>
- [4] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003, 13(11):2498-2504.

Βιβλιογραφία

Yuryev Anton. Pathway Analysis for Drug Discovery. Computational Infrastructure and Applications. Willey publications 2008.

Bernhard O. Palsson. Systems Biology: Properties of Reconstructed Networks. Cambridge University Press. January 2006.

Gene Regulatory Network, article in Wikipedia.
http://en.wikipedia.org/wiki/Gene_regulatory_network

KEGG: Kyoto Encyclopedia of Genes and Genomes. <http://www.genome.jp/kegg/>

KEGG Pathways for Homo Sapiens in KGML format.
<http://www.genome.jp/kegg/xml/hsa/index.html>

XGMML (eXtensible Graph Markup and Modeling Language).
<http://www.cs.rpi.edu/~puninj/XGMML/>

Cytoscape: Analyzing and Visualizing Network Data. <http://www.cytoscape.org/>

Kanterakis A, Potamias G. Decomposition of Gene Regulatory Networks into functional paths and their matching with microarray gene expression profiles. 6th European Symposium on Biomedical Engineering. June 2008, Chania, Greece