Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Σύστημα Αυτόματης Εξαγωγής Αλληλεπιδράσεων μεταξύ Μορίων microRNA και Γονιδίων από Επιστημονικές Δημοσιεύσεις στις Βιοεπιστήμες

Διπλωματική Εργασία

της

Ροδοθέας-Μυρσίνης Τσουπίδη

Επιβλέπων: Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων
Αθήνα, Ιούνιος 2014

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Σύστημα Αυτόματης Εξαγωγής Αλληλεπιδράσεων μεταξύ Μορίων microRNA και Γονιδίων από Επιστημονικές Δημοσιεύσεις στις Βιοεπιστήμες

Διπλωματική Εργασία

της

Ροδοθέας-Μυρσίνης Τσουπίδη

Επιβλέπων: Ιωάννης Βασιλείου
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 12η Ιουνίου 2014.

.....	.....	.....
Ιωάννης Βασιλείου	Κώστας Κοντογιάννης	Ιωάννης Σταύρακας
Καθηγητής Ε.Μ.Π.	Αν. Καθηγητής Ε.Μ.Π.	Ερευνητής Β' ΙΠΣΥ/Ε.Κ. "Αθηνά"

Αθήνα, Ιούνιος 2014

.....

Ροδοθέα-Μυρσίνη Τσουπίδη
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Στη γιαγιά μου, Μυρσίνη

Περίληψη

Η Εξαγωγή Πληροφορίας είναι η διαδικασία αυτόματης εξαγωγής δομημένης πληροφορίας από μη-δομημένα δεδομένα που είναι διαθέσιμα σε μορφή κατάλληλη για μηχανιστική επεξεργασία. Μία εφαρμογή της είναι η Εξαγωγή Συσχετίσεων μεταξύ διαφορετικών οντοτήτων από κείμενο φυσικής γλώσσας.

Στόχος της παρούσας εργασίας ήταν η κατασκευή ενός στατιστικού μοντέλου για την Εξαγωγή των Αλληλεπιδράσεων μεταξύ βιολογικών μορίων microRNA και γονιδίων, από επιστημονικές δημοσιεύσεις. Η μέθοδος περιλαμβάνει την αναγνώριση των Οντοτήτων που αναφέρονται σε microRNA και Γονίδια στο κείμενο, την εξαγωγή γλωσσικής πληροφορίας σχετικά με τους δύο όρους και την εκπαίδευση ενός στατιστικού μοντέλου χρησιμοποιώντας δεδομένα από επικυρωμένες αλληλεπιδράσεις microRNA-Γονιδίων. Το μοντέλο αυτό μπορεί στη συνέχεια να εφαρμοστεί σε νέα κείμενα δημοσιεύσεων και να εντοπίσει τις αλληλεπιδράσεις που αναφέρονται. Τα δεδομένα εκπαίδευσης προέρχονται από ειδικές βάσεις δεδομένων που περιλαμβάνουν αλληλεπιδράσεις microRNA και Γονιδίων, οι οποίες έχουν επιβεβαιωθεί πειραματικά. Τέτοιες βάσεις δεδομένων είναι το TarBase, το miRTarBase και το miRecords και τα δεδομένα που παρέχουν έχουν επιμεληθεί από ανθρώπους. Παράλληλα, στα πλαίσια της διπλωματικής αυτής υλοποιήθηκαν εργαλεία για την επίλυση υποπροβλημάτων της Επεξεργασίας Φυσικής Γλώσσας, τα οποία είναι απαραίτητα στην κατασκευή του μοντέλου.

Λέξεις Κλειδιά

microRNA, miRNA, γονίδιο, πρωτεΐνη, Εξαγωγή Συσχετίσεων, Κατηγοριοποίηση, Επεξεργασία Φυσικής Γλώσσας, Ανάκτηση Πληροφορίας

Abstract

Information Extraction is the task of automatically extracting structured information from unstructured machine-readable data. The idea can be applied to multiple tasks including Relation Extraction between Named Entities from Natural Language texts.

The aim of this thesis was the construction of a model for automatic extraction of microRNA-Gene interactions from biomedical publications. The methodology used includes identifying the Named Entities of microRNA and Gene, extracting important natural language information about the terms and training a statistical Binary Maximum Entropy Model on human curated microRNA-Gene interactions. The model can then be applied to new publications, identify undocumented interactions, and evaluate based on the training data. The curated data was provided by databases such as TarBase, miRTarBase and miRecords, which include miRNA-Gene interactions that were experimentally validated. Two models were trained, one for the identification of interactions described in full-text publications and the second for the identification of the interactions mentioned in the publication abstract and title. Additionally, software and tools associated with Natural Language Processing Subtasks were developed to aid the main classification task.

Keywords

microRNA, miRNA, genes, proteins, Relation Extraction, Classification, Natural Language Processing, Information Retrieval

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

Η παρούσα εργασία έχει σκοπό την αυτοματοποίηση της διαδικασίας εξαγωγής πληροφορίας που σχετίζεται με βιομόρια από επιστημονικές δημοσιεύσεις.

Συγκεκριμένα, ασχολείται με τα microRNA, βιολογικά μόρια που έχουν ανακαλυφθεί σχετικά πρόσφατα και τα οποία αλληλεπιδρούν με άλλα μεταγραφικά προϊόντα παρεμποδίζοντας την παραγωγή πρωτεϊνών. Οι αλληλεπιδράσεις αυτές, όταν ανιχνεύονται, καταγράφονται σε επιστημονικές δημοσιεύσεις τις οποίες μπορούν να αναζητήσουν βιοεπιστήμονες.

Η καταγραφή και η διαχείριση της πληροφορίας από τις δημοσιευμένες εργασίες μπορεί να γίνει χειρωνακτικά, με την επιμέλεια ανθρώπων που γνωρίζουν το αντικείμενο, όμως η διαδικασία αυτή είναι πολύ χρονοβόρα, ενώ με την μεγάλη αύξηση του αριθμού των δημοσιεύσεων που σχετίζονται με το θέμα γίνεται σχεδόν αδύνατη.

Έτσι, δημιουργείται η ανάγκη αυτοματοποίησης της διαδικασίας. Η αυτοματοποίηση χρησιμοποιεί τα ήδη υπάρχοντα καταγεγραμμένα αποτελέσματα για να εξάγει στοιχεία με βάση τα οποία θα αναγνωρίζει νέες αλληλεπιδράσεις από άγνωστες δημοσιεύσεις.

Στο κεφάλαιο αυτό γίνεται μία σύντομη εισαγωγή στους βιολογικούς όρους που χρησιμοποιούνται στο κείμενο. Ταυτόχρονα περιγράφεται το πρόβλημα της εξαγωγής αλληλεπιδράσεων μεταξύ μορίων microRNA και γονιδίων και συνοψίζονται οι βασικές συνεισφορές της παρούσας εργασίας.

1.1 Βιολογική Ορολογία

Η βιολογία είναι μία επιστήμη που συνεχώς εξελίσσεται. Στοιχεία που θεωρούνταν δεδομένα αμφισβητούνται και νέα ανακαλύπτονται βοηθώντας στην

κατανόηση των βιολογικών διεργασιών και της λειτουργίας των οργανισμών. Αυτά τα ευρήματα δίνουν νέες κατευθύνσεις στην έρευνα σε επιστήμες όπως η ιατρική, όπου δημιουργούνται νέες προοπτικές για την θεραπεία γενετικών και επίκτητων ασθενειών.

Οι βιολογικές διεργασίες είναι ιδιαίτερα πολύπλοκες, με πολλές επιμέρους διαδικασίες και ρυθμιστικά στάδια, που κάνουν τη μελέτη τους ιδιαίτερα δύσκολη. Σε κάθε διαδικασία συμμετέχουν πολλά μόρια και χημικές ουσίες και το αποτέλεσμα τους εξαρτάται από όλους τους παράγοντες αυτούς, όπως και από περιβαλλοντικά στοιχεία. Στα κύτταρα κάθε απορρύθμιση μπορεί να οφείλεται σε εγγενείς ατέλειες της γενετικής πληροφορίας, σε εξωτερικούς παράγοντες ή στο συνδυασμό τους.

1.1.1 Το DNA

Η γενετική πληροφορία που καθορίζει σε μεγάλο βαθμό τη λειτουργία των κυτταρικών μορφών ζωής και των περισσότερων ιών βρίσκεται αποθηκευμένη σε ένα μόριο, το δεοξυριβονουκλεϊκό οξύ ή DNA με τη μορφή ακολουθίας νουκλεοτιδίων, των δεοξυριβονουκλεοτιδίων. Τα δεοξυριβονουκλεοτίδια εμφανίζονται με τέσσερις τύπους, ανάλογα με την αζωτούχο βάση που περιέχουν, η οποία μπορεί να είναι η θυμίνη (T), η αδενίνη (A), η γουανίνη (G), ή η κυτοσίνη (C). Οι τέσσερις βάσεις συνδέονται με 3’-5’ φωσφοδιεστερικούς δεσμούς ανά δύο, η γουανίνη με την κυτοσίνη και η αδενίνη με τη θυμίνη. Αυτός ο τρόπος σύνδεσης που παρουσιάζουν τα νουκλεοτίδια του DNA λέγεται συμπληρωματικότητα.

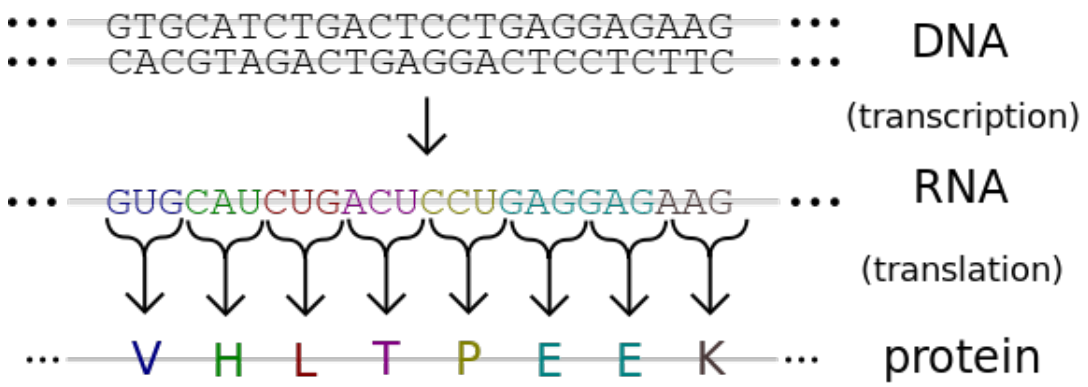
Το μόριο του DNA θεωρείται ως το κύριο μέσο μεταβίβασης χαρακτηριστικών από κάθε οργανισμό στους απογόνους του.

1.1.2 Το RNA

Το DNA είναι ένα μακρομόριο μεγάλου μήκους. Τμήματα του, τα γονίδια, αποτελούν περιοχές που κωδικοποιούν γενετική πληροφορία. Τα γονίδια μεταγράφονται και παράγουν ένα άλλο μακρομόριο, το ριβονουκλεϊκό οξύ ή RNA. Τα δομικά στοιχεία του είναι τα ριβονουκλεοτίδια που περιέχουν το σάκχαρο ριβόζη αντί της δεοξυριβόζης που περιέχει το DNA.

Το RNA, αναπαριστά την γενετική πληροφορία με τρόπο ανάλογο με το DNA, με τη διαφορά ότι στη θέση της θυμίνης υπάρχει η ουρακίλη (U) η οποία είναι συμπληρωματική με την αδενίνη. Παράλληλα, το RNA παίζει συχνά ρυθμιστικό ή ενζυμικό ρόλο στις διεργασίες του κυττάρου.

Η διαδικασία σχηματισμού του RNA από κάποιο γονίδιο του DNA ονομάζεται



Σχήμα 1.1: Διαδικασία Μετάφρασης της Γενετικής Πληροφορίας σε Πρωτεΐνη

μεταγραφή. Η μεταγραφή πραγματοποιείται μέσω της αντιστοίχισης των δεοξυριβονουκλεοτιδίων της μίας αλυσίδας του DNA με τα αντίστοιχα ριβονουκλεοτίδια. Η αντιστοίχιση αυτή γίνεται με βάση την συμπληρωματικότητα των αζωτούχων βάσεων (C->G, G->C, A->U, T->A) λέγεται **μεταγραφή**. Τα προϊόντα από τη διαδιακασία της μεταγραφής παίζουν διαφορετικό ρόλο στη λειτοργία του κυττάρου και μπορεί να μεταφέρουν τη γενετική πληροφορία, να λειτουργούν ως ένζυμα ή να έχουν κάποια άλλη λειτουργία. Πολλά από αυτά συμμετέχουν στην σύνθεση των πρωτεϊνών.

1.1.3 Οι Πρωτεΐνες

Οι πρωτεΐνες αποτελούνται από μία αλυσίδα δομικών μονάδων, τα αμινοξέα που συνδέονται μεταξύ τους με πεπτιδικούς δεσμούς. Οι πρωτεΐνες είναι πολύ σημαντικά βιομόρια και συμμετέχουν στις περισσότερες διεργασίες του κυττάρου. Λειτουργούν ως ένζυμα για την κατάλυση άλλων χημικών αντιδράσεων, συμμετέχουν στην διακυτταρική επικοινωνία, ή αποτελούν δομικά στοιχεία των κυττάρων.

Η διαδικασία σύνθεσης μιας πρωτεΐνης σε ένα κύτταρο, με βάση την γενετική πληροφορία που έχει μεταβιβαστεί από το DNA σε κάποιο RNA, λέγεται **μετάφραση**. Το RNA αυτό λέγεται mRNA και περιλαμβάνει την πληροφορία κάποιου γονιδίου σε μορφή ακολουθίας αζωτούχων βάσεων. Κάθε τριπλέτα βάσεων του mRNA, αντιστοιχεί σε ένα αμινοξύ και η αλληλουχία όλων των τριπλετών αντιστοιχεί στο μακρομόριο της πρωτεΐνης. Η αντιστοίχιση των αμινοξέων με τις τριπλέτες των βάσεων του mRNA γίνεται μέσω του tRNA, ενός άλλου είδους RNA, το οποίο περιέχει μία αντι-κωδική τριπλέτα, συμπληρωματική με την κωδική τριπλέτα του mRNA και την αντιστοιχίζει με ένα αμινοξύ. Όλη η διαδικασία γίνεται με τη βοήθεια των ριβοσωμάτων, σωματιδίων του κυττάρου τα οποία αποτελούνται κατά 60% από RNA (rRNA) και

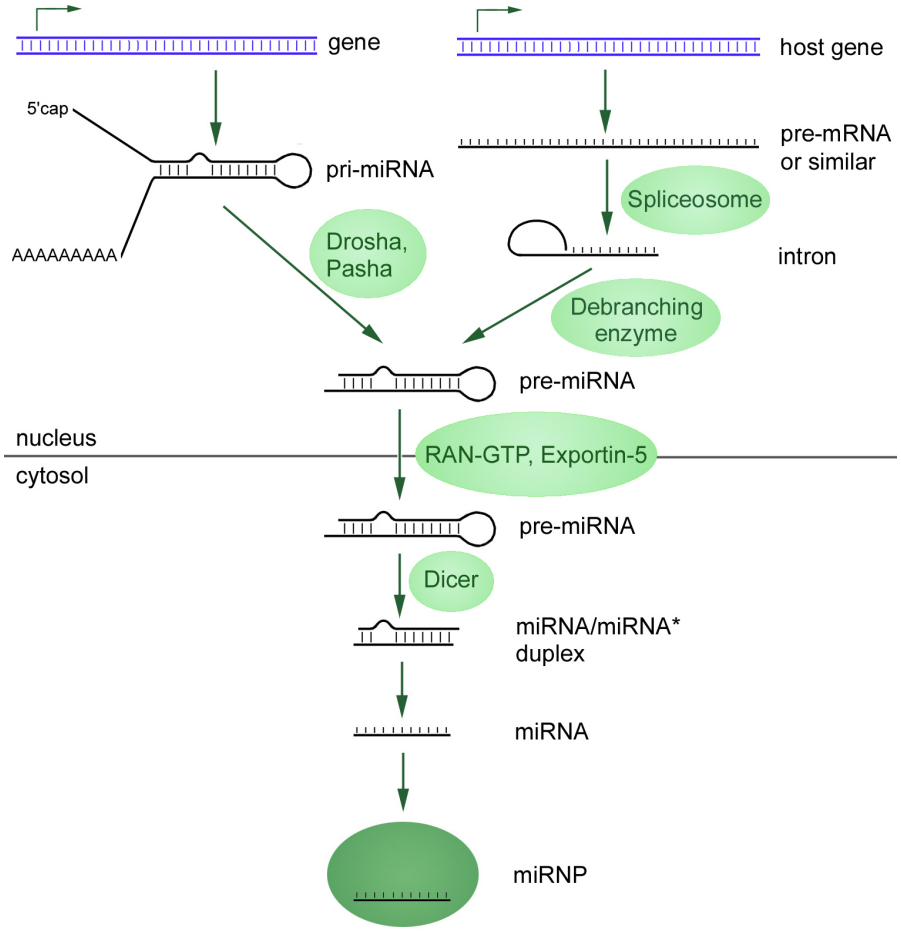
κατά 30% από ένζυμα.

Η διαδικασία της μετάφρασης της γενετικής πληροφορίας σε πρωτεΐνη φαίνεται στο Σχήμα 1.1.

1.1.4 Άλλα είδη RNA

Τα τελευταία χρόνια, έχει ανακαλυφθεί μεγάλος αριθμός νέων ειδών RNA που παίζουν διαφορετικό ρόλο στη λειτουργία των κυττάρων και των οργανισμών συνολικά. Πολλά από αυτά έχουν ρυθμιστικό ρόλο στις διάφορες λειτουργίες του κυττάρου. Ένα από τα πιο γνωστά τέτοια είδη είναι το microRNA.

1.1.4.1 microRNA



Σχήμα 1.2: Διαδικασία σχηματισμού microRNA

Τα microRNA είναι μόρια ριβονουκλεϊκού οξέως με μικρό μήκος που κυμαίνεται από 18 μέχρι 25 νουκλεοτίδια και παίζουν ρυθμιστικό ρόλο στην πρωτεϊνοσύνθεση.

Η ανακάλυψη του πρώτου microRNA, του lin-4, έγινε το 1993 σε έναν σχετικά απλό πολυκύτταρο ευκαρυωτικό οργανισμό, το Caenorhabtitis Elegans. Αρχικά, θεωρήθηκε ως μια ειδική μορφή ρύθμισης της έκφρασης των πρωτεϊνών που εμφανιζόταν μόνο στον συγκεκριμένο οργανισμό. Το 2000 έγινε η ανακάλυψη ενός δεύτερου μορίου microRNA στο Caenorhabtitis Elegans, του let-7, το οποίο όμως είχε διατηρηθεί εξελικτικά και στον άνθρωπο. Έτσι, μετά από μικρό χρονικό διάστημα, έγινε η αναγνώριση των microRNA, ως ένα νέο είδος RNA με ρυθμιστικό ρόλο στην μετάφραση των mRNA σε πρωτεΐνες και άρχισε να γίνεται μαζική αναζήτηση νέων μορίων microRNA και των mRNA-στόχων τους σε διαφορετικούς οργανισμούς.

Τα microRNA μεταγράφονται είτε από ανεξάρτητα γονίδια είτε από τμήματα των γονιδίων που τελικά μεταγράφονται σε mRNA και αποτελούν στόχους των αντίστοιχων microRNA (Σχήμα 1.2).

Η βασική λειτουργία του microRNA είναι η καταστολή της διαδικασίας της μετάφρασης του mRNA σε πρωτεΐνη, μέσω της πρόσδεσης του σε κάποιο σημείο του mRNA, με βάση τη συμπληρωματικότητα των αζωτούχων βάσεων. Το σημείο πρόσδεσης βρίσκεται συνήθως σε κάποιο ρυθμιστικό άκρο του mRNA και έχει ως αποτέλεσμα την καταστολή ή μείωση της έκφρασης του. Το mRNA στο οποίο συνδέεται κάποιο microRNA καταστέλοντας ή μειώνοντας την έκφραση του, λέγεται στόχος (target) του microRNA. Συχνά χρησιμοποιείται ο όρος στόχος και για το γονίδιο από το οποίο μεταγράφεται το mRNA.

Για να είναι λειτουργικό ένα microRNA αρκεί μόνο 7 με 8 βάσεις να είναι συμπληρωματικές με το mRNA-στόχο. Επομένως, κάθε microRNA έχει πολλά mRNA-στόχους και κάθε mRNA συνήθως έχει περισσότερα από ένα σημεία πρόσδεσης για διαφορετικά microRNA.

Για την καταχώρηση των νέων δεδομένων, σχετικά με τα microRNA, δημιουργήθηκαν ειδικές βάσεις δεδομένων, όπως η mirBase (`http://www.mirbase.org/`). Η mirBase καταγράφει τα microRNA που έχουν αναγνωριστεί και επιπλέον δίνει οδηγίες για την ονοματολογία των νέων microRNA, και των ομολόγων τους σε διαφορετικούς οργανισμούς, όπως και τον τρόπο αναφοράς τους (με ειδικές προσθήκες στην κατάληξη στο όνομα) στα διάφορα στάδια του μετασχηματισμού τους μέχρι να γίνουν λειτουργικά (σχήμα 1.2).

1.2 Το πρόβλημα του εντοπισμού των γονιδίων-στόχων των microRNA από τη Βιβλιογραφία

Την τελευταία δεκαετία, έχουν δημιουργηθεί βάσεις δεδομένων που καταχωρούν αλληλεπιδράσεις μεταξύ microRNA και γονιδίων-στόχων που έχουν επι-

βεβαιωθεί με κάποια πειραματική μέθοδο. Τέτοιες εφαρμογές είναι το TarBase (`http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index`), το miRTarBase (`http://mirtarbase.mbc.nctu.edu.tw/`) και το miRecords (`http://mirecords.umn.edu/miRecords/`). Τα δεδομένα αυτά προέρχονται από κείμενα δημοσιεύσεων ή από συμπληρωματικό υλικό που αναφέρει ή περιγράφει αποτελέσματα πειραμάτων.

Η διαδικασία της συγκέντρωσης των δεδομένων γίνεται συνήθως από ειδικούς που αναλαμβάνουν την ανάγνωση των δημοσιεύσεων, την επιλογή των στοιχείων που έχουν προκύψει ως αποτέλεσμα κάποιου πειράματος και την καταγραφή τους.

Καθώς με το πέρασμα του χρόνου, παρατηρείται σημαντική αύξηση του πλήθους των δημοσιεύσεων που σχετίζονται με τα microRNA, η καταγραφή αυτών των στοιχείων χειρωνακτικά γίνεται όλο και δυσκολότερη.

Συγκεκριμένα, στην βιβλιογραφική βάση δεδομένων, PubMed (Ενότητα 2.1.3.1), το πλήθος των δημοσιεύσεων που επιστρέφουν στο ερώτημα ”microRNA OR miRNA” ανά χρονιά από το 2000 μέχρι τον Μάιο του 2014 είναι:

Έτος	**# δημοσιεύσεων**
2000	0
2001	5
2002	43
2003	115
2004	229
2005	429
2006	732
2007	1114
2008	1817
2009	2659
2010	4051
2011	5353
2012	6971
2013	8478
2014*	3575

Πίνακας 1.1: Δημοσιεύσεις από την NCBI που επιστρέφουν στο ερώτημα ”mirna or microrna”, Για το 2014 οι δημοσιεύσεις είναι μέχρι 5/2014.

Στον πίνακα παρατηρείται, ειδικά στα πρώτα χρόνια μετά την ανακάλυψη της λειτουργίας των microRNA, εκθετική αύξηση των δημοσιεύσεων. Αυτό οφείλεται στο ενδιαφέρον που συγκεντρώνει το αντικείμενο, καθώς η γνώση για τη λειτουργία τους αυξάνεται και οι μεθοδολογίες για τον εντοπισμό των microRNA και των γονιδίων-στόχων τους βελτιώνονται. Η έρευνα για τα microRNA σχετίζεται με διάφορα πεδία, όπως η γενετική, η εξέλιξη των ειδών, η διαδικασία διαφοροποίησης των κυττάρων και η ιατρική. Ειδικά στην ιατρική θεωρείται ότι τα αποτελέσματα αυτά μπορούν να οδηγήσουν στην καλύτερη κατανόηση και πιθανώς στη θεραπεία πολλών ασθενειών.

Η εκτενής βιβλιογραφία καθιστά τη διαδικασία καταγραφής των microRNA και των στόχων τους δυσκολότερη, ενώ ταυτόχρονα γίνεται πιο επιτακτική ως βοηθητικό εργαλείο για την έρευνα.

Σε συγκεκριμένα πεδία, όπως για τις δημοσιεύσεις που αφορούν στον καρκίνο, έχουν ήδη αναπτυχθεί συστήματα αυτόματης εξαγωγής αλληλεπιδράσεων (miRCancer). [7]

Η συγκεκριμένη μέθοδος που ακολουθείται για την εξαγωγή των αλληλεπιδράσεων στη miRCancer, περιλαμβάνει τον ορισμό κανόνων με βάση τους οποίους γίνεται αναγνώριση των αλληλεπιδράσεων στο κείμενο. Αυτές οι μέθοδοι παρουσιάζουν συνήθως μεγάλη ακρίβεια στα αποτελέσματα, αλλά απαιτούν συστηματική ενασχόληση από ανθρώπους που γνωρίζουν το αντικείμενο. Επιπλέον, η ευαισθησία είναι συνήθως αρκετά μικρή. Τα αποτελέσματα που προκύπτουν σε άγνωστες δημοσιεύσεις αξιολογούνται από ανθρώπους και δημοσιεύονται στη σελίδα (`http://mircancer.ecu.edu/`).

1.3 Συνεισφορά

Η συνεισφορά της παρούσας εργασίας συνίσταται στην κατασκευή ενός μοντέλου για την αυτόματη αναγνώριση αλληλεπιδράσεων μεταξύ μορίων microRNA και των γονιδίων-στόχων τους από επιστημονικά κείμενα. Η εκπαίδευση του μοντέλου γίνεται στατιστικά, με βάση γνωστές αλληλεπιδράσεις που προέρχονται από ειδικές βάσεις δεδομένων, οι οποίες είναι επιμελημένες από ανθρώπους. Για την εξαγωγή της πληροφορίας του κειμένου, χρησιμοποιούνται μέθοδοι σχετικές με την Επεξεργασία Φυσικής Γλώσσας, ενός τομέα που ασχολείται με την αυτοματοποίηση και μηχανική παράσταση της γλωσσικής πληροφορίας.

Η συνεισφορά της εργασίας συνοψίζεται στα ακόλουθα σημεία:

- Μελέτη και μοντελοποίηση του προβλήματος αναγνώρισης αλληλεπιδράσεων γονιδίων και μορίων microRNA σε επιστημονικές δημοσιεύσεις.

- Ανάπτυξη Μοντέλου για την αυτόματη εξαγωγή των αλληλεπιδράσεων από κείμενα.
- Εφαρμογή του Μοντέλου σε βιβλιογραφικές αναφορές και πλήρη κείμενα δημοσιεύσεων.
- Υλοποίηση Βοηθητικών Εργαλείων για την Ανάλυση της Φυσικής Γλώσσας
- Διανομή του Κώδικα και των Μοντέλων με δυνατότητα επέκτασης και βελτίωσης.
- Δυνατότητα χρήσης των μοντέλων για εξαγωγή αλληλεπιδράσεων από μεγάλο πλήθος νέων δημοσιεύσεων.

Κεφάλαιο 2

Πηγές Δεδομένων

Η διαδικασία εξαγωγής αλληλεπιδράσεων μεταξύ των μορίων microRNA και των γονιδίων από κείμενα δημοσιεύσεων, προϋποθέτει την ανάκτηση συγκεκριμένων δεδομένων, όπως στοιχεία για τα γονίδια, τα κείμενα των δημοσιεύσεων και επαληθευμένες συσχετίσεις μεταξύ μορίων microRNA και γονιδίων. Τα στοιχεία αυτά είναι διαθέσιμα από ανοιχτές βάσεις δεδομένων που συγκεντρώνουν στοιχεία γύρω από συγκεκριμένα πεδία (πχ. το γονιδίωμα συγκεκριμένων οργανισμών).

Το κεφάλαιο αυτό περιλαμβάνει πληροφορίες για τις βάσεις δεδομένων από τις οποίες αντλήθηκαν τα απαραίτητα δεδομένα για την εργασία, δηλαδή τα δεδομένα για τις επιστημονικές δημοσιεύσεις, για τα γονίδια και τις αλληλεπιδράσεις με τα μόρια microRNA που έχουν καταγραφεί και επιβεβαιωθεί από ανθρώπους.

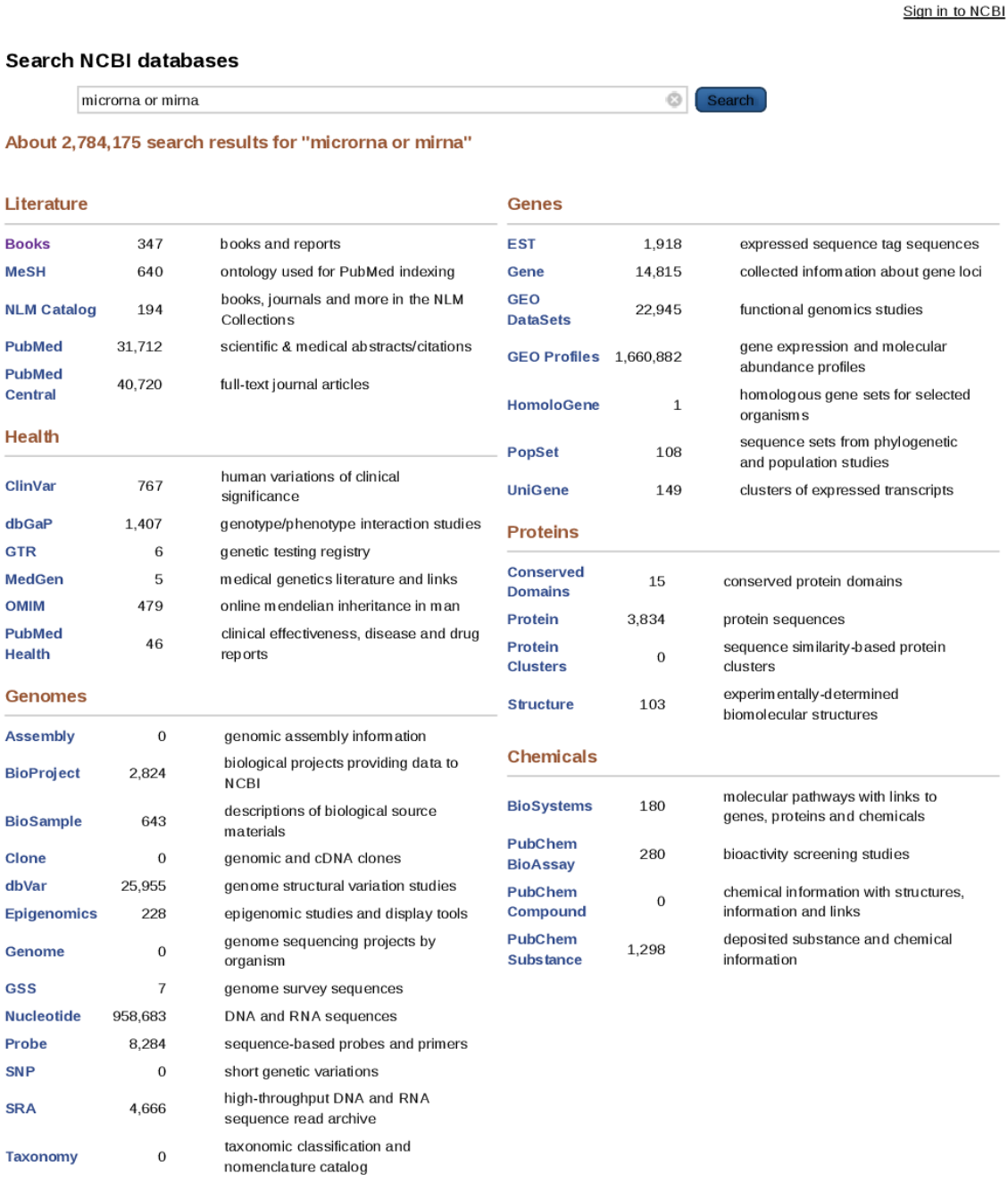
2.1 Η NCBI

Η National Center for Biotechnology Information (NCBI) είναι ένας οργανισμός που φιλοξενεί μεγάλο αριθμό βάσεων δεδομένων, που είναι σχετικές με τη βιοϊατρική και τη βιοτεχνολογία και την έρευνα που γίνεται στους τομείς αυτούς. Είναι τμήμα της National Library of Medicine (NLM) και βρίσκεται στη Maryland των ΗΠΑ. Η NCBI παρέχει πλήθος δεδομένων η πρόσβαση των οποίων μπορεί να γίνει μέσω της ιστοσελίδας `http://www.ncbi.nlm.nih.gov/`. Η ιστοσελίδα παρέχει τη μηχανή ENTREZ, η οποία δίνει τη δυνατότητα αναζήτησης στα δεδομένα όλων των βάσεων δεδομένων που φιλοξενεί. Επιπλέον, για μαζικές ανακτήσεις δεδομένων παρέχεται υπηρεσία FTP στη διεύθυνση `ftp.ncbi.nlm.nih.gov`. Ορισμένες από τις βάσεις δεδομένων που περιλαμβάνει είναι η H Taxonomy, η H PubMed, η H PubMed Central, η H Gene και η H RefSeq.

2.1.1 Η ENTREZ

Η ENTREZ είναι η μηχανή αναζήτησης της NCBI. Κάθε χρήστης μπορεί να έχει πρόσβαση στην ENTREZ μέσω της ιστοσελίδας της NCBI από όπου μπορεί να αναζητήσει στοιχεία από όλες τις επιμέρους βάσεις δεδομένων.

Στην παρακάτω εικόνα φαίνονται τα αποτελέσματα που επιστρέφει η ENTREZ στο ερώτημα ”microrna or mirna”:



Sign in to NCBI

Search NCBI databases

microrna or mirna — Search

About 2,784,175 search results for "microrna or mirna"

Literature

Books	347	books and reports
MeSH	640	ontology used for PubMed indexing
NLM Catalog	194	books, journals and more in the NLM Collections
PubMed	31,712	scientific & medical abstracts/citations
PubMed Central	40,720	full-text journal articles

Health

ClinVar	767	human variations of clinical significance
dbGaP	1,407	genotype/phenotype interaction studies
GTR	6	genetic testing registry
MedGen	5	medical genetics literature and links
OMIM	479	online mendelian inheritance in man
PubMed Health	46	clinical effectiveness, disease and drug reports

Genomes

Assembly	0	genomic assembly information
BioProject	2,824	biological projects providing data to NCBI
BioSample	643	descriptions of biological source materials
Clone	0	genomic and cDNA clones
dbVar	25,955	genome structural variation studies
Epigenomics	228	epigenomic studies and display tools
Genome	0	genome sequencing projects by organism
GSS	7	genome survey sequences
Nucleotide	958,683	DNA and RNA sequences
Probe	8,284	sequence-based probes and primers
SNP	0	short genetic variations
SRA	4,666	high-throughput DNA and RNA sequence read archive
Taxonomy	0	taxonomic classification and nomenclature catalog

Genes

EST	1,918	expressed sequence tag sequences
Gene	14,815	collected information about gene loci
GEO DataSets	22,945	functional genomics studies
GEO Profiles	1,660,882	gene expression and molecular abundance profiles
HomoloGene	1	homologous gene sets for selected organisms
PopSet	108	sequence sets from phylogenetic and population studies
UniGene	149	clusters of expressed transcripts

Proteins

Conserved Domains	15	conserved protein domains
Protein	3,834	protein sequences
Protein Clusters	0	sequence similarity-based protein clusters
Structure	103	experimentally-determined biomolecular structures

Chemicals

BioSystems	180	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	280	bioactivity screening studies
PubChem Compound	0	chemical information with structures, information and links
PubChem Substance	1,298	deposited substance and chemical information

Η ENTREZ δίνει τη η δυνατότητα εισαγωγής επιπλέον κριτηρίων και φίλτρων στην αναζήτηση, όπως η επιλογή αναζήτησης σε κάποια συγκεκριμένη βάση, η

αναζήτηση σε συγκεκριμένο πεδίο από κάποια βάση (πχ. στο πεδίο ArticleTitle στην PubMed (Ενότητα 2.1.3.1)) και η αναζήτηση σε υποσύνολα των βάσεων (πχ. στο OA Subset στην PubMed Central (Ενότητα 2.1.3.2)).

Τα περισσότερα από τα δεδομένα που παρέχονται μπορούν αν ανακτηθούν και σε μορφή XML.

2.1.1.1 Τα eUtils

Τα Entrez Programming Utilities (eUtils) είναι ένα σύνολο εργαλείων για πιο άμεση πρόσβαση στα αποτελέσματα των ερωτημάτων στην NCBI, ειδικά με προγραμματιστικό τρόπο. Τα ερωτήματα γίνονται άμεσα στον διακομιστή της NCBI, με την χρήση ειδικά μορφοποιημένων διευθύνσεων. Το αποτέλεσμα ενός ερωτήματος επιστρέφεται σε μορφή XML ή κειμένου.

Για παράδειγμα το ερώτημα (term = ”miRNA OR microRNA”) στην PubMed (db=pubmed) με μέγιστο αριθμό αποτελεσμάτων (retmax=100) και μορφή του αποτελέσματος σε XML (retmode = xml) με χρήση των eUtils είναι:

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?
db=pubmed
&term=mirna+OR+microrna
&retmode=xml
&retmax=100

Στο πεδίο term ορίζεται το ερώτημα, και μπορεί να περιλαμβάνει φίλτρα, όπως τα παρακάτω, τα οποία καθορίζουν τα πεδία στα οποία γίνεται η αναζήτηση.

mirna[All Fields] AND mirna[Abstract]

Το αποτέλεσμα του παραπάνω ερωτήματος είναι σε μορφή XML:

```
<?xml version="1.0" ?>
<!DOCTYPE eSearchResult PUBLIC "-//NLM//DTD
 esearch 20060628//EN"
"http://eutils.ncbi.nlm.nih.gov/eutils/
dtd/20060628/esearch.dtd">
<eSearchResult>
 <Count>31787</Count>
 <RetMax>10</RetMax>
 <RetStart>0</RetStart>
 <IdList>
  <Id>24822185</Id>
```

```
 <Id>24821854</Id>
 <Id>24821701</Id>
 <Id>24821460</Id>
 <Id>24821435</Id>
 <Id>24821285</Id>
 <Id>24820430</Id>
 <Id>24820117</Id>
 <Id>24820027</Id>
 <Id>24819892</Id>
</IdList>
<TranslationSet>
 <Translation>
   <From>mirna</From>
   <To>"micrornas"[MeSH Terms]
         OR "micrornas"[All Fields]
           OR "mirna"[All Fields]</To>
 </Translation>
 <Translation>
   <From>microrna</From>
   <To>"micrornas"[MeSH Terms]
           OR "micrornas"[All Fields]
             OR "microrna"[All Fields]</To>
 </Translation>
</TranslationSet>
<TranslationStack>
 <TermSet>
  <Term>"micrornas"[MeSH Terms]</Term>
  <Field>MeSH Terms</Field>
  <Count>22790</Count>
  <Explode>Y</Explode>
 </TermSet>
 <TermSet>
  <Term>"micrornas"[All Fields]</Term>
  <Field>All Fields</Field>
  <Count>27414</Count>
  <Explode>N</Explode>
 </TermSet>
 <OP>OR</OP>
 <TermSet>
  <Term>"mirna"[All Fields]</Term>
```

```
    <Field>All Fields</Field>
    <Count>15442</Count>
    <Explode>N</Explode>
   </TermSet>
   <OP>OR</OP>
   <OP>GROUP</OP>
   <TermSet>
    <Term>"micrornas"[MeSH Terms]</Term>
    <Field>MeSH Terms</Field>
    <Count>22790</Count>
    <Explode>Y</Explode>
   </TermSet>
   <TermSet>
    <Term>"micrornas"[All Fields]</Term>
    <Field>All Fields</Field>
    <Count>27414</Count>
    <Explode>N</Explode>
   </TermSet>
   <OP>OR</OP>
   <TermSet>
    <Term>"microrna"[All Fields]</Term>
    <Field>All Fields</Field>
    <Count>19752</Count>
    <Explode>N</Explode>
   </TermSet>
   <OP>OR</OP>
   <OP>GROUP</OP>
   <OP>OR</OP>
</TranslationStack>
<QueryTranslation>
   ("micrornas"[MeSH Terms] OR "micrornas"[All Fields]
   OR "mirna"[All Fields]) OR ("micrornas"[MeSH Terms]
   OR "micrornas"[All Fields] OR "microrna"[All Fields])
</QueryTranslation>
</eSearchResult>
```

Στα αποτελέσματα επιστρέφεται μία λίστα με αναγνωριστικά της PubMed, με άνω όριο την τιμή του retmax. Επιπλέον, στην ετικέτα ”QueryTranslation” του δέντρου φαίνεται η μορφή στην οποία μετασχηματίζεται το αρχικό ερώτημα.

Η ανάκτηση των δεδομένων της αναζήτησης μπορεί να γίνει με τη χρήση ενός

άλλου εργαλείου, του EFetch. Στο ερώτημα μπορεί να γίνει επιλογή της βάσης (db), της μορφής του αποτελέσματος (retmode) και των αναγνωριστικών των στοιχείων (id) της αντίστοιχης βάσης, τα οποία μπορεί να είναι ένα ή περισσότερα:

```
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?
db=pubmed
&id=24821460,2482143
&retmode=xml
```

Το αποτέλεσμα του ερωτήματος αυτού είναι ένα XML δέντρο που περιλαμβάνει στοιχεία για τη δημοσίευση, όπως το έτος και το περιοδικό που έγινε η δημοσίευση. Σε κάθε δημοσίευση περιλαμβάνεται ο τίτλος, ενώ στις περισσότερες περιπτώσεις περιλαμβάνεται και η περίληψη του άρθρου. Ένα παράδειγμα παρουσιάζεται στην ενότητα 2.1.3.1.

2.1.2 Η Taxonomy

Η NCBI Taxonomy, αποτελεί μια ταξινομία, δηλαδή κατάταξη των οργανισμών, από τις βάσεις δεδομένων που είναι διαθέσιμες και περιλαμβάνουν στοιχεία για το γενετικό υλικό των οργανισμών αυτών.

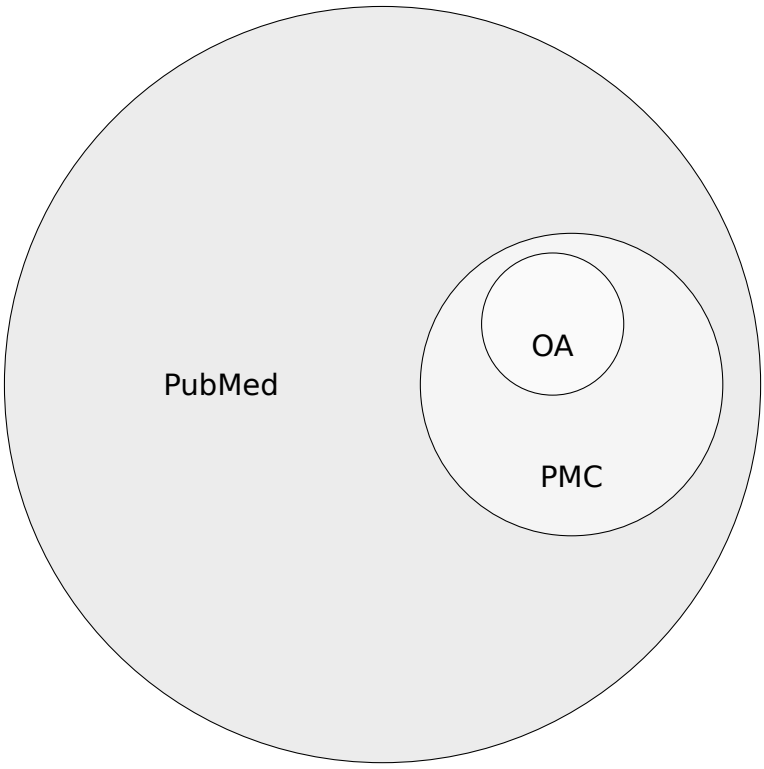
Σε κάθε οργανισμό αντιστοιχεί ένα διαφορετικό αναγνωριστικό, το taxid. Παρακάτω παρουσιάζονται ορισμένοι οργανισμοί με το taxid τους:

οργανισμός	taxid
homo sapiens	9606
mus musculus	10090
rattus norvegicus	10116
drosophila melanogaster	7227
gallus gallus	9031
danio rerio	7955
arabidopsis thaliana	3702
caenorhabtitis elegans	6239
k. s. herpesvirus	37296

Πίνακας 2.1: Οργανισμός - Taxid

Η ταξινομία αυτή χρησιμοποιείται και από άλλες βάσεις της NCBI όπως η Entrez Gene (βλέπε και Ενότητα 2.1.4).

2.1.3 Βάσεις Δεδομένων για Δημοσιεύσεις



Σχήμα 2.1: Διάγραμμα Venn για τις Βάσεις Δεδομένων της NCBI

Η NCBI περιλαμβάνει δύο βάσεις δεδομένων, την PubMed και την PMC, που σχετίζονται με τις δημοσιεύσεις σε περιοδικά από τα πεδία των βιοεπιστημών. Η PubMed περιέχει βιβλιογραφικές αναφορές και η PMC ολόκληρα τα κείμενα των δημοσιεύσεων. Ένα υποσύνολο της PMC, το OpenAccess Subset, περιλαμβάνει κείμενα που διανέμονται με ανοιχτές άδειες (Creative Commons ή παρόμοιες) και είναι διαθέσιμα για πιο ελεύθερη επεξεργασία. Τα κείμενα αυτά μπορούν να ανακτηθούν μαζικά για εφαρμογές εξόρυξης δεδομένων.

Γενικά, με λίγες ίσως εξαιρέσεις, η PMC είναι υποσύνολο της PubMed. Το OpenAccess είναι υποσύνολο της PMC. Στο διάγραμμα 2.1 φαίνεται η ποσοτική και ποιοτική σχέση μεταξύ τους.

2.1.3.1 Η PubMed

Η PubMed, είναι μια βιβλιογραφική βάση δεδομένων για τις βιοϊατρικές επιστήμες. Περιλαμβάνει περίπου 23 εκατομμύρια αναφορές σε κείμενα από τη MEDLINE (U.S. National Library of Medicine) και κανείς μπορεί να έχει πρόσβαση ηλεκτρονικά στα δεδομένα της βάσης μέσω της ιστοσελίδας `http://www.ncbi.nlm.nih.gov/pubmed/`.

Κάθε αναφορά χαρακτηρίζεται από ένα αναγνωριστικό, το PubMed ID (PMID), το οποίο είναι μοναδικό.

Οι βιβλιογραφικές αναφορές στην PubMed ακολουθούν συγκεκριμένη δομή. Κάποια πεδία συμπεριλαμβάνονται υποχρεωτικά, ενώ υπάρχει μεγάλος αριθμός προαιρετικών πεδίων, μερικά από τα οποία εμφανίζονται στις περισσότερες δημοσιεύσεις.

Ενα παράδειγμα μίας αναφοράς (citation) με τα υποχρεωτικά μόνο πεδία είναι η παρακάτω:

```
<MedlineCitation Owner="NLM" Status="MEDLINE">
  <PMID Version="1">10540283</PMID>
  <DateCreated>
    <Year>1999</Year>
    <Month>12</Month>
    <Day>17</Day>
  </DateCreated>
  <DateCompleted>
    <Year>1999</Year>
    <Month>12</Month>
    <Day>17</Day>
  </DateCompleted>
  <Article PubModel="Print">
    <Journal>
      <ISSN IssnType="Print">0950-382X</ISSN>
      <JournalIssue CitedMedium="Print">
        <PubDate>
          <Year>1999</Year>
        </PubDate>
      </JournalIssue>
      <Title>Molecular microbiology</Title>
    </Journal>
    <ArticleTitle>
      Transcription regulation of the nir gene
      cluster encoding nitrite reductase of Paracoccus
      denitrificans involves NNR and NirI, a novel type of
      membrane protein.
    </ArticleTitle>
    <Pagination>
      <MedlinePgn>24-36</MedlinePgn>
    </Pagination>
```

```
    <Language>eng</Language>
    <PublicationTypeList>
      <PublicationType>
        Journal Article
      </PublicationType>
      <PublicationType>
        Research Support, Non-U.S. Gov't
      </PublicationType>
    </PublicationTypeList>
  </Article>
</MedlineCitation>
```

Κάποια από τα προαιρετικά πεδία που περιλαμβάνονται στις περισσότερες βιβλιογραφικές αναφορές είναι η περίληψη (abstract) και κάποια MeSH terms που τη χαρακτηρίζουν.

Αναζήτηση βιβλιογραφικών αναφορών με βάση το pmid ή άλλα στοιχεία μπορεί να γίνει μέσω της ENTREZ και προγραμματιστικά με τα eUtils (Ενότητα 2.1.1). Επιπλέον, η σελίδα της κάθε αναφοράς περιέχει, σε περίπτωση που υπάρχουν, το σύνδεσμο προς το κείμενο της δημοσίευσης που παρέχει ο εκδότης και συνδέσμους σε άλλες πηγές που παρέχουν το κείμενο του άρθρου, όπως η PubMed Central.

2.1.3.2 H PubMed Central

H PubMed Central (PMC) είναι ένα ελεύθερο ψηφιακό αποθετήριο (repository) με κείμενα από άρθρα που έχουν δημοσιευτεί σε περιοδικά από τις βιοϊατρικές επιστήμες. H κάθε δημοσίευση χαρακτηρίζεται από ένα μοναδικό αναγνωριστικό, το PubMed Central ID (PMCID) και συνολικά περιλαμβάνει περίπου 3 εκατομμύρια κείμενα.

Τα κείμενα είναι διαθέσιμα σε μορφή HTML στη σελίδα της, `http://www.ncbi.nlm.nih.gov/pmc/` και ενσωματώνουν μεγάλη ποικιλία μεταδεδομένων, όπως λεξικά όρων, λέξεις κλειδιά και αναφορές σε βιοϊατρικές οντότητες. Επιπλέον συνήθως υπάρχει δυνατότητα ανάκτησης του PDF εγγράφου από τον εκδότη. Στη περίπτωση του OA Subset (υποσύνολου ελεύθερης πρόσβασης), τα κείμενα είναι διαθέσιμα και σε μορφή XML.

H ανάκτηση των κειμένων μπορεί να γίνει μέσω της μηχανής αναζήτησης, ENTREZ ή με τη χρήση των προγραμματιστικών εργαλείων eUtils (Ενότητα 2.1.1).

2.1.3.2.1 Το υποσύνολο Open Access

Το Open Access είναι υποσύνολο της PMC, στο οποίο ανήκουν τα άρθρα τα οποία είναι διαθέσιμα με άδειες Creative Commons, ή παρόμοιες, που δίνουν μεγαλύτερη ελευθερία στην αναδιανομή (redistribution), επαναχρησιμοποίηση (reuse) και επεξεργασία τους σε σχέση με τα άρθρα που διανέμονται με τις παραδοσιακές άδειες.

Περιλαμβάνει περίπου 800 χιλιάδες άρθρα, τα οποία μπορούν να ανακτηθούν μαζικά μέσω ftp, από τη διεύθυνση (`ftp://ftp.ncbi.nlm.nih.gov/pub/pmc`).

2.1.4 Η Gene

Η Gene είναι μια βάση δεδομένων της NCBI, που βρίσκεται στην ιστοσελίδα `http://www.ncbi.nlm.nih.gov/gene` και περιέχει δεδομένα και πληροφορίες για τα γονίδια αρκετών οργανισμών. Η ανάκτηση μεγάλης ποικιλίας δεδομένων σχετικά με τα γονίδια μπορεί να γίνει μέσω του (`ftp://ftp.ncbi.nih.gov/gene/`).

Συγκεκριμένα, τα αρχεία gene_info περιέχουν στοιχεία όπως το σύμβολο και τα ονόματα των γονιδίων όπως, και αναφορές σε άλλες βάσεις.Πιο αναλυτικά, μερικά από τα πεδία που περιλαμβάνονται είναι:

tax_id	Το αναγν. που παρέχεται από την NCBI Taxonomy (Ενότητα 2.1.2) για τα είδη (species) και είναι μοναδικό.
GeneID	Το μοναδικό αναγνωριστικό για κάθε γονίδιο.
Symbol	Το σύμβολο του γονιδίου.
LocusTag	LocusTag, χρησιμοποιείται ως εναλλακτικό όνομα για ένα γονίδιο, στην περίπτωση που δεν υπάρχει επίσημο σύμβολο για το γονίδιο ή ως αναφορά σε κάποια άλλη βάση (`http://www.ncbi.nlm.nih.gov/books/NBK3841/`).
Synonyms	Ανεπίσημα σύμβολα για το γονίδα. Χωρίζονται με |.
dbXrefs	Αναγνωριστικά από άλλες βάσεις. Χωρίζονται με |.

chromosome	Το χρωμόσωμα στο οποία βρίσκεται το γονίδιο, για τα μιτοχονδριακά γονίδια, η τιμή είναι 'MT'.
description	Περιγραφικό όνομα για το γονίδιο.
type of gene	Ο τύπος του γονιδίου (ανήκει σε μία λίστα: `http://www.ncbi.nlm.nih.gov/IEB/ToolBox/CPP_DOC/lxr/source/src/objects/entrezgene/entrezgene.asn`)

Πίνακας 2.2: Πεδία από τα αρχεία gene_info της Entrez Gene

2.1.5 H RefSeq

H Reference Sequence (RefSeq) είναι μια συλλογή από ακολουθίες γενετικού υλικού των γονιδίων, ακολουθίες RNA (προϊόντα μεταγραφής) και πρωτεΐνες.

Τα αναγνωριστικά των στοιχείων αποτελούνται από ένα πρόθεμα ακολουθούμενο από έναν αριθμό. Για παράδειγμα τα προθέματα NM_, NR_ και NP_ αντιπροσωπεύουν αναγνωριστικά mRNA, προϊόντα μεταγραφής και πρωτεΐνες, για τις γνωστές και επικυρωμένες ακολουθίες. Αντίστοιχα, τα προθέματα XM_, XR_ και XP_ προκύπτουν από μη επικυρωμένα στοιχεία όπως υπολογιστικές μέθοδοι.

Στον παρακάτω πίνακα παρουσιάζονται όλα τα προθέματα μαζί με σχόλια για την σημασία τους (`http://www.ncbi.nlm.nih.gov/books/NBK21091/table/ch18.T.refseq_accession_numbers_and_mole/?report=objectonly`):

Πρόθεμα	**Είδος Μορίου**	**Σχόλιο**
AC_	Γονίδιο	Ολοκληρωμένο γονιδιακό μόριο, συνήθως εναλλακτική ακολουθία
NC_	Γονίδιο	Ολοκληρωμένο γονιδιακό μόριο, συνήθως η επιλεγμένη ακολουθία
NG_	Γονίδιο	Μη ολοκληρωμένη γονιδιωματική περιοχή

Πρόθεμα	Είδος Μορίου	Σχόλιο
NT_	Γονίδιο	προκύπτει από τις διαδικασίες Contig ή scaffold, clone-based ή Whole Genome Shotgun (WGS)
NW_	Γονίδιο	Contig ή scaffold, κυρίως WGS
NS_	Γονίδιο	Μεταγονιδιακές ακολουθίες ή ακολουθίες που προέρχονται από αποικίες.
NZ_	Γονίδιο	Μη ολοκληρωμένη ακολουθία WGS
NM_	mRNA	
NR_	RNA	
XM_	mRNA	Μοντέλο που έχει προβλεφθεί με υπολογιστικές μεθόδους
XR_	RNA	Μοντέλο που έχει προβλεφθεί με υπολογιστικές μεθόδους
AP_	Πρωτεΐνη	Σχετίζεται με την εναλλακτική ακολουθία AC_
NP_	Πρωτεΐνη	Σχετίζεται με κάποιο αριθμό προσχώρησης NM_ ή NC_
YP_	Πρωτεΐνη	
XP_	Πρωτεΐνη	Μοντέλο πρόβλεψης, σχετίζεται με κάποια προσχώρηση XM_
ZP_	Πρωτεΐνη	Μοντέλο πρόβλεψης, επισημειωμένα στις γονιδιακές εγγραφές NZ_

2.2 Η Ensembl

Η Ensembl δημιουργήθηκε το 1999 με σκοπό την αυτοματοποίηση της διαδικασίας της υποσημείωσης των γονιδίων από δύο ομάδες, το European Bioinformatics Institute (EBI) και το European Molecular Biology Laboratory (EMBL). Μέχρι τότε δεν είχε ολοκληρωθεί το πρόγραμμα χαρτογράφησης του ανθρώπινου γονιδιώματος, αλλά ήταν ήδη γνωστό ότι το μέγεθος των δεδομένων ήταν πολύ μεγάλο (περίπου 3 δισεκατομμύρια ζεύγη βάσεων) και θα ήταν αδύνατη η διαχείριση του από ανθρώπους.

Επιπλέον, η Emsembl ενσωματώνει και άλλα βιολογικά δεδομένα που είναι προσβάσιμα από τη σελίδα της, που είναι διαθέσιμη στη ηλεκτρονική διεύθυνση http://www.ensembl.org/.

Τα αναγνωριστικά των γονιδίων που αντιστιχούν σε σταθερά αναγνωριστικά, ξεκινουν με το πρόθεμα ENSG για τα γονίδια, ENST για τα προϊόντα μεταγραφής και ENSP για τις πρωτεΐνες. Αυτά τα προθέματα αναφέρονται στα γονίδια του ανθρώπου, ενώ για άλλους οργανισμούς προστίθεται κάποιο μεσόθεμα, όπως ENSMUSG για το ποντίκι (Mus Musculus) ή ENSDARG για το Danio Rerio και ENSGALG για το Gallus Gallus.

Η Ensembl είναι ένα ανοιχτό project, και ο κώδικας της είναι διαθέσιμος στο https://github.com/Ensembl.

2.3 Το TarBase

Το TarBase 6.0 είναι μία βάση δεδομένων που είναι προσβάσιμη μέσω της διεύθυνσης: http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index.

Περιλαμβάνει πληροφορίες για αλληλεπιδράσεις μορίων microRNA με γονίδια-στόχους που έχουν επιβεβαιωθεί πειραματικά. Ένα τμήμα των δεδομένων προέρχεται από το TarBase 5.0, ενώ περιλαμβάνονται και νέες αλληλεπιδράσεις στο TarBase 6.0 που έχουν επιμεληθεί από ειδικούς ερευνητές.

Παράλληλα, συγκεντρώνει δεδομένα από άλλες εφαρμογές και βάσεις δεδομένων, όπως τα miR2Disease, miRecords και miRTarBase.

2.4 Αρχές Ονοματολογίας και Βάσεις Δεδομένων για Γονίδια

Οι υπευθυνες αρχές ονοματολογίας γονιδίων (ή Nomenclatures), αναλαμβάνουν τον καθορισμό οδηγιών και κανόνων για την ονοματολογία των γονιδίων, παρέχουν πληροφορίες και οδηγίες σχετικά με τα γονίδια και την ονοματολογία τους και εγκρίνουν νέα γονίδια.

Για τους πρότυπους οργανισμούς, υπάρχουν βάσεις δεδομένων που αναλαμβάνουν το ρόλο αυτό και επιπρόσθετα παρέχουν δεδομένα για τα γονίδια, τα προϊόντα τους και άλλες βιολογικές δομές τους.

2.4.1 Η HGNC

Η HUGO Gene Nomenclature Commitee (HGNC) είναι η υπεύθυνη επιτροπή για την ονοματολογία των γονιδίων του ανθρώπου (Homo Sapiens), για να ορίζει τις αρχές ονοματολογίας, να παρέχει πληροφορίες για την ονοματολογία των γονιδίων και να αποδίδει νέα και μοναδικά ονόματα και σύμβολα (συντομογραφίες) για τα γονίδια που ανακαλύπτονται. Επίσης, περιλαμβάνει και επιπλέον στοιχεία για τα γονίδια του ανθρώπου, όπως τα ομόλογα γονίδια από άλλους πρότυπους οργανισμούς (πχ. το ποντίκι Mus Musculus).

Η ιστοσελίδα της είναι η: `http://www.genenames.org/`.

Δεδομένα για τα ονόματα γονιδίων, τα συνώνυμα και τα αναγνωριστικά διαφορετικών βάσεων μπορεί να αντλήσει κανείς επιλέγοντας διαφορετικά πεδία από τη διεύθυνση: `http://www.genenames.org/cgi-bin/download`.

Όνομα Πεδίου	**Επεξήγηση**
HGNC ID	Αναγνωριστικό της HGNC
Approved Symbol	Εγκεκριμένο Σύμβολο
Approved Name	Εγκεκριμένο Όνομα
Status	Κατάσταση, λαμβάνει μία από τις τιμές: Approved, Entry Withdrawn, Status, Symbol Withdrawn
Locus Type	
Locus Group	
Previous Symbols	Προηγούμενα Σύμβολα
Previous Names	Προηγούμενα Όνομα
Synonyms	Συνώνυμα
Name Synonyms	Συνώνυμα Ονόματα
Chromosome	Χρωμόσωμα
Date Approved	Ημερομηνία Έγκρισης
Date Modified	Ημερομηνία Τροποποίησης
Date Symbol Changed	Ημερομηνία Τροποποίησης του Συμβόλου
Date Name Changed	Ημερομηνία Αλλαγής του Ονόματος
Accession Numbers	
Enzyme IDs	
Entrez Gene ID	Αναγνωριστικό της Entrez
Ensembl Gene ID	Αναγνωριστικό της Ensembl

Όνομα Πεδίου	Επεξήγηση
Mouse Genome Database ID	Αναγνωριστικό του ομόλογου γονιδίου στο MGI
Specialist Database Links	
Specialist Database IDs	
Pubmed IDs	
RefSeq IDs	Αναγνωριστικά του RefSeq
Gene Family Tag	
Gene family description	
Record Type	
Primary IDs	
Secondary IDs	
CCDS IDs	Αναγνωριστικά CCDS
VEGA IDs	Αναγνωριστικά VEGA
Locus Specific Databases	

2.4.2 Η MGI

Η Mouse Genome Informatics (MGI) είναι η υπεύθυνη αρχή για την ονοματοδοσία του Ποντικιού (Mus Musculus) και του αρουραίου (Rattus Norvegicus). Η ιστοσελίδα της είναι `http://www.informatics.jax.org/`.

Το αρχείο MRK_List2.rpt (report) περιέχει κάποια στοιχεία για τα γονίδια του Mus Musculus. Στο παρακάτω πίνακα εμφανίζονται τα πεδία του αρχείου:

Όνομα Πεδίου	Επεξήγηση	Παράδειγμα
MGI Accession ID	Αναγνωριστικό	MGI:1919525
Chr	Χρωμόσωμα	7
cM Position	Θέση	16.94
genome coordinate start	Αρχή του γονιδίου	29246561
genome coordinate end	Τέλος του γονιδίου	29248432
strand	αλυσίδα	
Marker Symbol	σύμβολο	2200002D01Rik
Status	Κατάσταση	0

Όνομα Πεδίου	Επεξήγηση	Παράδειγμα
Marker Name	Όνομα	RIKEN cDNA 2200002D01 gene
Marker Type	Τύπος	Gene
Feature Type	Χαρακτηριστικό	protein coding gene
Marker Synonyms (pipe-separated)	Συνώνυμα	HAI-2 related small protein|H2RSP

2.4.3 Η Flybase

Η Flybase δημιουργήθηκε από ομάδα ερευνητών για τη μύγα των φρούτων ή δροσόφιλα (Drosophila Melanogaster) και περιλαμβάνει δεδομένα για το γονιδίωμα και τα βιολογικά χαρακτηριστικά του οργανισμού και άλλων συγγενικών του. Είναι επίσης υπεύθυνη για την ονοματοδοσία των γονιδίων για τους οργανισμούς αυτούς.

Η σελίδα της βρίσκεται στη διεύθυνση `http://flybase.org/`.

2.4.4 Η WormBase

Η WormBase είναι μια κοινοπραξία μεταξύ επιστημόνων με σκοπό την παροχή πληροφοριών για το Caenorhabtitis Elegans και άλλους οργανισμούς από την οικογένεια των νηματωδών. Παρέχει επίσης οδηγίες για την ονοματοδοσία των γονιδίων για τους οργανισμούς αυτούς.

Η ιστοσελίδα της είναι η: `http://www.wormbase.org/`

2.4.5 Η TAIR

Η The Arabidopsis Information Resource (TAIR) είναι υπεύθυνη, για την ονοματοδοσία του Arabidopsis Thaliana, οργανισμό που αποτελεί πρότυπο για την έρευνα της βιολογίας των ανώτερων φυτών. Διατηρεί βάση δεδομένων με το πλήρες γονιδίωμα, τη δομή των γονιδίων, των πρωτεϊνών και των άλλων γονιδιακών προϊόντων και άλλες πληροφορίες.

Η ιστοσελίδα της βρίσκεται στο `http://www.arabidopsis.org/`.

2.4.6 Η ZFIN

Η Zebrafish Model Organism Database (ZFIN) είναι υπεύθυνη ονοματοδοσίας για το Danio Rerio ή Zebrafish, το οποίο έχει χρησιμοποιηθεί ως πρότυπος οργανισμός σε πολλές έρευνες.

Η ιστοσελίδα του βρίσκεται στο `http://zfin.org/`.

2.4.7 Η CGNC

Το Chicken Gene Nomenclature Consortium (CGNC) είναι μία διεθνής ομάδα επιστητημόνων, με σκοπό τον ορισμό των κανόνων ονοματολογίας για τον οργανισμό Gallus Gallus, με taxid 9031 (Ενότητα 2.1.2).

Η ιστοσελίδα του βρίσκεται στο `http://www.agnc.msstate.edu/`.

Από τη σελίδα `http://www.agnc.msstate.edu/Downloads.aspx` μπορεί να γίνει επιλογή πεδίων με χρήσιμα στοιχεία για τα γονίδια:

Όνομα Πεδίου	**Παράδειγμα**
CGNC ID	48941
Entrez ID	373885
Entrez Gene Version	20110818
Ensembl ID	ENSGALG00000002652
Ensembl Gene Version	
Gene Symbol	FZD10
Gene Name	frizzled family receptor 10
Gene Synonym	cFz-10|frizzled 10|seven transmembrane spanning receptor|frizzled homolog 10|frizzled-10|fz-10
Species	Gallus gallus
Curation Status	3
Last Edit Date	11/26/2011

2.4.8 Η RGD

Η Rat Genome Database (RGD) είναι μια βάση που έχει ως στόχο να συγκεντρώσει και να ενοποιήσει τα στοιχεία και τα ερευνητικά αποτελέσματα για τον αρουραίο, Rattus Norvegicus, taxid = 10116 (Ενότητα 2.1.2).

Η πρόσβαση στα δεδομένα της RGD μπορεί να γίνει μέσω της ιστοσελίδας `http://rgd.mcw.edu/` και μέσω FTP από τη διεύθυνση `ftp://rgd.mcw.edu/pub/`.

Στη διεύθυνση `ftp://rgd.mcw.edu/pub/data_release/` μπορεί να βρει κανείς στοιχεία για τα γονίδια του ανθρώπου (Homo Sapiens), του ποντικιού (Mus Musculus) και του αρουραίου (Rattus Norvegicus).

Το αρχείο *GENES_RAT_5.0.txt* περιέχει στοιχεία για τα ονόματα και συνώνυμα των γονιδίων του αρουραίου. Τα πεδία του είναι:

No	**Όνομα Πεδίου**	**Περιγραφή**
1	GENE_RGD_ID	Το αναγνωριστικό RGD_ID
2	SYMBOL	Επίσημο Σύμβολο για το γονίδιο
3	NAME	Όνομα γονιδίου
4	GENE_DESC	Περιγραφή Ονόματος (Προαιρετικό)
5	CHROMOSOME_CELERA	Χρωμόσωμα από το λογισμικό Celera Assembly
6	CHROMOSOME_3.4	Χρωμόσωμα από παλία αναφορά, έκδοση 3.4
7	CHROMOSOME_5.0	Χρωμόσωμα από την τρέχουσα έκδοση 5.0
8	FISH_BAND	fish band information
9	START_POS_CELERA	Θέση εκκίνησης για το Celera Assembly
10	STOP_POS_CELERA	Θέση τέλους για το Celera Assembly
11	STRAND_CELERA	Πληροφορίες για την αλυσίδα για το Celera Assembly
12	START_POS_3.4	θέση εκκίνησης για την έκδοση 3.4
13	STOP_POS_3.4	θέση τέλους για την έκδοση 3.4
14	STRAND_3.4	Πληροφορίες για την αλυσίδα για την έκδοση 3.4
15	START_POS_5.0	θέση εκκίνησης για την τρέχουσα έκδοση 5.0
16	STOP_POS_5.0	θέση τέλους για την τρέχουσα έκδοση 5.0
17	STRAND_5.0	Πληροφορίες για την αλυσίδα για την έκδοση 5.0

No	Όνομα Πεδίου	Περιγραφή
18	CURATED_REF_RGD_ID	RGD_ID των δημοσιεύσεων που χρησιμοποιήθηκαν για την επιμέλεια του γονιδίου
19	CURATED_REF_PUBMED_ID	PMID της δημοσίευσης που σχετίζεται με την επιμέλεια του γονιδίου
20	UNCURATED_PUBMED_ID	Αναγνωριστικά (PUBMED IDs) δημοσιεύσεων που σχετίζονται με το γονίδιο, αλλά δεν έχουν χρησιμοποιηθεί στην επιμέλεια του γονιδίου.
21	ENTREZ_GENE	Αναγνωριστικό της EntrezGene
22	UNIPROT_ID	Αναγνωριστικό(ά) της UniProtKB (http://www.uniprot.org/)
23	(UNUSED)	Κενό πεδίο
24	GENBANK_NUCLEOTIDE	Αναγνωριστικό της GenBank Nucleotide (http://www.ncbi.nlm.nih.gov/genbank/)
25	TIGR_ID	Αναγνωριστικό(ά) από τη TIGR (τώρα JCVI (http://www.jcvi.org/cms/home/))
26	GENBANK_PROTEIN	Αναγνωριστικό(ά) της GenBank Protein
27	UNIGENE_ID	Αναγνωριστικό της UniGene (http://www.ncbi.nlm.nih.gov/unigene)
28	SSLP_RGD_ID	Αναγνωριστικά SSLPs στην RGD που σχετίζονται με το συγκεκριμένο γονίδιο.
29	SSLP_SYMBOL	σύμβολο SSLP
30	OLD_SYMBOL	παλία σύμβολα
31	OLD_NAME	παλιά ονόματα

No	**Όνομα Πεδίου**	**Περιγραφή**
32	QTL_RGD_ID	Αναγνωριστικά του RGD των Quantitative trait locus (QTL) που σχετίζονται με το συγκεκριμένο γονίδιο
33	QTL_SYMBOL	σύμβολο QTL
34	NOMENCLATURE_STATUS	Κατάσταση Γονιδίου στην υπεύθυνη Αρχή Ονοματολογίας
35	SPLICE_RGD_ID	Αναγνωριστικά RGD_ID για τα σύμπλοκα του γονιδίου
36	SPLICE_SYMBOL	Σύμβολο του συμπλόκου
37	GENE_TYPE	τύπος γονιδίου
38	ENSEMBL_ID	Αναγνωριστικό Ensembl Gene
39	GENE_REFSEQ_STATUS	Κατάσταση του γονιδίου στη RefSeq
40	(UNUSED)	Κενό Πεδίο

Κεφάλαιο 3

Ορισμοί και Εργαλεία για την Επεξεργασία Φυσικής Γλώσσας

Η εξαγωγή των αλληλεπιδράσεων μεταξύ μορίων microRNA και γονιδίων-στόχων από κείμενο γραμμένο σε φυσική γλώσσα περιλαμβάνει την αξιολόγηση της γλωσσικής σχέσης μεταξύ των δύο όρων, ώστε να είναι δυνατή η αναγνώριση των σημαντικών στοιχείων της και η αποτίμηση της. Επομένως, για την διαδικασία αυτή, απαραίτητη είναι η αναγνώριση των γραμματικών, συντακτικών και σημασιολογικών δομών του κειμένου τα οποία αποτελούν υποπροβλήματα της Επεξεργασίας Φυσικής Γλώσσας.

Η Επεξεργασία Φυσικής Γλώσσας (ή Natural Language Processing (NLP)) είναι ένα πεδίο της επιστήμης των υπολογιστών, της τεχνιτής νοημοσύνης και της γλωσσολογίας και ασχολείται με την ανάλυση και κατανόηση της φυσικής γλώσσας για την χρήση από υπολογιστές.

Στο κεφάλαιο αυτό περιλαμβάνονται ορισμοί σημαντικών υποπροβλημάτων της Επεξεργασίας Φυσικής Γλώσσας, γίνεται αναφορά σε αλγορίθμους, εργαλεία και επισημειωμένα κείμενα που έχουν αποτελέσει σημείο αναφοράς στο πεδίο και περιγράφονται τα εργαλεία που χρησιμοποιήθηκαν και αυτά που αναπτύχθηκαν για την επίλυση του συγκεκριμένου προβλήματος, δηλαδή της αναγνώρισης των συσχετίσεων μεταξύ μορίων microRNA και γονιδίων.

3.1 Το μοντέλο Maximum Entropy Classifier

3.1.1 Θεωρητικό Μέρος

Το μοντέλο Maximum Entropy Classifier, είναι ένα πιθανοτικό στατιστικό μοντέλο για κατηγοριοποίηση στοιχείων σε κλάσεις και αποτελεί γενίκευση του Logistic Regression για προβλήματα με περισσότερες από δύο κλάσεις. Το ίδιο μοντέλο αναφέρεται με διαφορετικά ονόματα όπως Log-Linear, Gibbs, Maximum Entropy ή Logistic. Χρησιμοποιείται συχνά για την επίλυση προβλημάτων Επεξεργασίας Φυσικής Γλώσσας, όπου παρουσιάζει διαφορά από το αντίστοιχο στατιστικό μοντέλο, λόγω της διάστασης που έχει συνήθως το διάνυσμα των δεδομένων.

Η μοντελοποίηση του προβλήματος γίνεται ως εξής:

Έχουμε ένα πλήθος από n μεταβλητές:

$$d_1, d_2, .., d_n \in D$$

και m κλάσεις:

$$c_1, c_2, .., c_m \in C$$

Κάθε μία από τις μεταβλητές ανήκει σε μία κλάση.

Ορίζουμε l features:

$$f_1, f_2, .., f_l$$

τα οποία συνήθως στα NLP προβλήματα παίρνουν τις τιμές 0 ή 1.

Ένα παράδειγμα χαρακτηριστικού (feature) θα μπορούσε να είναι:

d==’the’ and c==’Άρθρο’

Η πιθανότητα κάποια μεταβλητή d να ανήκει στην κλάση c_i είναι:

$$P(c_i|d) = \frac{e^{(w \cdot f(c_i))}}{\sum_{c' \in C} e^{(w \cdot f(c'_i))}}$$

Το διάνυσμα w αντιστοιχεί στο βάρος κάθε χαρακτηριστικού (feature).

Ο εκθέτης επομένως είναι το άθροισμα από τα βάρη για τα χαρακτηριστικά που είναι παρόντα σε κάθε στοιχείο.

Προσαρμογή στο σύνολο εκπαίδευσης

Για την εύρεση του w, πρέπει να βρεθεί το διάνυσμα που δίνει τη βέλτιστη προσαρμογή στο σύνολο εκπαίδευσης, D.

Η λογαριθμική πιθανότητα κάθε κλάσης δεδομένου του στοιχείου d είναι:

$$log(P(C|D,w)) = log\left(\prod_{(c,d)\in(C,D)} P(c|d,w)\right) = \sum_{(c,d)\in(C,D)} log(P(c|d,w))$$

Και για να βελτιστοποιηθεί πρέπει:

$$\frac{\partial log(P(C|D,w))}{\partial w} = 0$$

Το οποίο μπορεί να προσεγγιστεί με διάφορες μεθόδους και έχει μοναδική λύση αφού η συνάρτηση πιθανότητας είναι κοίλη.

3.1.2 Δυαδικό Πρόβλημα

Το Δυαδικό Πρόβλημα, δηλαδή η περίπτωση που υπάρχουν μόνο δύο κλάσεις απλοποιείται θεωρώντας ως αναφορά μία από τις δύο κλάσεις.

Αν οι δύο κλάσεις είναι οι **True** και **False**, και η κλάση False θεωρηθεί αναφορά οι πιθανότητες κάθε κλάσης θα είναι:

$$P(True|d) = \frac{e^{(w\cdot f(True))}}{e^{(w\cdot f(True))}+1}$$

και

$$P(False|d) = \frac{1}{1+e^{(w\cdot f(True))}} = 1 - P(True|d)$$

3.1.3 Το λογισμικό MegaM

Ο MegaM είναι λογισμικό εύρεσης του βέλτιστου διανύσματος βάρους του μοντέλου Maximum Entropy για δυαδικά ή multiclass προβλήματα. Δηλαδή, προσαρμόζει το μοντέλο στο σύνολο εκπαίδευσης.

Ο κώδικας και πληροφορίες για το λογισμικό και τον αλγόριθμο υπάρχουν στην ιστοσελίδα του, `http://www.umiacs.umd.edu/~hal/megam/`.

3.2 Το διάνυσμα των χαρακτηριστικών

Το διάνυσμα των χαρακτηριστικών (features) όπως ορίζεται σε πολλά στατιστικά μοντέλα για την επίλυση προβλημάτων Επεξεργασίας Φυσικής Γλώσσας, είναι ένα διάνυσμα από χαρακτηριστικά, το καθένα από τα οποία παίρνει τις τιμές 0 ή 1. Σε κάθε στοιχείο ή δεδομένο του προβλήματος, αντιστοιχεί ένα τέτοιο διάνυσμα και κάθε συντεταγμένη του είναι η τιμή του αντίστοιχου χαρακτηριστικού για το συγκεκριμένο στοιχείο.

Η διάσταση του διανύσματος είναι ίση με τον αριθμό των χαρακτηριστικών που συνήθως είναι πολύ μεγάλος, επειδή εξαρτάται από το λεξιλόγιο του προβλήματος. Συχνά, τα χαρακτηριστικά δεν ορίζονται μεμονωμένα, αλλά σε ομάδες.

Πρακτικά τα χαρακτηριστικά είναι συνθήκες, που παίρνουν την τιμή 1, ή 0, ανάλογα με το αν είναι αληθείς, ή ψευδείς.

Για παράδειγμα έχουμε τις παρακάτω προτάσεις:

Πρόταση 1 Τα γονίδια είναι αλληλουχίες βάσεων του DNA.

Πρόταση 2 Κάποια γονίδια κωδικοποιούν πρωτεΐνες.

Πρόταση 3 Το DNA περιέχει γονίδια.

Πρόταση 4 Οι πρωτεΐνες είναι βιομόρια.

Αν θεωρήσουμε χαρακτηριστικά την ύπαρξη κάθε λέξης (unigram) από το λεξιλόγιο στην πρόταση (**Π(**λέξη**)** = Η πρόταση περιέχει τη λέξη), δηλαδή το διάνυσμα:

(**Π(**Τα**)**, **Π(**γονίδια**)**, **Π(**είναι**)**, **Π(**αλληλουχίες**)**, **Π(**βάσεων**)**, **Π(**του**)**, **Π(**DNA**)**, **Π(**Κάποια**)**, **Π(**κωδικοποιούν**)**, **Π(**πρωτεΐνες**)**, **Π(**Το**)**, **Π(**περιέχει**)**, **Π(**βιομόρια**)**, **Π(**Οι**)**, **Π(.)**)

Η καθεμία από τις προτάσεις θα έχει διάνυσμα χαρακτηριστικών:

Πρόταση 1 (1,1,1,1,1,1,1,0,0,0,0,0,0,0,1)

Πρόταση 2 (0,1,0,0,0,0,0,1,1,1,0,0,0,0,1)

Πρόταση 3 (0,1,0,0,0,0,1,0,0,0,1,1,0,0,1)

Πρόταση 4 (0,0,1,0,0,0,0,0,0,1,0,0,1,1,1)

Το μέγεθος του διανύσματος είναι συνήθως πολύ μεγάλο και περιέχει πολλούς μηδενικούς όρους. Επιπλέον εκτός από τις λέξεις μπορούν να υπάρχουν χαρακτηριστικά όπως η αλληλουχία δύο λέξεων (bigrams) και η αλληλουχία τριών λέξεων (trigrams) όπως και πιο συγκεκριμένα χαρακτηριστικά για κάθε στοιχείο, όπως η ίδια η πρόταση στη συγκεκριμένη περίπτωση. Αυτό αυξάνει περισσότερο την διάσταση του διανύσματος που γίνεται πιο αραιό.

3.3 Η βιβλιοθήκη NLTK

Η Natural Language Toolkit (NLTK) είναι μία βιβλιοθήκη της Python που ενσωματώνει πληθώρα από επισημειωμένα κείμενα, αλγόριθμους και εργαλεία χρήσιμα για την Επεξεργασία Φυσικής Γλώσσας.

Η σελίδα όπου περιλαμβάνονται οδηγίες, παραδείγματα και πληροφορίες είναι: `http://www.nltk.org/`.

3.4 Genia

Το GENIA είναι ένα πρόγραμμα του εργαστηρίου Tsuji του Πανεπιστημίου του Τόκιο και περιλαμβάνει επισημειωμένα κείμενα βιοϊατρικού περιεχομένου.

Το Genia Corpus περιλαμβάνει 1999 περιλήψεις (abstracts) από την PubMed (Ενότητα 2.1.3.1) που σχετίζονται με το ερώτημα ”human”, ”blood cells” και ”transcription factors”. Τα κείμενα είναι επισημειωμένα σε συντακτικό και σημασιολογικό επίπεδο.

Η σελίδα του Genia είναι: `http://www.nactem.ac.uk/genia/`

3.5 Το Penn TreeBank

Το Penn Treebank αποτελείται από ένα μεγάλο σύνολο προτάσεων της αγγλικής γλώσσας, που έχει επισημειωθεί (annotate) με γλωσσικές έννοιες οι οποίες παρέχουν σημασιολογική και συντακτική πληροφορία για τη πρόταση.

Για κάθε πρόταση έχει καταγραφεί η συντακτική και η σημασιολογική δομή, όπως και άλλα στοιχεία όπως το μέρος του λόγου στο οποίο αντιστοιχεί κάθε λέξη.

Είναι το πρώτο μεγάλο αποθετήριο γλωσσικών δέντρων (TreeBank) που δημιουργήθηκε και έχει παίξει σημαντικό ρόλο στην ανάπτυξη τόσο της υπολογιστικής γλωσσολογίας, όσο και άλλων πεδίων που σχετίζονται με τη γλώσσα.

Οι ετικέτες των Μερών του Λόγου (Part-of-Speech (POS) (Πίνακας 3.1)) που χρησιμοποιήθηκαν στο Penn TreeBank χρησιμοποιούνται σε πολλά μοντέλα και δεδομένα εκπαίδευσης και αποτελούν αναφορά στο πρόβλημα της αναγνώρισης του μέρους του λόγου.

Πίνακας 3.1: POS Tags

Σύμβολο	Επεξήγηση
CC	Σύνδεσμοι που συνδέουν κύριες προτάσεις (πχ. and, or)
CD	Απόλυτος αριθμός
DT	Άρθρο
EX	Υπαρξιακό ”there“
FW	Ξένη (μη αγγλική) λέξη
IN	Πρόθεση ή Σύνδεσμοι που συνδέουν δευτερεύουσες προτάσεις (πχ as, although)
JJ	Επίθετο
JJR	Επίθετο, συγκριτικός βαθμός
JJS	Επίθετο, υπερθετικός βαθμός
LS	Στοιχείο Λίστας
MD	Βοηθητικό Ρήμα
NN	Ουσιαστικό, ενικός αριθμός
NNS	Ουσιαστικό, πληθυντικός αριθμός
NNP	Κύριο όνομα, ενικός αριθμός
NNPS	Κύριο όνομα, πληθυντικός αριθμός
PDT	Λέξη πριν από ένα άρθρο (πχ. all στη φράση ”all this time“
POS	Κτητική κατάληξη
PRP	Προσωπική Αντωνυμία
PRP$	Κτητική Αντωνυμία
RB	Επίρρημα
RBR	Επίρρημα, συγκριτικός βαθμός
RBS	Επίρρημα, υπερθετικός βαθμός
RP	Άκλιτες λέξεις, συχνά χρησιμοποιούνται στη προφορική γλώσσα (πχ. um, etc)
SYM	Σύμβολο
TO	Η λέξη ”to“ (να)
UH	Επιφώνημα
VB	Ρήμα, βασική μορφή
VBD	Ρήμα, παρελθοντικός χρόνος
VBG	Ρήμα, γερούνδιο ή μετοχή
VBN	Ρήμα, μετοχή παρελθόντος
VBP	Ρήμα, ενεστώτας εκτός από το 3ο πρόσωπο
VBZ	Ρήμα, τρίτο πρόσωπο ενεστώτα

Σύμβολο	Επεξήγηση
WDT	Άρθρο που αρχίζει με ”Wh“
WP	Αντωνυμία που αρχίζει με ”Wh“
WP$	Κτητική αντωνυμία που αρχίζει με ”Wh“
WRB	Επίρρημα που αρχίζει με ”Wh“

3.6 Γραμματική Επισημείωση

Η Γραμματική Επισημείωση (ή Part-of-Speech tagging), είναι ένα σημαντικό υποπρόβλημα στην Επεξεργασία Φυσικής Γλώσσας. Η αυτόματη επίλυση του με χρήση στατιστικών μοντέλων είναι πρόβλημα κατηγοριοποίησης (classification) των λέξεων σε κλάσεις που αντιστοιχούν στα μέρη του λόγου.

Συνήθως εκτός από την ίδια τη λέξη χρησιμοποιούνται και άλλα χαρακτηριστικά (Ενότητα 3.2), όπως οι γειτονικές λέξεις, αλλά και τα μέρη του λόγου των προηγούμενων ή και επόμενων λέξεων για την βελτίωση στην απόδοση του. Η χρήση προηγούμενων και επόμενων ετικετών κάνει τη διαδικασία επιλογής της κατάλληλης ετικέτας πιο πολύπλοκη. Στην περίπτωση χρήσης μόνο της προηγούμενης ετικέτας ως χαρακτηριστικό, η επίλυση μπορεί να γίνει με κάποιο δυναμικό αλγόριθμο.

Για το πρόβλημα της γραμματικής επισημείωσης έχουν αναπτυχθεί διάφορα αυτόματα εργαλεία, όπως ο Stanford και ο Brill Tagger.

3.6.1 Stanford Tagger

Ο Stanford Tagger ανήκει στο λογισμικό του Stanford Natural Language Processing (NLP). Είναι ένας κατηγοριοποιητής που κατατάσσει τις λέξεις σε μέρη του λόγου με βάση τις ετικέτες από το PennTreeBank (Ενότητα 3.5).

Χρησιμοποιεί το Log Linear (Ενότητα 3.1) μοντέλο.

Η ιστοσελίδα με τον κώδικα, τα μοντέλα και περισσότερες πληροφορίες είναι η: `http://nlp.stanford.edu/software/tagger.shtml`

3.6.2 Brill Tagger

Ο Brill Tagger είναι μία μέθοδος για κατηγοριοποίηση λέξεων σε μέρη του λόγου (POS tagging). Αναπτύχθηκε από τον Eric Brill, το 1995.

Ο αλγόριθμος συνοπτικά μπορεί να περιγραφεί από δύο στάδια:

Βήμα 1 Θέτει σε κάθε λέξη το POS Tag με τη μεγαλύτερη συχνότητα για τη λέξη αυτή.

Βήμα 2 Διορθώνει τα λάθη με βάση κανόνες.

Χρησιμοποιεί τα tags από το Penn TreeBank (Πίνακας 3.1).

3.7 Απομόνωση Θέματος

Η απομόνωση του Θέματος (ή Stemming) είναι η διαδικασία απομόνωσης τμήματος μιας λέξης που μπορεί να είναι το θέμα, το οποίο ορίζεται με διαφορετικούς τρόπους, είτε ως η απομάκρυνση της κλιτικής κατάληξης είτε η απομάκρυνση των προσφυμάτων (κατάληξη που αλλάζει τη σημασία της λέξης όπως η κατάληξη -ότητα).

Η διαδικασία αυτή είναι πολύ χρήσιμη για την Επεξεργασία Φυσικής Γλώσσας, αφού έτσι απομονώνεται το σημασιολογικά σημαντικότερο τμήμα της λέξης και οι αλγόριθμοι μηχανικής μάθησης συχνά αποδίδουν καλύτερα. Συγκεκριμένα, επιτυγχάνεται βελτίωση της ευαισθησίας (recall) χωρίς αυτό να συνεπάγεται μείωση στην ακρίβεια (precision).

Για την επίλυση του προβλήματος αυτού, έχουν αναπτυχθεί αρκετά εργαλεία, όπως ο Morpha Stemmer και ο Porter Stemmer.

3.7.1 Morpha Stemmer

Ο Morpha Stemmer είναι ένα μορφολογικός αναλυτής για την αγγλική γλώσσα. Επιστρέφει το λήμμα (lemma) και τη κλιτική κατάληξη (inflection) της λέξης με είσοδο τη λέξη και το POS-tag (Πίνακας 3.1). Μπορεί να λειτουργήσει και χωρίς POS-tags, αλλά η απόδοση του είναι πολύ χαμηλότερη.

Ενα παράδειγμα εκτέλεσης παρουσιάζεται παρακάτω,

οι φράσεις:

... miRNA-10 regulate**s** GENE ...
... miRNA-10 and miRNA-12 regulate GENE ...
... GENE regulation by miRNA-10 ...

γίνονται:

... mirna-10 regulate gene ...
... mirna-10 and mirna-12 regulate gene ...
... gene regulation by mirna-10 ...

3.7.2 Porter Stemmer

Ο Porter Stemmer είναι ένας αλγόριθμος για απομόνωση θέματος (Ενότητα 3.7) της αγγλικής γλώσσας που δημοσιεύθηκε το 1980 από τον Martin Porter. Χρησιμοποιείται ευρέως και αποτελεί αναφορά στην έρευνα και μηχανική Ανάλυση και Επεξεργασία της Φυσικής Γλώσσας.

Ο αλγόριθμος επιστρέφει ως αποτέλεσμα την λέξη χωρίς τις καταλήξεις της και αποτελείται από 6 διαδοχικά στάδια:

βήμα 1 Αποκόπτονται πληθυντικοί και καταλήξεις -ed ή -ing.

βήμα 2 Το τελικό y μετατρέπεται σε i όταν υπάρχει άλλο φωνήεν στο stem

βήμα 3 αντιστοιχίζει διπλές καταλήξει σε μονές, για παράδειγμα το -ization (= -ize και -ation) αντιστοιχίζει σε -ize.

βήμα 4 εξετάζει τα -ic-, -full, -ness και άλλα με παρόμοιο τρόπο όπως στο βήμα 3

βήμα 5 αφαιρεί τα -ant, -ence κλπ.

βήμα 6 αφαιρεί το τελικό -e.

Στο παρακάτω παράδειγμα εκτέλεσης, οι φράσεις:

... miRNA-10 regul**ates** GENE ...
... miRNA-10 and miRNA-12 regul**ate** GENE ...
... GENE regul**ation** by miRNA-10 ...

γίνονται:

... miRNA-10 regul GENE ...
... miRNA-10 and miRNA-12 regul GENE ...
... GENE regul by miRNA-10 ...

3.8 Εξαγωγή Ορισμών για Συντομογραφίες

Η εξαγωγή συντομογραφιών από το κείμενο (Abbreviation Definition Extraction) είναι η αναγνώριση συντμήσεων που χρησιμοποιούνται στο κείμενο και ορίζονται σε κάποιο σημείο του κειμένου αυτού. Ο ορισμός αυτός γίνεται συνήθως στη πρώτη αναφορά του όρου.

Η διαδικασία αυτή είναι σημαντική λόγω της πολυσημίας που εμφανίζουν οι συντομογραφίες, ιδιαίτερα σε τεχνικά ή επιστημονικά κείμενα όπου για τους όρους συχνά χρησιμοποιούνται σύντομες μορφές. Συγκεκριμένα, στην περίπτωση των βιοϊατρικών κειμένων, όροι όπως οι ιστοί, οι ασθένειες και οι χη-

μικές ουσίες εμφανίζονται στο κείμενο με σύντομη μορφή και πολύ συχνά συγχέονται με άλλους όρους όπως γονίδια.

Για παράδειγμα, ο όρος *SDS* είναι το επίσημο σύμβολο για το γονίδιο με όνομα *serine dehydratase*, όμως ταυτόχρονα η ίδια λέξη μπορεί να αναφέρεται σε αρκετούς άλλους όρους (`http://en.wikipedia.org/wiki/SDS`):

Safety data sheet	a form with data regarding the properties of a particular substance
Sodium dodecyl sulfate	an anionic surfactant used in many cleaning and hygiene products
Shwachman–Diamond syndrome	a rare genetic disorder chiefly affecting the blood and pancreas

Συγκεκριμένα, στο άρθρο με PMCID=PMC2532718 (`http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2532718/`) ο όρος *SDS* είναι συνομογραφία του *sodium dodecyl sulfate* που είναι οργανική ένωση και συχνά χρησιμοποιείται στην τεχνική western blot.

Ο ορισμός του στο κείμενο είναι ο ακόλουθος:

> ...buffer with 0.5% sodium dodecyl sulfate (SDS) in the presence of...

Η μορφή των ορισμών που εμφανίζονται στα κείμενα είναι παρόμοια με το παραπάνω παράδειγμα και υπάρχουν πολλοί αλγόριθμοι για την αυτόματη αναγνώριση και εξαγωγή τους.

3.8.1 Abbreviation Definition Recognition Software

Το Abbreviation Definition Recognition Software είναι ένα εργαλείο που εξάγει συντομογραφίες και τους ορισμούς τους από βιοϊατρικά κείμενα.

Περισσότερες πληροφορίες και ο κώδικας υπάρχουν στην σελίδα `http://biotext.berkeley.edu/software.html`.

3.9 CONLL

Το Conference on Computational Natural Language Learning (CONLL) είναι συνέδριο που διοργανώνεται κάθε χρόνο με θέμα την υπολογιστική επεξεργασία φυσικής γλώσσας.

Πολύ συχνά οργανώνει shared tasks στα οποία οι συμμετέχοντες προσπαθούν να επιλύσουν κάποιο πρόβλημα από το πεδίο της επεξεργασίας φυσικής γλώσ-

σας με δεδομένο κάποιο σημειωμένο σύνολο εκπαίδευσης (corpus) που πολύ συχνά επισημειώνεται με αφορμή το συνέδριο.

3.9.1 CONLL-2000 Shared Task

Το 2000, οργανώθηκε το shared task "Chunking" (`http://www.cnts.ua.ac.be/conll2000/chunking/`), που αφορούσε την αναγνώριση φράσεων όπως οι ονοματικές, οι ρηματικές, οι προθετικές φράσεις και άλλες κατηγορίες.

Το corpus αποτελείται από προτάσεις από το Wall Street Journal (WSJ) corpus. Τα μέρη του λόγου ή POS tags (Πίνακας 3.1) είναι τα αποτελέσματα του Brill Tagger (Ενότητα 3.6.2). Δεν χρησιμοποιήθηκαν τα tags του WSJ corpus, για να είναι πιο ρεαλιστικά τα αποτελέσματα.

Γενικά, οι προτάσεις του corpus δεν ανήκουν στο πεδίο της βιοϊατρική ή της βιολογίας, αλλά περιέχονται παραδείγματα από διάφορους τομείς. Το παρακάτω παράδειγμα ανήκει στο πεδίο της βιολογίας:

Word	POS	Chunk
The	DT	B-NP
molecule	NN	I-NP
is	VBZ	B-VP
the	DT	B-NP
mouse	NN	I-NP
version	NN	I-NP
of	IN	B-PP
a	DT	B-NP
protein	NN	I-NP
called	VBD	B-VP
the	DT	B-NP
interleukin-4	NN	I-NP
receptor	NN	I-NP
,	,	O
which	WDT	B-NP
directs	VBZ	B-VP
the	DT	B-NP
growth	NN	I-NP
and	CC	I-NP
function	NN	I-NP
of	IN	B-PP
white	JJ	B-NP
blood	NN	I-NP

Word	POS	Chunk
cells	NNS	I-NP
.	.	O

3.10 Αναγνώριση Ονοματικών, Ρηματικών και Προθετικών Φράσεων

Η Αναγνώριση Φράσεων (Phrase Chunking) είναι η διαδικασία χωρισμού μιας πρότασης σε φράσεις. Οι πιο σημαντικές είναι οι ονοματικές, οι ρηματικές και οι προθετικές φράσεις.

3.10.1 Ο Maxent Phrase Chunker

Για τις ανάγκες της εργασίας, υλοποιήθηκε ένας κατηγοριοποιητής φράσεων (Phrase Chunker) που χρησιμοποιεί το στατιστικό μοντέλο Maximum Entropy (Ενότητα 3.1) για την κατηγοριοποίηση των λέξεων σε τμήματα ονοματικών, ρηματικών ή προθετικών φράσεων.

Η εκπαίδευση του έγινε στο corpus από το CONLL 2000 (Ενότητα 3.9.1) και χρησιμοποιήθηκε ο αλγόριθμος MegaM (Ενότητα 3.1.3) για την εύρεση του διανύσματος βάρους που ταιριάζει με βέλτιστο τρόπο στο σύνολο εκπαίδευσης.

Επιπλέον χρησιμοποιήθηκε το module maxent της βιβλιοθήκης NLTK (Ενότητα 3.3).

Από το CONLL 2000 χρησιμοποιήθηκαν μόνο οι ονοματικές, οι ρηματικές και οι προθετικές φράσεις, ενώ όλες τα υπόλοιπες κατατάσσονται στην κατηγορία 'O' (Other).

Στον παρακάτω πίνακα φαίνονται οι ετικέτες για τις κλάσεις που αντιστοιχούν στις ονοματικές, τις ρηματικές και τις προθετικές φράσεις:

Πίνακας 3.3: Phrase Chunk Tags

Symbol	Επεξήγηση
B-NP	Αρχή ονοματικής Φράσης
I-NP	Μέσο ονοματικής Φράσης
B-VP	Αρχή ρηματικής Φράσης
I-VP	Μέσο ρηματικής Φράσης
B-PP	Αρχή προθετικής Φράσης

Πίνακας 3.3: Phrase Chunk Tags

Symbol	Επεξήγηση
I-PP	Μέσο προθετικής Φράσης
O	Άλλο

Τα χαρακτηριστικά (Ενότητα 3.2) που χρησιμοποιήθηκαν για το μοντέλο είναι:

```
featureset = {
  "pos": pos,
  "word": word,
  "prevpos": prevpos,
  "nextpos": nextpos,
  "prevpos+pos": "%s+%s" % (prevpos, pos),
  "pos+nextpos": "%s+%s" % (pos, nextpos),
  "prevch": prevch,
  "prevch+pprevch": "%s+%s" % (prevch, pprevch),
  "pprevch": pprevch }
```

Η εκπαίδευση του μοντέλου έγινε στις προτάσεις εκπαίδευσης του CONLL 2000 και η αποτίμηση των αποτελεσμάτων στο σύνολο ελέγχου είναι:

scores/class	B-NP	B-PP	B-VP	I-NP	I-PP	I-VP	O
precision	0.96	0.96	0.95	0.96	0.71	0.94	0.93
recall	0.96	0.97	0.95	0.95	0.50	0.95	0.93
f1 score	0.96	0.97	0.95	0.96	0.59	0.94	0.93

Ο Phrase Chunker εφαρμόζεται στο κείμενο αφού πρώτα γίνει ο χωρισμός σε προτάσεις (sentence tokenization) και λέξεις (word tokenization) και έχει εφαρμοστεί ο POS Tagger. Επομένως, στην πραγματική περίπτωση οι ετικέτες POS δεν είναι ίδιες με τις ετικέτες από τον Brill Tagger (Ενότητα 3.6.2) με τον οποίο έχουν αναγνωριστεί οι λέξεις των προτάσεων στις οποίες εκπαιδεύτηκε το μοντέλο (CONLL 2000 corpus). Παρόλα αυτά τα αποτελέσματα για κάθε κλάση όταν αντικατασταθούν στο σύνολο ελέγχου τα POS tags του corpus από τα αποτελέσματα του Stanford POS Tagger (Ενότητα 3.6.1) είναι παρόμοια:

scores/class	B-NP	B-PP	B-VP	I-NP	I-PP	I-VP	O
precision	0.96	0.96	0.95	0.96	0.71	0.96	0.93
recall	0.96	0.97	0.95	0.95	0.50	0.95	0.94
f1 score	0.96	0.96	0.95	0.95	0.59	0.95	0.93

Η κατηγορία I-PP εμφανίζει πολύ χαμηλό recall. Μερικά παραδείγματα προθετικών φράσεων με περισσότερες από μία λέξεις είναι:

..	B-PP	I-PP	..	
..	rather	than	..	
..	because	of	..	
..	as	well	as	..

Ένας λόγος για την χαμηλή επίδοση του είναι η μικρή συχνότητα εμφάνισης του. Συγκεκριμένα στο training set ισχύει:

ετικέτα	# εμφανίσεων στο σύνολο εκπαίδευσης
I-NP	63307
B-NP	55081
O	38297
B-VP	21467
B-PP	21281
I-VP	12003
I-PP	291

3.11 Αναγνώριση Οντοτήτων

Αναγνώριση Οντοτήτων (Named-Entity Recognition (NER)), λέγεται το πρόβλημα εντοπισμού και κατηγοριοποίησης λέξεων σε ένα προκαθορισμένο σύνολο από κατηγορίες. Στο τομέα των βιοεπιστημών τέτοιες κατηγορίες είναι οι Ασθένειες, τα Γονίδια και οι Πρωτεΐνες, τα Φάρμακα, οι Χημικές Ουσίες και άλλα.

Αποτελεί συχνά υποπρόβλημα κάποιου προβλήματος εξαγωγής πληροφορίας.

3.11.1 Αναγνώριση Οντοτήτων Γονιδίων

Η αναγνώριση των αναφορών σε γονίδια μέσα σε κείμενο αποτελεί μία ιδιαίτερα δύσκολη εργασία. Σημαντικά προβλήματα είναι η μη κανονικότητα των ονομάτων, οι αμφισημίες στα ονόματα και τα σύμβολα των γονιδίων και η ύπαρξη πολλαπλών συνωνύμων.

Τα εργαλεία αυτόματης εξαγωγής γονιδίων που υπάρχουν δεν δίνουν ικανοποιητικά αποτελέσματα και η ανάπτυξη ενός τέτοιου εργαλείου είναι ιδιαίτερα δύσκολη και απαιτεί την εκπαίδευση του μοντέλου σε υποσημειωμένο (annotated) αντιπροσωπευτικό δείγμα από το συγκεκριμένο πεδίο.

Για τις ανάγκες της εργασίας, η αναζήτηση των όρων έγινε με χρήση λεξικών, τα οποία δημιουργήθηκαν με βάση δεδομένα που ανακτήθηκαν από συγκεκριμένες βάσεις, που συνήθως είναι υπεύθυνες και για την ονοματολογία (nomenclature) των γονιδίων. Επιπλέον χρησιμοποιήθηκαν στοιχεία από την ENTREZ Gene (Ενότητα 2.1.4), που περιέχει μεγάλο πλήθος πληροφοριών, όπως και συνώνυμα γονιδίων που εμφανίζονται στη βιβλιογραφία.

Τα στοιχεία που συλλέχθηκαν αναφέρονται σε οργανισμούς που εμφανίζονται στο Tarbase (Ενότητα 2.3).

Συγκεκριμένα:

οργανισμός	**Αναγν. taxid**	**ΒΔ ή Αρχή Ονοματολογίας**	**Προέλευση Δεδομένων**
Homo Sapiens	9606	HGNC	HGNC,ENTREZ
Mus Musculus	10090	MGI	MGI,ENTREZ
Rattus Norvegicus	10116	RGD	RGD,ENTREZ
Drosophila Melanogaster	7227	FLYBASE	FLYBASE,ENTREZ
Gallus Gallus	9031	CGNC	CGNC,ENTREZ
Danio Rerio	7955	ZFIN	ENTREZ
Arabidopsis Thaliana	3702	TAIR	ENTREZ
Caenorhabtitis Elegans	6239	WORMBASE	ENTREZ
k. s. herpesvirus	37296	-	HGNC, ENTREZ

Τα δεδομένα από τις παραπάνω βάσεις περιέχουν, ανάλογα με την περίπτωση, το όνομα, το σύμβολο (σύντομη μορφή του ονόματος), συνώνυμα σύμβολα και ονόματα, παλιά σύμβολα και ονόματα, το id ομόλογων γονιδίων από άλλες πηγές και άλλες πληροφορίες όπως το είδος του γονιδίου (protein coding, non coding RNA, pseudo, και λοιπά).

Συγκεκριμένα, τα πεδία που χρησιμοποιήθηκαν από κάθε βάση [8, 11–14] είναι τα ακόλουθα:

	Πεδία Δεδομένων	**αναγν. από άλλες ΒΔ**	**Τύπος γονιδίου**
HGNC			
	* Approved Symbol * Approved Name * Previous Symbols * Previous Names * Synonyms * Name Synonyms	* Entrez Gene ID * Ensembl Gene ID * MGI ID * RefSeq IDs	**Locus Group:** * gene with protein product * pseudogene
MGI			
	* MGI Accession ID * Marker Symbol * Marker Name * Marker Synonyms		**Feature Type:** * protein coding gene * pseudogene
RGD			
	* Gene RGD ID * Symbol * Name * Old Symbol * Old Name	* Entrez Gene * Ensembl ID	**Gene Type:** * protein-coding * pseudo * gene
CGNC			
	* CGNC ID * Gene Symbol * Gene Name * Gene Synonym	* Entrez ID * Ensembl ID	
ENTREZ			
	* GeneID * Symbol * LocusTag * Synonyms * description	* dbXrefs	**Type of Gene:** * protein-coding * pseudo

Η επεξεργασία τους περιλαμβάνει αφαίρεση χαρακτήρων όπως η παύλα ”-” και αντικατάσταση ελληνικών χαρακτήρων και λέξεων τους αντίστοιχους λατινικούς.

Για παράδειγμα, το γονίδιο BCL-2 (B-cell lymphoma 2) μπορεί να εμφανιστεί στο κείμενο ως ”BCL 2” και ως ”BCL2”. Ενώ η πρωτεΐνη TGF-beta (Transforming growth factor beta), εμφανίζεται και ως ”TGF-b” και ”TGF-β”.

3.11.2 Αναγνώριση Οντοτήτων microRNA

Τα μόρια miRNA έχουν κανονικοποιημένη ονομασία, όμως εμφανίζουν κάποιου είδους πολυμορφία στα ονόματα. Το πρόθεμα ή μεσόθεμα τους εκτός από μερικές περιπτώσεις (let-7, lsy-6, lin-4) είναι η λέξη microRNA, η οποία όμως μπορεί να γραφτεί με διάφορους τρόπους: mirna, miR, miRNA, microrna, mirn και mir.

Έτσι, αρχικά πρέπει να γίνει μετονομασία των miRNA, ώστε να έχουν μια πιο ενιαία μορφή. Ο κώδικας σε python είναι:

```
def mirna2mir(tok):
    return tok.lower().replace('mirna','mir'
                      ).replace('microrna','mir'
                        ).replace('mirn','mir')
```

Στη βιβλιογραφία τα miRNA εμφανίζονται εκτός από την πλήρη μορφή τους και σε συμπτυγμένη μορφή όπως φαίνεται στα παρακάτω παραδείγματα:

mir-100, -200, -3a, and -88-3p

microRNA10/11/13b

miRNAs -33, -2b and -3a-5p

miR 10 or 11

let-7a, 7b and 7c

Αν για κάθε πλήρες όνομα θέσουμε το tag <MIRNA> και για κάθε σύντομη μορφή το tag <EXT>, οι παραπάνω αναφορές μπορεί να τυποποιηθούν ως εξής:

<MIRNA>, <EXT>, <EXT>, and <EXT>

<MIRNA>/<EXT>/<EXT>

miRNAs <EXT>, <EXT> and <EXT>

miR <EXT> or <EXT>

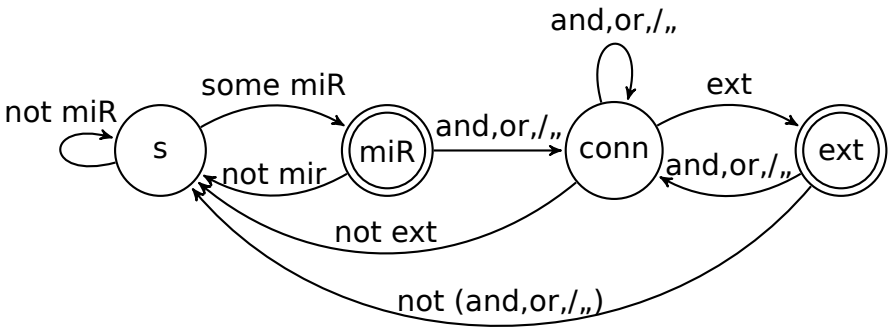
<MIRNA>, <EXT> and <EXT>

Επιπλέον οι λέξεις miRNA και miRNAs πιθανώς να ακολουθούνται από κάποιο αναγνωριστικό και να αναφέρονται σε κάποιο miRNA.

Έτσι, για τον εντοπισμό των σύντομων αναφορών, πρέπει μετά τον εντοπισμό κάποιου miRNA, να γίνει μετάβαση σε μια κατάσταση που θα επιτρέπει την αναγνώριση των miRNA που ακολουθούν.

Οι λέξεις που συνδέουν δύο mirna, είναι συνδετικές λέξεις ή σημεία στίξης, όπως (”,”,”/”,”and”,”or”).

Η γραμματική που ορίστηκε για τον εντοπισμό των σύντομων αναφορών περιγράφεται από το ακόλουθο αυτόματο:



Σχήμα 3.1: miRNA ER grammar

Τα tokens του tagger που ανήκει στη κλάση nltk.tag.RegexpTagger:

```
mirs = r'mirs$'
mir = r'mir$'
conn = r'(and|or|\\|/|,)$'
num = r'(3|5)$'
symbol = r'(\xc3\xa2|`|\xe2\x80\xb2|\')$'
mirna = r'.*((lsy-?6($|[\^0-9]))|(lin-?4($|[^0-9]))|'
         r'(bantam)|(let-?7($|[\^0-9]))'
         r'|(mir-?(((tar|lat)($|[^a-z]))|m01))|
         r'(mir-?[a-df-z]*-?[1-9][0-9a-z\-]*))'
ext = r'(((-[1-9][0-9a-z]{0,5})|([1-9][0-9]{0,5}[a-z]?))'
         r'(-(([1-9][0-9a-z]{0,3})|([a-z])))?\*?$)'

tokens = [
  (mirna, 'MIRNA'),
  (mirs, 'MIRs'),
  (mir, 'MIR'),
  (conn, 'CON'),
  (num, 'NUM'),
  (ext, 'EXT'),
```

```
  (other, 'LAST'),
  (symbol, 'SYM'),
  (r'.*', 'OTHER') ]
```

Ενώ η γραμματική για τον parser της κλάσης nltk.chunk.RegexpParser είναι:

```
grammar = '''
  UTR: {<NUM><SYM>}
  NMR: {(<CON>)+(<EXT>|<NUM>)<LAST>?}
  MR: {<MIRNA> (<NMR>)*}
  MR: {<MIRs>(<EXT>|<NUM>)<LAST>?(<NMR>)*}
  MR: {<MIR>(<EXT>|<NUM>)<LAST>?(<NMR>)*}
  O: {<OTHER>|<CON>|<LAST>|<MIRs>|<MIR>|<EXT>|<NUM>|<SYM>}
'''
```

Με την παραπάνω διαδικασία, υπάρχει περίπτωση να εντοπιστεί κάποιο όρος που δεν είναι miRNA. Για παράδειγμα, το ”let-7a, -78” θα παράξει το let-78, το οποίο δεν είναι miRNA. Για να αποφευχθούν κάποια από αυτά τα λάθη, τα αποτελέσματα των καινούριων όρων ελέγχονται από το regular expression που χρησιμοποιήθηκε για την αναγνώριση των πλήρων ονομάτων και αν ταιριάζουν τότε αναγνωρίζονται ως miRNA, ενώ σε διαφορετική περίπτωση όχι.

3.11.3 MyGene.info

Το MyGene.info είναι μία web εφαρμογή που παρέχει υπηρεσίες για την ανάκτηση στοιχείων σχετικά με τα γονίδια.

Για να λάβουμε πληροφορίες για το όνομα, το σύμβολο και συνώνυμα σύμβολα του γονιδίου με Ensembl Gene Id, ENSG00000167470 (Ενότητα 2.2) μπορούμε να κάνουμε το παρακάτω ερώτημα:

http://mygene.info/v2/query?q=ensemblgene:ENSG00000123374
&fields=symbol,name,alias

Στη ιστοσελίδα `http://mygene.info/` υπάρχουν οδηγίες για την λειτουργία και τη χρήση του.

3.12 Πιθανοτική Γραμματική Χωρίς Συμφραζόμενα

Η Πιθανοτική Γραμματική Χωρίς Συμφραζόμενα (ή Probabilistic Context Free Grammar (PCFG)) επεκτείνει μία Γραμματική Χωρίς Συμφραζόμενα, αναθέτο-

ντας μία πιθανότητα σε κάθε παραγωγή. Η συνολική πιθανότητα του δέντρου είναι το γινόμενο των πιθανοτήτων των παραγωγών από τις οποίες προκύπτει.

Στην Επεξεργασία Φυσικής Γλώσσας, χρησιμοποιείται συχνά για την συντακτική ανάλυση προτάσεων. Οι πιθανότητες προκύπτουν από την εκπαίδευση του μοντέλου σε μεγάλο πλήθος επισημειωμένων προτάσεων.

Για περισσότερες πληροφορίες, βλ. Collins [15].

3.13 Εξαγωγή Εξαρτήσεων

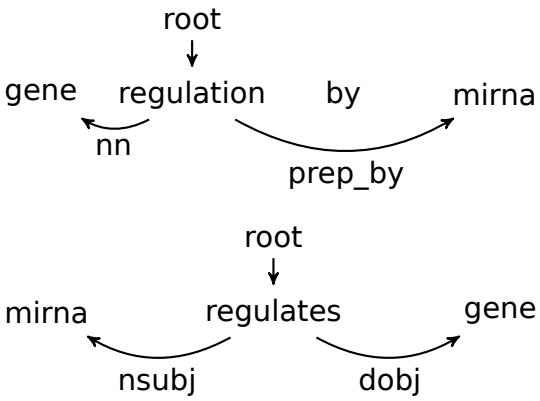
Η Εξαγωγή Εξαρτήσεων (Dependency Parsing) είναι το πρόβλημα εξαγωγής γραμματικών σχέσεων μεταξύ λέξεων σε μία πρόταση. Μερικά παραδείγματα εξαρτήσεων είναι η σχέση υποκειμένου με το ρήμα, η σχέση του άμεσου και του έμμεσου αντικειμένου με το ρήμα και η ονοματική σχέση. Για την αυτόματη εξαγωγή εξαρτήσεων, οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούν treebanks, όπως το Penn Treebank (Ενότητα 3.5).

Δύο απλά παραδείγματα εξαρτήσεων είναι:

gene regulation by miRNA.

miRNA regulates gene.

Και οι εξαρτήσεις είναι της μορφής:



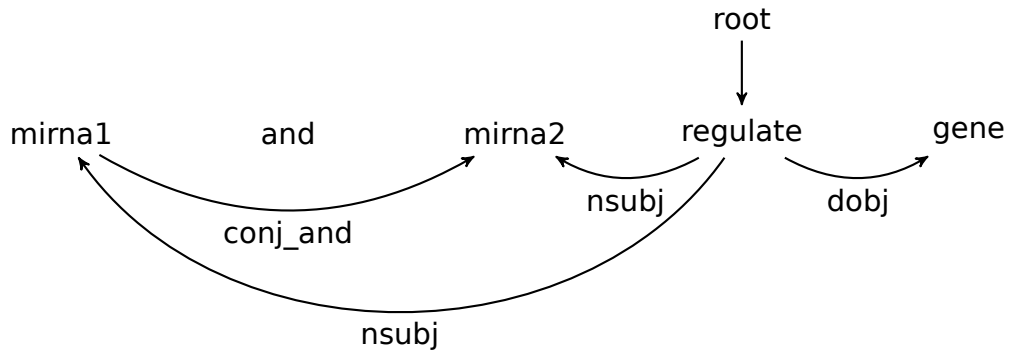
3.13.1 Stanford Dependency Parser

Ο Stanford Dependency Parser ανήκει στα εργαλεία του Stanford NLP και χρησιμοποιεί τροποποιημένες εξαρτήσεις, που στη γενική περίπτωση δεν σχηματίζουν δέντρο, αλλά γράφο. Το αποτέλεσμα για κάθε πρόταση είναι μια τριπλέτα με τις δύο λέξεις και τη σχέση που τις συνδέει.

Ένα απλό παράδειγμα είναι:

mirna1 and mirna2 regulate gene.

Οι εξαρτήσεις με το μοντέλο PCFG (Ενότητα 3.12), με χρήση των POS tags από τον Stanford POS Tagger είναι:



Όπως φαίνεται οι εξαρτήσεις δεν σχηματίζουν δεντρο αλλά γράφο.

Τα μοντέλα PCFG και Factored έχουν εκπαιδεύτει πάνω στο WSJ και στο Genia (Ενότητα 3.4). Η ιστοσελίδα του, όπου υπάρχει ο κώδικας, τα μοντέλα και πληροφορίες είναι: `http://nlp.stanford.edu/software/lex-parser.shtml`

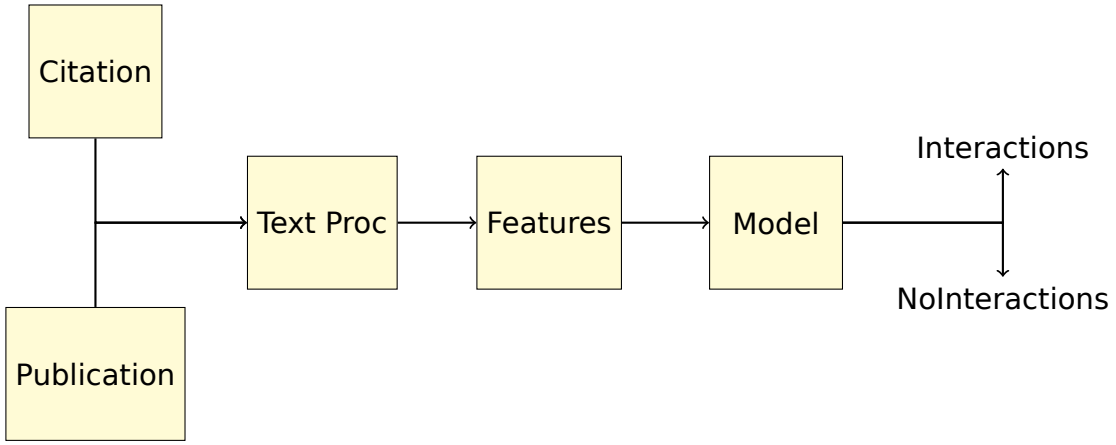
3.13.1.1 Αναπαράσταση Γράφου

Ο γράφος των εξαρτήσεων είναι κατευθυνόμενος, αλλά για τη εύρεση της βέλτιστης διαδρομής μεταξύ των δύο όρων miRNA και γονίδιο, μπορεί να αναπαρασταθεί με απλό γράφο.

Για την εξαγωγή της βέλτιστης διαδρομής μεταξύ των δύο όρων, εντοπίζονται τα συντομότερα μονοπάτια με χρήση της συνάρτησης all_shortest_paths της βιβλιοθήκης NetworkX (`https://networkx.github.io/`) της Python.

Κεφάλαιο 4

Αναγνώριση συσχετίσεων μεταξύ miRNA και γονιδίων



Σχήμα 4.1: Classifier

Σκοπός της διπλωματικής είναι η σχεδίαση ενός συστήματος για την αυτόματη εξαγωγή αλληλεπιδράσεων μεταξύ μορίων microRNA και γονιδίων που έχουν καταγραφεί σε κείμενα επιστημονικών δημοσιεύσεων. Το είδος της σχέσης είναι η ρυθμιστική ικανότητα του microRNA σε κάποιο στάδιο της έκφρασης του γονιδίου.

Τα κείμενα των δημοσιεύσεων προέρχονται από την NCBI (Ενότητα 2.1) και μπορούν να ανακτηθούν σε μορφή HTML ή XML. Για τις περισσότερες δημοσιεύσεις είναι διαθέσιμη η βιβλιογραφική αναφορά, που περιλαμβάνει τον τίτλο και με ελάχιστες εξαιρέσεις την περίληψη, ενώ για κάποια άρθρα είναι διαθέσιμο και το πλήρες κείμενο. Εκτός από τις λέξεις του κειμένου, στα αρχεία αυτά περιλαμβάνονται και άλλα στοιχεία σε δομημένη μορφή, όπως αναφορές σε άλλα άρθρα, λεξικά, λέξεις κλειδιά και ειδικοί όροι. Τα στοιχεία αυτά μπορούν να δώσουν πολλές πληροφορίες για το κείμενο.

Για την εξαγωγή των αλληλεπιδράσεων, απαραίτητη είναι η αναγνώριση των όρων miRNA και γονιδίου, ενώ για την αξιολόγηση της κάθε συσχέτισης (πραγματική Αλληλεπίδραση ή μη πραγματική), πρέπει να γίνει σημασιολογική επεξεργασία του κειμένου, για την εξαγωγή γλωσσικών δομών όπως οι προτάσεις, οι λέξεις, τα μέρη του λόγου και σε πιο υψηλό επίπεδο οι φράσεις και το σημασιολογικό δέντρο.

Όλες οι αναφορές σε κάποιο ζεύγος (microRNA, γονίδιο) ομαδοποιούνται και από τα δεδομένα αυτά εξάγονται κάποια χαρακτηριστικά (Ενότητα 3.2).

Ο κατηγοριοποιητής χρησιμοποιεί το διάνυσμα των χαρακτηριστικών σε καθένα από τα οποία αντιστοιχεί κάποιο βάρος, για να κατατάξει την κάθε αλληλεπίδραση σε Πραγματική ή Μη. Το βάρος που δίνεται σε κάθε διάνυσμα προκύπτει κατά τη διαδικασία της εκπαίδευσης (Κεφάλαιο 5).

Για την εξαγωγή των αλληλεπιδράσεων εκτελούνται τα βήματα που φαίνονται στα Σχήματα 4.1 και 4.2.

4.1 Μεθοδολογία αναγνώρισης αλληλεπιδράσεων

4.1.1 Εξαγωγή Αλληλεπιδράσεων

Το είδος της σχέσης μεταξύ των δύο όρων, miRNA και γονιδίου, που θέλουμε να εξάγουμε αφορά στην ρυθμιστική ικανότητα που έχει το microRNA στο γονίδιο. Αυτή μπορεί να είτε θετική είτε, όπως συμβαίνει στις περισσότερες περιπτώσεις, αρνητική. Επιπλέον, οι αλληλεπιδράσεις αυτές πρέπει να έχουν επαληθευτεί με κάποια πειραματική μέθοδο που μπορεί να είναι βιοχημική ή υπολογιστική.

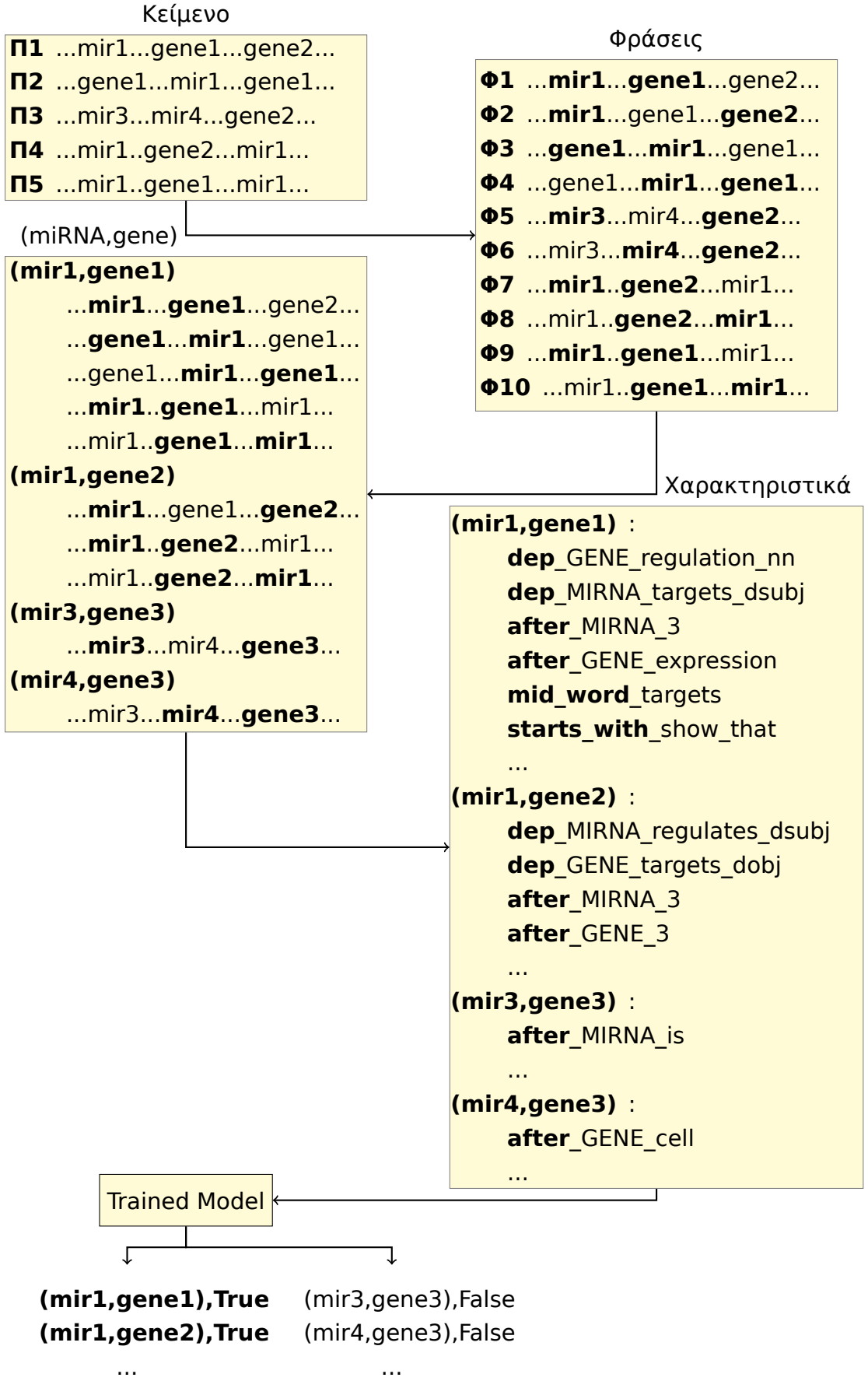
Το αποτέλεσμα, επομένως, που πρέπει να εξάγει ο κατηγοριοποιητής είναι οι επαληθευμένες συσχετίσεις μορίων microRNA με γονίδια που αναφέρονται σε κάθε δημοσίευση.

4.1.1.1 Πρόταση-Φράση

Θεωρούμε ότι η μεγαλύτερη δομική μονάδα στην οποία δύο όροι που εμφανίζονται μπορεί να συσχετίζονται είναι η πρόταση.

Λαμβάνονται υπόψιν όλα τα ζεύγη με δύο προϋποθέσεις:

- Να μην εμφανίζεται κανένας από τους δύο όρους στις ενδιάμεσες λέξεις.
- Να μην εμφανίζονται ανάμεσα στους δύο όρους περισσότερες από μία αλλαγές όρων.



Σχήμα 4.2: Αναπαράσταση της διαδικασίας της κατηγοριοποίησης με λεπτομέρεια στο μετασχηματισμό των δεδομένων

Για παράδειγμα, σε μία πρόταση μπορεί να εμφανίζονται ακρετοί όροι γονιδίων και microRNA:

...miR1...gene1...miR2...miR3...miR4...miR2...gene2...gene3...miR5...gene4...

Από την παραπάνω πρόταση τα ζεύγη που εξάγονται για να κατηγοριοποιηθούν στη συνέχεια είναι:

Πιθανές Αλληλεπιδράσεις:

(miR1,gene1), (miR2_1,gene1), (miR2_2,gene2), (miR2_2,gene3),
(miR3,gene1), (miR3,gene2), (miR3,gene3), (miR4,gene1),
(miR4,gene2), (miR4,gene3), (miR5,gene2), (miR5,gene3),
(miR5,gene4)

ή σχηματικά:

Μία σημαντική παράμετρος του προβλήματος είναι το γεγονός ότι τα δεδομένα που προσπαθούμε να κατηγοριοποιήσουμε δεν είναι αρκετό να χαρακτηρίζονται από την συγκεκριμένη συσχέτιση, αλλά πρέπει η επαλήθευση της να γίνεται στο κείμενο με κάποια πειραματική μέθοδο.

Δηλαδή, η παρακάτω φράση, για τους όρους miRNA1 και gene1 δεν είναι ικανή να καθορίσει απόλυτα το αποτέλεσμα του κατηγοριοποιητή (Πραγματική αλληλεπίδραση), παρότι η σημασία της είναι γενικά ξεκάθαρη.

...miRNA1 **target(s)** gene1...

Συγκεκριμένα, αν θεωρηθεί ο κανόνας της ύπαρξη της λέξης "targets" ανάμεσα στους δύο όρους miRNA και γονίδιο ως μοναδικός παράγοντας ώστε τοαποτέλεσμα του κατηγοριοποιητή να κρίνεται ως θετικό:

```
def classify(sentence, term1, term2):
    if (sentence[term1:term2] == 'target') or \
           (sentence[term1:term2] == 'targets' ):
            return True
    else:
            return False
```

Τα αποτελέσματα του κανόνα αυτού, με βάση τα δεδομένα εκπαίδευσης (Κεφάλαιο 5) είναι:

true_pos	true_neg	false_pos	false_neg	precision	recall
113	22	4027	860	0.90	0.11

αντίστοιχα, για την φράση:

...miRNA1 **regulate(s)** gene1...

και τα αποτελέσματα είναι:

true_pos	true_neg	false_pos	false_neg	precision	recall
116	42	4007	857	0.73	0.12

Αν θεωρήσουμε ως μοναδικούς κανόνες ύπαρξης αλληλεπίδρασης μία τουλάχιστον από τις δύο αυτές φράσεις τα αποτελέσματα θα είναι:

true_pos	true_neg	false_pos	false_neg	precision	recall
210	63	3986	763	0.77	0.21

Στα αποτελέσματα παραπάνω, χρησιμοποιούνται οι δείκτες recall και precision όπως ορίζονται στην Ενότητα 5.3, ενώ οι όροι true_pos, true_neg, false_pos και false_neg ορίζονται ως εξής:

true_pos είναι ο αριθμός των ζευγών (Ενότητα 4.4.1.4), που περιέχουν τουλάχιστον μία φράση (Ενότητα 4.4.1.3) που επαληθεύει τους ανάλογους κανόνες που ορίστηκαν και ταυτόχρονα είναι καταχωρημένα ως True (Ενότητα 5).

true_neg είναι ο αριθμός των ζεύγων που δεν εμφανίζουν καμία φράση που επαληθεύει τους κανόνες και δεν βρίσκονται καταχωρημένα ως True.

false_pos είναι ο αριθμός των ζευγών εμφανίζουν τουλάχιστον μία φράση που επαληθεύει του κανόνες και δεν βρίσκονται καταχωρημένα ως True.

false_neg είναι ο αριθμός των ζευγών εμφανίζουν δεν εμφανίζουν καμία μία φράση που επαληθεύει του κανόνες και βρίσκονται καταχωρημένα ως True.

Τα παραπάνω αποτελέσματα δείχνουν ότι ακόμα και για φράσεις πολύ απλές και σχετικά ξεκάθαρες στη σημασία τους, υπάρχουν περιπτώσεις όπου η αναφορά τους δεν χαρακτηρίζει μία πραγματική (για τον κατηγοριοποιητή) αλληλεπίδραση μέσα σε ένα κείμενο. Τέτοιες περιπτώσεις είναι, να γίνεται αναφορά σε προηγούμενα αποτελέσματα, η σειρά των όρων να είναι αντίστροφη από αυτή που φαίνεται στο παράδειγμα, το γονίδιο (ειδικά στην περίπτωση του regulates) να αντιστοιχεί σε κάποιο διαφορετικό όρο που δεν είναι γονίδιο (πχ. ασθένεια) και τέλος, ένας από τους δύο όρους να αναφέρεται γενικά σε καποια οικογένεια γονιδίων ή miRNA.

4.1.1.2 Ζεύγη microRNA-Γονιδίων

Σε ένα κείμενο η κάθε φράση δεν παίζει από μόνη της καθοριστικό ρόλο, ακόμη και στην περίπτωση μία πολύς ισχυρής φράσης όπως αυτές που αναφέρθηκαν στην προηγούμενη παράγραφο.

Επομένως, πρέπει να ληφθούν υπόψιν τόσο η κάθε φράση και τα συμφραζόμενα της, όσο και οι αναφορές που υπάρχουν στους συγκεκριμένους όρους στο υπόλοιπο κείμενο. Έτσι, οι φράσεις οι οποίες δεν υποδηλώνουν ξεκάθαρα την ύπαρξη συσχέτισης μεταξύ των όρων δεν κυριαρχούν και το συμπέρασμα προκύπτει από το σύνολο του κειμένου που περιλαμβάνει συμπεράσματα, υποθέσεις, προϋποθέσεις, πειραματικά δεδομένα και αποτελέσματα για την κάθε αλληλεπίδραση.

Έτσι, τα δεδομένα ομαδοποιήθηκαν ανάλογα με το ζεύγος miRNA-γονιδίου στο οποίο αντιστιχούν. Τα ζεύγη αυτά χρησιμοποιούνται ως είσοδος για τον κατηγοριοποιητή.

Για παράδειγμα, οι παρακάτω φράσεις είναι ένα υποσύνολο των φράσεων της δημοσίευσης με PMCID: PMC1764209 που αναφέρονται στους όρους (**mir-430**, **nanos1**):

Φράση α Here we report that the microRNA **mir-430** targets the 3′ untranslated region (UTR) of **nanos1** during zebrafish embryogenesis.

Φράση β A **mir-430** target site within the **nanos1** 3′ UTR reduces poly (A) tail length , mRNA stability and translation.

Φράση γ The differential regulation of the **nanos1** 3 ′ UTR by **mir-430** contrasts with previous studies of **mir-430** targets.

Φράση δ Second , it is unlikely that the **nanos1** mRNA is sequestered from **mir-430** , because extra copies of the mir-430 target site make the nanos1 reporter susceptible to repression in PGCs (Fig.4H).

Φράση ε Second , it is unlikely that the nanos1 mRNA is sequestered from mir-430 , because extra copies of the **mir-430** target site make the **nanos1** reporter susceptible to repression in PGCs (Fig.4H).

Όπως φαίνεται από τις παραπάνω προτάσεις, η **Φράση α**, αναφέρει ρητά ότι βρέθηκε ότι το mir-430 στοχεύει το nanos1. Ενώ οι **Φράση δ**, **Φράση ε**, είναι δύο φράσεις μέσα στην ίδια πρόταση.

Με τη μοντελοποίηση αυτή, τα ζεύγη για τα οποία γίνεται εκτενής περιγραφή μέσα στο κείμενο έχουν πλεονέκτημα έναντι των σύντομων αναφορών. Αυτό μπορεί να επηρεάσει αρνητικά δημοσιεύσεις στις οποίες προκύπτουν συμπεράσματα για πολλές αλληλεπιδράσεις και η αναφορά σε καθεμία είναι σύντομη.

4.1.1.3 Επεξεργασία Κειμένου

Στο πρώτο παράδειγμα στην πρώτη παράγραφο, εμφανίζεται ακόμη ένα πρόβλημα που είναι ιδιαίτερα συχνό στην αγγλική γλώσσα. Η λέξη ”targets” μπο-

ρεί να είναι τόσο ρήμα όσο και ουσιαστικό, δηλαδή η φράση αυτή παρουσιάζει αμφισημία στην γραμματική της:

...We demonstrated that **miRNA1 targets gene1, gene2**...

...**miRNA1 targets gene1, gene2** are...

Παρόλο που το συμπέρασμα των δύο φράσεων δεν είναι σημασιολογικά διαφορετικό, πολύ συχνά η δεύτερη φράση εμφανίζεται για αναφορές σε ήδη γνωστά ζεύγη, ενώ η πρώτη υποδηλώνει στις περισσότερες περιπτώσεις κάποια νέα αλληλεπίδραση.

Αυτές οι πολυσημίες στη γλώσσα, είναι δύσκολο να αντιμετωπιστούν χωρίς να γίνει σημασιολογική ανάλυση του κειμένου. Έτσι είναι αναγκαία μια αρχική προεπεξεργασία του κειμένου για την αναγνώριση κάποιων γλωσσικών δομών.

Όπως αναφέρθηκε και παραπάνω, η πρόταση θεωρείται ως η μέγιστη δομή, στο εσωτερικό της οποίας δύο όροι, miRNA και γονίδιο μπορούν να συσχετίζονται. Άλλες σημαντικές κατηγοροποιήσεις που βοηθούν την σημασιολογική ανάλυση ενός κειμένου είναι ο χωρισμός σε φράσεις, όπως ονοματικές και ρηματικές φράσεις, η αναγνώριση των μερών του λόγου και η σημασιολογική ανάλυση με τον εντοπισμό σχέσεων υποκειμένου, άμεσου και έμμεσου αντικειμένου και άλλων εξαρτήσεων μεταξύ των λέξεων.

4.1.1.4 Εξωτερικές Αναφορές

Σε πολλά σημεία των κειμένων των δημοσιεύσεων υπάρχουν βιβλιογραφικές αναφορές σε παλαιότερες εργασίες, ευρήματα και μεθοδολογίες. Η ύπαρξη μίας βιβλιογραφικής αναφοράς σε κάποιο σημείο ενός άρθρου, υποδηλώνει ότι κάποια από τα συμφραζόμενα του αναφέρονται σε κάποια διαφορετική δημοσίευση.

Για παράδειγμα η παρακάτω φράση δεν δίνει στοιχεία για τα αποτελέσματα που έχουν προκύψει από την συγκεκριμένη δημοσίευση σχετικά με τα miRNA1 και gene1.

....miRNA1 targets gene1 **[1]**...

Επιπλέον, η παραπάνω φράση που κατά την εκπαίδευση του κατηγοριοποιητή θα έχει τιμή False, θα δώσει αρνητικό βάρος στη λέξη ”targets” όταν βρίσκεται ανάμεσα στους δύο όρους.

Μία προσέγγιση για την επίλυση του προβλήματος αυτού είναι η απόρριψη κάθε πρότασης που περιέχει τουλάχιστον μία βιβλιογραφική αναφορά.

4.1.2 Χαρακτηριστικά

Τα χαρακτηριστικά ή features (Ενότητα 3.2) είναι ένας από τους πιο σημαντικούς παράγοντες που καθορίζουν την επίδοση ενός κατηγοριοποιητή σε ένα συγκεκριμένο πρόβλημα.

Δημιουργούνται κατά τη διαδικασία της εκπαίδευσης (Κεφάλαιο 5) με βάση τα δεδομένα εκπαίδευσης. Το πλήθος τους, όπως και το διάνυσμα με τα βάρη που αντιστοιχούν σε καθένα παράγονται στην ίδια διαδικασία.

Τα καινούρια χαρακτηριστικά που παράγονται για ένα καινούριο ζεύγος miRNA, γονιδίου σε ένα καινούριο κείμενο δεν λαμβάνονται καθόλου υπόψη.

4.1.3 Κατηγοριοποιητής

Για την εξαγωγή των αλληλεπιδράσεων από το κείμενο σχεδιάστηκε ένας δυαδικός κατηγοριοποιητής με κλάσεις Αληθής (True) και Ψευδής (False). Αληθή είναι τα ζεύγη όρων microRNA, γονιδίου που αποτελούν πραγματικές αλληλεπιδράσεις και Ψευδή όλα τα υπόλοιπα ζεύγη που εντοπίζονται στο κείμενο.

Ο κατηγοριοποιητής, Binary Maximum Entropy Classifier, (Ενότητα 3.1) κατατάσσει κάθε στοιχείο με βάση τα χαρακτηριστικά (Ενότητα 3.2) που έχει σε μία από τις δύο κλάσεις.

Τα χαρακτηριστικά όπως και το βάρος που αντιστοιχεί σε καθένα από αυτά έχει προκύψει κατά τη διαδικασία της εκπαίδευσης (Κεφάλαιο 5).

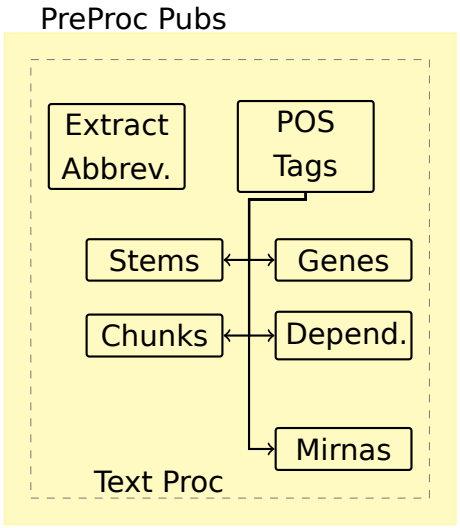
Η τιμή που επιστρέφει ο κατηγοριοποιητής είναι μία πιθανότητα ανάμεσα στις τιμές 0 και 1, αν η τιμή είναι μεγαλύτερη από 0.5 η αλληλεπίδραση θεωρείται Αληθής (True), ενώ σε διαφορετική περίπτωση, θεωρείται Ψευδής (False).

4.2 Διαδικασία Επεξεργασίας Κειμένου

Η επεξεργασία του κειμένου είναι απαραίτητη διαδικασία για την εξαγωγή συμπερασμάτων από κείμενα που είναι γραμμένα σε φυσική γλώσσα.

Η διαδικασία περιλαμβάνει, την αναγνώριση των δύο όρων, των γονιδίων και των microRNA, τον χωρισμό του κειμένου σε προτάσεις, τον χωρισμό των προτάσεων σε λέξεις, την αναγνώριση των μερών του λόγου, την απομόνωση του θέματος της λέξης, την εξαγωγή φράσεων και την εύρεση των εξαρτήσεων μεταξύ των λέξεων. Επιπλεόν από το κείμενο εξάγονται και οι συντομογραφίες.

Αναγνώριση όρων microRNA και γονιδίου



Σχήμα 4.3: Διαδικασία για την επεξεργασία του Κειμένου

Η αναγνώριση των microRNA γίνεται με τη χρήση γραμματικής όπως περιγράφεται στην Ενότητα 3.11.2.

Τα γονίδια προκύπτουν από αναζήτηση στο κείμενο λέξεων που αντιστοιχούν σε σύμβολα, ονόματα, καθώς και σε ονόματα και σύμβολα που έχουν αποσυρθεί.

Τα λεξικά που χρησιμοποιήθηκαν, διαχωρίζονται ανάλογα με τον οργανισμό αναφοράς. Προέρχονται από τα δεδομένα της Entrez Gene (Ενότητα 2.1.4), της HGNC (Ενότητα 2.4.1), της MGI (Ενότητα 2.4.2), της CGNC (Ενότητα 2.4.7), της RGD (Ενότητα 2.4.8) και τέλος, κάποια δεδομένα προέρχονται από την FLYBASE (Ενότητα 2.4.3). Η μέθοδος που χρησιμοποιήθηκε για την εξαγωγή των γονιδίων περιγράφεται στην Ενότητα 3.11.1.

Γλωσσικά Στοιχεία

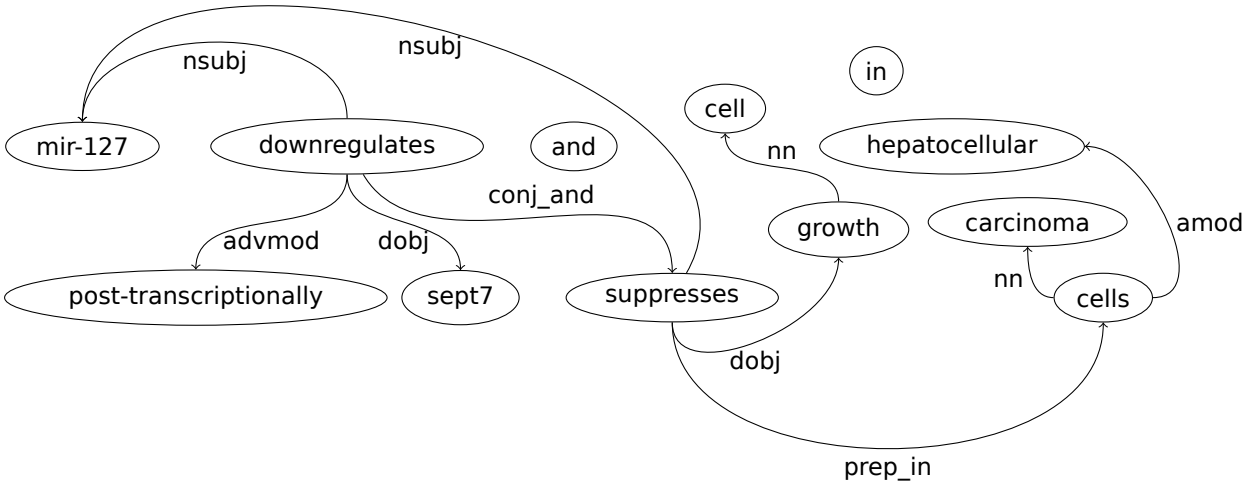
Αρχικά το κείμενο χωρίζεται σε προτάσεις, στη συνέχεια σε λέξεις και σε κάθε λέξη ανατίθεται ένα μέρος του λόγου (Ενότητα 3.6.1) [16]. Ακολουθούν η εξαγωγή των εξαρτήσεων (Ενότητα 3.13.1) [17], η απομάκρυνση των γραμματικών καταλήξεων (Ενότητα 3.7.1) [18] και ο χωρισμός σε ονοματικές, ρηματικές και προθετικές φράσεις (Ενότητα 3.10.1). Μετά την ανάλυση αυτή, κάθε πρόταση αποτελείται από ένα σύνολο στοιχείων που αναφέρονται ανεξάρτητα σε καθε λέξη (token). Ένα παράδειγμα δίνεται για τον τίτλο του άρθρου PMID:24854842:

> *mir-127 Post-Transcriptionally Downregulates Sept7 and Suppresses Cell Growth in Hepatocellular Carcinoma Cells.*

Η κάθε λέξη της πρότασης περιλαμβάνει τα παρακάτω στοιχεία:

Λέξη	**Θέμα**	**Μέρος του Λόγου**	**Φράση**
mir-127	mir-127	NN	B-NP
Post-Transcriptionally	post-transcriptionally	RB	O
Downregulates	downregulate	VBZ	B-VP
Sept7	sept7	NN	B-NP
and	and	CC	O
Suppresses	suppress	VBZ	B-VP
Cell	cell	NN	B-NP
Growth	growth	NN	B-NP
in	in	IN	B-PP
Hepatocellular	hepatocellular	JJ	B-NP
Carcinoma	carcinoma	NN	I-NP
Cells	cell	NN	I-NP
.	.	.	O

Ο Γράφος των εξαρτήσεων είναι:



Εξαγωγή Συντμήσεων

Για κάθε δημοσίευση γίνεται εξαγωγή των ορισμών των συντομογραφιών που περιέχονται στο κείμενο (Ενότητα 3.8.1) [19].

Για παράδειγμα για το παρακάτω τμήμα του άρθρου PMCID: PMC3325277:

> Principal components analysis (PCA) and unsupervised, hierarchical clustering analysis (HCL) of the samples and the hybridization levels of the 463 microRNAs showing detectable and variable expression revealed a clear segregation pattern of the samples. Both PCA and HCL are multivariate

methods commonly used to analyze the behaviour of multiple variables at the same time.

Οι ορισμοί που περιέχονται είναι:

Συντμήσεις	**Ορισμός**
PCA	Principal components analysis
HCL	hierarchical clustering analysis

Τα συγκεκριμένα στοιχεία χρησιμεύουν στον αποκλεισμό ορισμένων συντομεύσεων από την αναγνώριση τους ως γονίδια.

Αυτό βασίζεται στον απλό κανόνα ότι η τελευταία λέξη μίας ονοματικής φράσης καθορίζει τη σημασία του ορισμού. Οι λέξεις οι οποίες χρησιμοποιήθηκαν ανήκουν στο παρακάτω σύνολο:

'disease', 'diseases', 'disorder', 'disorders', 'syndrome',

'cancer', 'carcinoma', 'carcinomas', 'tumor', 'tumors', 'tumour', 'tumours', 'melanoma', 'melanomas', 'leukemia', 'lymphoma', 'lipoma', 'sarcoma',

'stress', 'arthritis', 'diabetes', 'infection',

'tissue', 'tissues', 'cell', 'cells', 'stem',

'neurons', 'heart', 'kidney',

'fibroblasts','fibroblast','cytotrophoblast',

'pcr', 'assay',

'biosystems', 'systems',

'deviation', 'rate', 'threshold',

'element', 'elements', 'sequences', 'sequence',

'virus',

'column','recombination','transition','instability'

Υπάρχουν πολλές περιπτώσεις στις οποίες το όνομα μίας ασθένειας ή ενός φαινότυπου δίνει το όνομα του στο μεταλλαγμένο γονίδιο που προκαλεί την κατάσταση αυτή. Στις περιπτώσεις αυτές είναι δύσκολη η διάκριση των όρων που αναφέρονται σε γονίδια.

Ένα παράδειγμα είναι όρος MDR που αναφέρεται στη αντίσταση στη θεραπεία του καρκίνου. Μία από τις πρωτεΐνες που σχετίζονται με την κατάσταση αυτή είναι η MDR-1. Παρακάτω φαίνεται το τμήμα του άρθρου (PMCID: PMC3763502) που περιλαμβάνει τον ορισμό του όρου *MDR*:

The multidrug resistance (MDR) refers to the ability of tumor cells to resist several unrelated drugs after exposure to a single chemotherapy drug (1). Nearly all initially responsive breast tumors will eventually acquire an **MDR** phenotype (2). P-glycoprotein (**MDR-1**) (3), multidrug resistance associated protein (MRP-1) (4), and breast cancer resistance protein (BCRP) (2,5) have been considered as critical **MDR**-related factors.

Η πρωτεΐνη MDR-1 εμφανίζεται και με την συνώνυμη ονομασία MDR σε κείμενα.

4.3 Χαρακτηριστικά

Τα χαρακτηριστικά είναι τα στοιχεία τα οποία χρησιμοποιεί ο κατηγοριοποιητής για να κατατάξει κάθε ζεύγος microRNA-Γονιδίου σε Αλληλεπίδραση ή Όχι Αλληλεπίδραση.

Αρχικά, όλοι οι όροι microRNA και Γονίδιο παίρνουν τις τιμές MIRNA και GENE σε όλο το κείμενο, επειδή το ακριβές όνομα τους δεν έχει σημασία για τον κατηγοριοποιητή και μπορεί να μπερδέψει το μοντέλο και τα ακριβή αναγνωριστικά και ονόματα να γίνουν τα κυρίαρχα χαρακτηριστικά.

Όλοι οι αριθμοί, δεκαδικοί και ακέραιοι αντιμετωπίζονται ενιαία και παίρνουν την τιμή 0. Ειδική περίπτωση είναι οι αριθμοί 3 και 5 που αντιστοιχούν στα άκρα των νουκλεϊκών οξέων και αντιστοιχίζονται με τα ψηφία 35.

Επιπλέον, εκτός της επεξεργασίας που έχει ήδη γίνει στις λέξεις για την απομάκρυνση των γραμματικών καταλήξεων σε πολλά χαρακτηριστικά έχει χρησιμοποιηθεί και ο Porter Stemmer (Ενότητα 3.7.2) [20], η χρήση του οποίου απλοποιεί περισσότερο τα αποτελέσματα.

Για τις άγνωστες λέξεις έχει χρησιμοποιηθεί λεξικό που αποτελείται από τη συχνότητα εμφάνισης λέξεων στα κείμενα εκπαίδευσης του μοντέλου (Κεφάλαιο 5). Οι λέξεις αυτές αποκόπτονται με βάση κάποια συχνότητα που ορίζεται κατά την εκπαίδευση και είναι γενικά μικρή.

Αντίστοιχα, οι πολύ συχνές λέξεις (stopwords) απορρίπτονται. Η λίστα των λέξεων που χρησιμοποιείται, για ένα υποσύνολο των χαρακτηριστικών είναι:

it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those, be, have, do, a, b, c, an, the, and, or, because, until, while, at, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, out, on, off, over, under, again, further,

then, once, here, there, when, where, why, how, all, any, each, few, more, most, other, some, such, only, same, so, than, too, very, s, t, can, may, might, would, could, also, will, just, don, should, now, cell, mrna, mirna, gene, tissue, protein

Τέλος, έχει χρησιμοποιηθεί μία λίστα με τροποποιημένες λέξεις στις οποίες έχει γίνει αφαίρεση των καταλήξεων με τον αλγόριθμο Porter και σε περιπτώσεις που χρειάζεται έχει διατηρηθεί κάποιο μέρος της κατάληξης. Για παράδειγμα για τη λέξη “activation” ο αλγόριθμος επιστρέφει το θέμα “activ” και αυτό συγχέεται με τις λέξεις “active” ή “activity” για τις οποίες ο αλγόριθμος δίνει το ίδιο αποτέλεσμα.

Οι ακόλουθες θεωρούνται σημαντικές για την αλληλεπίδραση των ρυθμιστικών μορίων microRNA με τα γονίδια:

target	upregulat	downregulat	regulat	bind
repress	suppress	inhibit	abolish	diminish
induc	reduc	increas	decreas	enhanc
overexpress	underexpress	stimulat	cleav	silenc
deadenyl	degrad	activat	posttranscr	transcr
control	mediator	mediat	knockdown	

Πίνακας 4.1: Σημαντικές Λέξεις για τις αλληλεπιδράσεις microRNA-γονιδίου

4.3.1 Ομάδες Χαρακτηριστικών

Τα χαρακτηριστικά ανήκουν στις παρακάτω ομάδες:

4.3.1.1 Τελευταία λέξη της Ονοματικής Φράσης

Κάθε όρος γονίδιο ή miRNA θα ανήκει σε κάποια ονοματική φράση. Η τελευταία λέξη της φράσης αυτής δίνει στοιχεία για το είδος του όρου και τη σημασία του στη συγκεκριμένη φράση.

Ορίζονται δύο ομάδες χαρακτηριστικών, της μορφής:

’**CH1_LAST**:(word)_(MIRNA|GENE,GENE|MIRNA)’

’**CH2_LAST**:(word)_(MIRNA|GENE,GENE|MIRNA)’

Από τους παραπάνω κανόνες κατασκευάζονται χαρακτηριστικά με βάση τη σειρά των όρων microRNA (MIRNA) και γονίδιο (GENE) και την λέξη (”word“).

Επομένως για παράδειγμα στην παρακάτω φράση ο πρώτος όρος δεν θα επιστρέψει κανένα χαρακτηριστικό:

*"**microRNA-34a** Inhibits the Growth, Invasion and Metastasis of Gastric Cancer by Targeting PDGFR and **MET** Expression."*

'**CH2_LAST**:(expression)_(MIRNA,GENE)'

Αντίθετα η παρακάτω φράση δίνει δύο χαρακτηριστικά:

*"**miR-34a** levels in human gliomas inversely correlated to **c-Met** levels measured in the same tumors."*

'**CH1_LAST**:(level)_(MIRNA,GENE)'

'**CH2_LAST**:(level)_(MIRNA,GENE)'

4.3.1.2 Σημασιολογικές Εξαρτήσεις - Μονοπάτι

Οι σημασιολογικές εξαρτήσεις είναι σημαντικά χαρακτηριστικά και προκύπτουν από την εξαγωγή εξαρτήσεων (Ενότητα 3.13).

Για τον έλεγχο της σχέσης των όρων αρχικά εντοπίζονται τα συντομότερα μονοπάτια μεταξύ των δύο όρων από τον γράφο των εξαρτήσεων (Ενότητα 3.13.1.1) και στη συνέχεια λαμβάνονται στοιχεία από αυτά.

Παρακάτω ακολουθούν οι ορισμοί των τριών διαφορετικών ομάδων χαρακτηριστικών που σχετίζονται με τις σημασιολογικές εξαρτήσεις και παραδείγματα που προέρχονται από την ακόλουθη φράση:

*In the present study, we found that **mir-19a** inhibitor decreased the expression of MDR-1, **MRP-1**, and BCRP.*

SPCH1,SPCH2: Η πρώτη ομάδα χαρακτηριστικών είναι οι λέξεις που ανήκουν τόσο στο μονοπάτι όσο και σε μία από τις δύο ονοματικές φράσεις των δύο όρων.

Στο συγκεκριμένο παράδειγμα, μόνο ο πρώτος όρος περιέχει μία λέξη που ανήκει σε κάποιο συντομότερο μονοπάτι μεταξύ των δύο όρων και ταυτόχρονα είναι μέρος της ονοματικής φράσης "mir-19a inhibitor".

Έτσι, παράγεται το παρακάτω χαρακτηριστικό:

'**SPCH1**:(inhibit)_(MIRNA,GENE)'

SPATH: Η δεύτερη ομάδα χαρακτηριστικών είναι οι λέξεις που βρίσκονται στο μονοπάτι και δεν ανήκουν σε κάποια από τις δύο ονοματικές φράσεις των δύο όρων.

Για παράδειγμα:

'**SPATH**:(decreas_express)_(MIRNA,GENE)'

REL: Η τρίτη ομάδα χαρακτηριστικών, αποτελείται από ζεύγη λέξεων μαζί με τη σχέση που τις συνδέει. Από αυτα, λαμβάνονται τα ζεύγη για τα οποία το ένα μέλος είναι ένας από τους όρους MIRNA ή GENE ή ανήκει στις λέξεις του του Πίνακα 4.1.

Για παράδειγμα:

'**REL**:(inhibit_decreas_subj)_(MIRNA,GENE)'

4.3.1.3 Τοπικές Εξαρτήσεις - Πριν τον πρώτο όρο

*In the present study, we found that **mir-19a** inhibitor decreased the expression of MDR-1, **MRP-1**, and BCRP.*

Για την παραπάνω φράση, οι λέξεις που προηγούνται του όρου microRNA είναι οι "in the present study, we found that". Αφαιρώντας, όμως, τις κοινές λέξεις (stopwords), στις οποίες ανήκει η λέξη the και that, οι δύο λέξεις που προηγούνται του πρώτου όρου είναι:

'**FRONT1**:(we_find)_(MIRNA,GENE)'

4.3.1.4 Τοπικές Εξαρτήσεις - Πριν τον δεύτερο όρο

Όμοια με την παραπάνω περίπτωση, ορίζεται η ομάδα χαρακτηριστικών για τις δύο λέξεις πριν τον δεύτερο όρο.

Ένα παράδειγμα είναι:

'**FRONT2**:(express_of)_(MIRNA,GENE)'

4.3.1.5 Τοπικές Εξαρτήσεις - Μετά τον πρώτο όρο

Ορίζεται η ομάδα χαρακτηριστικών για τις δύο λέξεις μετά τον πρώτο όρο.

Ένα παράδειγμα είναι:

'**LAST1**:(target_for)_(GENE,MIRNA)'

Εκτός από τις δύο λέξεις μετά τον όρο, ορίζεται και η ομάδα χαρακτηριστικών για μία λέξη μετά τον δεύτερο όρο.

Ένα παράδειγμα είναι:

'**AFTER1**:(GENE)_(pathway)'

Στην περίπτωση αυτή, η λέξη ακολουθεί το είδος του όρου (GENE ή MIRNA).

4.3.1.6 Τοπικές Εξαρτήσεις - Μετά τον δεύτερο όρο

mir-19a *modelates expression of MDR-1,* ***MRP-1*** *and BCRP in MDR cells.*

Στο παραπάνω παράδειγμα μετά τον δεύτερο όρο ακολουθούν η κοινή λέξη 'and' (stopwords), το γονίδιο BCRP και το λάθος γονίδιο MDR (φαινότυπος multidrug resistance) (αναγνωρίζεται και είναι συνώνυμο του MDR-1 που αντιστοιχεί με την μετάλλαξη που προκαλεί τον φαινότυπο MDR). Έτσι, απομένει η φράση "in cells".

'**LAST2**:(in_cell)_(MIRNA,GENE)'

Εκτός από τις δύο λέξεις μετά τον δεύτερο όρο ορίζεται και η ομάδα χαρακτηριστικών για μία λέξη μετά τον δεύτερο όρο. Ένα παράδειγμα είναι:

'**AFTER2**:(GENE)_(gene)'

4.3.1.7 Τοπικές Εξαρτήσεις - Ενδιάμεσοι Όροι

Για τους ενδιάμεσους όρους ορίζονται διαφορετικές ομάδες χαρακτηριστικών.

T: Στη ομάδα αυτή ανήκουν ομάδες 2 ή 3 λέξεων που περιέχουν κάποια από τις λέξεις του Πίνακα 4.1.

Μερικά παραδείγματα είναι:

'**T**:direct_target_of_(GENE,MIRNA)'

'**T**:famili_suppress_(MIRNA,GENE)'

MID: Στην ομάδα αυτή ανήκουν το πολύ 4 λέξεις, ξεκινώντας από τον δεύτερο όρο. Ένα παράδειγμα είναι:

'**MID**:(neg_regulat_by)_(GENE,MIRNA)'

MID_ALL: Στην ομάδα αυτή ανήκουν όλες οι ενδιάμεσες λέξεις.

Μερικά Παραδείγματα είναι:

'**MID_ALL**:(predict_target_of)_(GENE,MIRNA)'

MIDR: Στην ομάδα αυτή περιλαμβάνονται λέξεις ανάμεσα στους δύο όρους. Όσες λέξεις ανήκουν στον Πίνακα 4.1 αντικαθίστανται από την ετικέτα του μέρους του λόγου (Πίνακας 3.1).

Μερικά παραδείγματα είναι:

'**MIDR**:(via_VB)_(MIRNA,GENE)'

'**MIDR**:(indirect_VB)_(MIRNA,GENE)'

MIDT: Το αντίστροφο από την ομάδα **MIDR**, δηλαδή όλες οι ενδιάμεσες λέξεις αντικαθίστανται από την ετικέτα του μέρους του λόγου τους (Πίνακας 3.1), εκτός από αυτές που ανήκουν στον Πίνακα 4.1.

Τα παραδείγματα που χρησιμοποιήθηκαν παραπάνω είναι:

'**MIDT**:(IN_target)_(MIRNA,GENE)'

'**MIDT**:(JJ_target)_(MIRNA,GENE)'

4.3.1.8 Συνδυασμοί

Μαζί με τις ομάδες χαρακτηριστικών που παρουσιάστηκαν παραπάνω, χρησιμοποιήθηκαν και αρκετοί συνδυασμοί. Κάποιοι από αυτούς είναι οι ακόλουθοι:

'**FRONT1_LAST2**:(analysi_of)_(express_in)_(MIRNA,GENE)'

'**FRONT1_LAST1**:(these_includ)_(target_for)_(GENE,MIRNA)'

'**FMID**:(function_of)_(indirect_target)_(MIRNA,GENE)'

4.4 Δομή κώδικα και χρήση

Οι κλάσεις που σχετίζονται άμεσα με την κατηγοριοποίηση του κειμένου. είναι οι "ArticleI", "MGTupleD", "Phrase", "Gene", "Sent", "Citation", "Publication" και "BinMGMaxentClassifier". Παρακάτω, ακολουθεί πιο λεπτομερής περιγραφή των μεθόδων τους.

4.4.1 Δομή Κώδικα

4.4.1.1 Άρθρο - Article

Το άρθρο αποτελείται από το pubid (PMID ή PMCID), από προτάσεις και τους ορισμούς των συντομογραφιών (Ενότητα 3.8.1) που υπάρχουν και ορίζονται σε κάποιο σημείο του κειμένου.

Η κλάση Article λαμβάνει ως παραμέτρους, τα δεδομένα, "data" και τη μεταβλητή "is_parsed" που δηλώνει αν από τα δεδομένα έχουν εξαχθεί οι εξαρτήσεις:

data: Δομή tuple της Python που αποτελείται από 4 στοιχεία:

pubid: Το PMID ή το PMCID ανάλογα με την περίπτωση

orgs: Μία δομή ”Dictionary“ της Python που περιλαμβάνει μόνο τους οργανισμούς που εντοπίστηκαν στο κείμενο, ή σε περίπτωση που δεν βρέθηκαν οργανισμοί, την τιμή ’all’

abbrev: Οι ορισμοί των συντομογραφιών που εντοπίστηκαν στο κείμενο.

sents: Το κείμενο ως μία λίστα από προτάσεις και καθεμία από αυτές περιλαμβάνει τις λέξεις, τα μέρη του λόγου, τα microRNA και τα γονίδια που εντοπίστηκαν.

is_parsed: Παίρνει τις τιμές True, αν στα δεδομένα περιλαμβάνεται ο γράφος των εξαρτήσεων και False αν όχι.

Citation

Το Citation είναι υποκλάση του Article και μπορεί να αρχικοποιηθεί με δεδομένα το κείμενο του τίτλου και το κείμενο της περίληψης.

Συγκεκριμένα, περιλαμβάνει την μέθοδο ”parse“ που λαμβάνει ως παραμέτρους, τον τίτλο (title) και την περίληψη (abstract), το αναγνωριστικό (pubid) και τη μεταβλητή που δηλώνει αν θα γίνει εξαγωγή των εξαρτήσεων ή όχι.

Citation.parse: Αρχικοποιεί την κλάση (classmethod)

title: Ο τίτλος της δημοσίευσης σε μορφή κειμένου (αλληλουχία χαρακτήρων).

abstract: Η περίληψη της δημοσίευσης σε μορφή κειμένου (αλληλουχία χαρακτήρων).

pubid: Το αναγνωριστικό τη βιβλιογραφικής αναφοράς (PMID) στο άρθρο.

parse: True αν θα γίνει εξαγωγή των εξαρτήσεων, False στην αντίθετη περίπτωση.

Publication

Το Publication είναι υποκλάση του Article και μπορεί να αρχικοποιηθεί με το HTML κείμενο της δημοσίευσης και την συνάρτηση parse.

Publication.parse: Αρχικοποιεί την κλάση (classmethod)

text: Το κείμενο της δημοσίευσης σε μορφή HTML.

pubid: Το αναγνωριστικό της δημοσίευσης (PMCID) του άρθρου, μπορεί επίσης να είναι κάποιο άλλο αναγνωριστικό.

parse: True αν θα γίνει εξαγωγή των εξαρτήσεων, False στην αντίθετη περίπτωση.

4.4.1.2 Πρόταση - Sent

Η πρόταση προκύπτει από τη διαδικασία του Χωρισμού των Προτάσεων ή Sentence Tokenization (Ενότητα 3.6.1) και περιλαμβάνει ένα σύνολο από στοιχεία για την κάθε λέξη, αλλά και καθολικά στοιχεία πρότασης.

Για κάθε λέξη της πρότασης υπάρχουν τέσσερα στοιχεία, η ίδια η λέξη όπως χωρίζεται από τον Word Tokenizer (Ενότητα 3.6.1), η ετικέτα του μέρους του λόγου, POS tag, (Ενότητα 3.6.1), η ετικέτα των γραμματικών φράσεων (Πίνακας 3.3) και το θέμα (Ενότητα 3.7).

Επιπλέον, το πεδίο psent είναι ο γράφος των εξαρτήσεων (Ενότητα 3.13.1) που περιλαμβάνει τα ζεύγη των λέξεων που εξαρτώνται σημασιολογικά.

Ακόμα, η κάθε πρόταση περιλαμβάνει το sent_no που είναι η σειρά της πρότασης μέσα στο κείμενο, το ptag που είναι την ετικέτα του HTML αρχείου στο οποίο ανήκει η πρόταση και το xref που είναι η παρουσία ή μη βιβλιογραφικής αναφοράς στην συγκεκριμένη πρόταση.

Συγκεκριμένα η κλάση αρχικοποιείται με τις παρακάτω παραμέτρους και μεθόδους.

Παράμετροι:

parsed_sent: Η πρόταση προεπεξεργασμένη. Περιλαμβάνει τις λέξεις, τα μέρη του λόγου, τα γονίδια τα microRNA και προαιρετικά τον γράφο των εξαρτήσεων.

sent_no: Ο αύξων αριθμός της πρότασης.

organisms: Προαιρετικό, δομή ”Dictionary“ που περιλαμβάνει τους οργανισμούς που βρέθηκαν στο κείμενο.

Μέθοδοι:

toks: Οι λέξεις στις οποίες χωρίζεται η πρόταση

tags: Τα μέρη του λόγου που αντιστοιχούν οι λέξεις

chunks: Οι γραμματικές φράσεις στις οποίες χωρίζονται οι λέξεις

stems: Τα θέματα των λέξεων

mirnas: Τα microRNA που περιλαμβάνει η πρόταση

genes: Τα γονίδια που περιλαμβάνει η πρόταση

psent: Ο γράφος των εξαρτήσεων σε περίπτωση που υπάρχει, σε διαφορετική περίπτωση επιστρέφεται η κενή λίστα ([]).

xref: True, αν υπάρχει εξωτερική βιβλιογραφική αναφορά στην πρόταση, αλλιώς False.

4.4.1.3 Φράση - Phrase

Μια ”Φράση“ αποτελείται από μία πρόταση (Ενότητα 4.4.1.2), δηλαδή τις λέξεις και τα υπόλοιπα στοιχεία της πρότασης, και από δύο ζεύγη σημείων που είναι οι θέσεις των δύο όρων, του miRNA και του γονιδίου. Επιπλέον, περιλαμβάνει και τα υπόλοιπα στοιχεία της πρότασης, όπως η ύπαρξη εξωτερικής αναφοράς και ο αριθμός της πρότασης από την οποία έχει προέλθει.

Αρχικοποιείται με τα στοιχεία της ”Φράσης“ και περιλαμβάνει τις παρακάτω μεθόδους:

Μέθοδοι:

sent: Συνάρτηση που επιστρέφει την πρόταση.

psent: Επιστρέφει τον γράφο των Εξαρτήσεων.

gm: Επιστρέφει το ζεύγος (microRNA, γονίδιο). Περιλαμβάνει τις θέσεις στην πρόταση.

xref: Επιστρέφει True αν η πρόταση στην οποία ανήκει η φράση περιέχει κάποια εξωτερική βιβλιογραφική αναφορά και False σε διαφορετική περίπτωση.

4.4.1.4 Ζεύγος (miRNA,γονίδιο)

Το κάθε ζεύγος (miRNA, γονίδιο) αποτελείται από τρία στοιχεία, το miRNA, το γονίδιο και τις φράσεις (Ενότητα 4.4.1.3) που αναφέρονται στα δύο πρώτα στοιχεία.

Το miRNA, αναγνωρίζεται από το όνομα του (πχ. miR-100). Το γονίδιο αποτελείται από δύο στοιχεία, όλα τα αναγνωριστικά (gids) και όλα τα ονόματα με τα οποία εμφανίζεται στο κείμενο (gtoks). Οι φράσεις είναι όλες οι αναφορές του ζεύγους μέσα στο κείμενο. Κάποιες προτάσεις περιλαμβάνονται δύο φορές (η διεύθυνση τους).

Η κλάση αποτελεί υποκλάση του dictionary της python και περιλαμβάνει μία μέθοδο την add για την προσθήκη νέας φράσης.

Παράμετροι:

pubid: Αναγνωριστικό PMCID για τα πλήρη άρθρα και PMID για τις βιβλιογραφικές αναφορές

mir: Το όνομα και η θέση του microRNA της πρότασης.

gen: Το γονίδιο, δομή Gene (Ενότητα 4.4.1.5)

phrases: Λίστα από φράσεις (Ενότητα 4.4.1.3)

Μέθοδοι:

add: Προσθήκη νέας φράσης και ενημέρωση των γονιδίων με τη μέθοδο Gene.update

4.4.1.5 Γονίδιο - Gene

Το γονίδιο είναι μία απλή δομή που περιέχει τα αναγνωριστικά κάποιου γονιδίου και τα ονόματα με τα οποία εμφανίζεται μέσα στο κείμενο.

Παράμετροι:

gids: Λίστα με τα αναγνωριστικά γονιδίων που έχουν βρεθεί.

gen: Η λέξη από το κείμενο που αναγνωρίστηκε ως γονίδιο

Μέθοδοι:

update: Συνάρτηση που προσθέτει νέο γονίδιο στη δομή, ανανεώνοντας τόσο τα αναγνωριστικά όσο και τα ονόματα.

4.4.1.6 BinMGMaxentClassifier

Ο BinMGMaxentClassifier είναι υποκλάση της nltk.classify.ClassifierI της βιβλιοθήκης της Python, NLTK (Ενότητα 3.3).

Η αρχικοποίηση γίνεται κατά την εκπαίδευση (Κεφάλαιο 5) και η περιγραφή της κλάσης παρουσιάζεται με περισσότερες λεπτομέρειες στην ενότητα 5.4.1.

Για την κατηγοριοποίηση μίας νέας δημοσίευσης ή πολλών μπορεί να χρησιμοποιηθεί καποιο ήδη εκπαιδευμένο μοντέλο και η κατηγοριοποίηση γίνεται με τις συναρτήσεις ”classify“ και ”prob_classify“ για ένα μόνο ζεύγος και ”batch_classify“ και ”batch_prob_classify“ για μία λίστα από στοιχεία.

4.5 Χρήση

Υλοποιήθηκε η βιβλιοθήκη ”mg_classifier“ και περιλαμβάνει τον κώδικα που είναι απαραίτητος για την κατηγοριοποίηση των αλληλεπιδράσεων νέων δημοσιεύσεων ή περιλήψεων με βάση ήδη εκπαιδεύμένα μοντέλα. Η διαδικασία εκπαίδευσης νέων μοντέλων περιγράφεται στην ενότητα 5.4.2.

Η ανάκτηση κειμένων της NCBI για να την κατηγοριοποίηση των αλληλεπιδράσεων που περιλαμβάνουν, μπορεί να γίνει με τη χρήση του module ”find_pubs“ της βιβλιοθήκης.

Η αναζήτηση μπορεί να γίνει με βάση κάποιο αναγνωριστικό της PubMed ή της PMC.

Για παράδειγμα η ανάκτηση του τίτλου (ArticleTitle) και της περίληψης (Abstract) της δημοσίευσης με PMID:24887517 γίνεται ως εξής:

```
>>> from mg_classifier import find_pubs
>>> title,abstract = find_pubs.fetch_citation('24887517')
>>> title
'miR-21 Overexpression Enhances TGF-\xce\xb21-induced epithelial-to-mesenchymal Transition by Target smad7 and Aggravates Renal Damage in Diabetic Nephropathy.'
>>> abstract[:305]
'Epithelial-to-mesenchymal transition (EMT) plays an important role in renal interstitial fibrosis (RIF) with diabetic nephropathy (DN). Smad7(a inhibitory smad), a downstream signaling molecules of TGF-\xce\xb21, represses the EMT. The physiological function of miR-21 is closely linked to EMT and RIF. However,'
```

Η επεξεργασία του κειμένου, δηλαδή η εξαγωγή των συντομογραφιών, ο χωρισμός σε προτάσεις και λέξεις, η αναγνώριση των μερών του λόγου, η αναγνώριση των microRNA και των γονιδίων μπορεί να γίνει με τη συνάρτηση "parse_pub" και "parse_citation" του module "text_proc":

```
>>> from mg_classifier import text_proc
>>> pubid,orgs,abbrev,sents = text_proc.parse_citation(pubid = '24887517', title = title, abstract = abstract, parse = True)
>>> pubid
'24887517',
>>> orgs
{'hmr': True}
>>> ab,ab_def = zip(*abbrev)
>>> zip(*ab)[0]
('EMT', 'RIF', 'DN')
>>> zip(*ab_def)[0]
('Epithelial-to-mesenchymal transition',
'renal interstitial fibrosis',
'diabetic nephropathy')
>>> ab[1]
```

```
('RIF', {'ENTREZ:110685', 'ENTREZ:23912', 'ENTREZ:54509', 'ENTREZ:80
196', 'Ensembl:ENSG00000139725', 'Ensembl:ENSG00000170633', 'Ense
mbl:ENSMUSG00000029449', 'HGNC:15703', 'HGNC:17297', 'HPRD:1050
9', 'HPRD:15246', 'MGI:1345629', 'MGI:2153340', 'MGI:87953', 'MIM:608
299', 'RefSeq:NM_194271', 'Vega:OTTHUMG00000169077', 'Vega:OTTHU
MG00000171524', 'Vega:OTTMUSG00000030767'})
>>> ab_def[1]
('renal interstitial fibrosis', set())
>>> len(sents)
11
```

Το αποτέλεσμα που επιστρέφει η συνάρτηση είναι το αναγνωριστικό που πήρε ως είσοδο, οι οργανισμοί που αναγνωρίστηκαν, οι ορισμοί συντομογραφιών στο κείμενο και τέλος οι προτάσεις που περιλαμβάνουν τις λέξεις, τις ετικέτες για τα μέρη του λόγου (Πίνακας 3.1), τα microRNA που έχουν αναγνωριστεί και τα γονίδια, καθώς και ο γράφος των εξαρτήσεων, αν επιλεγεί.

Τα σύνολα που επιστρέφονται στις συντομογραφίες είναι τα αναγνωριστικά των γονιδίων που εντοπίστηκαν Στην συγκεκριμένη περίπτωση ο όρος "RIF" αντιστοιχεί στη συντομογραφία κάποιου γονιδίου, αλλά από τον ορισμό προκύπτει ότι πρόκειται για ασθένεια (διάμεση νεφρική ίνωση) και στην περιγραφή του δεν αναγνωρίζεται κάποιο γονίδιο.

Η μορφή που επιστρέφει η συνάρτηση είναι χρήσιμη για την μαζική επεξεργασία πολλών κειμένων ή περιλήψεων και είναι η μορφή στην οποία βρίσκονται αποθηκευμένα τα αρχεία των δημοσιεύσεων.

Η εξαγωγή των υπόλοιπων στοιχείων, δηλαδή των θεμάτων και των γραμματικών φράσεων γίνεται στην αρχικοποίηση της κλάσης "Article" (ή της υποκλάσης "Citation" στην περίπτωση αυτή):

```
>>> from mg_classifier import article
>>> cit = article.Citation((pubid,orgs,abbrev,sents), is_parsed=True)
```

Επιπλέον, μπορεί να γίνει άμεσα όλη η επεξεργασία με την μέθοδο της κλάσης Citation, parse:

```
>>> from mg_classifier import article
>>> cit = article.Citation.parse(title = title, abstract = abstract, pubid =
'24887517', parse = True)
```

Για την εκτύπωση των προτάσεων:

```
>>> for sent in cit2.sents()[:2]:
....        print sent
....
mir-21 Overexpression Enhances TGF-β1-induced epithelial-to-mesenchy
mal Transition by Target smad7 and Aggravates Renal Damage in Diabetic
Nephropathy .
Epithelial-to-mesenchymal transition ( EMT ) plays an important role in
renal interstitial fibrosis ( RIF ) with diabetic nephropathy ( DN ) .
```

Τέλος η ομαδοποίηση σε ζεύγη γονιδίων-microRNA μπορεί να γίνει με τη συνάρτηση ”article2tuples“ του module ”phrase“:

```
>>> from mg_classifier import phrase
>>> pubid,tups = phrase.article2tuples(cit)
>>> print len(tups)
6
>>> print tups[4]
pubid: 24887517
mir: mir21
genes:
gtoks: smad3
gids: MGI:1201674, ENTREZ:17127, Ensembl:ENSRNOG00000008620, En
sembl:ENSMUSG00000032402, Vega:OTTMUSG00000021500, RGD:3032
, ENTREZ:25631

Moreover , mir-21 over-expression enhanced TGF-β1-induced EMT ( up
regulation of a-SMA and downregulation of E-cadherin ) by directly down-
regulating smad7 / p-smad 7 and indirectly up-regulating smad3 / p-sma
d 3 , accompanied by the decrease of Ccr and the increase of col-IV , FN ,
the content of collagen fibers , RTBM , RTIAW and ACR .
```

Για την αξιολόγηση των αλληλεπιδράσεων πρέπει να χρησιμοποιηθεί κάποιο εκπαιδευμένο μοντέλο, ως εξής:

```
>>> import cPickle as pickle
>>> import os
>>> cl = pickle.load(open(os.path.join(mg_classifier_path, 'data/models'
```

```
, 'abstract_classifier.pickle')))
>>> map(cl.classify, tups)
[False, True, False, False, True, False]
>>> cl.prob_classify(tups[4]).prob(True)
0.99998434491121391
```

Για την ανάκτηση και κατηγοριοποίηση δημοσιεύσεων από την PMC (Ενότητα 2.1.3.2), μπορεί να ακολουθηθεί η αντίστοιχη διαδικασία, με τη διαφορά ότι η ανάκτηση θα γίνει με βάση το κείμενο και στη συνέχεια θα αρχικοποιηθεί μέσω της κλάσης "Publication":

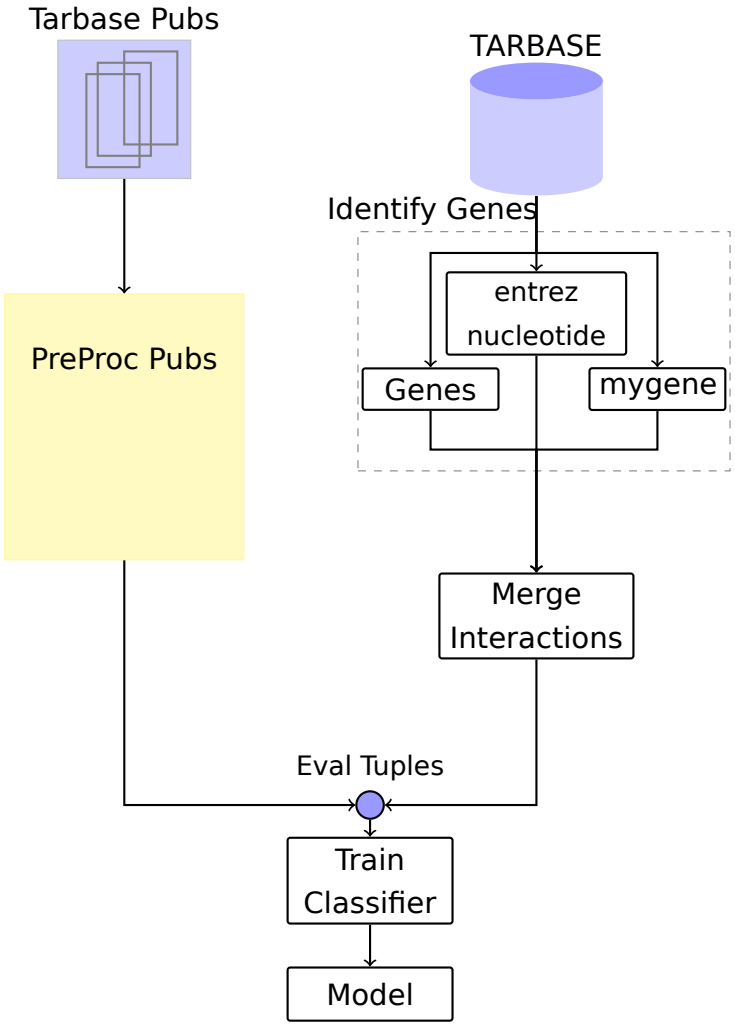
```
>>> from mg_classifier import find_pubs
>>> from mg_classifier.article import Publication
>>> html_text = find_pubs.fetch_html('PMC3510537')
>>> pub = Publication.parse(text = html_text, pubid = 'PMC3510537',
parse = True)
```

Η εύρεση των αναγνωριστικών PMCID και PMID που αντιστοιχούν στην ίδια δημοσίευση, σε περίπτωση που υπάρχουν και τα δύο μπορεί να γίνει με τη συνάρτηση "fetch_ids" του module "find_pubs":

```
>>> from mg_classifier import find_pubs
>>> find_pubs.fetch_ids('PMC3510537')
[{'doi': '10.1128/MCB.01002-12',
'iscurrent': 1,
'islive': 1,
'mid': '',
'msg': '',
'pmcid': 'PMC3510537',
'pmid': '23045399',
'releasedata': '',
'version': 'PMC3510537.1'}]
```

Κεφάλαιο 5

Εκπαίδευση μοντέλου για αναγνώριση συσχετίσεων



Σχήμα 5.1: Εκπαίδευση Classifier

Το Μοντέλο με βάση το οποίο γίνεται η κατηγοριοποίηση των αλληλεπιδράσεων μεταξύ microRNA και γονιδίων (Κεφάλαιο 4), προϋποθέτει την εκπαίδευση του Στατιστικού Μοντέλου Maximum Entropy (Ενότητα 3.1) σε ήδη γνωστά και επικυρωμένα δεδομένα, τα οποία προέρχονται από ειδικές βάσεις δεδομένων.

Στη συνέχεια, από το κείμενο εξάγονται ζεύγη όρων microRNA-γονίδιο (Κεφάλαιο 4) και το κάθε ζεύγος λαμβάνει την τιμή True, αν ανήκει στις επικυρωμένες αλληλεπιδράσεις και False σε διαφορετική περίπτωση.

Με βάση τις δύο κλάσεις True και False γίνεται εκπαίδευση του κατηγοριοποιητή και το αποτέλεσμα είναι ένα μοντέλο που περιλαμβάνει το διάνυσμα με τα χαρακτηριστικά (Ενότητα 3.2) που προέκυψαν και μία τιμή-βάρος για το κάθενα. Η αποτίμηση μίας νέας αλληλεπίδρασης γίνεται με βάση τα χαρακτηριστικά της και τις τιμές βάρους τους.

5.1 Συλλογή δεδομένων εκπαίδευσης

Η εκπαίδευση του κατηγοριοποιητή προϋποθέτει αρχικά την ανάκτηση και μορφοποίηση των επιβεβαιωμένων αλληλεπιδράσεων και στη συνέχεια την ανάκτηση των δημοσιεύσεων στις οποίες αναφέρονται.

Τα κείμενα των δημοσιεύσεων πρέπει να υποστούν την κατάλληλη επεξεργασία, ώστε οι όροι των γονιδίων και των microRNA, να αναγνωριστούν, ώστε να είναι δυνατή η σύγκριση τους με τις επιβεβαιωμένες αλληλεπιδράσεις. Το ίδιο πρέπει να γίνει και για τα δεδομένα από τις βάσεις, οι εγγραφές των οποίων περιλαμβάνουν μόνο ορισμένα στοιχεία για τα γονίδια και τα microRNA. Έτσι, πρέπει να διασταυρωθούν με άλλες βάσεις (Ενότητα 4.2) για τη διευκόλυνση της σύγκρισης.

5.1.1 Επιβεβαιωμένες Αλληλεπιδράσεις

Η εκπαίδευση του κατηγοριοποιητή βασίζεται σε επιβεβαιωμένες από ανθρώπους αλληλεπιδράσεις. Ο μεγαλύτερος όγκος των δεδομένων προέρχεται από το TarBase 6.0 (Ενότητα 2.3), το οποίο περιέχει αλληλεπιδράσεις microRNA-γονιδίου και αναφορά στην αντίστοιχη δημοσίευση, μέσω του PMID, δηλαδή τη βιβλιογραφική αναφορά της δημοσίευσης στο PubMed (Ενότητα 2.1.3.1). Επιπλέον κάποια δεδομένα προέρχονται από το miRTarBase και το miRecords και αντίστοιχα περιλαμβάνουν το PMID, ένα αναγνωριστικό του γονιδίου και το όνομα microRNA. [4, 21, 22]

Δεδομένου ότι το TarBase περιλαμβάνει και στοιχεία από τα miRTarBase και miRecords και τα επιπρόσθετα δεδομένα αφορούν σε ενημερώσεις που έχουν

γίνει στις συγκεκριμένες βάσεις, στο υπόλοιπο κεφάλαιο η αναφορά στο TarBase και TarBase 6.0 θα αντιστοιχεί στο σύνολο των επικυρωμένων αλληλεπιδράσεων που συλλέχθηκαν.

5.1.1.1 Αναγνώριση Όρων

Τα δεδομένα που μας ενδιαφέρουν στη παρούσα διπλωματική από τις βάσεις TarBase, miRTarBase και miRecords είναι μόνο τα αναγνωριστικά των δύο όρων miRNA και γονίδιο.

Η αναγνώριση των miRNA από τα δεδομένα των βάσεων, είναι σχετικά εύκολη διαδικασία. Αυτό οφείλεται στο ότι οι εγγραφές περιέχουν τα ονόματα των miRNA με ομοιόμορφο τρόπο και στο γεγονός ότι τα miRNA γενικά ακολουθούν ενιαίους κανόνες ονοματολογίας.

Η αναγνώριση των γονιδίων απαιτεί περισσότερη επεξεργασία για να γίνει δυνατή η σύγκριση τους με τα αποτελέσματα από την επεξεργασία των κειμένων. Αυτό οφείλεται στις διαφορετικές πηγές από τις οποίες προέρχονται, τις διαφορετικές βάσεις που έχουν καταχωρήσει δεδομένα για το γονιδίωμα διαφορετικών οργανισμών και στην ονοματολογία των γονιδίων και των πρωτεϊνών που για ιστορικούς λόγους δεν είναι ομοιόμορφη.

5.1.1.1.1 Αναγνώριση miRNA

Στις εγγραφές του TarBase (Ενότητα 2.3), παρατηρούνται οι παρακάτω μορφές για τα miRNA:

cel-miR-76-3p
dme-miR-79
ath-miR164a

Αποτελούνται από τρία τμήματα:

(οργανισμός)-(miR-?id)-((3|5)p)

Για την ταυτοποίηση τους με τις αναφορές στο κείμενο συγκρίνεται μόνο το όνομα, δηλαδή η λέξη mir (με μικρούς χαρακτήρες) μαζί με τον αριθμό-ταυτότητα του microRNA, ενώ αποκόπτονται τα προθέματα που χαρακτηρίζουν τον οργανισμό και τα επιθέματα που δίνουν πληροφορίες για το στάδιο της ωρίμανσης του microRNA ή άλλες πληροφορίες για τις ιδιότητες του (πχ -3p, -5p).

Αυτή η απλοποίηση των miRNA, δίνει ικανοποιητικά αποτελέσματα στην εύρεση όρων, αλλά μπορεί να οδηγήσει σε λάθη ταυτοποιώντας διαφορετικούς όρους.

5.1.1.1.2 Αναγνώριση Γονιδίων

Στο TarBase 6.0 (Ενότητα 2.3), για κάθε αλληλεπίδραση δίνεται ένα σύνολο από πληροφορίες για τα γονίδια. Τα στοιχεία που χρησιμοποιούνται για την αναγνώριση τους είναι τα gene_symbol, gene_id και gene_trid.

Το gene_symbol είναι το επίσημο σύμβολο του γονιδίου. Το gene_id είναι το αναγνωριστικό του συγκεκριμένου γονιδίου από κάποια βάση και το gene_trid είναι το αναγνωριστικό του συγκεκριμένου μεταγραφικού προϊόντος για το συγκεκριμένο στοιχείο.

Επειδή τα δεδομένα αυτά προέρχονται από διαφορετικές βάσεις, τα αναγνωριστικά που περιλαμβάνονται δεν είναι ίδια. Έτσι, υπάρχουν αναγνωριστικά από την Ensembl (Ενότητα 2.2), υπάρχουν εγγραφές με αναγνωριστικά RefSeq (Ενότητα 2.1.5) και υπάρχουν και αναγνωριστικά από άλλες βάσεις όπως η FlyBase (Ενότητα 2.4.3) και η TAIR (Ενότητα 2.4.5). Τέλος υπάρχουν και εγγραφές που περιέχουν μόνο το σύμβολο και τον οργανισμό στον οποίο ανήκει το γονίδιο. Η αντιμετώπιση του κάθε αναγνωριστικού μεμονωμένα δεν είναι εύκολο να γίνει γιατί δεν παρέχονται από όλες τις παραπάνω βάσεις στοιχεία κατάλληλα, όπως αναφορές στις βάσεις που χρησιμοποιήθηκαν για την εξαγωγή των γονιδίων από τα κείμενα (HGNC, ENREZ Gene, MGI, RGD και CGNC (Ενότητα 2.4)). Επιπλέον, δεν είναι άμεση η αναγνώριση της προέλευσης του κάθε αναγνωριστικού.

Για τους λόγους αυτούς, χρησιμοποιήθηκε ένα εργαλείο, το mygene (Ενότητα 3.11.3) [23], το οποίο επιστρέφει αποτελέσματα σε ερωτήματα με βάση το αναγνωριστικό ή το σύμβολο από τις περισσότερες από τις παραπάνω βάσεις και επιπλέον, επιστρέφει τα αντίστοιχα αναγνωριστικά από όλες τις βάσεις αυτές, όπως και συνώνυμα σύμβολα και ονόματα.

Παράλληλα, για τα αναγνωριστικά από τη RefSeq (Ενότητα 2.1.5) έγινε και χρήση των προγραμματιστικών εργαλείων της NCBI, eUtils (Ενότητα 2.1.1.1).

5.1.2 Δημοσιεύσεις

Οι δημοσιεύσεις που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων προέρχονται από τα δεδομένα που παρέχουν οι βάσεις, TarBase 6.0, mirTarBase και miRecords και αναφέρονται με το PMID τους. Το πλήθος των άρθρων είναι 1851 και από αυτά τα 1167 βρίσκονται στην PMC.

Επομένως για την πλειοψηφία των άρθρων, 1829 , είναι διαθέσιμα η περίληψη και ο τίτλος, ενώ για τα 1167 υπάρχει διαθέσιμη και είναι σχετικά εύκολη η ανάκτηση του κειμένου από την NCBI.

Έτσι δημιουργήθηκαν δύο ομάδες δεδομένων, τα κείμενα από την περίληψη και τον τίτλο για τα 1829 άρθρα και τα κείμενα από ολόκληρο το σώμα της δημοσίευσης για τα 1167 άρθρα από την PMC.

Η επεξεργασία των κειμένων περιγράφεται στο Κεφάλαιο 4.

5.1.2.1 Βιβλιογραφικές Αναφορές

Οι βιβλιογραφικές αναφορές προέρχονται από την PubMed (Ενότητα 2.1.3.1) και περιέχουν σε όλες τις περιπτώσεις τον τίτλο, ενώ στις περισσότερες περιλαμβάνεται και η περίληψη (abstract).

Ο τίτλος και η περίληψη δεν περιγράφουν με λεπτομέρεια τη μεθοδολογία και τα αποτελέσματα των πειραμάτων που έχουν πραγματοποιηθεί, με αποτέλεσμα σε πολλές περιπτώσεις να μην υπάρχουν αναφορές στα γονίδια και τα microRNA, τα οποία μελετά η δημοσίευση και των οποίων καταγράφονται αλληλεπιδράσεις στη βάση.

Έτσι, το πλήθος των συσχετίσεων που εντοπίζονται στα κείμενα αναφοράς του TarBase είναι μικρότερο από το πλήθος αυτών που καταγράφει. Για το λόγο αυτό, το ποσοστό των αλληλεπιδράσεων που εντοπίζονται κατά την αναγνώριση τους, αυξάνει καθώς μειώνεται το συνολικό πλήθος των αλληλεπιδράσεων που περιέχει μια δημοσίευση (Πίνακας Α′.2).

Το πλήθος των ζευγών που δεν είναι πραγματικές αλληλεπιδράσεις είναι επίσης αρκετά μεγάλο (Πίνακας Α′.3).

5.1.2.2 Πλήρεις Δημοσιεύσεις

Τα πλήρη κείμενα προέρχονται από δημοσιεύσεις της PMC (Ενότητα 2.1.3.2) και περιέχουν τον τίτλο, την περίληψη, αλλά και το πλήρες κείμενο της δημοσίευσης. Στο σώμα του κειμένου, περιγράφονται οι μέθοδοι και τα αποτελέσματα, όπως και πειραματικά στοιχεία για τις επαληθευμένες αλληλεπιδράσεις. Έτσι στα πλήρη κείμενα (άρθρα) εντοπίζεται μεγάλο ποσοστό των επιβεβαιωμένων αλληλεπιδράσεων. Όπως και στη περίπτωση των βιβλιογραφικών αναφορών, το ποσοστό αυτό αυξάνει καθώς μειώνεται το πλήθος των συνολικών αλληλεπιδράσεων του κειμένου (Πίνακας Α′.7).

Παράλληλα, μεγάλος αριθμός αλληλεπιδράσεων που εντοπίζονται δεν αποτελεί πραγματικές αλληλεπιδράσεις microRNA–γονιδίου-στόχου. (Πίνακας Α′.8).

Σε μεγάλο ποσοστό η πολύ χαμηλή ακρίβεια, οφείλεται σε όρους που αναγνωρίζονται ως γονίδια.

Από αυτά, υπάρχουν κάποια που είναι γονίδια, αλλά έχουν διαφορετικό ρόλο όπως δείκτες για την επιβεβαίωση κάποιου πειράματος (πχ. γονίδιο-δείκτης), ή έχουν διαφορετικό ρόλο σε κάποια διαδικασία που περιγράφεται. Μία ακόμη περίπτωση, είναι οι αναφορές σε γονίδια που σχετίζονται με το γονίδιο-στόχο ή με τη ασθένεια και το φαινότυπο που μελετάται. Στις περιπτώσεις αυτές το microRNA μπορεί να ρυθμίζει τελικά την έκφραση του γονιδίου, αλλά όχι άμεσα.

Σε άλλες περιπτώσεις αλληλεπιδράσεων, τα γονίδια είναι πραγματικοί στόχοι του αντίστοιχου microRNA, το γεγονός όμως αυτό επαληθεύεται σε προηγούμενα αποτελέσματα ή πειράματα και η αναφορά τους στο κείμενο γίνεται για την επεξήγηση κάποιας μεθοδολογίας ή για ιστορικούς λόγους.

Επιπλέον, υπάρχουν περιπτώσεις τυπογραφικών σφαλμάτων στις βάσεις δεδομένων, λάθη στην αναγνώριση των αναγνωριστικών που οδηγούν στην κατάταξη κάποιων πραγματικών αλληλεπιδράσεων στην κλάση False. Επιπρόσθετα, κάποιες φορές γίνεται αναγνώριση τμήματος του όρου με αποτέλεσμα να μην ταυτίζεται με τον πραγματικό όρο, αλλά να εντοπίζεται στο κείμενο ως διαφορετικός που επίσης κατατάσσεται στην κλάση False.

Σε πολλές περιπτώσεις, όμως, οι όροι που αναγνωρίζονται ως γονίδια αντιστοιχούν σε διαφορετικές οντότητες, όπως ασθένειες, ομάδες κυττάρων, χημικές ουσίες, προγραμματιστικά εργαλεία, ή χημικές μεθόδους. Το πρόβλημα αυτό οφείλεται στην πολυσημία των συντομεύσεων που χρησιμοποιούνται ή ορίζονται στα βιοϊατρικά κείμενα (Ενότητα 3.8). Σε άλλες περιπτώσεις είναι κοινές αγγλικές λέξεις που τυχαίνει να ταυτίζονται με το σύμβολο κάποιου γονιδίου. Ένα παράδειγμα είναι το γονίδιο ”OF“ που ταυτίζεται με την αγγλική πρόθεση ”of“ η οποία εμφανίζεται πολύ συχνά σε αγγλικά κείμενα.

Έτσι, το πλήθος των αλληλεπιδράσεων που εξάγονται από το κείμενο και δεν είναι επαληθευμένες είναι πολύ μεγάλο. Αυτό δημιουργεί μεγάλη ανισότητα στις δύο κλάσεις κατά την διαδικασία της εκπαίδευσης και το μοντέλο αποδίδει άνισες πιθανότητες σε κάθε κλάση.

5.2 Εκπαίδευση Μοντέλου

Μετά την συλλογή των δεδομένων των δημοσιεύσεων, γίνεται επεξεργασία τους με τη διαδικασία που περιγράφεται στην Ενότητα 4. Τελικά, από κάθε δημοσίευση εξάγεται ένα σύνολο δεδομένων της μορφής (microRNA,γονίδιο) (Ενότητα 4.4.1.4).

Στη συνέχεια, γίνεται κατάταξη των δεδομένων στις κλάσεις True και False, εκπαίδευση του μοντέλου και έλεγχος.

5.2.1 Επιλογή Πραγματικών Συσχετίσεων

Για την κατάταξη των δεδομένων σε κλάσεις, χρησιμοποιούνται τα δεδομένα των βάσεων δεδομένων (Ενότητα 2.3). Το κάθε ζεύγος (microRNA, γονίδιο) σε κάθε δημοσίευση, κατατάσσεται στην κατηγορία True αν εμφανίζεται στις εγγραφές του TarBase και False στην αντίθετη περίπτωση.

Η αποτίμηση της κάθε τριπλέτας (PMID,microRNA,γονίδιο) περιλαμβάνει τη σύγκριση και των τριών όρων με τις εγγραφές του TarBase.

PMID: Ο έλεγχος για το PMID είναι άμεσος. Η ανάκτηση των δημοσιεύσεων έχει προκύψει μέσω του κωδικού αυτού. Στις σπάνιες περιπτώσεις που έχει γίνει κάποια αλλαγή στο κωδικό της δημοσίευσης, ο προηγούμενος κωδικός δεν είναι έγκυρος και επομένως η δημοσίευση αυτή δεν έχει ανακτηθεί.

microRNA: Ο έλεγχος των microRNA γίνεται με βάση μια προεπεξεργασία που πραγματοποιείται τόσο στους όρους microRNA του TarBase όσο και στους όρους που εμφανίζονται στο κείμενο μιας δημοσίευσης. Η προεπεξεργασία αυτή περιλαμβάνει απομάκρυνση συμβόλων όπως η παύλα και ειδικοί χαρακτήρες, και των προθεμάτων και επιθεμάτων του microRNA. Με τον τρόπο αυτό επιτυγχάνεται όσο το δυνατόν καλύτερη ταύτιση των όρων που αναφέρονται στο ίδιο βιομόριο και προέρχονται από διαφορετικές πηγές.

Γονίδιο: Στην περίπτωση των γονιδίων, για τη σύγκριση χρησιμοποιούνται κυρίως τα αναγνωριστικά από τις βάσεις δεδομένων και κατά δεύτερο λόγο τα σύμβολα, για τα οποία γίνεται σύγκριση της λέξης αφού αφαιρεθούν ειδικοί χαρακτήρες όπως παύλες, κενά διαστήματα και άλλα.

5.2.2 Σύνολο Εκπαίδευσης και Ελέγχου

Για την αξιολόγηση ενός μοντέλου, πρέπει να γίνει εκπαίδευση του κατηγοριοποιητή σε ένα υποσύνολο των δεδομένων που ονομάζεται Σύνολο Εκπαίδευσης (Training Set) και έλεγχος σε ένα ξένο με αυτό υποσύνολο, που ονομάζεται Σύνολο Ελέγχου (Test Set).

5.2.3 Φιλτράρισμα Ζευγών

Στα πλήρη κείμενα των δημοσιεύσεων, υπάρχουν πολλές περιπτώσεις αναφορών σε γονίδια και microRNA που δεν αποτελούν αλληλεπιδράσεις, όπως αναφέρθηκε και στην Ενότητα 5.1.2.2.

Για το λόγο αυτό, υλοποιήθηκαν τα ακόλουθα φίλτρα που βελτιώνουν τη δυνατότητα του μοντέλου να αναγνωρίζει πραγματικές αλληλεπιδράσεις.

Φίλτρο Πλήθους Φράσεων

Το φίλτρο πλήθους φράσεων, απορρίπτει τα ζεύγη microRNA-γονιδίου, τα οποία περιέχουν μικρότερο από n αριθμό φράσεων.

Ειδικά στις πλήρεις δημοσιεύσεις, όπως φαίνεται στον Πίνακα Α'.8, η ακρίβεια στην αναγνώριση αλληλεπιδράσεων είναι πολύ μικρή, στη χειρότερη περίπτωση 8%. Αντίθετα, εφαρμόζοντας το φίλτρο, η ακρίβεια αυξάνει και η χειρότερη περίπτωση γίνεται 13%.

Με τον τρόπο αυτό μειώνεται αρκετά η η δυσαναλογία False/True.

Φίλτρο Εξωτερικών Αναφορών

Το φίλτρο αυτό εφαρμόζεται μόνο όταν το κείμενο είναι διαθέσιμο σε μορφή HTML. Στη περίπτωση αυτή, οι εξωτερικές αναφορές μπορούν να αναγνωριστούν ευκολότερα και οι προτάσεις που περιέχουν τουλάχιστον μία εξωτερική αναφορά απορρίπτονται.

Τα ζεύγη που προέρχονται από τις προτάσεις αυτές, συχνά αναφέρονται σε αποτελέσματα από διαφορετικά κείμενα και συνεπώς, κατατάσσονται με βάση την πληροφορία του TarBase στην κατηγορία False. Αυτές οι περιπτώσεις, επηρεάζουν αρνητικά τον αλγόριθμο, γιατί εμφανίζουν περίπου την ίδια σύνταξη και συμφραζόμενα με τις πραγματικές αλληλεπιδράσεις.

Φίλτρα Γονιδίων Τα φίλτρα αυτά χρησιμοποιούνται για την βελτίωση στην απόδοση των όρων που αναφέρονται σε γονίδια. Στον πίνακα που ακολουθεί, φαίνονται κάποιες από τις ενέργειες που έγιναν για την επίλυση των προβλημάτων που σχετίζονται με την αναγνώριση όρων-γονιδίων (Ενότητα 5.1.2.2):

Όροι	Προβλημα	Επίλυση
Γονίδια	Άλλες Λειτουργίες (γονίδιο-δείκτης)	Χαρακτηριστικά*
	Τμήμα όρου	-
	Αναφορά σε άλλη δημοσίευση	Φίλτρο xref
Όχι Γονίδια	Ορισμοί Συντομεύσεων	Εξαγωγή ορισμών
	Κοινές Αγγλικές Λέξεις	Χρήση Λεξικού
	Συντομογραφίες που δεν ορίζονται	Χαρακτηριστικά*

5.3 Αξιολόγηση Μοντέλου

Υλοποιήθηκαν δύο μοντέλα, ένα για την αναγνώριση των αλληλεπιδράσεων από πλήρη κείμενα και το δεύτερο για την αναγνώριση των αλληλεπιδράσεων μόνο από το κείμενο των περιλήψεων. Και στις δύο περιπτώσεις χρησιμοποιούνται τρεις δείκτες, η ευαισθησία (recall), η ακρίβεια (precision) και το F1-score ως μέτρα της αξιοπιστίας του μοντέλου.

Οι τρεις δείκτες ορίζονται ως εξής:

$$recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$F1 - score = \frac{2 \cdot recall \cdot precision}{recall + precision}$$

Οι όροι TruePositive, TrueNegative, FalsePositive και FalseNegative αναφέρονται στα δεδομένα ελέγχου ενός δυαδικού μοντέλου, ως εξής:

TruePositive: Τα αποτελέσματα που ανήκουν στην κατηγορία True και ο κατηγοριοποιητής τους αποδίδει την τιμή True.

TrueNegative: Τα αποτελέσματα που ανήκουν στην κατηγορία False και ο κατηγοριοποιητής τους αποδίδει την τιμή False.

FalsePositive: Τα αποτελέσματα που ανήκουν στην κατηγορία False και ο κατηγοριοποιητής τους αποδίδει την τιμή True.

FalseNegative: Τα αποτελέσματα που ανήκουν στην κατηγορία True και ο κατηγοριοποιητής τους αποδίδει την τιμή False.

Αυτά τα στοιχεία, υπολογίζονται αρχικά για τα αποτελέσματα του κατηγοριοποιητή με βάση τα δεδομένα ελέγχου και στη συνέχεια τα αποτελέσματα αυτά σε σύγκριση με τα δεδομένα του TarBase, που συνδυάζουν την συνολική απόδοση του μοντέλου τόσο στο επίπεδο του κατηγοριοποιητή όσο και στο επίπεδο της αξιοπιστίας στην αναγνώριση των όρων.

Τέλος, μία παράμετρος από την οποία εξαρτώνται τα τελικά αποτελέσματα είναι οι περιπτώσεις στις οποίες τα αποτελέσματα για τις αλληλεπιδράσεις δεν αναφέρονται στο κείμενο. Αντίθετα, αναφέρονται σε διαγράμματα, εικόνες, πίνακες ή συμπληρωματικά αρχεία. Τέτοιες περιπτώσεις αυξάνονται καθώς αυξάνεται το πλήθος των αλληλεπιδράσεων που περιέχει μία δημοσίευση (Πίνακες Α′.2 και Α′.7).

5.3.1 Αξιολόγηση Μοντέλου Άρθρου

Χωρίς Φίλτρο Φράσεων

Στον Πίνακα Β′.5 φαίνονται τα αποτελέσματα του Κατηγοριοποιητή χωρίς φίλτρο στον αριθμό των φράσεων για 10 επαναλήψεις του αλγορίθμου με σύνολο εκπαίδευσης 85% και σύνολο ελέγχου το υπόλοιπο 15% των δημοσιεύσεων. Τα

αποτελέσματα για την ευαισθησία, την ακρίβεια και το F1-score προκύπτουν ως μέσος όρος των 10 επαναλήψεων.

Τα αποτελέσματα για τα δεδομένα που προέρχονται από το TarBase, αναφέρονται στη σχέση του αριθμού των αλληλεπιδράσεων που βρέθηκαν από τον κατηγοριοποιητή, σε σχέση με όσα πραγματικά καταγράφονται στο TarBase. Προκύπτουν από τους Πίνακες Β′.5 και Α′.7.

Η ακρίβεια είναι η ακρίβεια του κατηγοριοποιητή, δηλαδή τα δεδομένα του Πίνακα Β′.5 και η ευαισθησία προκύπτει ως το γινόμενο της ευαισθησίας του κατηγοριοποιητή πολλαπλασιασμένης με την ευαισθησία στην αναγνώριση των όρων από τον Πίνακα Α′.7, δηλαδή ως το αποτέλεσμα των δύο σταδίων που μεσολαβούν.

Στον Πίνακα Β′.6 παρουσιάζονται τα αποτελέσματα.

Αριθμός Φράσεων > 1

Με τη χρήση του φίλτρου στον αριθμό των φράσεων, για να γίνει η εκπαίδευση, ο αριθμός των ψευδών αλληλεπιδράσεων μειώνεται και το αποτέλεσμα της κατηγοριοποίησης βελτιώνεται, αφού αλλάζει η αναλογία False/True.

Τα αποτελέσματα προκύπτουν με τον ίδιο τρόπο όπως στην περίπτωση, όπου δεν γίνεται χρήση φίλτρου.

Τα αποτελέσματα του κατηγοριοποιητή παρουσιάζονται στον Πίνακα Β′.7, ενώ τα συνολικά αποτελέσματα σε σχέση με τα δεδομένα που καταγράφει το TarBase στον Πίνακα Β′.8. Τα δεδομένα από την αναγνώριση των αλληλεπιδράσεων προέρχονται από τον Πίνακα Α′.9.

5.3.2 Αξιολόγηση Μοντέλου Περιλήψεων

Αντίστοιχα με την περίπτωση της κατηγοριοποίησης του μοντέλου Άρθρων, το μοντέλο Περιλήψεων αξιολογείται με τους ίδιους δείκτες: ευαισθησία, ακρίβεια και F1-score.

Χωρίς Φίλτρο Φράσεων

Στον Πίνακα Β′.1 φαίνονται τα αποτελέσματα του Κατηγοριοποιητή χωρίς φίλτρο στον αριθμό των φράσεων για 10 επαναλήψεις του αλγορίθμου. Το σύνολο εκπαίδευσης είναι 85% και το σύνολο ελέγχου το υπόλοιπο 15% των περιλήψεων. Τα αποτελέσματα τόσο για την ευαισθησία, την ακρίβεια και το F1-score προκύπτουν ως μέσος όρος των 10 επαναλήψεων.

Τα αποτελέσματα συνολικά για τα δεδομένα που προέρχονται από το TarBase για τις αλληλεπιδράσεις συγκεκριμένων δημοσιεύσεων προκύπτουν από τους Πίνακες Β′.1 και Α′.2.

Όπως και στην περίπτωση του μοντέλου άρθρων, η ακρίβεια ισούται με την ακρίβεια του κατηγοριοποιητή, δηλαδή τα δεδομένα του Πίνακα Β′.1 και η ευαισθησία προκύπτει ως το γινόμενο της ευαισθησίας του κατηγοριοποιητή πολλαπλασιασμένης με την ευαισθησία στην αναγνώριση των όρων από τον Πίνακα Α′.2. Όπως είναι φανερό, τα συνολικά αποτελέσματα είναι μικρά, αφού το κείμενο της περίληψης πολύ συχνά δεν περιέχει τα ονόματα των γονιδίων και των microRNA που μελετώνται. Παράλληλα εμφανίζεται μεγάλο ποσοστό όρων που αναγνωρίζονται μόνο σε ένα μέρος τους, αφού στην περίληψη συχνά γίνεται περιγραφική περιγραφή με το πλήρες όνομα των γονιδίων που είναι δυσκολότερο να αναγνωριστεί.

Στον Πίνακα Β′.2 παρουσιάζονται τα τελικά αποτελέσματα.

Αριθμός Φράσεων > 1

Αντίστοιχα με το μοντέλου των πλήρων κειμένων, η χρήση φίλτρου φράσεων βελτιώνει την απόδοση του μοντέλου.

Τα αποτελέσματα του κατηγοριοποιητή παρουσιάζονται στον Πίνακα Β′.3 και τα συνολικά αποτελέσματα για το TarBase στον Πίνακα Β′.4.

Τα αποτελέσματα αυτά προέρχονται από τους Πίνακες Β′.3 και Α′.4.

5.3.3 Σχολιασμός

Από τα πειράματα που εκτελέστηκαν, προέκυψαν διάφορα συμπεράσματα. Παρακάτω ακολουθεί σύντομος σχολιασμός των αποτελεσμάτων και αναφορά στις περιπτώσεις που παρουσιάζουν την βέλτιστη επίδοση.

5.3.3.1 Αναγνώριση Όρων

Στους Πίνακες Α′.2 και Α′.7 υπάρχει σημαντική μείωση της ευαισθησίας των μεθόδων εξαγωγής αλληλεπιδράσεων σε σχέση με τα δεδομένα που είναι καταχωρημένα στο TarBase, καθώς αυξάνει το πλήθος των αλληλεπιδράσεων που περιέχει το άρθρο. Ο λόγος για αυτό είναι η πιθανή απουσία των όρων αυτών από το κείμενο, κάτι αναμενόμενα για δημοσιεύσεις που περιλαμβάνουν 10, 20 ή 50 αλληλεπιδράσεις, πλήθος το οποίο είναι δύσκολο να περιγραφεί σε ένα κείμενο.

Παρόλα αυτά, κάποιες από τις δημοσιεύσεις αυτές, που γενικά έχουν καταχωρήσει τα αποτελέσματά τους σε συμπληρωματικό υλικό ή πίνακες, μπορεί να περιγράφουν κάποιο περιορισμένο αριθμό αλληλεπιδράσεων και πειραμάτων που εκτελέστηκαν στο σώμα του κειμένου. Έτσι, με την απόρριψη των δημοσιεύσεων αυτών χάνεται κάποια από την διαθέσιμη πληροφορία σχετικά τις

δημοσιεύσεις και τους τρόπους περιγραφής της σχέσης των microRNA με τα γονίδια-στόχους τους.

5.3.3.2 Βιβλιογραφικές αναφορές

Τα αποτελέσματα των πειραμάτων κατηγοριοποίησης για τις βιβλιογραφικές αναφορές περιλαμβάνονται στο Παράρτημα Β′.

Παρατηρείται μείωση της απόδοσης του κατηγοριοποιητή στις περιπτώσεις που συμπεριλαμβάνονται άρθρα που περιέχουν πολλές αλληλεπιδράσεις. Περισσότερο όμως μειώνεται η ευαισθησία σε σχέση με τα δεδομένα του TarBase κάτι που οφείλεται στην απουσία αναφορών στις αλληλεπιδράσεις αυτές, όπως περιγράφεται στην προηγούμενη υποενότητα.

Παράλληλα, με τη χρήση του φίλτρου για τον αριθμό των φράσεων η ακρίβεια και η ευαισθησία του κατηγοριοποιητή βελτιώνονται. Όμως σε σχέση με τα δεδομένα του TarBase η ευαισθησία μειώνεται πολύ με αποτέλεσμα η συνολική αποτίμηση να είναι λίγο μικρότερη από την περίπτωση που δεν χρησιμοποιηθεί φίλτρο.

Αυτό το αποτέλεσμα οφείλεται στο μικρό μέγεθος της περίληψης και στην συνοπτική παρουσίαση των αποτελεσμάτων της κάθε εργασίας. Έτσι, συχνά τα αποτελέσματα αναφέρονται μόνο σε μία πρόταση ή φράση.

Τελικά, η καλύτερη απόδοση παρατηρείται στην περίπτωση των άρθρων μόνο μία αλληλεπίδραση να εξάγεται από το κείμενο (μέγιστος αριθμός αλληλεπιδράσεων < 2). Στην περίπτωση αυτή έχουμε (βλ. Πίνακα Β′.4):

Μεγ. # Αλληλ.	Ακρίβεια	Ευαισθησία	F1-score
2	0.676	0.579	0.623

5.3.4 Πλήρη Κείμενα Δημοσιεύσεων

Τα αποτελέσματα των πειραμάτων που έγιναν περιλαμβάνονται στο Παράρτημα Β′.

Αντίθετα με τις βιβλιογραφικές αναφορές, στην περίπτωση των πλήρων κειμένων η συχνότητα εμφάνισης κάποιου ζεύγους microRNA-γονιδίου είναι σημαντικός παράγοντας. Αφαιρώντας τα ζεύγη που περιλαμβάνουν μία φράση μόνο, παρατηρείται βελτίωση της απόδοσης τόσο του κατηγοριοποιητή όσο και των συνολικών αποτελεσμάτων για τις αλληλεπιδράσεις που είναι καταχωρημένες στο TarBase.

Η καλύτερη απόδοση παρατηρείται στην περίπτωση που έχουν μόνο μία αλληλεπίδραση να εξάγεται από το κείμενο (μέγιστος αριθμός αλληλεπιδράσεων <

2) και αριθμό φράσεων μεγαλύτερο από 1. Στην περίπτωση αυτή έχουμε (βλ. Πίνακα Β'.8):

Μεγ. # Αλληλ.	Ακρίβεια	Ευαισθησία	F1-score
2	0.769	0.691	0.726

5.4 Δομή κώδικα και χρήση

5.4.1 Δομή Κώδικα

Η διαδικασία της κατηγοριοποίησης περιγράφεται στο Κεφάλαιο 4.

Ο κατηγοριοποιητής είναι υποκλάση της κλάσης nltk.classify.ClassifierI και δέχεται ως είσοδο τα ζεύγη που αποτελούν τα δεδομένα εκπαίδευσης στην μορφή MGTuple (Ενότητα 4.4.1.4), τη συνάρτηση που επιστρέφει τα χαρακτηριστικά για κάθε ζεύγος, ένα λεξικό με τη συχνότητα εμφάνισης κάθε λέξης στα δεδομένα εκπαίδευσης, τη μικρότερη συχνότητα που μπορεί να εμφανίζει μία λέξη ώστε να θεωρηθεί σπάνια, τον αλγόριθμο εύρεσης του διανύσματος βάρους για την προσαρμογή στο σύνολο εκπαίδευσης, και τη συχνότητα αποκοπής των χαρακτηριστικών.

Το αντικείμενο που επιστρέφεται περιέχει έναν κατηγοριοποιητή και συναρτήσεις για κατηγοριοποίηση, μαζική κατηγοριοποίηση και βοηθητικές συναρτήσεις.

Η κλάση BinMGMaxentClassifier και οι παράμετροι αρχικοποίησης, τα πεδία και οι μέθοδοι είναι:

Παράμετροι:

tups: Τα δεδομένα εκπαίδευσης. Είναι μία δομή MGTupleD και περιλαμβάνει εκτός από τις υπόλοιπες τιμές και την τιμή 'eval' που είναι True, αν το συγκεκριμένο ζεύγος θεωρείται πραγματική Αλληλεπίδραση και False σε διαφορετική περίπτωση.

features: Η συνάρτηση που επιστρέφει τα χαρακτηριστικά (features) που χρησιμοποιούνται από τον κατηγοριοποιητή.

vocabulary: Το λεξικό που περιλαμβάνει τις λέξεις που εμφανίζονται στα κείμενα από τα οποία προέρχονται τα tups. Οι λέξεις είναι τα θέματα (stems) των λέξεων και οι καταλήξεις έχουν περαιτέρω απομακρυνθεί με τον Porter Stemmer (Ενότητα 3.7.2). Προεπιλεγμένη τιμή είναι None, δηλαδή χωρίς λεξικό.

min_value: Η ελάχιστη συχνότητα εμφάνισης μίας λέξεις στο λεξικό για να μην λαμβάνεται υπόψη. Προεπιλεγμένη τιμή είναι 4.

algo: Ο αλγόριθμος για την εκπαίδευση, προεπιλεγμένη τιμή είναι ο αλγόριθμος ”MEGAM“ και είναι ο αλγόριθμος με τον οποίο έχει γίνει η εκπαίδευση των μοντέλων.

count_cutoff: Η τιμή αποκοπής για τα χαρακτηριστικά (features) που περνούν στον κατηγοριοποιητή. Τα χαρακτηριστικά που εμφανίζονται λιγότερες από ”count_cutoff“ φορές δεν λαμβάνονται υπόψη. Προεπιλεγμένη τιμή είναι 2

Μέθοδοι:

features: Η μέθοδος αυτή καλεί την συνάρτηση features με την οποία έχει αρχικοποιηθεί η κλάση.

label: Επιστρέφει τις ετικέτες. Οι τιμές που επιστρέφονται είναι True και False.

classify: Η συνάρτηση δέχεται ως όρισμα ένα στοιχείο MGTupleD (Ενότητα 4.4.1.4) και επιστρέφει την τιμή που δίνει ο κατηγοριοποιητής που μπορεί να είναι True ή False.

batch_classify: Δέχεται ως όρισμα μία λίστα από στοιχεία MGTupleD (Ενότητα 4.4.1.4) και επιστρέφει μία λίστα με τα αποτελέσματα του κατηγοριοποιητή.

prob_classify: Λειτουργεί όπως η συνάρτηση classify, αλλά επιστρέφει τις τιμές της πιθανότητας που δίνει το μοντέλο για κάθε κλάση. Οι τιμές αυτές είναι πραγματικοί αριθμοί και ανήκουν στο σύνολο [0-1].

batch_prob_classify: Εφαρμόζει την prob_classify σε μία λίστα από δεδομένα (MGTupleD).

explain: Δέχεται ως είσοδο ένα MGTupleD και τυπώνει τα πιο σημαντικά χαρακτηριστικά που οδηγούν στο αποτέλεσμα του κατηγοριοποιητή.

get_features: Δέχεται ως είσοδο ένα στοιχείο MGTupleD και επιστρέφει τα χαρακτηριστικά.

evaluate: Δέχεται ως είσοδο μία λίστα από στοιχεία MGTupleD και επιστρέφει την τριάδα δεικτών (precision, recall, f_measure)

5.4.2 Χρήση

Η εκπαίδευση του κατηγοριοποιητή απαιτεί την επεξεργασία των κειμένων, την απόδοση τιμών True η False με βάση επικυρωμένα δεδομένα, τον ορισμό της συνάρτησης απόδοσης χαρακτηριστικών και τέλος την εκπαίδευση του στατιστικού μοντέλου στο σύνολο εκπαίδευσης.

Τα δεδομένα των δημοσιεύσεων, τα οποία χρησιμοποιούνται για την εκπαίδευση του μοντέλου πρέπει να μετασχηματιστούν στην μορφή MGTupleD (Ενότητα 4.4.1.4).

Παρακάτω ακολουθεί ένα παράδειγμα εκπαίδευσης ενός μοντέλου που βασίζεται στα δεδομένα του TarBase.

Αρχικά φορτώνονται τα δεδομένα που έχουν ήδη υποστεί επεξεργασία (εναλλακτικά, ακολουθείται η διαδικασία της ενότητας 4.5 για την επεξεργασία.)

```
>>> import gzip, os
>>> import cPickle as pickle
>>> data = pickle.load(gzip.open(os.path.join(mg_classifier_path,'data/
pubs_data/html_parse_part1.2.3.pickle.gz'))
```

Στη συνέχεια, μετασχηματίζονται σε "Article" και "MGTupleD", όπως περιγράφηκε και στην ενότητα 4.5. Έπειτα, γίνεται η αξιολόγηση της κάθε τριπλέτας (pubid,miRNA,γονίδιο) με βάση τα δεδομένα του TarBase. Οι απαραίτητες συναρτήσεις περιέχονται στο module db_eval_tuples:

```
>>> from mg_classifier.article import Publication
>>> from mg_classifier.phrase import article2tuple
>>> from mg_classifier.db_eval_tuples import load_tarbase, eval_mgtupl
es
>>> load_tarbase()
>>> all_tups = []
>>> for pubid in data:
>>>     pub = Publication(data = data[pubid], is_parsed = True)
>>>     _,tups = article2tuples(pub, min_no_phrases = 2, no_xref = True)
>>>     eval_mgtuples(pubid,tups)
>>>     all_tups.append(tups)
```

Στη μεταβλητή "all_tups" βρίσκεται όλες οι τριπλέτες που περιλαμβάνουν ένα πεδίο 'eval' το οποίο είναι True ή False ανάλογα με την αποτίμηση από το TarBase.

Στη συνέχεια φορτώνεται το module "classifier" της βιβλιοθήκης που περιλαμβάνει την κλάση BinMGMaxentClassifier και γίνεται η εκπαίδευση του κατηγοριοποιητή.

```
>>> from mg_classifier import classifier
>>> cl = BinMGMaxentClassifier(all_tups,features=features_mg2, vocab
ulary=vocabulary, min_value=4, count_cutoff=2)
```

Το λεξικό ”vocabulary“ των λέξεων που περιέχονται στις δημοσιεύσεις έχει δημιουργηθεί ως εξής:

```
>>> from mg_classifier import create_vocabulary
>>> voc = create_vocabulary(data)
>>> vocabulary = Counter()
>>> for art in data:
>>>     vocabulary.update(voc[art])
```

Η αξιολόγηση του μοντέλου πρέπει να γίνει σε διαφορετικά δεδομένα, ”data2“, για τα οποία ακολουθείται η ίδια διαδικασία που περιγράφεται παραπάνω μέχρι την διασταύρωση των δεδομένων από το TarBase. Η αποτίμηση του μοντέλου γίνεται με την μέθοδο ”evaluate“ , η οποία επιστρέφει τους δείκτες recall, precision και f_measure. Η διαδικασία είναι η παρακάτω:

```
>>> data2 = pickle.load(gzip.open(os.path.join(mg_classifier_path,'data
/pubs_data/html_parse_partr2.2.3.pickle.gz'))
>>> all_tups2 = []
>>> for pubid in data2:
>>>     pub = Publication(data = data2[pubid], is_parsed = True)
>>>     _,tups = article2tuples(pub, min_no_phrases = 2, no_xref = True)
>>>     eval_mgtuples(pubid,tups)
>>>     all_tups2.append(tups)
>>> precision, recall, f_measure = cl.evaluate(all_tups2)
```

Κεφάλαιο 6

Επίλογος

Στις ενότητες που ακολουθούν συνοψίζονται τα συμπεράσματα από την εργασία αυτή και αναφέρονται μελλοντικές επεκτάσεις, εφαρμογές και εργασίες.

6.1 Σύνοψη

Συνοψίζοντας, στόχος της παρούσας εργασίας ήταν η μοντελοποίηση ενός συστήματος για την αυτόματη εξαγωγή αλληλεπιδράσεων μεταξύ βιομορίων από επιστημονικά κείμενα. Στα πλαίσια της, έγινε μελέτη διαφορετικών τεχνικών για την επίλυση του προβλήματος και υλοποιήθηκαν δύο μοντέλα ανάλογα με το είδος του κειμένου που είναι διαθέσιμο.

Στις περιπτώσεις που είναι διαθέσιμη η περίληψη, η κάθε πρόταση είναι σημαντική και μπορεί να καθορίσει το αποτέλεσμα του κατηγοριοποιητή. Αντίθετα, στο σώμα των δημοσιεύσεων υπάρχουν πολλές αναφορές σε διαφορετικά βιολογικά μόρια και όρους και έτσι το αποτέλεσμα της κατηγοριοποίησης βασίζεται σε περισσότερες από μία αναφορές στους όρους. Έτσι ενώ για την περίπτωση των άρθρων το φίλτρο στο πλήθος των φράσεων βελτιώνει την επίδοση, στην περίπτωση των βιβλιογραφικών αναφορών (περιλήψεις) η επίδοση είναι καλύτερη χωρίς τη χρήση κάποιου φίλτρου.

Για την ανάλυση του κειμένου, χρησιμοποιήθηκαν ή υλοποιήθηκαν κάποια βοηθητικά εργαλεία, που σχετίζονται με δομές της φυσικής γλώσσας.

Το αποτέλεσμα είναι η δημιουργία ενός μοντέλου το οποίο αξιολογείται με βάση κάποιες μετρικές και παρουσιάζει ακρίβεια της τάξης του 70% και ευαισθησία στην τάξη του 69%. Η μικρή τάξη στην ευαισθησία οφείλεται σε διάφορες παραμέτρους και στα πολλά μηχανοποιημένα στάδια που μεσολαβούν

μέχρι το αποτέλεσμα. Επιπλέον, η πολυπλοκότητα του προβλήματος και ο αυτόματος τρόπος ορισμού των κλάσεων επηρεάζουν και τις δύο μετρικές.

6.2 Μελλοντικές Εργασίες

Παρακάτω, ακολουθούν πιθανές μελλοντικές εργασίες που μπορούν να συμπληρώσουν, να χρησιμοποιήσουν ή να βελτιώσουν τα αποτελέσματα της εργασίας αυτής:

- Μία άμεση εφαρμογή είναι η χρήση των μοντέλων που κατασκευάστηκαν για την εξαγωγή αλληλεπιδράσεων από άγνωστα κείμενα στα οποία γίνεται αναφορά σε κάποιο microRNA. Μια τέτοια συλλογή περιλαμβάνεται στο miRPub, μια εφαρμογή που συγκεντρώνει δημοσιεύσεις που περιέχουν κάποιο όνομα μορίου microRNA. Η ενσωμάτωση των αποτελεσμάτων του μοντέλου σε ένα τέτοιο σύστημα, μπορεί να περιλαμβάνει τη δυνατότητα θετικής ή αρνητικής βαθμολόγησης της κάθε αλληλεπίδρασης από τους χρήστες
- Επίσης, τα μοντέλα που κατασκευάστηκαν θα μπορούσαν να χρησιμοποιηθούν για την διευκόλυνση των ειδικών που έχουν επιμεληθεί τις αλληλεπιδράσεις που περιλαμβάνουν οι βάσεις, όπως οι TarBase και miRTarBase, στην εργασία εντοπισμού νέων αλληλεπιδράσεων από κείμενα.
- Επέκταση των προηγούμενων εργασιών θα μπορούσε να είναι η συγκεντωτική μελέτη των αλληλεπιδράσεων microRNA και γονιδίων-στόχων και η συγκέντρωση των επιστημονικών εργασιών που αναφέρονται σε καθεμία. Συγκεκριμένα, θα μπορούσε να γίνει κατασκευή γράφου εξαρτήσεων για όλες τις γνωστές αλληλεπιδράσεις, η μελέτη συγκρουόμενων αποτελεσμάτων ή επαληθεύσεων καθεμιάς.
- Παράλληλα, θα μπορούσαν να γίνουν επεκτάσεις στη μέθοδο για την εξαγωγή συσχετίσεων μεταξύ microRNA και διαφορετικών Οντοτήτων, όπως ασθένειες, ιστοί και κύτταρα. Στην περίπτωση των ασθενειών, υπάρχουν καταγεγραμμένα δεδομένα στις βάσεις miR2Disease και miRCancer. Σε μια τέτοια εργασία, θα ήταν χρήσιμη η αναγνώριση του συνόλου των βιοϊατρικών οντοτήτων που περιέχονται στο κείμενο.
- Όπως αναφέρθηκε στο Κεφάλαιο 5, μεγάλο ποσοστό των αλληλεπιδράσεων δεν περιγράφονται στο κείμενο, αλλά σε Πίνακες και συμπληρωματικά αρχεία. Στις περιπτώσεις που τα δεδομένα αυτά είναι διαθέσιμα σε επεξεργάσιμη μορφή, μπορεί να αναπτυχθεί αυτόματος τρόπος εξαγωγής τους.

Παραρτήματα

Παράρτημα Α′

Αποτελέσματα Αναγνώρισης Αλληλεπιδράσεων

Άνω Όριο Αλ-λήλ./Δημοσ.	# Βιβλιογραφικών Αναφορών
2	852
3	1249
5	1543
10	1728
20	1796
50	1829

Πίνακας Α′.1: Πλήθος Βιβλιογραφικών Αναφορών/Αριθμός Αλληλεπιδράσεων

Άνω Όριο Αλλήλ./Δημοσ.	True στις Βιβλ. Αναφ.	στο TarBase	True/TarBase
2	655	832	0.787
3	1171	1620	0.723
5	1702	2602	0.654
10	2086	3745	0.557
20	2256	4655	0.485
50	2274	5609	0.405

Πίνακας Α′.2: Βιβλιογραφικές Αναφορές: Ευαισθησία Αναγνώρισης Αλληλεπιδράσεων

Άνω Όριο Αλλήλ./Δημοσ.	False στις Βιβλ. Αναφ.	True στις Βιβλ. Αναφ.	True/Σύνολο
2	841	483	0.365
3	1375	898	0.395
5	1819	1280	0.413
10	2229	1623	0.421
20	2307	1750	0.431
50	2328	1770	0.432

Πίνακας Α′.3: Βιβλιογραφικές Αναφορές: Ακρίβεια Αναγνώρισης Αλληλεπιδράσεων

Άνω Όριο Αλλήλ./Δημοσ.	True στις Βιβλ. Αναφ.	στο TarBase	True/TarBase
2	550	832	0.661
3	960	1620	0.593
5	1304	2602	0.501
10	1535	3745	0.410
20	1602	4655	0.344
50	1611	5609	0.287

Πίνακας Α′.4: Βιβλιογραφικές Αναφορές: Ευαισθησία Αναγνώρισης Αλληλεπιδράσεων για Ζεύγη με αριθμό φράσεων μεγαλύτερο από 1.

Άνω Όριο Αλλήλ./Δημοσ.	False στις Βιβλ. Αναφ.	True στις Βιβλ. Αναφ.	True/Σύνολο
2	365	401	0.523
3	590	724	0.550
5	755	963	0.561
10	866	1163	0.573
20	887	1220	0.579
50	893	1229	0.579

Πίνακας Α′.5: Βιβλιογραφικές Αναφορές: Ακρίβεια Αναγνώρισης Αλληλεπιδράσεων με πλήθος φράσεων μεγαλύτερο από 1

Άνω Όριο Αλλήλ./Δημοσ.	# Άρθρων
2	533
3	795
5	990
10	1104
20	1147
50	1167

Πίνακας Α′.6: Πλήθος Άρθρων/Αριθμός Αλληλεπιδράσεων

Άνω Όριο Αλλήλ./Δημοσ.	True στα Άρθρα	in TarBase	True/TarBase
2	490	533	0.919
3	968	1057	0.916
5	1539	1709	0.901
10	2128	2430	0.876
20	2485	3025	0.821
50	2627	3604	0.729

Πίνακας Α′.7: Άρθρα: Ευαισθησία στην Αναγνώριση Αλληλεπιδράσεων

Άνω Όριο Αλ-λήλ./Δημοσ.	False στα Άρθρα	True στα Άρθρα	True/Σύνολο
2	7422	636	0.079
3	11646	1206	0.094
5	15494	1873	0.108
10	17675	2554	0.126
20	18592	2940	0.137
50	18994	3091	0.174

Πίνακας Α΄.8: Άρθρα: Ευαισθησία Αναγνώρισης Αλληλεπιδράσεων

Άνω Όριο Αλ-λήλ./Δημοσ.	True στα Άρθρα	στο TarBase	True/TarBase
2	489	533	0.917
3	953	1057	0.902
5	1508	1709	0.882
10	2025	2430	0.833
20	2319	3025	0.767
50	2423	3604	0.672

Πίνακας Α΄.9: Άρθρα: Ακρίβεια Αναγνώρισης Αλληλεπιδράσεων για ζεύγη με πλήθος φράσεων μεγαλύτερο από 1

Άνω Όριο Αλ-λήλ./Δημοσ.	False στα Άρθρα	True στο Άρθρο	True/Σύνολο
2	4056	593	0.128
3	6229	1110	0.151
5	8243	1707	0.172
10	9283	2256	0.196
20	9666	2536	0.208
50	9839	2641	0.212

Πίνακας Α΄.10: Άρθρα: Ακρίβεια στην Αναγνώριση Αλληλεπιδράσεων για ζεύγη με αριθμό φράσεων μεγαλύτερο από 1.

Παράρτημα Β′

Αποτελέσματα Αξιολόγησης Μοντέλου

Άνω Όριο Αλ-λήλ./Δημοσ.	Ακρίβεια	Ευαισθησία	F1-score
2	0.676	0.736	0.704
3	0.720	0.721	0.720
5	0.730	0.681	0.704
10	0.718	0.665	0.690
20	0.725	0.657	0.689
50	0.710	0.677	0.692

Πίνακας Β′.1: Βιβλιογραφικές Αναφορές: Αποτελέσματα Κατηγοριοποιητή χωρίς φίλτρο φράσεων.

Άνω Όριο Αλλήλ./Δημοσ.	Ακρίβεια	Ευαισθησία	F1-score
2	0.676	0.579	0.623
3	0.720	0.521	0.604
5	0.730	0.445	0.553
10	0.718	0.371	0.488
20	0.725	0.319	0.442
50	0.710	0.274	0.395

Πίνακας Β′.2: Βιβλιογραφικές Αναφορές: Αποτελέσματα για τα δεδομένα του TarBase.

Άνω Όριο Αλλήλ./Δημοσ.	Ακρίβεια	Ευαισθησία	F1-score
2	0.784	0.781	0.781
3	0.801	0.792	0.796
5	0.815	0.796	0.805
10	0.794	0.744	0.767
20	0.804	0.757	0.779
50	0.809	0.751	0.778

Πίνακας Β′.3: Βιβλιογραφικές Αναφορές: Αποτελέσματα Κατηγοριοποιητή με αριθμό φράσεων μεγαλύτερο από 1.

Άνω Όριο Αλλήλ./Δημοσ.	Ακρίβεια	Ευαισθησία	F1-score
2	0.784	0.516	0.621
3	0.801	0.470	0.592
5	0.815	0.399	0.535
10	0.794	0.305	0.440
20	0.804	0.260	0.393
50	0.809	0.215	0.340

Πίνακας Β′.4: Βιβλιογραφικές Αναφορές: Αποτελέσματα για τα δεδομένα του TarBase για αλληλεπιδράσεις με αριθμό φράσεων μεγαλύτερο από 1.

Άνω Όριο Αλ-λήλ./Δημοσ.	Ακρίβεια	Ευαισθησία	F1-score
2	0.729	0.662	0.691
3	0.695	0.670	0.681
5	0.697	0.643	0.667
10	0.661	0.542	0.594
20	0.714	0.528	0.606
50	0.668	0.532	0.591

Πίνακας Β′.5: Άρθρα: Αποτελέσματα Κατηγοριοποιητή χωρίς φίλτρο φράσεων.

Άνω Όριο Αλ-λήλ./Δημοσ.	Ακρίβεια	Ευαισθησία	F1-score
2	0.729	0.608	0.660
3	0.695	0.614	0.651
5	0.697	0.579	0.631
10	0.661	0.474	0.552
20	0.714	0.433	0.538
50	0.668	0.388	0.490

Πίνακας Β′.6: Άρθρα: Αποτελέσματα σε σχέση με τα δεδομένα του TarBase.

Άνω Όριο Αλ-λήλ./Δημοσ.	Ακρίβεια	Ευαισθησία	F1-score
2	0.769	0.754	0.759
3	0.799	0.734	0.764
5	0.767	0.686	0.723
10	0.718	0.633	0.672
20	0.702	0.610	0.650
50	0.715	0.592	0.646

Πίνακας Β′.7: Άρθρα: Αποτελέσματα Κατηγοριοποιητή με αριθμό φράσεων μεγαλύτερο από 1.

Άνω Όριο Αλ-λήλ./Δημοσ.	Ακρίβεια	Ευαισθησία	F1-score
2	0.769	0.691	0.726
3	0.799	0.662	0.723
5	0.767	0.605	0.676
10	0.718	0.559	0.628
20	0.702	0.501	0.583
50	0.715	0.431	0.537

Πίνακας Β′.8: Άρθρα: Αποτελέσματα σε σχέση με τα δεδομένα του TarBase για ζεύγη με αριθμό φράσεων μεγαλύτερο από 1.

Κατάλογος πινάκων

Κατάλογος σχημάτων

Βιβλιογραφία

[1] Zardo G1, Ciolfi A, Vian L, Starnes LM, Billi M, Racanicchi S, Maresca C, Fazi F, Travaglini L, Noguera N, Mancini M, Nanni M, Cimino G, Lo-Coco F, Grignani F, and Nervi C. Polycombs and microrna-223 regulate human granulopoiesis by transcriptional control of target gene expression. *Blood*, 119(3):4034–4046, 2012.

[2] Anna M. Krichevsky. Introduction to microrna, January 2014. URL `http://serious-science.org/videos/437`.

[3] Anna M. Krichevsky. microrna studies, March 2014. URL `http://serious-science.org/videos/68`.

[4] Thanasis Vergoulis, Ioannis S. Vlachos, Panagiotis Alexiou, George Georgakilas, Manolis Maragkakis, Martin Reczko, Stefanos Gerangelos, Nectarios Koziris, Theodore Dalamagas, and Artemis G. Hatzigeorgiou. Tarbase 6.0: capturing the exponential growth of mirna targets with experimental support. *Nucleic Acids Research*, 2011.

[5] Bird Steven, Edward Loper, and Ewan Klein. *Natural Language Processing with Python*. OŔeilly Media Inc., 2009.

[6] Hal Daumé III. Notes on CG and LM-BFGS optimization of logistic regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`, August 2004.

[7] Boya Xie, Qin Ding, Hongjin Han, and Di Wu. mircancer: a microrna-cancer association database constructed by text mining on literature. *Bioinformatics*, 29:638–644, 2013.

[8] Laulederkind SJ, Hayman GT, Wang SJ, Smith JR, Lowry TF, Nigam R, Petri V, De Pons J, Dwinell MR, and Shimoyama M. The rat genome database 2013–data tools and users. *Brief Bioinform.*, Feb 2013.

[9] Α. Φλιάτουρας. Μορφολογική και Λεξιλογική Ανάλυση της Ελληνικής: Βιβλιογραφικά και Ηλεκτρονικά Εργαλεία. URL `http://repository.edulll.gr/edulll/retrieve/1389/246.pdf`.

[10] Da Klein and Christopher Manning. Maxent models and discriminative estimation, May 2003. URL `http://www.cs.berkeley.edu/~klein/papers/maxent-tutorial-slides.pdf`.

[11] Blake JA, Bult CJ, Eppig JT, Kadin JA, Richardson JE, and The Mouse Genome Database Group. The mouse genome database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res.*, 42:D810–D817, 2014.

[12] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Res.*, 33:D54–D58, Jan 2005.

[13] Burt DW, Carrë W, Fell M, Law AS, Antin PB, Maglott DR, Weber JA, Schmidt CJ, Burgess SC, and McCarthy FM. The chicken gene nomenclature committee report. *BMC Genomics*, 10, 2009.

[14] Gray KA, Daugherty LC, Gordon SM, Seal RL, Wright MW, and Bruford EA. genenames.org: the hgnc resources in 2013. *Nucleic Acids Res.*, 41:D545–52, 2013.

[15] Michael Collins. Probabilistic context-free grammars (pcfgs), 2011. URL `http://www.cs.columbia.edu/~mcollins/courses/nlp2011/notes/pcfgs.pdf`.

[16] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. *HLT-NAACL*, pages 252–259, 2003.

[17] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pages 423–430, 2003.

[18] Guido Minnen, John Carroll, and Darren Pearce. Applied morphological processing of english. *Natural Language Engineering*, 7(3):207–223, 2001.

[19] Ariel Schwartz and Marti Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. 2003.

[20] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14:130–137, 1980.

[21] Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel Anas, Chou CH, Chu CF, Huang HY, Lin CM, Ho SY, Jian TY, Lin FM, Chang TH, Weng SL, Liao KW, Liao IE, Liu CC, and Huang HD. mirtarbase update 2014: an information

resource for experimentally validated mirna-target interactions. *Nucleic acids research.*, 2014.

[22] Xiao F, Zuo Z, Cai G, Kang S, Gao X, and Li T. mirecords: an integrated resource for microrna-target interactions. *Nucleic Acids Res.*, 37:D105–D110, 2009.

[23] Wu C, MacLeod I, and Su AI. Biogps and mygene.info: organizing online, gene-centric information. *Nucl. Acids Res.*, 41(D1):D561–D565, 2013.

Γλωσσάρι

Arabidopsis Thaliana Είναι ένα φυτό που έχει σχετικά μικρό κύκλο ζωής και έχει χρησιμοποιηθεί στην έρευνα στη βιολογία και τη γενετική ως πρότυπος οργανισμός για τα φυτά.

Caenorhabtitis Elegans Ανήκει στην κατηγορία των νηματωδών και είναι ένας μη παρασιτικός και διάφανος σκώληκας. Είναι ο πρώτος οργανισμός για τον οποίο έγινε πλήρης καταγραφή του γονιδιώματος και αποτελεί πρότυπο οργανισμό για τη βιολογία και τη μελέτη της εξέλιξης.

Celera Assembly Ακολουθία από το Celera Assembler (`http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Main_Page`) , επιστημονικό λογισμικό για την έρευνα στη βιολογία. Χρησιμοποιεί την τεχνική WGS.

Danio Rerio Η επίσημη ονομασία του zebrafish, που είναι πρότυπος οργανισμός. Είναι τροπικό ψάρι του γλυκού νερού, ανήκει στην οικογένεια των κυπρινιδών.

Drosophila Melanogaster Επίσημη ονομασία για το είδος της μύγας των φρούτων ή δροσόφιλας. Είναι ένα είδος μύγας ανήκει στην τάξη των διπτέρωνΑποτελεί πρότυπο οργανισμό για την έρευνα στη Βιολογία.

F1-score Το F1-score είναι ένα μέτρο αξιοπιστίας ενός δυαδικού στατιστικού ελέγχου. Προκύπτει από τις τιμές της ακρίβειας και της ευαισθησίας.

Gallus Gallus Επίσημη ονομασία για το είδος της όρνιθας, ανήκει στη τάξη των Ορνιθόμορφων, στην οικογένεια των Φασιανίδων. Στη Έρευνα στη Βιολογία χρησιμοποιείται ως πρότυπος οργανισμός.

Homo Sapiens Επίσημη ονομασία για το είδος του ανθρώπου.

MeSH Η MeSH είναι το βιοϊατρικό λεξιλόγιο της βιβλιοθήκης της Medline (National Library of Medicine). Αποτελείται από όρους που σχηματίζουν

μία ιεραρχική δομή, η οποία επιτρέπει την αναζήτηση σε διάφορα επίπεδα ακρίβειας.

Mus Musculus Επίσημη ονομασία για το είδος του ποντικιού. Αποτελεί πρότυπο οργανισμό στην έρευνα στη Βιολογία.

QTL Τμήματα DNA που σχετίζονται με κάποιο συγκεκριμένο φαινότυπο.

Rattus Norvegicus Επίσημη ονομασία για το είδος του αρουραίου.

SSLP Συντομογραφία του Simple Sequence Length Polymorphism. Αποτελεί μίας μορφή πολυμορφισμού ανάμεσα σε οργανισμούς ενός είδους. Είναι επαναλαμβανόμενες ακολουθίες που διαφέρουν σε μήκος και βρίσκονται στις περιόχές των DNA που δεν μεταφράζονται σε πρωτεΐνες, αλλά αποκόπτονται (εσόνια). plural

western blot Αναλυτική μέθοδος που χρησιμοποιείται συχνά για τον εντοπισμό συγκεκριμένων πρωτεϊνών σε δείγμα ιστών ή κυττάρων. Λέγεται και ανοσοαποτύπωση (Immunoblot).

WGS Είναι μία ημιαυτόματη τεχνική ανάλυσης ακολουθίας (Sequencing) γενετικού υλικού.

Γραμματική Χωρίς Συμφραζόμενα Είναι μία τυπική γραμματική, δηλαδή ένα σύνολο κανόνων παραγωγής για συμβολοσειρές σε μία τυπική γλώσσα. Σε κάθε κανόνα περιέχει μόνο ένα μη-τελικό σύμβολο.

γονίδιο-δείκτης Το γονίδιο-δείκτης ή reporter gene χρησιμοποιείται για αναφορά άλλων γονιδίων και ακολουθιών σε πειράματα.

θέμα Το θέμα ή stem αναφέρεται συχνά σε διαφορετικές έννοιες:

- Είναι το τμήμα της λέξεις από το οποίο αφαιρούνται όλες οι κλιτικές καταλήξεις (κλιτική κατάληξη). Για παράδειγμα τος stem της λέξης regulations είναι regulation
- Είναι το τμήμα της λέξεις από το οποίο αφαιρούνται όλα τα προσφύματα (πχ: -ότητα). Για παράδειγμα το θέμα της λέξης regulations είναι regul, όπως και το θέμα της λέξης regulates

κλιτική κατάληξη Κλιτική κατάληξη ή inflection είναι η κατάληξη της λέξεις που δεν αλλάζει τη σημασίας της αλλά προσθέτει πληροφορία σχετικά με το χρόνο, τον αριθμό το πρόσωπο και άλλα. Για παράδειγμα για το ρήμα target:

- target-s

- target-ed
- target-ing

προϊόν μεταγραφής Είναι το προϊόν της μεταγραφής του DNA σε RNA που στη συνέχεια μπορεί να μεταφραστεί σε πρωτεΐνη. Ένα γονίδιο μπορεί να μεταγράφεται σε περισσότερες από μία τέτοιες ακολουθίες.

Ακρωνύμια

CGNC Chicken Gene Nomenclature Consortium

CONLL Conference on Computational Natural Language Learning

EBI European Bioinformatics Institute

EMBL European Molecular Biology Laboratory

eUtils Entrez Programming Utilities

HGNC HUGO Gene Nomenclature Commitee

MGI Mouse Genome Informatics

NCBI National Center for Biotechnology Information

NER Named-Entity Recognition

NLM National Library of Medicine

NLP Natural Language Processing

NLTK Natural Language Toolkit

PCFG Probabilistic Context Free Grammar

PMC PubMed Central

PMCID PubMed Central ID

PMID PubMed ID

POS Part-of-Speech

QTL Quantitative Trait Locus

RefSeq Reference Sequence

RGD Rat Genome Database

TAIR The Arabidopsis Information Resource

WGS Whole Genome Shotgun

WSJ Wall Street Journal

ZFIN Zebrafish Model Organism Database