ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ

ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

"ΥΠΟΛΟΓΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΣΤΟΧΩΝ Mrna ΚΑΙ ΕΚΦΡΑΣΗ ΓΟΝΙΔΙΩΝ"

Μεταπτυχιακή Εργασία

Δήμου Μαρία

Επιβλέπουσα:

Χατζηγεωργίου Άρτεμις

Καθηγήτρια Πανεπιστημίου Θεσσαλίας

Βόλος, Ιούνιος 2016

Πανεπιστήμιο Θεσσαλίας
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

"ΥΠΟΛΟΓΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΣΤΟΧΩΝ Mrna ΚΑΙ ΕΚΦΡΑΣΗ ΓΟΝΙΔΙΩΝ"

Μεταπτυχιακή Εργασία

Δήμου Μαρία

Επιβλέπουσα:
Χατζηγεωργίου Άρτεμις
Καθηγήτρια Πανεπιστημίου Θεσσαλίας

Εγκρίθηκε από την τριμελή επιτροπή

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Χατζηγεωργίου Άρτεμις Καθηγήτρια Π.Θ.	Μποζάνης Παναγιώτης Καθηγητής Π.Θ.	Σταμούλης Γεώργιος Καθηγητής Π.Θ.

Βόλος, Ιούνιος 2016

ΕΥΧΑΡΙΣΤΙΕΣ

Με την ολοκλήρωση της παρούσας διπλωματικής εργασίας θα ήθελα να ευχαριστήσω ιδιαίτερα την επιβλέπουσα καθηγήτριά κυρία Άρτεμις Χατζηγεωργίου για την αμέριστη συμπαράσταση και τη βοήθεια της κατά της διάρκεια της συνεργασίας μας. Ιδιαίτερες ευχαριστίες απευθύνω στην υποψήφια διδάκτορα του τμήματος Καραγκούνη Δήμητρα οι συμβουλές της οποία και οι παρατηρήσεις της ήταν πολύ χρήσιμες καθ όλη την διάρκεια εκπόνησης της εργασίας μου.

Περιεχόμενα

Ευρετήριο Εικόνων

Ευρετήριο Πινάκων

Περίληψη

Η ανακάλυψη της λειτουργίας των microRNAs, αλλά και η επιρροή που αυτά ασκούν στην έκφραση των γονιδίων και τελικά στην έκφραση των ασθενειών έστρεψε όλη την επιστημονική κοινότητα στο να μελετήσει τον τρόπο με τον οποίο αυτά λειτουργούν και την επιρροή που ασκούν στον ανθρώπινο οργανισμό.

Το παραπάνω οδήγησε στη δημιουργία τεραστίων βάσεων δεδομένων που περιλαμβάνουν σημαντικά στοιχεία σχετικά με τη δράση των microRNAs. Η σημαντικότερη πληροφορία, η οποία και περιλαμβάνεται φυσικά στις βάσεις αυτές, είναι η αλληλεπίδρασή τους με τα διάφορα γονίδια η οποία οδηγεί τελικά στην εξάπλωση των ασθενειών.

Ο εντοπισμός της πληροφορίας που σχετίζεται με την αλληλεπίδραση των microRNAs και των γονιδίων γίνεται στα διάφορα επιστημονικά άρθρα και τεκμηριώνεται από τις πειραματικές μεθόδους που περιλαμβάνονται σε αυτά. Για να γίνει αυτό, χρειάζεται η συμβολή των curators οι οποίοι είναι υπεύθυνοι για τον εντοπισμό, την αξιολόγηση και τελικά την καταχώρηση της πληροφορίας.

Με στόχο τη διευκόλυνση των curators αλλά και τη σημαντική εξοικονόμηση χρόνου δημιουργήθηκαν οι αλγόριθμοι εξόρυξης κειμένου οι οποίοι εντοπίζουν την αλληλεπίδραση σε ένα επιστημονικό κείμενο και υποδεικνύουν στον curator το άρθρο στο οποίο βρίσκεται καταχωρημένη η πληροφορία. Ένας τέτοιος αλγόριθμος είναι και ο TarMiner, ο βέλτιστος έως σήμερα αλγόριθμος στην κατηγορία αυτή.

Με στόχο την εκπαίδευση του αλγορίθμου, δημιουργήθηκε ένα σετ δεδομένων με επιστημονικά άρθρα που προέρχονται από την βιβλιοθήκη TarBase και περιέχουν όλη την απαραίτητη πληροφορία η οποία θα πρέπει να εντοπιστεί από τον αλγόριθμο. Η σημαντικότητα της ύπαρξης τέτοιων σετ δεδομένων είναι μεγάλη καθώς βοηθά τους αλγορίθμους ώστε αυτοί να βελτιστοποιούνται.

Abstract

The discovery of the microRNA function, and the impact they exert on gene expression and ultimately in the expansion of various diseases turned the scientific community into studying their behavior and their influenceon the human body.

The aforementioned led to the creation of huge databases that contain important information about the microRNA activity. The most important information included in the databases is their interaction with various genes which ultimately leads to the spread of different type of diseases. This can be traced into scientific articles that contain both the interaction as well as the method used in order to trace and validate it.

In order for the reaction among microRNAs and genes to be indexed, curators are occupied on identifying, evaluating and ultimately performing the information entry. The above process requires a lot of time spend and brought up the necessity of text mining – time saving algorithms that would detect the necessary information and indicate it to the curator. TarMiner is a text mining algorithm that functions beyond potential and traces any possible interaction referred to a scientific article.
Aiming to train even more the TarMiner algorithm, we created a new dataset based on entries given by TarBase, the most accurate base in its category.

Κεφάλαιο 1

1.1 Διάφοροι Οργανισμοί και Κύτταρο

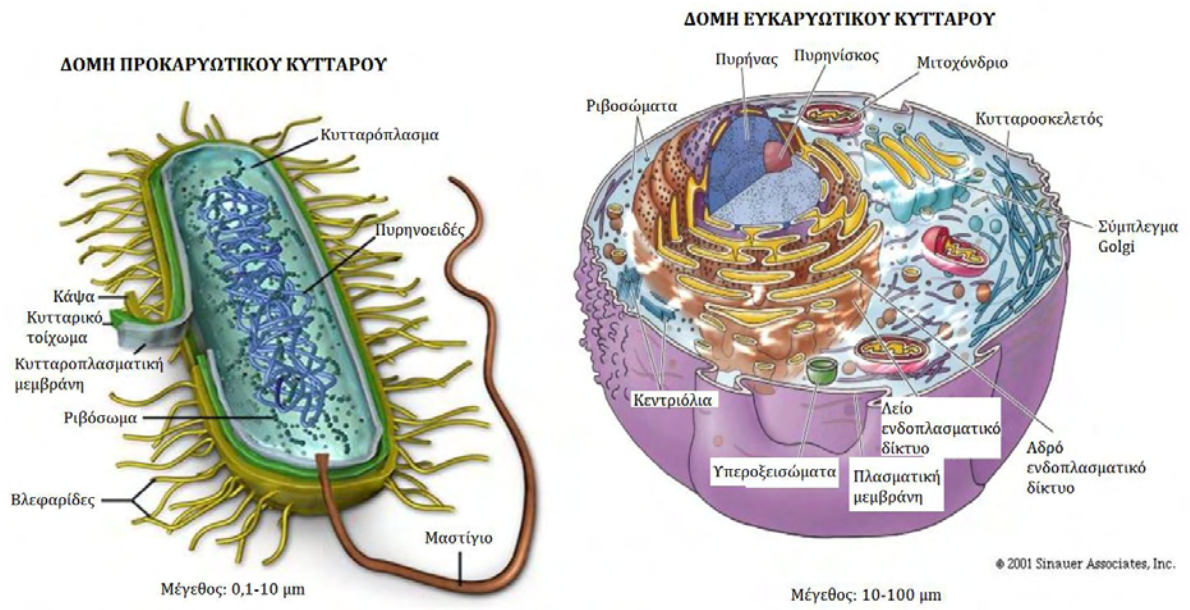
Στοιχείο εκκίνησης για την επιστήμη της βιολογίας αποτελεί η μελέτη του κυττάρου. Ο όρος κύτταρο, αναφέρεται στην ύπαρξη μιας συστηματικά οργανωμένης ομάδας μορίων, βασικό χαρακτηριστικό της οποίας είναι η συνεχής αλληλεπίδραση των μορίων αυτών μεταξύ τους (1). Κάθε κύτταρο διαθέτει μορφολογική, φυσική και χημική οργάνωση αλλά και τις ικανότητες της αφομοίωσης, της ανάπτυξης και της αναπαραγωγής.

Η μορφολογία των κυττάρων που εντοπίζονται στους διάφορους οργανισμούς διαφέρει κατά το μέγεθος, τις διαστάσεις, και την ικανότητα της εξελικτικής προσαρμογής στα διάφορα περιβάλλοντα (2). Η διάμετρός τους ποικίλει από δέκατα του μm, έως μερικά εκατοστά ανάλογα πάντα από το είδος από το οποίο προέρχονται (1) (3).

Ως οργανισμός το κύτταρο, διαθέτει την ικανότητα της επιβίωσης, ακόμη και εάν αυτό βρίσκεται μόνο του σε ένα περιβάλλον. Η παραπάνω ιδιότητα προϋποθέτει την ύπαρξη μιας μεταβολικής μηχανής που μπορεί να αντλήσει ενέργεια από το περιβάλλον και να τη χρησιμοποιήσει σε ουσιώδεις βιοχημικές διεργασίες, που περιλαμβάνουν την κίνηση ουσιών, την εκλεκτική μεταφορά μορίων μέσα και έξω από το κύτταρο και την ικανότητα αλλαγής και διαμόρφωσής τους, δηλαδή της προσαρμογής τους στις περιβάλλουσες φυσικές και χημικές συνθήκες (2). Πέραν της μεταβολικής μηχανής του, κάθε κύτταρο, διαθέτει ομάδες γονιδίων που καθορίζουν τη σύνθεση ουσιών και μια διακριτή δομή την κυτταρική ή πλασματική μεμβράνη που τα απομονώνει από το εξωτερικό περιβάλλον.

Η κατηγοριοποίηση που γίνεται σήμερα στα κύτταρα τα διαχωρίζει σε προκαρυωτικά και ευκαρυωτικά εξαιρώντας από αυτά τους ιούς και τους φάγους, μια ιδιαίτερη κατηγορία οργανισμών που έχουν τη δυνατότητα να παρεμβαίνουν στις κυτταρικές λειτουργίες (4). Τα

ζωντανά κύτταρα, αποτελούνται από περιορισμένο αριθμό χημικών ουσιών με βασικότερα αυτών τον Άνθρακας (C), το Υδρογόνο (H), το Οξυγόνο (O), το Άζωτο(N) και το Θείο (S) (4).



Εικόνα 1 : Προκαρυωτικά και ευαρυωτικά κύτταρα (5)

1.2 Μικρομόρια και Μακρομόρια

Τα μόρια τα οποία σχετίζονται με το φαινόμενο της ζωής καλούνται βιομόρια και διαχωρίζονται σε δύο βασικές κατηγορίες, αυτή των μικρών μορίων και αυτή των μακρομορίων. Η κατηγορία των μικρών μορίων, εμφανίζεται σε δύο βασικές μορφές. Αφενός μπορεί να αποτελεί δομική μονάδα των μακρομορίων και αφετέρου να διαδραματίζει ανεξάρτητουτς ρόλους όπως είναι η μετάδοση των σημάτων. Τα σημαντικότερα γνωστά μόρια, τα οποία και περιγράφονται παρακάτω, είναι τα λιπαρά οξέα, τα αμινοξέα και τα νουκλεοτίδια (6).

- ***Νουκλεοτίδια***

Τα νουκλεοτίδια, αποτελούν χημικές ενώσεις που απαρτίζονται από τα παρακάτω τρία μέρη :

1. Ετεροκυκλική βάση
2. Σάκχαρο
3. Μία ή περισσότερες φωσφορικές ομάδες

Στα πιο συχνά απαντώμενα νουκλεοτίδια η βάση είναι ένα παράγωγο είτε της πουρίνης ή της πυριμιδίνης και το σάκχαρο είναι μία πεντόζη, ένα σακχαρο με 5 άτομα άνθρακα δεοξυριβόζη ή ριβόζη (7) . Κάθε νουκλεοτίδιο, αποτελεί μία μονομερή δομική μονάδα των DNA και RNA, ενώ η ένωση τριών ή και περισσοτέρων νουκλετιδίων μεταξύ τους σηματοδοτεί το σχηματισμό ενός νουκλεϊκού οξέως. Τα νουκλεϊκά οξέα που έχουν ως βάση τη δεοξυριβόζη (7) ονομάζονται DNA ενώ αυτά που έχουν ως βάση ριβόζη RNA. Ο ρόλος των νουκλεοτιδίων στην επιστήμη της βιολογίας είναι ιδιαίτερα σημαντικός καθώς αυτά διαδραματίζουν ιδιαίτερα σημαντικούς ρόλους στην επικοινωνία των κυττάρων και στο μεταβολισμό (7).

- ***Αμινοξέα***

Πρόκειται για μόρια αποτελούμενα από ένα κεντρικό άτομο άνθρακα, που ονομάζεται α-άνθρακας, ενωμένο με μια αμινομάδα (NH3 +), μια καρβοξυλομάδα (COO-) και μια πλευρική ομάδα η οποία είναι διαφορετική για κάθε αμινοξύ, συμβολίζεται με R και προσδίδει στο κάθε αμινοξύ τα ιδιαίτερα χαρακτηριστικά του. Στα κύτταρα των ζωντανών οργανισμών συναντώνται περίπου είκοσι είδη αμινοξέων, κάθε ένα από τα οποίοα αποτελεί δομικό λίθο για την κάθε πρωτείνη (8).

1.3 Πρωτεΐνες

Τα μεγάλα αυτά μόρια, αποτελούνται από αμινοξέα τα οποία σχηματίζουν μεταξύ τους μία γραμμική αλυσίδα και ενώνονται με τους πεπτιδικούς δεσμούς. Η ακολουθία αμινοξέων σε μια πρωτεΐνη καθορίζεται από ένα γονίδιο και κωδικοποιείται κατά τον γενετικό κώδικα (9). Παρόλο που ο γενετικός κώδικας κωδικοποιεί 20 αμινοξέα, τα αμινοξέα που συνιστούν την πρωτεΐνη συχνά υφίστανται χημικές αλλαγές κατά τη μετα-μεταφραστική τροποποίηση: είτε προτού να μπορέσει η πρωτεΐνη να λειτουργήσει στο κύτταρο, είτε ως τμήμα των μηχανισμών ελέγχου (9). Ο συνδυασμός περισσότερων της μίας πρωτεϊνών, έχει ως στόχο την επίτευξη κάποιας συγκεκριμένης λειτουργίας εντός των οργανισμών ή αφορά στη συσσωμάτωση με στόχο τη δημιουργία συμπλοκών.

Οι πρωτεΐνες αποτελούν τα μακρομόρια εκείνα που είναι απαραίτητα για όλους τους ζωντανούς οργανισμούς καθώς έχουν ενεργό ρόλο σε κάθε διαδικασία των κυττάρων. Πολλές πρωτεΐνες δρουν ως ένζυμα που καταλύουν τις βιοχημικές αντιδράσεις, και είναι ζωτικής σημασίας στο μεταβολισμό (3) (9). Άλλες πάλι, έχουν μηχανικές λειτουργίες και πολλές φορές συμβάλλουν στη διατήρηση της μορφής των κυττάρων. Σημαντικός είναι και ο ρόλος τους στη διακυτταρική επικοινωνία, τη δράση του ανοσοποιητικού συστήματος, τον σχηματισμό κυτταρικών ιστών, και τον κυτταρικό κύκλο.

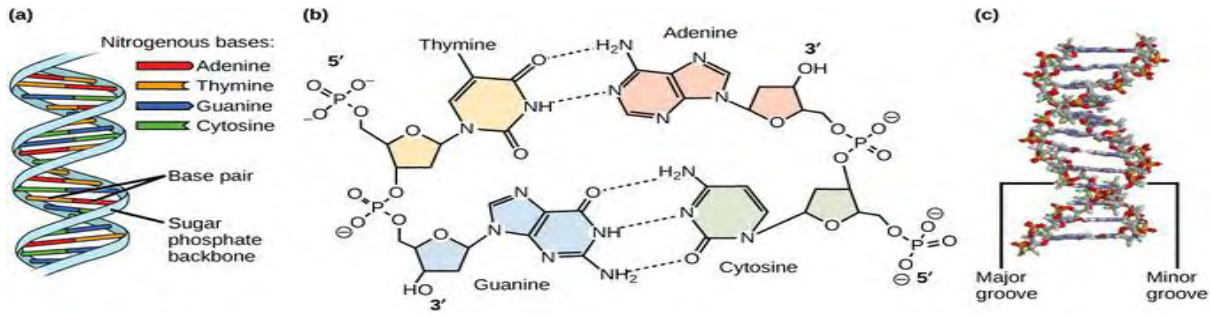
1.4 Το DNA

Η γνωστή διπλή έλικα ή αλλιώς δε(σ)οξυριβο(ζο)νουκλεϊ(νι)κό οξύ (Deoxyribonucleic acid - DNA) περιέχει όλες εκείνες τις γενετικές πληροφορίες που καθορίζουν την ανάπτυξη των κυττάρων σε οποιαδήποτε μορφή ζωής (10). Πρόκειται για μια μεγαλομοριακή ένωση που αποτελείται από αζωτούχες-πρωτεϊνικές βάσεις, φωσφορικές ρίζες και ένα σάκχαρο με πέντε άτομα άνθρακα (πεντόζη) (10). Ο φορέας αυτός των γενετικών πληροφοριών, εντοπίζεται κυρίως στον πυρήνα του κυττάρου και είναι υπεύθυνος για τη ρύθμιση της φυσιογνωμίας τους κυττάρου αλλά και τις ιδιαίτερες λειτουργίες που επιτελεί αυτό (10).

Η διαμόρφωση των μεγάλων μορίων του DNA στο χώρο έχει τη μορφή δύο επιμηκών αλύσεων, οι οποίες συστρέφονται ελικοειδώς μεταξύ τους. Οι πρωτεϊνικές βάσεις του είναι τέσσερις (10):

- Κυτοσίνη C
- Αδενίνη A
- Γουανίνη G
- Θυμίνη T

Χωρίζονται σε τριάδες και κωδικοποιούν το κάθε μήνυμα μεταφοράς των αμινοξέων του κυττάρου ανάλογα με τη σειρά με την οποία θα εμφανιστούν στην αλληλουχία τους. Μάλιστα, η σειρά με την οποία μεταφέρονται τα αμινοξέα στο ριβόσωμα είναι και αυτή που θα οδηγήσει στη σύνθεση των διαφορετικών πρωτεϊνών.



Εικόνα 2 : Δομή του DNA (11)

1.5 Το RNA

Το RNA ή **Ριβονουκλεικό οξύ** λειτουργεί ως αγγελιοφόρος του DNA και των πρωτεϊνικών συμπλεγμάτων και αποτελείται από μονομερή νουκλεοτίδια που παίζουν σημαντικό ρόλο στη διαδικασία της μετάφρασης από το δεοξυριβονουκλεικό οξύ (DNA) σε πρωτεϊνικά προϊόντα (12). Η δομή του, είναι παρόμοια με αυτή του DNA, με διαφορές που εντοπίζονται σε κάποια δομικά στοιχεία. Για παράδειγμα, τα μόρια RNA περιέχουν ριβόζη αντί για δεοξυριβόζη ως κύριο σάκχαρο και επίσης σε αυτέ περιέχεται η βάση ουρακίλη αντί της θυμίνης που απαντάται στο DNA (12).

Το RNA μεταγράφεται από το DNA με τη βοήθεια κυρίως ενός ενζύμου που ονομάζεται RNA πολυμεράση και στη συνέχεια επεξεργάζεται με έναν αριθμό άλλων δευτερευόντων ενζύμων (12).

Εικόνα 3 : Διαφορές ανάμεσα στο DNA και το RNA (13)

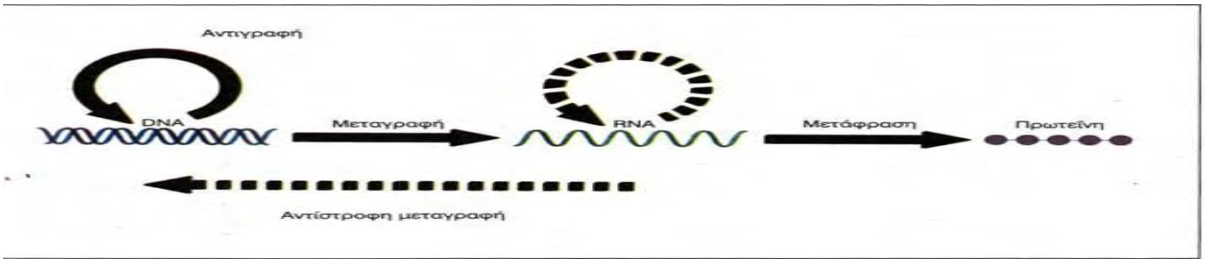
1.6 Γονίδια & Γονιδίωμα

Ως μονάδα κληρονομικότητας των ζωντανών οργανισμών, τα γονίδια αποτελούν τη σύνθεση του γονιδιώματος ενός οργανισμού που δεν είναι τίποτα άλλο από DNA και RNA. Η συντριπτική πλειοψηφία αυτών, κωδικοποιεί πρωτεΐνες, δηλαδή τα βιολογικά μακρομόρια που αποτελούνται από γραμμικές αλυσίδες των αμινοξέων και επηρεάζουν τις περισσότερες από τις χημικές αντιδράσεις που πραγματοποιούνται από τα κύτταρα (10).

Υπάρχουν κάποιες περιοχές οι οποίες δεν κωδικοποιούνται από τα γονιδιακά προϊόντα, έχουν όμως την ικανότητα να ρυθμίζουν τη γονιδιακή έκφραση. Μία τέτοια περιοχή μη-κωδικοποίησης είναι το γονίδιο-υποκινητής, μια σύντομη ακολουθία DNA που είναι απαραίτητη για την έναρξη της γονιδιακής έκφρασης (10).

1.6.1 Γενετική πληροφορία και έκφραση αυτής

Οι ακριβείς πληροφορίες σχετικά με τον ανθρώπινο οργανισμό βρίσκονται αποθηκευμένες ως γνωστόν στο DNA το οποίο είναι σε θέση να καθορίζει τόσο τη δομή όσο και τη λειτουργία αυτού. Η έκφραση της πληροφορίας ξεκινά με την μεταφορά αυτής προς το RNA, με τη γνωστή διαδικασία της αντιγραφής (14). Αφού αντιγραφεί η πληροφορία, το RNA είναι υπεύθυνο για τη μεταφορά της πληροφορίας στις πρωτεΐνες που με τη σειρά τους ρυθμίζουν τη δομή και τη λειτουργία των κυττάρων και κατ' επέκταση και των οργανισμών (14).



Εικόνα 4 : Δόγμα της σύγχρονης βιολογίας

Το παραπάνω σχήμα, αποτελεί τη βασική αρχή πάνω στην οποία στηρίζεται η σύγχρονη βιολογία, και με βάση αυτό παρατηρείται ότι η γενετική πληροφορία δεν αποτελεί τίποτε παραπάνω από μία προκαθορισμένη σειρά βάσεων (15). Αντίστοιχα, σε κάθε κομμάτι DNA, υπάρχει μία καθορισμένη σειρά γονιδίων τα οποία διαμέσου των διαδικασιών μεταγραφής και μετάφρασης καθορίζουν τη σειρά που θα καταλάβουν τα αμινοξέα στην πρωτεΐνη (15). Οι πορείες της μεταγραφής και της μετάφρασης των γονιδίων αποτελούν τη γονιδιακή έκφραση.

Λαμβάνοντας υπόψη τα όσα περιγράφηκαν παραπάνω μπορεί κανείς να συνοψίσει τις λειτουργίες του DNA στα παρακάτω (15):

- ***Αντιγραφή*** *: Διαιώνιση της γενετικής πληροφορίας*
- ***Μετάφραση*** *: Χρήση της πληροφορίας ώστε να κατασκευαστεί πολυπεπτίδιο*
- ***Μεταγραφή*** *: Καθορισμός της έκφρασης των γονιδίων σε συγκεκριμένους ιστούς και στάδια ανάπτυξης*

Τα γονίδια που περιέχονται στον ανθρώπινο οργανισμό διακρίνονται στις δύο παρακάτω κατηγορίες (16):

• Στα γονίδια που μεταγράφονται σε mRNA και μεταφράζονται στη συνέχεια σε πρωτεΐνες

• Στα γονίδια που μεταγράφονται και παράγουν tRNA, rRNA ,snRNA

Το απλοειδές ανθρώπινο γονιδίωμα έχει μήκος 3*109 ζεύγη βάσεων. Από αυτό, μικρό ποσοστό μεταγράφεται σε RNA, δηλαδή αποτελεί τα γονίδια.

Υπάρχουν τέσσερα διαφορετικά είδη μορίων RNA που παράγονται με μεταγραφή (16):

1. **Το αγγελιοφόρο RNA (mRNA)** : Μεταφέρει την πληροφορία που απαιτείται για τη δημιουργία της πολυπεπτικής αλυσίδας

2. **Το μεταφορικό RNA (tRNA)** : Συνδέεται με ένα συγκεκριμένο αμινοξύ και το μεταφέρει στη θέση της πρωτεινοσύνθεσης
3. **Το ριβοσωμικό RNA (rRNA)** : Συνδέεται με πρωτεΐνες και σχηματίζει τα ριβοσώματα τα οποία είναι απαραίτητα για την πρωτεινοσύνθεση
4. **Το μικρό πυρηνικό RNA (snRNA)** : Συνδέεται με την πρωτεΐνη και σχηματίζει μικρά ριβονουκλεοπρωτεϊνικά σωματίδια.

1.6.2Μεταλλάξεις και Πολυμορφισμοί του DNA

Η πρόκληση των διαφόρων γενετικών ανωμαλιών αλλά και η έλευση των διαφόρων ασθενειών προέρχεται από αλλαγές στην ακολουθία – αλληλουχία του DNA, οι οποίες επιστημονικά καλούνται μεταλλάξεις. Οι μεταλλάξεις αυτές δημιουργούν ένα διαφορετικό φαινότυπο DNA ο οποίος οφείλεται για όποια ενδεχόμενη ανωμαλία. Το αποτέλεσμα της μετάλλαξης αυτής επηρεάζει άμεσα στο γονιδιακό προϊόν, δηλαδή την πρωτεΐνη, με πολλούς και διαφορετικούς τρόπους (17).

Οι μεταλλάξεις έχουν διαχωριστεί από τους επιστήμονες σε δύο κατηγορίες, τις γονιδιακές που αφορούν σε προσθήκη, αντικατάστασή ή έλλειψη του αριθμού των βάσεων, και στις χρωμοσωμικές που αφορούν σε αλλαγές σε μεγάλο τμήμα του χρωμοσώματος και καλούνται χρωμοσωμικές ανωμαλίες (17).

Οι μεταλλάξεις στα γενετικά κύτταρα μπορούν να μεταβιβαστούν σε επόμενες γενικές και ευθύνονται αποκλειστικά και μόνο για την κληρονομικότητα πολλών ασθενειών. Υπεύθυνοι για τις μεταλλάξεις είναι πολλοί παράγοντες τόσο του εσωτερικού όσο και του εξωτερικού περιβάλλοντος.

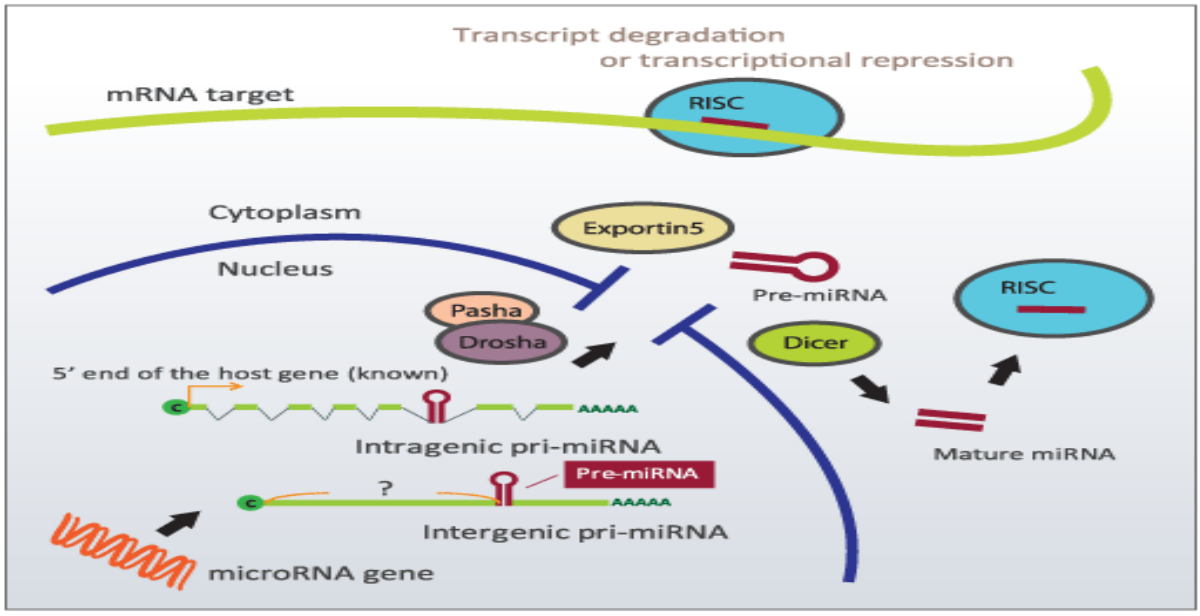
1.7 Το microRNA

Κάθε microRNA έχει μήκος 21-24 νουκλεοτίδια και είναι υπεύθυνο για τη γονιδιακή έκφραση στα ευκαρυωτικά κύτταρα διαμέσου της πρόσδεσης σε μία μη κωδικοποιημένη περιοχή των mRNA που καλείται 3’ UTR (3’ Αμετάφραστη Περιοχή) (18). Ο μηχανισμός λειτουργίας των microRNAs, τα βοηθά ώστε αυτά να είναι σε θέση να ρυθμίζουν λειτουργίες των κυττάρων όπως η αυτό – ανανέωση, η διαφοροποίηση αλλά και η διαίρεση αυτών διαμέσου της μεταγραφικής σίγησης των γονιδίων (18).

Τα πρώτα χρόνια εντοπισμού των microRNAs η ύπαρξη αυτών θεωρούνταν ελάσσονος σημασίας, τα τελευταία χρόνια όμως, έχουν συσχετιστεί άμεσα με τις διάφορες μορφές πολλών σοβαρών ασθενειών και ιδιαίτερα του καρκίνου (18). Ο αριθμός των mRNA που κωδικοποιούν τα microRNAs ποικίλει ανάλογα με τον οργανισμό στον οποίο αναφέρεται κανείς, συνήθως όμως είναι γύρω στα 300.

Η διαδικασία του σχεδιασμού των microRNAs λειτουργεί με τον παρακάτω τρόπο. Αρχικά, δημιουργείται ένα πρώιμο μετάγραφο miRNA, το οποίο με τη βοήθεια της Drosha, μετατρέπεται μέσα στον πυρήνα σε pre-miRNA (19). Το pre-miRNA έχει μορφή φουρκέτας-θηλιάς με μέγεθος περίπου 70-100 νουκλεοτιδίων που αναδιπλώνεται σε μια δομή με πολλαπλές προεκβολές (buldges) και εμφανίζει ατελή συμπληρωματικότητα (mismatch) σε πολλά σημεία (19). Το pre-miRNA μεταφέρεται στο κυτταρόπλασμα μέσω της εξπορτίνης 5 (exportin 5) και η μεταφορά του εξαρτάται από τη Ran-GTP (19) που ανήκει στις Ras-GTPάσες και συμμετέχει μεταξύ άλλων στις μεταφορές μορίων μέσα και έξω από την πυρηνική μεμβράνη. Φτάνοντας στο κυτταρόπλασμα, το pre-miRNA μετατρέπεται από το ένζυμο Dicer σε miRNA που έχει μέγεθος 21-23 νουκλεοτίδια.

Η βιογένεση των miRNAs είναι επίσης πολύ εξειδικευμένη διαδικασία, αφού εξαρτάται από το εκάστοτε miRNA που εκφράζεται. Αποκλίσεις στο επίπεδο της μεταγραφής περιλαμβάνουν τα εξής: Πρώτον, τα γονίδια των miRNAs μπορεί να μεταγράφονται είτε από την RNA πολυμεράση II ή την RNA πολυμεράση III, αφού κάθε μια αναγνωρίζει ειδικούς υποκινητές και περιοχές λήξης και υφίσταται ειδική ρύθμιση (19). Δεύτερον, η έκφραση μπορεί επιπρόσθετα να ρυθμιστεί από μεταγραφικούς παράγοντες όπως το c-Myc ή το p53. Τέλος, υπάρχουν αποκλίσεις και στη διαδικασία πέψης (19).



Εικόνα 5 : Το Biogenesis του MicroRNA (20)

Η δράση του συμπλόκου των Drosha-DGCR8 (αντίστοιχος της Pasha στην Drosophila melanogaster) δίνει γένεση στο pre-miRNA, ένα μόριο μήκους 60−70 νουκλεοτιδίων με δομή μίσχου-θηλιάς που εξάγεται ταχύτατα στο κυτταρόπλασμα από την εξπορτίνη-5 με μια εξαρτώμενη από GTP διαδικασία (19). Το ώριμο miRNA βρίσκεται στην 5' ή στην 3' πλευρά του μίσχου του pre-miRNA. Κάποιες φορές και τα δύο τμήματα παράγουν ώριμα miRNAs. Αφού το premiRNA βρεθεί στο κυτταρόπλασμα, μια δεύτερη RNάση III, η Dicer, δρα στο pre-miRNA για την απελευθέρωση ενός δίκλωνου miRNA 22 νουκλεοτιδίων, στο

οποίο το ώριμο miRNA είναι εν μέρει συνδεδεμένο με το συμπληρωματικό του κλώνο (miRNA*) μαζί με τον οποίο συνιστά το μίσχο (19). Συνήθως, μόνο ο κλώνος miRNA (ώριμο miRNA) του δίκλωνου μορίου miRNA-miRNA* είναι ενεργός και εισέρχεται σε ένα εξειδικευμένο πρωτεϊνικό σύμπλοκο, το RISC, για να καταστείλει τη γονιδιακή έκφραση (19).

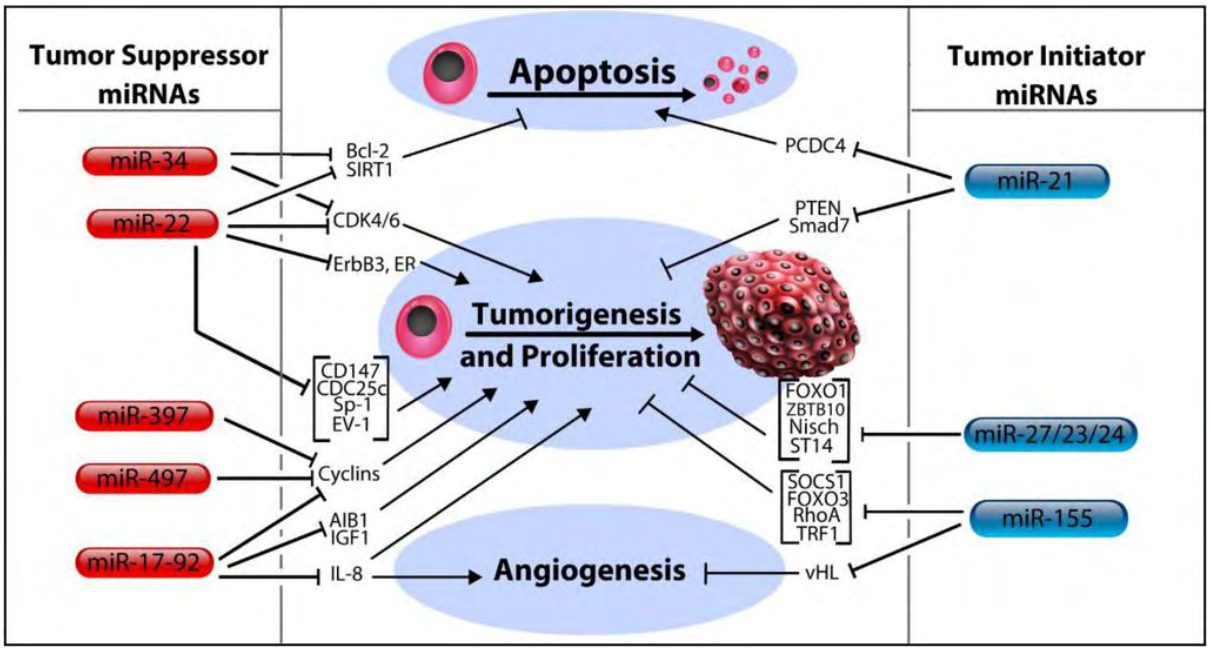
Στο σημείο αυτό εισάγεται η λειτουργία των micoRNAs που έχουν ως κύριο στόχο να καταστείλουν τη μεταγραφή και το επιτυγχάνουν διαμέσου τη πρόσδεσής του στο mRNA του υπό καταστολή γονιδίου. Αποτέλεσμα της λειτουργίας αυτή είναι η περαιτέρω μετάφραση του mRNA σε πρωτείνη. Η αλληλεπίδραση αυτή λαμβάνει χώρα στην 3΄ μη μεταφραζόμενη περιοχή (3΄UTR) του mRNA του γονιδίου ενώ ο ακριβής μηχανισμός με τον οποίο παρεμποδίζεται η μετάφραση, εξακολουθεί να παραμένει άγνωστος.

Έτσι λοιπόν, όπως συμπεραίνεται, τα microRNAs είναι οι βασικοί υπεύθυνοι για τη ρύθμιση της γονιδιακής έκφρασης καθώς είτε αναστέλλουν τη μετάφραση, είτε προωθούν την αποδόμηση συγκεκριμένων mRNAs.

Δεδομένων των παραπάνω ανακαλύψεων σχετικά με τη λειτουργία των microRNAs, τα τελευταία χρόνια έχει γίνει ιδιαίτερα εντατική μελέτη γύρω από την αποσαφήνιση του ρόλου τους. Ιδιαίτερα σημαντικός φαίνεται να είναι ο ρόλος των microRNAs σε ότι αφορά στον κυτταρικό θάνατο. Ο αριθμός των microRNAs που ανακαλύπτονται καθημερινά αυξάνεται διαρκώς ενώ ιδιαίτερα εντατική είναι και η προσπάθεια εντοπισμού της λειτουργίας τους σε κάθε περίπτωση (15).

Τα microRNAs συναντώνται μονάχα σε πολυκύτταρους οργανισμούς γεγονός που αυξάνει τη σημαντικότητά τους σε ότι αφορά τη διαφοροποίηση των τύπων και των ιστών. Τα αδιαφοροποίητα κύτταρα ή τα κύτταρα χαμηλής διαφοροποίησης δεν απαιτούν σε καμία περίπτωση την ύπαρξη των microRNAs ώστε να είναι σε θέση να επιβιώσουν (21).

Από την περιγραφή της λειτουργίας των microRNAs δε θα μπορούσε να απουσιάζει ο σημαντικός ρόλος που αυτά διαδραματίζουν στο μετασχηματισμό κακοηθών όγκων και στην εμφάνιση του καρκίνου στον ανθρώπινο οργανισμό. Τα γονίδια τα οποία εμπλέκονται στον καρκίνο είναι συνήθως μεταλλαγμένες μορφές φυσιολογικών γονιδίων που έχουν ενεργοποιηθεί ή απενεργοποιηθεί και είναι γνωστά ως ογκογονίδια ή ογκοκατασταλτικά γονίδια, αντίστοιχα. Κάθε καρκινικός όγκος, αποτελεί έναν ετερογενή πληθυσμό κυττάρων με διαφορές σε ότι αφορά στα στάδια της διαφοροποίησής τους (21).



Εικόνα 6 : MicroRNAs και καρκίνος (22)

Στον πίνακα που ακολουθεί, εμφανίζονται κάποια από τα πλέον γνωστότερα microRNAs αλλά και ο τύπος του καρκίνου στον οποίο αυτά διαδραματίζουν σημαντικό ρόλο.

MicroRNA	Τύπος Καρκίνου
let-7-a-2	Καρκίνος μαστού, πνεύμονα, ήπατος
let-7-a-3	Καρκίνος μαστού, ήπατος
let-7d	Καρκίνος μαστού, ήπατος
let-7f	Καρκίνος μαστού, ήπατος, θυρεοειδούς

miR-101	Καρκίνος πνεύμονα, μαστού, αδένωμα
miR-102	Καρκίνος μαστού, θυρεοειδούς
miR-124a	Καρκίνος πνεύμονα, ήπατος, αδένωμα
miR-125b-1	Καρκίνος μαστού
miR-125b-1	Καρκίνος θυρεοειδούς
miR-125b-2	Καρκίνος μαστού
miR-125b-2	Καρκίνος θυρεοειδούς
miR-140	Καρκίνος πνεύμονα, μαστού, θυρεοειδούς
miR-141	Καρκίνος ήπατος, θυρεοειδούς
miR-142	Καρκίνος ήπατος, αδένωμα υπόφυσης
miR-143	Καρκίνος μαστού, πνεύμονα, ήπατος
miR-145	Καρκίνος μαστού, πνεύμονα, ήπατος
miR-146	Καρκίνος πνεύμονα, θυρεοειδούς
miR-150	Καρκίνος πνεύμονα, αδένωμα υπόφυσης
miR-150	Καρκίνος πνεύμονα, αδένωμα υπόφυσης
miR-155	Καρκίνος πνεύμονα, μαστού, θυρεοειδούς, λέμφωμα
miR-15b	Καρκίνος θυρεοειδούς, αδένωμα υπόφυσης
miR-15b	Καρκίνος θυρεοειδούς, αδένωμα υπόφυσης
miR-181a	Καρκίνος ήπατος
miR-181a	Καρκίνος θυρεοειδούς
miR-181b	Γλοιοβλάστωμα, αδένωμα υπόφυσης
miR-181c	Καρκίνος ήπατος, γλοιοβλάστωμα
miR-181c	Καρκίνος θυρεοειδούς
miR-191	Καρκίνος πνεύμονα, μαστού, αδένωμα υπόφυσης
miR-192	Καρκίνος πνεύμονα, αδένωμα υπόφυσης
miR-198	Καρκίνος πνεύμονα, γλοιοβλάστωμα
miR-199b	Καρκίνος πνεύμονα, ήπατος
miR-202	Καρκίνος μαστού, θυρεοειδούς

miR-203	Καρκίνος πνεύμονα, μαστού
miR-21	Καρκίνος μαστού, πνεύμονα, θυρεοειδούς, γλοιοβλάστωμα
miR-210	Καρκίνος πνεύμονα, μαστού
miR-212	Καρκίνος πνεύμονα, αδένωμα υπόφυσης
miR-213	Καρκίνος μαστού, θυρεοειδούς
miR-219-1	Καρκίνος πνεύμονα, θυρεοειδούς
miR-220	Καρκίνος πνεύμονα
miR-220	Καρκίνος πνεύμονα
miR-220	Καρκίνος θυρεοειδούς
miR-221	Γλοιοβλάστωμα, καρκίνος ήπατος,
miR-222	Καρκίνος θυρεοειδούς
miR-24-2	Καρκίνος πνεύμονα, θυρεοειδούς

Πίνακας 1 : MicroRNAs και τύποι καρκίνου

Κεφάλαιο 2

2.1 Βιολογία και Βιοπληροφορική

Ο όρος βιολογία προέρχεται από τη σύνθεση των όρων *βίος* = ζωή και *λόγος* = διήγηση, εξήγηση, λογική και ως επιστήμη εξετάζει τη δημιουργία των διαφόρων ειδών αλλά και των μεμονωμένων μελών τους ενώ παράλληλα μελετά της αλληλεπιδράσεις αυτών με το περιβάλλον. Η επιστήμη αυτή, περιλαμβάνει ένα ευρύ φάσμα επιστημονικών πεδίων που συχνά αντιμετωπίζονται ως ανεξάρτητες ειδικεύσεις, στο σύνολό τους όμως εξετάζουν το φαινόμενο της ζωής σε όλα τα είδη των οργανισμών.

Τα τελευταία χρόνια, η επιστήμη αυτή αναπτύσσεται με φρενήρεις ρυθμούς. Απαρχή της ανάπτυξης αυτής αποτέλεσε η αναγνώριση του DNA ως γενετικό υλικό από τους ερευνητές Κρικ (Crick) και Γουότσον αλλά και η πλήρης κατανόηση της λειτουργίας του και των μηχανισμών γύρω από αυτό (23). Η παραπάνω εξέλιξη, συνδυάστηκε με την δημιουργία ενός νέου και ανερχόμενου επιστημονικού πεδίου που καλείται Βιοπληροφορική και δεν αποτελεί τίποτα περισσότερο από την συνεργασία δύο επιστημών, αυτής της Βιολογίας και αυτής της Πληροφορικής.

Ως Βιοπληροφορική λοιπόν, ορίζεται η επιστήμη εκείνη που ασχολείται με την εφαρμογή υπολογιστικών τεχνικών και μεθόδων στην προσπάθεια κατανόησης και οργάνωσης των δεδομένων και πληροφοριών που σχετίζονται με τα βιολογικά μακρομόρια (24)**.**

Ο τεράστιος πλέον όγκος δεδομένων που εξάγεται καθημερινά από την επιστήμη της βιολογίας, απαιτεί τη συνεργασία αυτής με την επιστήμη της πληροφορικής ώστε να μπορεί ο όγκος αυτός να επεξεργαστεί, να αναλυθεί, να αποθηκευθεί αλλά και να παράγει διάφορα χρήσιμα συμπεράσματα. Η συνεργασία των δύο αυτών επιστημονικών πεδίων, βασίζεται

κυρίως στο γεγονός ότι η ίδια η ζωή περιλαμβάνει έναν τεράστιο όγκο πληροφοριών που υπόκεινται σε καθημερινή επεξεργασία.

Η επιστήμη της Βιοπληροφορικής, αντιμετωπίζει τα διάφορα βιολογικά δεδομένα ως πληροφορία, και χρησιμοποιεί αλγόριθμους ώστε να επεξεργαστεί τα δεδομένα αυτά αλλά και να εξάγει χρήσιμα συμπεράσματα σχετικά με τη λειτουργία τους. Τα βιολογικά μόρια, όπως το DNA, το RNA και οι πρωτεΐνες μπορούν να θεωρηθούν ως ακολουθίες συμβόλων, δηλαδή συμβολοσειρές (24). Για παράδειγμα, το DNA μπορεί να θεωρηθεί ως μια ακολουθία χιλιάδων νουκλεοτιδίων ή βάσεων. Υπάρχουν τέσσερα είδη βάσεων και αντίστοιχα τέσσερα είδη νουκλεοτιδίων:

1. Αδενίνη
2. Θυμίνη
3. Γουανίνη
4. Κυτοσίνη

Εάν λοιπόν αντιστοιχηθεί κάθε μία από τις βάσεις αυτές με ένα σύμβολο, όπως για παράδειγμα «A» για την Αδενίνη, «T» για τη Θυμίνη, «G» για τη Γουανίνη και «C» για την Κυτοσίνη, τότε μπορούν να παραχθούν ακολουθίες συμβολοσειρών που θα είναι της μορφής AAGATCGGTACGGTAAAGAT, θα είναι ιδιαίτερα εύκολα επεξεργάσιμες και θα βοηθούν στην εξαγωγή χρήσιμων συμπερασμάτων και άρα ιδιαίτερα χρήσιμης πληροφορίας για τη λειτουργία του DNA (24).

Πέραν τούτου, η επιστήμη της βιοπληροφορικής δίνει στους επιστήμονες τη δυνατότητα της εκτέλεσης διαφόρων πειραμάτων, τα οποία διαφορετικά παρουσιάζουν δυσκολίες στο να εκτελεστούν, αλλά και τη δυνατότητα εξαγωγής ιδιαίτερα σημαντικών συμπερασμάτων από αυτά. Η εξέλιξη της επιστήμης μάλιστα, δίνει τη δυνατότητα απεικόνισης των πειραμάτων αυτών αλλά και των διαφόρων αποτελεσμάτων ώστε αυτά να είναι εμπεριστατωμένα και

πλήρως κατανοητά. Μάλιστα, η απεικόνιση των γονιδιακών μεταλλάξεων αναμένεται να συντελέσει σημαντικά στην εξέλιξη της εξατομικευμένης ιατρικής, στη στοχευόμενη παρέμβαση και τελικά στην αντιμετώπιση των διαφόρων ασθενειών με βάση πάντα το περιστατικό και όχι τη γενικευμένη γνώση.

Ιδιαίτερα σημαντικός τομέας της Βιοπληροφορικής δε θα μπορούσε παρά να είναι και αυτός της δημιουργίας τεραστίων βάσεων δεδομένων που περιέχουν πληροφορίες σημαντικές σχετικά με το ανθρώπινο γονιδίωμα αλλά και τις μεταλλάξεις αυτού.

Λαμβάνοντας υπόψη τα όσα περιγράφηκαν παραπάνω, θα μπορούσε κανείς να συνοψίσει τις άπειρες εφαρμογές της Βιοπληροφορικής στις παρακάτω (25):

- Ανεύρεση της λειτουργίας των πρωτεϊνών και ομαδοποίηση αυτών σε λειτουργικές ομάδες
- Ανεύρεση αλληλεπιδράσεων των πρωτεϊνών μεταξύ τους και κατανόηση της πολυπλοκότητας των βιολογικών συστημάτων
- Σύγκριση του γονιδιώματος διαφόρων ειδών
- Εύρεση των εξελικτικών σχέσεων των οργανισμών μεταξύ τους
- Απόκτηση γνώσης για το ρόλο των μη κωδικοποιημένων περιοχών του DNA στη μορφολογία και έκφραση των γονιδίων
- Προσπάθεια αντιμετώπισης διαφόρων ασθενειών με την ανάπτυξη νέων διαγνωστικών μέτρων και θεραπευτικών μεθόδων

Το μέλλον του επιστημονικού πεδίου της Βιοπληροφορικής αναμένεται ιδιαίτερα λαμπρό ενώ οι απαιτήσεις από αυτό αυξάνονται καθημερινά. Αναμένονται λοιπόν από τον τομέα της Βιοπληροφορικής όλα εκείνα τα λογισμικά, τα οποία θα παρέχουν πλήρη απεικόνιση της λειτουργίας των πρωτεϊνών αλλά και της έκφραση των γονιδίων ενώ ταυτόχρονα θα

προσφέρουν πληρέστερη διεξαγωγή συμπερασμάτων για τους τομείς του οποίους εκπροσωπούν.

2.2 Τομείς έρευνας που εντοπίζονται στο επιστημονικό πεδίο της Βιοπληροφορικής

- ***Ανάλυση της αλληλουχίας του γονιδιώματος***

Η εύρεση της αλληλουχίας του DNA συντελείται έπειτα από αναζήτηση αυτού στις τεράστιες βάσεις δεδομένων που έχουν δημιουργηθεί και αφού φυσικά γίνει ανάλυση των δεδομένων αυτών. Μονάχα μία απλή σύγκριση ανάμεσα σε γονίδια ιδίου ή διαφορετικών ειδών είναι αρκετή ώστε να αποδείξει ομοιότητες μεταξύ των πρωτεϊνικών λειτουργιών ή σχέσεις μεταξύ των ειδών.

- ***Επισημείωση γονιδιώματος***

Πρόκειται για τη διαδικασία συμπλήρωσης της σήμανσης αλλά και άλλων βιολογικών χαρακτηριστικών στην ακολουθία του DNA. Το πρώτο λογισμικό που δημιουργήθηκε για τη διαδικασία αυτή, σχεδιάστηκε το 1995 και έχει ως βασική λειτουργία τον εντοπισμό των γονιδίων που βρίσκονται τοποθετημένα σε μία ακολουθία DNA.

- ***Συγκριτική ανάλυση γονιδιώματος***

Πρόκειται για τη μέθοδο εκείνη η οποία μελετά την ανάλυση του γονιδιώματος διαμέσου της ανταλλαγής πληροφορίας αυτού με άλλα γονίδια. Η μέθοδος αυτή, αφορά στην ύπαρξη χαρτών, που καθιστούν δυνατό τον εντοπισμό των εξελικτικών διαδικασιών που είναι υπεύθυνες για την απόκλιση δυο γονιδιωμάτων.

- ***Ανάλυση της έκφρασης των γονιδίων***

Η έκφραση των γονιδίων εξαρτάται από πολλούς παράγοντες και καθορίζεται με χρήση πολλών τεχνικών που είναι ιδιαίτερα επιρρεπής στο θόρυβο και εξαρτώνται πάντα από τη

βιολογική μέτρηση. Έτσι λοιπόν, ένα σημαντικό κομμάτι του επιστημονικού πεδίου της Βιοπληροφορικής είναι ο διαχωρισμός του θορύβου αυτού από το πραγματικό σήμα που παράγεται από την έκφραση του DNA. Οι μελέτες του είδους αυτού αφορούν συνήθως στα γονίδια τα οποία εμπλέκονται σε διαταραχές, όπως για παράδειγμα ο καρκίνος, και στοχεύουν στην ανεύρεση των μεταλλάξεων εκείνων που οδήγησαν στην εμφάνιση της νόσου.

- ***Ανάλυση της έκφρασης των πρωτεϊνών και πρόβλεψη της δομής τους***

Ο τομέας της Βιοπληροφορικής ασχολείται ιδιαίτερα με τη σύνθεση της πρωτεΐνης μικροσυστοιχιών που απευθύνονται σε mRNA. Παράλληλα, καταβάλλει σημαντικές προσπάθειες ώστε να κατορθώσει να προβλέψει τη δομή αυτών.

2.3 Βάσεις δεδομένων που χρησιμοποιούνται στη Βιοπληροφορική

Όπως αναφέρθηκε στις παραπάνω ενότητες, η λειτουργία των microRNAs αλλά και η αλληλεπίδραση αυτών με τα mRNAs είναι ιδιαίτερα σημαντική σε ότι αφορά στην ανάπτυξη ή την καταστολή των διαφόρων ασθενειών καθώς από την αλληλεπίδραση αυτή είτε διακόπτεται η μετάφραση σε πρωτεΐνη είτε καταστρέφεται η μεταγραφή του mRNA. Έτσι, έχουν δημιουργηθεί αρκετές βάσεις δεδομένων που περιέχουν ιδιαίτερα σημαντικό όγκο δεδομένων σχετικά με τη γονιδιακή ακολουθία αλλά και τις μεταλλάξεις που συντελούνται.

Η πληροφορία που υπάρχει καταγεγραμμένη στις βάσεις αυτές έχει συλλεχθεί ύστερα από έντονη προσπάθεια και με πολλούς και διαφορετικούς τρόπους. Αρχικά, η συλλογή των δεδομένων ήταν μία αρκετά χρονοβόρα διαδικασία καθώς προερχόταν από ανάγνωση διαφόρων επιστημονικών άρθρων, εξαγωγή όλης της απαραίτητης πληροφορίας από αυτά και μετέπειτα καταχώρηση των δεδομένων αυτών στη βάση.

Η εξέλιξη των επιστημονικών πεδίων της βιολογίας και της πληροφορικής βοήθησε σημαντικά τη συλλογή και την καταχώρηση των δεδομένων καθώς εμφανίστηκαν στο προσκήνιο νέες μέθοδοι ανάκλησης πληροφορίας. Ιδιαίτερα γνωστή στο χώρο κατέστη η μέθοδος PAR – CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) η οποία ανακαλύφθηκε από τον Markus Hafner και τους συνεργάτες του (26). Η μέθοδος αυτή, βασίζεται στην ενσωμάτωση φωτοαντιδραστικών ριβονουκλεοτιδικών αναλόγων, όπως η 4-θειουριδίνη (4-SU) και η 6-θειογουανοσίνη (6-SG) σε RNA μεταγραφές ζωντανών κυττάρων. Τα κύτταρα αυτά υπόκεινται σε υπεριώδη ακτινοβολία 365nm η οποία κατορθώνει να ενεργοποιήσει τα φωτοαντιδραστικά ριβονουκλεοτιδικά RNA και να τα κάνει να αλληλεπιδράσουν με τα RBPs (RNA-Binding Proteins) (26). Η ανοσοκαταβύθηση των προς μελέτη RBPs, ακολουθείται από απομόνωση του διασταυρωμένου RNA το οποίο στη συνέχεια καταχωρείται σε μία cDNA βιβλιοθήκη αφού έχει υποστεί επεξεργασία με την τεχνολογία Solexa. Ιδιαίτερο χαρακτηριστικό των cDNA βιβλιοθηκών που προέρχονται από τη μέθοδο PAR CLIP αποτελεί το γεγονός ότι η ακριβής θέση της αλληλεπίδραση μπορεί να προσδιοριστεί από τις διάφορες μεταλλάξεις που υφίστανται στις αλληλουχίες cDNA (26). Όταν χρησιμοποιείται η 4-SU, η μετάβαση γίνεται από την θυμιδίνη στην κυτοσίνη ενώ όταν χρησιμοποιείται η 6-SG η γουνοασίνη οδηγεί σε μεταλλάξεις της αδενοσίνης. Η παρουσία των μεταλλάξεων διασταυρώνεται και έτσι είναι δυνατό να διαχωριστούν αυτές από τις ακολουθίες που προέρχονται από τα διάφορα κυτταρικά RNAs (26).

Μία ακόμη διάσημη μέθοδος στο χώρο αυτό είναι και η HITS – CLIP (High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation) (27). Πρόκειται για μία μέθοδο που έχει τη δυνατότητα χαρτογράφησης των σημείων πρόσδεσης των πρωτεϊνών και των RNA in vivo. Αρχικά χρησιμοποιήθηκε ώστε να διευκολύνει τη χαρτογράφηση της αλληλεπίδρασης της πρωτεΐνης με το RNA για συγκριμένα σημεία πρόσδεσης και

συγκεκριμένους παράγοντες διαχωρισμού όπως οι Nova1 και Nova2 (27). Στη συνέχεια, προστέθηκε στη μέθοδο αυτή η χαρτογράφηση ενός σημαντικού αριθμού παραγόντων διαχωρισμού, συμπεριλαμβανομένων αυτών που αφορούν σε PTB, RbFox2, SFRS1, και hnRNP C (27).

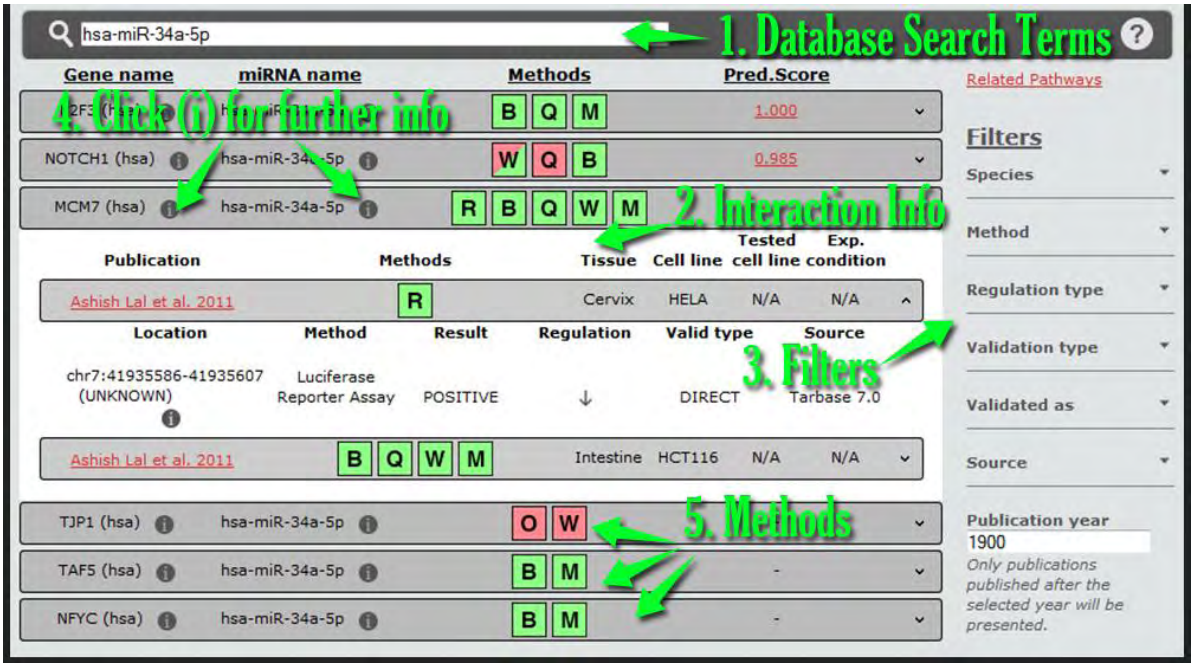
Η μέθοδος HITS CLIP χρησιμοποιήθηκε στην πρωτεΐνη Αργοναύτης με στόχο να αναγνωρίσει τους στόχους των microRNAs αποκωδικοποιώντας τους χάρτες αλληλεπίδρασης για microRNA-mRNA και RNA-πρωτεΐνης στον εγκέφαλο των ποντικιών και στη συνέχεια στην Caenorhabditis elegans, στα εμβρυικά βλαστικά κύτταρα και στα κύτταρα των ιστών (27).

Με βάση τις παραπάνω μεθόδους λοιπόν, όποιες από αυτές και εάν χρησιμοποιήθηκαν, δημιουργήθηκε ένας αριθμός βάσεων δεδομένων, στις οποίες υπάρχει σημαντική πληροφορία για τη δράση των microRNAs. Κάθε μία από τις βάσεις αυτές, περιέχει ξεχωριστό περιεχόμενο το οποίο έχει συλλέξει με τις παραπάνω μεθόδους και προσφέρει σημαντική πληροφορία για τη σύγχρονη επιστήμη. Η σημαντικότητα τους είναι ιδιαίτερα μεγάλη καθώς συντελούν στη διαφύλαξη ιδιαίτερα σημαντικής πληροφορίας η οποία χωρίς την ύπαρξή τους θα είχε απλά χαθεί. Πέραν τούτου, οι βάσεις αυτές δεδομένων, συντελούν σημαντικά στην εξέλιξη της επιστήμης καθώς αποτελούν το βάθρο πάνω στο οποίο στηρίζεται η επιστήμη ώστε να σχεδιάσει τα επόμενα βήματά της.

Η πλέον διαδεδομένη βάση δεδομένων στο χώρο της Βιοπληροφορικής είναι η DIANA – TarBase η οποία δημοσιεύθηκε για πρώτη φορά το 2005. Πρόκειται για την πρώτη βάση δεδομένων που είχε ως στόχο συμπεριλάβει όλες τις έως τότε γνωστές αλληλεπιδράσεις ανάμεσα σε microRNAs και γονίδια. Από τότε, ξεκίνησαν πολλές και σοβαρές προσπάθειες προς την κατεύθυνση αυτή οι οποίες προσπάθησαν διαμέσου της ανάγνωσης των επιστημονικών άρθρων να συλλέξουν την απαραίτητη πληροφορία, να την επεξεργαστούν και στη συνέχεια να την καταχωρήσουν σε βάση δεδομένων.

Η έκτη έκδοση της DIANA – TarBase άνοιξε νέους ορίζοντες για τον τομέα των βάσεων δεδομένων καθώς ήταν η πρώτη που εισήγαγε στη συλλογή των δεδομένων της καινοτόμες μεθοδολογίες πειραμάτων όπως η CLIP-Seq (28). Επιπλέον, μεγάλωσε σημαντικά το διαθέσιμο χώρο στο στόχο για 65000 αλληλεπιδράσεις που είχαν καταχωρηθεί και επεξεργαστεί χειροκίνητα (28). Η σημαντική αυτή αύξηση αποτέλεσε προοίμιο των νέων μεθόδων υψηλού throughput.

Στόχος της επόμενης έκδοσης της TarBase, της TarBase v7.0 είναι να θέσει τον πήχη ακόμη υψηλότερα, και να παρέχει για πρώτη φορά γνώση για χιλιάδες αλληλεπιδράσεις ανάμεσα στα microRNAs και τα γονίδια που συνοδεύονται από έναν μοναδικό αριθμό χαρακτηριστικών και τα πλέον αξιόπιστα δεδομένα (28). Στη βάση αυτή, συμπεριλαμβάνονται τη στιγμή αυτή περισσότερες από μισό εκατομμύριο καταχωρήσεις, γεγονός που την καθιστά από 9 έως και 250 φορές μεγαλύτερη από οποιαδήποτε άλλη βάση δεδομένων (28). Κάθε καταχώρηση στη βάση δεδομένων συνοδεύεται από σειρά πληροφοριών που μπορούν να χρησιμοποιηθούν ως φίλτρα αναζήτησης στο ιδιαίτερα φιλικό προς το χρήστη περιβάλλον χρήσης. Για παράδειγμα, η TarBase v7.0 είναι πλέον σε θέση να συλλέξει δεδομένα όπως η έκθεση σε στρεσσογόνους παράγοντες ή φάρμακα τα οποία μπορούν να τροποποιήσουν τα συνήθη δίκτυα microRNA (28). Ένα άλλο σημαντικό χαρακτηριστικό της βάσης αυτής είναι η ικανότητά της να συμπεριλαμβάνει λεπτομερή πληροφορία σχετικά με την πειραματική μεθοδολογία που ακολουθήθηκε ώστε να αναγνωριστεί η κάθε αλληλεπίδραση.

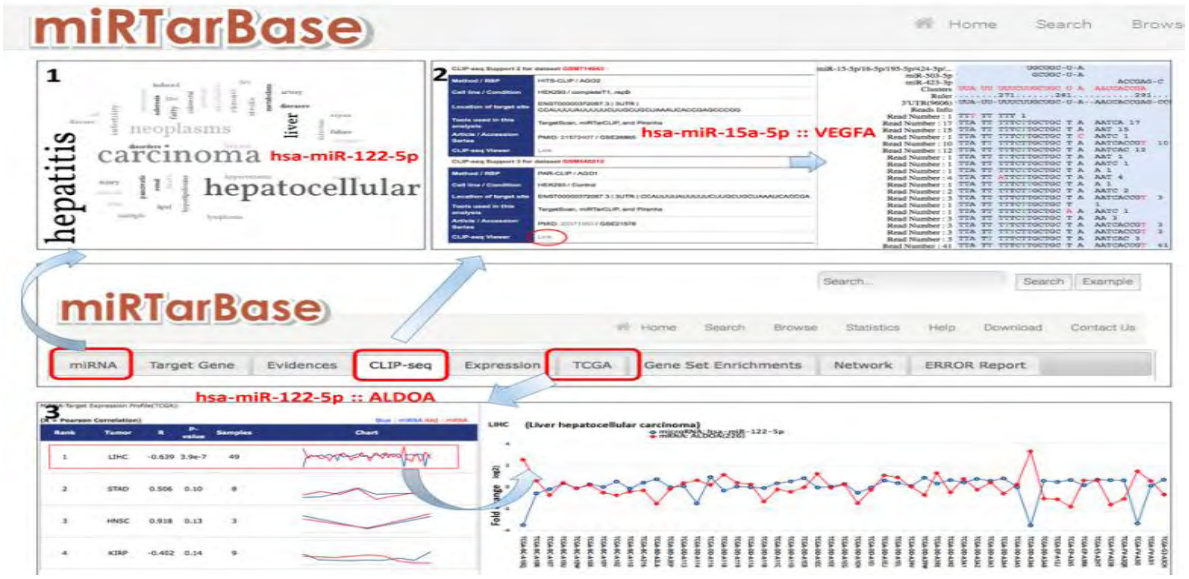


Εικόνα 7 : Η TarBAse **(28)**

Η βάση περιλαμβάνει περισσότερες από μισό εκατομμύριο αλληλεπιδράσεις που αφορούν σε 24 είδη. Τα δεδομένα που περιλαμβάνονται σε αυτή έχουν επαληθευθεί με 28 διαφορετικές πειραματικές μεθόδους, σε 356 τύπους κυττάρων και 59 διαφορετικούς ιστούς (28).

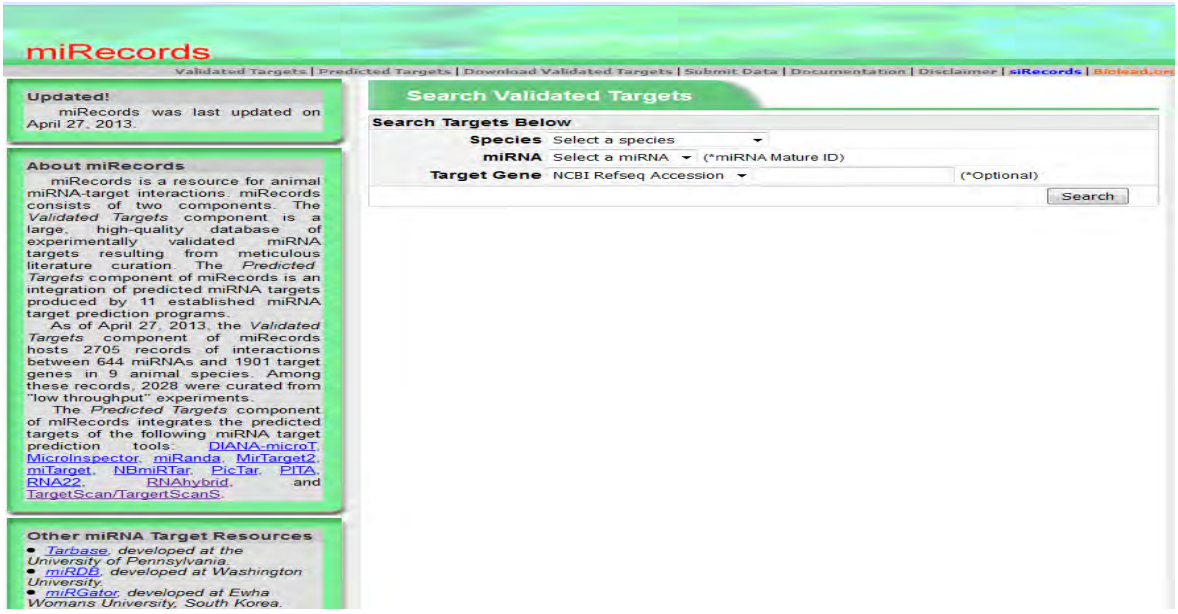
Μία άλλη ιδιαίτερα γνωστή βάση δεδομένων στο χώρο της Βιοπληροφορικής είναι η miRTarBase. Η τελευταία έκδοση της, η έκδοση 6, που κυκλοφόρησε το Σεπτέμβριο του 2015, περιλαμβάνει 366.181 αλληλεπιδράσεις ανάμεσα 3786 microRNAs και 22.563 γονίδια στόχους που έχουν συλλεγεί από 4966 επιστημονικά άρθρα (29). Ο αριθμός των αλληλεπιδράσεων που περιλαμβάνονται στη βάση έχει αυξηθεί σημαντικά από το 2004 και την έκδοση miRTarBase v4.5 (29).

Ο σημαντικά μεγαλύτερος αριθμός αλληλεπιδράσεων στη νέα έκδοση της βάσης οφείλεται στην CLIP-seQ ανάλυση των ομάδων δεδομένων και στις μεθόδους που αυτή χρησιμοποιεί. Οποιοδήποτε δεδομένο ήταν δυνατό να εξαχθεί από τα επιστημονικά άρθρα συλλέχθηκε και αμέσως καταχωρήθηκε στη βάση.



Εικόνα 8 : Η miRTarBase **(29)**

Η miRecords αποτελεί μία βάση δεδομένων η οποία αναφέρεται αποκλειστικά στις αλληλεπιδράσεις ανάμεσα στα microRNAs και τα γονίδια που συντελούνται στα ζώα. Η βάση αυτή περιλαμβάνει δεδομένα που έχουν εισαχθεί χειροκίνητα και έπειτα από εκτενή έρευνα στα διάφορα επιστημονικά άρθρα και στις μεθόδους που χρησιμοποιήθηκαν σε κάθε περίπτωση. Κάθε εγγραφή στη βάση, υποστηρίζεται από την αντίστοιχη μέθοδο που χρησιμοποιήθηκε και την πειραματική διαδικασία που ακολουθήθηκε. Η τρέχουσα έκδοση περιλαμβάνει 1135 καταχωρήσεις αλληλεπιδράσεων ανάμεσα σε 301 microRNAs και 902 γονίδια στόχους σε επτά είδη ζώων (30). Οι στόχοι των οποίων κάνει πρόβλεψη η miRecords έχουν επαληθευθεί από 11 διαφορετικά προγράμματα πρόβλεψης στόχων. Στόχος της βάσης αυτής είναι να αποτελέσει μία χρήσιμη πηγή πληροφοριών τόσο για τους βιολόγους όσο και για τους επιστήμονες της πληροφορικής που στοχεύουν στη δημιουργία νέων, ταχύτερων και πιο αποτελεσματικών προγραμμάτων πρόβλεψης στόχων για τα microRNAs (30).



Εικόνα 9 : Η miRecords (30)

Μία ακόμη βάση που χρησιμοποιείται στο χώρο της Βιοπληροφορικής είναι η miR2Disease που έχει ως στόχο να αποτελέσει μία πηγή που θα παρέχει χρήσιμη και ουσιαστική πληροφορία για την απορύθμιση των microRNAs στις διάφορες ασθένειες στον άνθρωπο. Η τρέχουσα έκδοση της miR2Disease περιλάμβανε 1939 αλληλεπιδράσεις ανάμεσα σε 299 microRNAs και 94 ασθένειες (31). Η καταχώρηση των δεδομένων έγινε ύστερα από έρευνα σε 600 δημοσιευμένα επιστημονικά άρθρα (31).

Κάθε καταχώρηση που γίνεται στην miR2Disease, περιέχει λεπτομερή πληροφορία αναφορικά με τη σχέση του microRNA και της εκδηλωμένης ασθένειας, συμπεριλαμβανομένου του microRNA ID, του ονόματος της ασθένειας, της σύντομης περιγραφής της σχέσης ανάμεσα στην ασθένεια και το microRNA, της μεθόδου ανίχνευσης της έκφρασης του microRNA, του γονιδίου στόχου και τέλος της βιβλιογραφικής αναφοράς (31). Το περιβάλλον του χρήστη είναι ιδιαίτερα φιλικό και προσφέρει τρεις επιλογές αναζήτησης που αφορούν στο όνομα της ασθένειας, στο ID του microRNA και στο γονίδιο στόχο. Επιπλέον, παρέχεται η δυνατότητα στους ερευνητές να καταχωρήσουν οποιαδήποτε

αλληλεπίδραση ανάμεσα σε microRNA και ασθένεια η οποία απουσιάζει από τη βάση. Φυσικά, για να πραγματοποιηθεί κάτι τέτοιο, θα πρέπει να υπάρξει σχετική έγκριση από την επιτροπή οι οποία ελέγχει τις εγγραφές και τις καταχωρεί στη βάση δεδομένων.

Εικόνα 10 : Η MiR2Disease (31)

2.4 Πειραματικές μέθοδοι

Η καταχώρηση των διαφόρων αλληλεπιδράσεων ανάμεσα σε microRNAs και γονίδια φυσικά υποστηρίζεται και από πειραματικές μεθόδους οι οποίες μάλιστα αποτελούν την πλέον αξιόπιστη πηγή επιβεβαίωσης της ύπαρξης της αλληλεπίδρασης αλλά και εξαγωγής συμπερασμάτων. Η πλέον γνωστή πειραματική μέθοδος είναι αυτή της PCR. Πρόκειται για μία ιδιαίτερα γρήγορη και αξιόπιστη μέθοδο που έχει ως στόχο των πολλαπλασιασμό των τμημάτων του DNA. Ο πολλαπλασιασμός αυτός είναι ιδιαίτερα σημαντικός καθώς αποτελεί τον τρόπο με τον οποίο γίνεται μετέπειτα ανάλυση των μεταλλάξεων ή των πολυμορφισμών που εντοπίζονται στο DNA (32). Ακριβώς επειδή οι μεταλλάξεις είναι πολλές, χρειάζεται

αντίστοιχα και μεγάλη ποσότητα DNA και έτσι απομονωμένα τμήματα αυτού δε θα ήταν δυνατό να μελετηθούν χωρίς την PCR .

Σε ιδιαίτερα μεγάλο βαθμό χρησιμοποιείται επίσης και η Real Time PCR ή αλλιώς Quantitave PCR η οποία αποτελεί και μία παραλλαγή της συνηθισμένης PCR. Στη μέθοδο αυτή, η τροποποίηση του DNA, γίνεται σε πραγματικό χρόνο, όπως άλλωστε σημαίνεται και από το ίδιο το όνομα της μεθόδου (33). Η μέθοδος αυτή, παρουσιάζει μεγαλύτερη ακρίβεια και αποδοτικότητα σε πολύ μικρότερο χρόνο από την PCR, από την οποία και προέρχεται. Επίσης, δίνει τη δυνατότητα παρακολούθησης της πορείας σύνθεσης του DNA σε κάθε κύκλο της αντίδρασης (33) .

Μία ακόμη μέθοδος που χρησιμοποιείται για την αναγνώριση των αλληλεπιδράσεων είναι η μέθοδος ELISA. Η μέθοδος αυτή πρωτοεμφανίστηκε το 1990 και από τότε χρησιμοποιείται κατά κόρον. Πρόκειται μία γρήγορη, ευέλικτη και αξιόπιστη μέθοδο που στηρίζεται στην ανάπτυξη και την εμφάνιση χρώματος από ένα αντίσωμα ή αντιγόνο που φέρει ένα ένζυμο-δείκτη (συνήθως υπεροξειδάση ή αλκαλική φωσφατάση), το οποίο προκαλεί υδρόλυση ενός χρωμογόνου υποστρώματος. Μπορεί να ανιχνεύσει ποσότητα αντιγόνου 10pg/ml, γεγονός, το οποίο εξαρτάται από τη σχέση μεταξύ των διαθέσιμων αντισωμάτων και τη διαμόρφωση της δοκιμασίας (34).

Τέλος, ιδιαίτερα διαδεδομένη μέθοδος σε ότι αφορά τον εντοπισμό των αλληλεπιδράσεων ανάμεσα σε microRNA και γονίδια είναι και αυτή των Microarrays. Η μέθοδος αυτή, είναι ιδιαίτερα χρήσιμη καθώς εισάγει μία ιδιαίτερα αποδοτική τεχνική σε ότι αφορά στη μέτρηση της ομοιομορφίας των εκφράσεων πολλών χιλιάδων γονιδίων σε ένα και μόνο πείραμα. Οι Microarrays έχουν τρεις διαφορετικές χρήσεις στον τομέα της μικροβιολογίας με την απόδοση αυτών να εξαρτάται από την ποσότητα νουκλεΐκών οξέων των παθογόνων μικροοργανισμών στα υπό διερεύνηση δείγματα.

Κεφάλαιο 3

3.1 Η σημαντικότητα των αλγορίθμων εξόρυξης κειμένου στη Βιοπληροφορική

Ο τομέας της εξόρυξης κειμένου αποτελεί έναν ιδιαίτερα ανερχόμενο τομέα. Η σημαντική άνοδος του τομέα αυτού πηγάζει από την ανάγκη ανάλυσης τεραστίου όγκου δεδομένων που προέρχεται από την ύπαρξη ενός σημαντικά μεγάλου αριθμού επιστημονικών κειμένων που πλέον δημοσιεύονται στον παγκόσμιο ιστό. Το παραπάνω πρόβλημα εντείνεται ίσως ακόμη περισσότερο όταν τα επιστημονικά αυτά κείμενα δεν ακολουθούν μία αποδεκτή δομή και μία πεπατημένη στον τρόπο της γραφής τους. Έτσι, παρόλο που η ποσότητα των δεδομένων κειμένου αυξάνει με ραγδαίο ρυθμό καθημερινά, η ικανότητα εξαγωγής χρήσιμης πληροφορίας από αυτά εξακολουθεί να παραμένει σταθερή.

Οι αλγόριθμοι που υπάρχουν διαθέσιμοι έως και σήμερα, βοηθούν τον curator ώστε να αναγνωρίσει μονάχα την εμφάνιση ενός καινούριου γεγονότος. Στη συνέχεια αυτός θα πρέπει να προβεί σε προσεκτική ανάγνωση του κειμένου ώστε να εξάγει από αυτό όλη την απαραίτητη πληροφορία. Αν αναλογιστεί κανείς το σύνολο της διαθέσιμης πληροφορίας, θα αναγνωρίσει πως είναι πρακτικά αδύνατο για έναν curator να συλλάβει τον όγκο των διαθέσιμων δεδομένων. Στο σημείο αυτό καθίσταται κατανοητή η ανάγκη για αυτόματη εξαγωγή χρήσιμης πληροφορίας από τεράστιο όγκο επιστημονικών κειμένων που θα έχει ως στόχο να βοηθήσει σημαντικά την ανάλυση των κειμένων αυτών από τους curators.

Η εξαγωγή της χρήσιμης και ουσιαστικής πληροφορίας από έναν μεγάλο αριθμό επιστημονικών άρθρων αποτελεί μία ιδιαίτερα χρονοβόρα, δύσκολή αλλά και πολυσύνθετη εργασία. Για να επιλεγούν λοιπόν τα άρθρα αυτά και στην πορεία να αξιολογηθούν ώστε να εξαχθεί χρήσιμη πληροφορία από αυτά απαιτείται η χρήση αλγορίθμων εξόρυξης κειμένου οι

οποίοι θα έχουν τη δυνατότητα να εντοπίζουν την πληροφορία που ενδιαφέρει και να συστήνουν το αντίστοιχο άρθρο στον curator ώστε αυτός με τη σειρά του να το καταχωρήσει στη βάση δεδομένων. Συχνά, χρησιμοποιούνται όροι – λέξεις «κλειδιά» που καταδεικνύουν στον αλγόριθμο πώς να εντοπίσει τη χρήσιμη πληροφορία. Παραδείγματα των λέξεων αυτών θα μπορούσαν να είναι οι όροι microRNA, miR, miRNA, gene κ.α.

Βασικός στόχος της ύπαρξης των αλγορίθμων αυτών είναι ο εντοπισμός όλων εκείνων των επιστημονικών άρθρων που μετέπειτα θα μεταβιβαστούν στους curators ώστε να εξάγουν όλη την απαραίτητη πληροφορία. Παράλληλα, αυτό το οποίο αναμένεται από τους αλγορίθμους αυτούς, είναι να συλλάβουν όλη εκείνη την απαραίτητη πληροφορία που αφορά στην εξέλιξη των επιστημονικών μεθόδων.

Φυσικά, οι αλγόριθμοι αυτοί δε θα πρέπει να παραμένουν στάσιμοι αλλά να εξελίσσονται και να βελτιώνονται ώστε να εντοπίζουν περισσότερη χρήσιμη πληροφορία σε μικρότερο χρονικό διάστημα και έτσι να εξυπηρετούν στο μέγιστο δυνατό βαθμό τους curators που ασχολούνται αποκλειστικά με την ενημέρωση της πληροφορίας. Μία τέτοια εξέλιξη είναι ιδιαίτερα σημαντική για τον τομέα της Βιοπληροφορικής καθώς θα σημαίνει για αυτόν πολύ μεγαλύτερο όγκο αξιόλογης και αξιόπιστης πληροφορίας σε πολύ μικρό χρονικό διάστημα.

Βέβαια, για να εκπαιδευτούν οι αλγόριθμοι αυτοί και τελικά να μεταβούν στο επόμενο βήμα της εξέλιξης τους, απαιτείται η ύπαρξη θετικών και αρνητικών σετ δεδομένων. Τα σετ αυτά, αφορούν αποκλειστικά στην εκπαίδευση και την εξέλιξη του αλγορίθμου καθώς έχουν τη δυνατότητα να του «μαθαίνουν» τα σημεία στα οποία πρέπει να εστιάσει και πώς να εξάγει τη χρήσιμη πληροφορία.

Στο σημείο αυτό υπεισέρχεται και η έννοια της μηχανικής μάθησης που έχει ως στόχο και αντικείμενο της μηχανικής μάθησης είναι η δημιουργία μηχανών που θα έχουν την ικανότητα να μαθαίνουν, να βελτιώνουν δηλαδή την απόδοσή τους στον τομέα που

δραστηριοποιούνται διαμέσου της προηγούμενης γνώσης και εμπειρίας. Έτσι λοιπόν, τα συστήματα τα οποία διαθέτουν την ικανότητα της μηχανικής μάθησης είναι σε θέση να μεταβάλλονται διαρκώς προς το καλύτερο, φυσικά σε ότι αφορά στις λειτουργίες τους, να μεταβάλλουν τη βάση γνώσης τους και τέλος να εκτελούν γενικεύσεις, δηλαδή, να αγνοούν χαρακτηριστικά και ιδιότητες που δεν είναι αντιπροσωπευτικά της έννοιας που οφείλουν να μάθουν.

Όπως είναι αντιληπτό, όσο ραγδαία και εάν είναι η εξέλιξη στον τομέα της τεχνολογίας, η επιστήμη δεν έχει φτάσει ακόμη στο σημείο εκείνο όπου θα παράγει μηχανές με αντίληψη αλλά και μαθησιακές ικανότητες όμοιες με αυτές του ανθρώπου. Για τον παραπάνω λόγο, αναπτύχθηκε μία σειρά αλγορίθμων οι οποίοι έχουν συντελέσει αλλά και βοηθήσει σημαντικά στην επίλυση του προβλήματος αυτού.

Ο Mitchell έδωσε ένα σύντομο ορισμό για τη μηχανική μάθηση και υποστήριξε ότι ένα πρόγραμμα υπολογιστή μαθαίνει από την εμπειρία Ε σχετικά με κάποια διεργασία που θα εκτελέσει Τ αλλά το μέτρο απόδοσης αυτής Ρ εάν και μόνο αν αυτό βελτιώσει την επίδοσή του Ρ σε μία συγκεκριμένη Τ ακολουθώντας την εμπειρία Ε (35). Εάν μεταφέρει κανείς τον παραπάνω ορισμό στο πρόβλημα της αυτόματης κατηγοριοποίησης κειμένου θα παρατηρήσει ότι οι όροι που χρησιμοποιούνται αντιστοιχούν σε (35):

1. *Έργο Τ* : Η κατάταξη κειμένων φυσικής γλώσσας σε ένα προκαθορισμένο σύνολο θεματικών κατηγοριών.
2. *Μέτρο απόδοσης Ρ* : Το ποσοστό των κειμένων που ταξινομήθηκαν σωστά.
3. *Εμπειρία Ε* : Ένα σύνολο από κείμενα με γνωστή κατηγοριοποίηση.

Η σχεδίαση ενός συστήματος μηχανικής μάθησης συνίσταται στον προσπορισμό της γνώσης που θα χρησιμοποιηθεί κατά την εκπαίδευση αυτού. Οποιαδήποτε αλλαγή, θα επιφέρει σημαντικά αποτελέσματα στη συνολική λειτουργία του συστήματος. Ακόλουθο στάδιο είναι

αυτό του τρόπου με τον οποίο το σύστημα θα διαχειριστεί τη γνώση που αποκομίζει ώστε να είναι αποδοτικότερη η λειτουργία του.

Ανάλογα με το είδος της παρεχόμενης γνώσης, η μηχανική μάθηση διαχωρίζεται σε δύο πολύ μεγάλες κατηγορίες, σε αυτή της μάθησης με επίβλεψη και σε αυτή της μάθησης χωρίς επίβλεψη οι οποίες διαθέτουν τα παρακάτω χαρακτηριστικά (35) :

- Μάθηση με επίβλεψη (Supervised Learning) : Στην περίπτωση αυτή η μάθηση θεωρείται ότι γίνεται επίβλεψη το σύστημα θα πρέπει επαγωγικά να αφομοιώσει μία συνάρτηση η οποία καλείται και συνάρτηση στόχος και αποτελεί έκφραση του μοντέλου που περιγράφει τα δεδομένα. Η συνάρτηση αυτή, χρησιμοποιείται κυρίως με στόχο την πρόβλεψης της τιμής μιας μεταβλητής που ονομάζεται εξαρτημένη ή μεταβλητή εξόδου, δεδομένων των τιμών ενός συνόλου μεταβλητών που καλούνται ανεξάρτητες ή εισόδου. Στην κατηγορία αυτή διακρίνονται δύο είδη προβλημάτων :
    - Η Ταξινόμηση που αφορά στη δημιουργία μοντέλων πρόβλεψης διακριτών τάξεων
    - Η παρεμβολή που αφορά στη δημιουργία μοντέλων πρόβλεψης αριθμητικών τιμών
- Μάθηση χωρίς επίβλεψη (Unsupervised Learning) : Στην περίπτωση αυτή, το σύστημα έχει ως στόχο να ανακαλύψεις τις συσχετίσεις εκείνες αλλά και τις ομάδες δεδομένων βασιζόμενο αποκλειστικά και μόνο στις ιδιότητες τις οποίες διαθέτει. Αποτέλεσμα του παραπάνω, είναι η δημιουργία προτύπων, κάθε ένα από τα οποία περιγράφει ένα μέρος των δεδομένων. Παραδείγματα τέτοιων προτύπων αποτελούν :
    - Οι κανόνες συσχέτισης
    - Οι ομάδες που προκύπτουν από τη διαδικασία της ομαδοποίησης

Η θεωρία της μηχανικής μάθησης βασίζεται πρωτίστως στην υπόθεση ότι η κατανομή των στιγμιότυπων εκπαίδευσης θα είναι αντιπροσωπευτική και ανάλογη της γενικής κατανομής

στο χώρο που πρόκειται να μοντελοποιηθεί. Άλλωστε, οι προβλέψεις ενός μοντέλου για τα στιγμιότυπα που θα προκύψουν από την εκτέλεση των αλγορίθμων θεωρούνται πιο αξιόπιστες εάν τα στιγμιότυπα εκπαίδευσης ακολουθούν παρόμοια κατανομή με αυτά που πρόκειται να προβλεφθούν. Βέβαια, καθώς το παραπάνω παρουσιάζεται να είναι ιδανικό, δυστυχώς στην πράξη ελάχιστες φορές ισχύει καθώς περάν της αβεβαιότητας ύπαρξης ενός τέτοιου ιδανικού μοντέλου, ακόμη και εάν αυτό υπάρχει δεν είναι βέβαιο ότι θα αποτελεί την ιδανική λύση.

3.2 Αξιολόγηση των αλγορίθμων εξόρυξης κειμένου

Η αξιολόγηση των αλγορίθμων εξόρυξης κειμένου γίνεται με βάση τις παρακάτω μετρικές (35):

1. **Ακρίβεια (precision) :** Η μετρική της *ακρίβειας (precision)* εκτιμά την ορθότητα των αποτελεσμάτων μιας εργασίας.
2. **Ανάκληση (recall) :** Η μετρική της *ανάκλησης (recall)* εκτιμά την *πληρότητα (completeness)* των αποτελεσμάτων μιας εργασίας.
3. **F-Measure :** Η μετρική F-Measure δίνει μια εκτίμηση της επίδοσης μιας εργασίας, συνδυάζοντας την ακρίβεια και την ανάκληση.
4. **Διασταυρωμένη επικύρωση (cross validation) :** Η *διασταυρωμένη επικύρωση (cross validation)* είναι ένας τρόπος να λάβουμε μια αξιόπιστη εκτίμηση για την επίδοση ενός συστήματος βασισμένου σε μηχανική μάθηση.

Κεφάλαιο 4

4.1 Ο αλγόριθμος TarMiner

Τα τελευταία χρόνια έχουν δημιουργηθεί πολλές βάσεις δεδομένων που έχουν ως στόχο την καταγραφή όλο και περισσότερων στοιχείων που αφορούν στις αλληλεπιδράσεις ανάμεσα στα microRNAs και τα γονίδια αλλά και στη δημιουργία και την εξάπλωση των ασθενειών. Στις περισσότερες των περιπτώσεων, τα δεδομένα αυτά, συλλέχθηκαν, καταγράφηκαν, αξιολογήθηκαν και στη συνέχεια καταχωρήθηκαν από curators που αφιέρωσαν αρκετές εργατοώρες ώστε να μπορέσουν να ανταπεξέλθουν στις παραπάνω απαιτήσεις.

Ο TarMiner, δεν αποτελεί τίποτε παραπάνω από έναν αλγόριθμο που δημιουργήθηκε με στόχο να διευκολύνει τους curators στη συλλογή των στοιχείων και στην καταχώρηση αυτών στη βάση δεδομένων (36). Ο αλγόριθμος αυτός, έχει τη δυνατότητα να αναγνωρίζει αυτόματα στο κείμενο στο οποίο περιλαμβάνεται στα επιστημονικά άρθρα προτάσεις που περιλαμβάνουν αλληλεπιδράσεις ανάμεσα σε microRNAs και γονίδια οι οποίες έχουν τεκμηριωθεί από πειραματικές μεθόδους (36).

Σε αντίθεση με άλλες αυτοματοποιημένες ο αλγόριθμος αυτός εισάγει τα παρακάτω μοναδικά πλεονεκτήματα (36):

- Χρησιμοποιεί ολόκληρο το κείμενο των επιστημονικών άρθρων και όχι μονάχα το abstract
- Χρησιμοποιεί έναν ταξινομητή «εκπαιδευμένο» σε ήδη καταχωρημένα δεδομένα που χρησιμοποιεί πολλά στοιχεία από το NLP (Natural Language Processing) ώστε να επιτύχει τη βέλτιστη ακρίβεια

- Αναγνωρίζει microRNAs αλλά και γονίδια που προέρχονται από πολλά και διαφορετικά είδη

Ο TarMiner, εξάγει από τη δημοσίευση όλες εκείνες τις προτάσεις που περιλαμβάνουν τουλάχιστον ένα microRNA και ένα γονίδιο. Στη συνέχεια, κάνει χρήση της μεθόδου NLP και δημιουργεί ένα διάνυσμα για κάθε ζεύγος microRNA – γονιδίου που εντοπίζεται μέσα στις προτάσεις (36). Τέλος, ένας ταξινομητής, αναγνωρίζει το κάθε ζεύγος και προβαίνει σε ταξινόμηση της αλληλεπίδρασης ή της μη αλληλεπίδρασης.

Η αναγνώριση των ονομάτων των microRNAs για τον αλγόριθμο αυτό γίνεται βασισμένη σε απλούς γραμματικούς κανόνες γεγονός το οποίο οφείλεται στην απλοϊκή περιγραφή αυτών στη βιβλιογραφία. Ειδικότερα, κάθε όνομα ενός microRNA, έχει την ακόλουθη δομή :

(species preffix)-mir-(miRNA suffix) (36). Το πρόθεμα του microRNA, το οποίο χαρακτηρίζει το είδος από το οποίο προέρχεται αυτό, αποτελείται από τρία γράμματα ενώ η κατάληξη είναι διαμορφωμένη με τρόπο τέτοιο ώστε να διαχωρίζει το ένα microRNA από το άλλο.

Ο TarMiner, λαμβάνει υπόψη του και τις διαφοροποιήσεις που εντοπίζονται στη βιβλιογραφία και αφορούν στον τρόπο με τον οποίο ονοματίζονται τα microRNAs. Έτσι, προθέματα όπως miRNA, microRNA και mir θεωρούνται όμοια του προθέματος mir. Πέραν τούτου, όταν κάποιος συγγραφέας αναφέρεται σε περισσότερα του ενός microRNA, τείνει να τα παρουσιάζει σε μία πιο σύντομη μορφή τους ώστε να εξοικονομήσει χώρο στο κείμενο. Για παράδειγμα, η πρόταση “mir-100, mir-200, και mir-300” θα μπορούσε πείσης να γραφεί με τους ακόλουθους τρόπους (36):

- mir-100,-200, και -300
- mir100/200/300
- miRNAs -100, -200, και -300

- mir 100, 200 και 300

Η αναγνώριση των γονιδίων μέσα στην πρόταση είναι μία πιο περίπλοκη διαδικασία καθώς η αναφορά τους στη βιβλιογραφία γίνεται με πολλούς και ποικίλους τρόπους. Η δομή των ονομάτων αυτών καθιστά πολλές φορές δύσκολή τη δημιουργία προτύπων που οδηγούν στην αναγνώρισή τους. Με στόχο να επιτευχθεί το βέλτιστο δυνατό αποτέλεσμα σε ότι αφορά στην αναγνώριση των γονιδίων, ο TarMiner, χρησιμοποιεί μία πολύ μεγάλη βάση δεδομένων που περιλαμβάνει τα γονίδια, την περιγραφική ονομασία τους αλλά και χαρακτηριστικούς δείκτες για πολλά είδη.

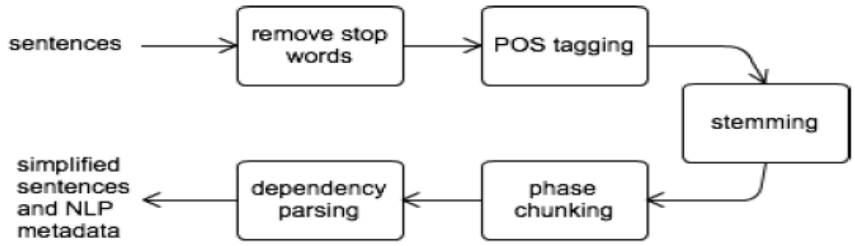
Ένα άλλο σημαντικό πρόβλημα που προκύπτει σε ότι αφορά στην αναγνώριση των γονιδίων, είναι η χρήση συντομογραφιών που τα περιγράφουν. Όροι που αφορούν σε ασθένειες, ιστούς ή χημικά συστατικά, συχνά εμφανίζονται στα επιστημονικά άρθρα με τη μορφή συντομογραφιών. Για να διαπιστώσει εάν οι συντομογραφίες αυτές αφορούν σε κάποιο γονίδιο, ο αλγόριθμος χρησιμοποιεί το κατάλληλο λογισμικό που αναγνωρίζει τις συντομογραφίες αυτές.

Αφού λοιπόν ο TarMiner, αναγνωρίσει όλα τα ζεύγη αλληλεπιδράσεων ανάμεσα σε microRNA και γονίδιο τα οποία συμπεριλαμβάνονται σε ένα επιστημονικό κείμενο, προχωρά στην εφαρμογή της NLP (Natural Language Processing) στα ζεύγη αυτά. Αποτέλεσμα της διαδικασίας αυτής είναι φυσικά η αναγνώριση των ζευγών και η παραγωγή ενός διανύσματος χαρακτηριστικών για κάθε ένα από αυτά.

Τα βήματα που ακολουθεί ο αλγόριθμος στο στάδιο της NLP επεξεργασίας είναι τα εξής (36):

1. *Αφαιρούνται οι stop words, οι λέξεις που χρησιμοποιούν κατά κόρον στις προτάσεις*
2. *Με χρήση του Stanford Tagger, τοποθετούνται POS ετικέτες στις λέξεις ενδιαφέροντος*
3. *Αποκόπτονται οι καταλήξεις*

4. *Γίνεται τμηματοποίηση των προτάσεων σε ρήμα, υποκείμενο, αντικείμενο ή προσδιορισμό*
5. *Τέλος, γίνεται εξαγωγή των γραμματικών σχέσεων ανάμεσα στις λέξεις των προτάσεων, όπως για παράδειγμα, ρήμα – υποκείμενο ή υποκείμενο αντικείμενο*



Εικόνα 11 : Η NLP διαδικασία του TarMiner (36)

Υπάρχει πάντα η περίπτωση, σε μία και μόνο πρόταση να εμφανίζονται πολλά ονόματα που αφορούν είτε σε microRNA είτε σε γονίδια. **Ο αλγόριθμος έχει τη δυνατότητα εξαγωγής όλων των ζευγών από την κάθε πρόταση, με έναν και μόνο περιορισμό που αφορά στη σειρά με την οποία παρουσιάζονται τα microRNAs και τα γονίδια μέσα σε αυτή** (36). Ειδικότερα, ο αλγόριθμος έχει τη δυνατότητα να εξάγει πληροφορία για ζεύγη που βρίσκονται σε διπλανά block μέσα στην πρόταση. Ως block θεωρείται μία αλληλουχία λέξεων που περιλαμβάνει μόνο έναν από τους δύο τύπους που περιλαμβάνονται στα ζεύγη, δηλαδή είτε τα microRNAs είτε τα γονίδια.

Ένα ακόμη ιδιαίτερο χαρακτηριστικό του αλγορίθμου αυτού είναι ότι προσφέρει τη δυνατότητα εμφάνισης ενός ζεύγους μέσα σε επιστημονικό κείμενο ακόμη και εάν αυτό εμφανίζεται πολύ λίγες φορές. Το χαρακτηριστικό αυτό εισάγει μία μοναδική ακρίβεια και μία σημαντική αποτελεσματικότητα στον αλγόριθμο ο οποίος περιλαμβάνει μία ξεχωριστή παράμετρο για τον ελάχιστο αριθμό των εμφανίσεων ενός ζεύγους μέσα στο κείμενο.

Ο TarMiner, αναγνωρίζει ως χαρακτηριστικά του ζεύγους που εντοπίζει στο κείμενο, τα χαρακτηριστικά που περιλαμβάνονται μέσα στην πρόταση που περιλαμβάνει το ζεύγος αυτό. Τα συνολικά χαρακτηριστικά του αλγορίθμου μπορούν να διαχωριστούν στις παρακάτω κατηγορίες (36):

- Χαρακτηριστικά που βασίζονται στα λεξικά : Στην κατηγορία αυτή περιλαμβάνονται χαρακτηριστικά που εξαρτώνται από τον εντοπισμό συγκεκριμένης αλληλεπίδρασης μέσα στην πρόταση.
- Χαρακτηριστικά που βασίζονται στην τμηματοποίηση των λέξεων : Στη κατηγορία αυτή συμπεριλαμβάνονται χαρακτηριστικά που αναγνωρίζουν την τελευταία λέξη σε οποιαδήποτε πρόταση περιλαμβάνει τους όρους microRNA και γονίδιο
- Χαρακτηριστικά που αφορούν στην εξάρτηση από το γράφο : Η κατηγορία αυτή διαχωρίζεται σε δύο υποκατηγορίες. Η πρώτη κατηγορία κάνει χρήση των κοντινότερων μονοπατιών ανάμεσα στους όρους microRNA και γονίδιο που εντοπίζονται στον γράφο εξάρτησης μιας πρότασης. Η δεύτερη κάνει χρήση ζευγών λέξεων, συμπεριλαμβανομένων των όρων microRNA, γονίδιο, ή των όρων που αφορούν στην αλληλεπίδραση και εξάγονται από το γράφο εξάρτησης. Το εξαγόμενο χαρακτηριστικό στην περίπτωση αυτή είναι ο συνδυασμός των λέξεων αλλά και η σχέση εξάρτησής τους.
- Χαρακτηριστικά που βασίζονται στην τοποθεσία της λέξης : Η κατηγορία αυτή περιλαμβάνει χαρακτηριστικά που αφορούν στη θέση των λέξεων μέσα στην πρόταση όποιες και αν είναι οι γραμματικές ιδιότητες αυτών. Στην κατηγορία αυτή η συμπεριλαμβάνονται τέσσερις υπό – κατηγορίες :
    - Χαρακτηριστικά που αφορούν στις λέξεις που προηγούνται των ονομάτων των microRNA ή των γονιδίων

- Χαρακτηριστικά που αφορούν στις λέξεις που ακολουθούν τα ονόματα των microRNA ή των γονιδίων
- Χαρακτηριστικά που αφορούν σε μεμονωμένες λέξεις που προηγούνται των ονομάτων των microRNA ή των γονιδίων
- Χαρακτηριστικά που αφορούν στις ενδιάμεσες των όρων microRNA και γονιδίου λέξεις

- Μικτά χαρακτηριστικά : Στην κατηγορία αυτή συμπεριλαμβάνονται χαρακτηριστικά από όλες τις παραπάνω κατηγορίες. Και η κατηγορία αυτή περιλαμβάνει τέσσερις υπό – κατηγορίες :
  - Χαρακτηριστικά που αφορούν σε λέξεις που εντοπίζονται στα κοντινότερα μονοπάτια των γράφων εξάρτησης και την ίδια στιγμή ανήκουν στις φράσεις που περιλαμβάνουν τα ονόματα των γονιδιών ή των microRNAs
  - Χαρακτηριστικά που αφορούν σε ακολουθίες λέξεων που περιλαμβάνουν έως και τρεις λέξεις στις οποίες αναφέρεται ένας από τους όρους της αλληλεπίδρασης και βρίσκονται ανάμεσα στα ονόματα των γονιδιών ή των microRNAs
  - Χαρακτηριστικά που αφορούν στις ακολουθίες των λέξεων που βρίσκονται ανάμεσα στα ονόματα των γονιδιών ή των microRNAs και περιλαμβάνουν ένα όρο που να αφορά στην αλληλεπίδραση και να έχει σημανθεί με ετικέτα
  - Χαρακτηριστικά που αφορούν στις ενδιάμεσες ακολουθίες λέξεων ανάμεσα στα ονόματα των γονιδιών ή των microRNAs οι οποίες περιλαμβάνουν όρους της αλληλεπίδρασης οι οποίοι όμως έχουν αντικατασταθεί από τις POS ετικέτες τους.

Ο ταξινομητής που χρησιμοποιεί ο αλγόριθμος TarMiner, διαβάζει το διάνυσμα που αναπαριστά το ζεύγος microRNA – γονιδίου και αποφασίζει εάν υπάρχει αλληλεπίδραση ή

όχι στο ζεύγος αυτό. Η έξοδος που παράγει ο ταξινομητής, είναι η πιθανότητα κάθε ζεύγος microRNA – γονιδίου να ανταποκρίνεται σε μία πραγματική αλληλεπίδραση.

Η εκπαίδευση του αλγορίθμου αυτού πραγματοποιείται διαμέσου μίας σειρά χαρακτηριστικών διανυσμάτων που προέρχονται από επιστημονικά άρθρα τα οποία υπάρχουν καταχωρημένα στη βιβλιοθήκη PCM. Κάθε χαρακτηριστικό διάνυσμα περιέχει ένα πεδίο το οποίο σηματοδοτεί εάν υπάρχει ή όχι αλληλεπίδραση με βάση πάντα τα δεδομένα της βάσης TarBase v7.0 η οποία περιέχει δεδομένα που έχουν επεξεργαστεί και καταχωρηθεί από curators (36).

Για να αξιολογηθεί η αποτελεσματικότητα του αλγορίθμου, χρησιμοποιήθηκε ένα σετ δεδομένων το οποίο περιλαμβάνει αλληλεπιδράσεις ανάμεσα σε microRNA και γονίδια. Μετά από σειρά πειραμάτων ο αλγόριθμος εμφανίστηκε να παρουσιάζει ιδιαίτερα υψηλή αξιοπιστία όσες φορές και εάν εκτελέστηκε το πείραμα. Ακόμη, ιδιαίτερα υψηλή ήταν και η δυνατότητα ανάκλησης που αυτός παρουσίασε ενώ την ίδια τάση ακολούθησε και η F – Μετρική (36). Ο λόγος για τον οποίο οι δύο τελευταίες μετρικές παρουσιάστηκαν λίγο χαμηλότερες από αυτή της ακρίβειας εντοπίζεται στο γεγονός ότι ο TarMiner εστιάζει κυρίως στις αλληλεπιδράσεις που αναφέρονται σε προτάσεις που περιλαμβάνουν τα ονόματα τόσο των microRNAs όσο και των γονιδίων. Ένα ακόμη σημαντικό στοιχείο για τον αλγόριθμο αυτό είναι ότι οι συγγραφείς πολλές φορές χρησιμοποιούν συγκεκριμένο τρόπο έκφρασης για τα γονίδια που εξετάζουν και έτσι ο αλγόριθμος αποτυγχάνει να τα εντοπίσει εάν αυτά παρουσιαστούν με διαφορετικό τρόπο (36).

Precision

minSent / maxInter	1	5	10
1	0.8168	0.7901	0.7610
2	0.8614	0.8196	0.8197
3	0.8461	0.8151	0.8096
4	0.8157	0.8443	0.8081

Recall

minSent / maxInter	1	5	10
1	0.7698	0.6103	0.5069
2	0.7618	0.5784	0.5184
3	0.7346	0.5941	0.5444
4	0.7547	0.5991	0.5317

F-measure

minSent / maxInter	1	5	10
1	0.7924	0.6883	0.6081
2	0.8080	0.6780	0.6345
3	0.7863	0.6871	0.6506
4	0.7834	0.7004	0.6411

Εικόνα 12 : Αξιολόγηση του TarMiner

Τέλος, όπως είναι αντιληπτό, σε κάποιες περιπτώσεις, όταν αυξάνεται το ελάχιστο κατώφλι των προτάσεων στις οποίες εμφανίζεται μία αλληλεπίδραση, επιτυγχάνονται υψηλότερες τιμές σε ότι αφορά στην ακρίβεια. Το παραπάνω οφείλεται στην επαναλαμβανόμενη αναφορά στην αλληλεπίδραση microRNa – γονιδίου στο ίδιο κείμενο η οποία αυξάνει σημαντικά την πιθανότητα η αλληλεπίδραση αυτή να έχει επιβεβαιωθεί από πειραματική μέθοδο και να ανταποκρίνεται σε μία πραγματική αλληλεπίδραση.

4.2 Άλλοι γνωστοί αλγόριθμοι εξόρυξης κειμένου

Πέραν του TarMiner, έχουν αναπτυχθεί και άλλοι αλγόριθμοι που αφορούν στην κατηγορία της εξόρυξης κειμένου από επιστημονικά άρθρα που πραγματατεύονται το θέμα της αλληλεπίδρασης ανάμεσα σε microRNA και γονίδιο και αναφέρονται στις επιπτώσεις που προκύπτουν από την αλληλεπίδραση αυτοί. Στην ενότητα αυτή περιγράφονται οι δύο γνωστότεροι αλγόριθμοι της κατηγορίας.

4.2.1 Ο MirSel

Η δημιουργία μίας βάσης δεδομένων που θα περιλαμβάνει τις αλληλεπιδράσεις ανάμεσα σε microRNA και γονίδια απαιτεί τη συλλογή ονομάτων λεξικά που αφορούν σε microRNA, γονίδια, πρωτεΐνες και τη συσχέτιση αυτών με αντίστοιχα αναγνωριστικά της βάσης δεδομένων. Η απόδοση της εξαγωγής της πληροφορίας αυτής, οφείλεται σε πολύ μεγάλο βαθμό στη μοναδικότητα αλλά και στην πληρότητα της καταχώρησης των οντοτήτων αυτών στα διάφορα λεξικά.

Ακόλουθο βήμα του παραπάνω, είναι η συντήρηση αλλά και η επέκταση των λεξικών αυτών. Για να γίνει το παραπάνω, ενημερώνονται οι λίστες των πρωτεϊνών και σε αυτές προστίθενται διάφορα ακρώνυμα, συντομεύσεις, αλλά και οι πλήρεις μορφές των ονομάτων. Στη συνέχεια, αφαιρούνται από τα λεξικά αυτά οποιαδήποτε συνώνυμα ή φράσεις είναι ανακριβή και ενδέχεται να οδηγήσουν τους αλγόριθμους στην εξαγωγή λάθος συμπερασμάτων.

Οι συντάκτες του MirSel, παρατήρησαν ότι υπάρχει ένας σημαντικός αριθμός από ονόματα microRNAs που αναφέρονται στη βιβλιογραφία, παρόλα αυτά όμως δεν περιέχονται ακόμη στις βάσεις δεδομένων (37). Έτσι, επικεντρώθηκαν στην αναζήτηση των ονομάτων των microRNAs κάνοντας χρήση μίας κανονικής έκφρασης που έχει κατασκευαστεί με τρόπο τέτοιο ώστε να αντιστοιχεί σε όλα τα συνώνυμα αλλά και τις περιγραφικές ονομασίες που δίνονται στα microRNAs και περιέχονται στις βάσεις. Η φράση αυτή καλύπτει επίσης και τις διάφορες μορφές με τις οποίες αναφέρονται στα microRNAs στα κείμενα (π.χ. . miR101b, miRNA-101b, microRNA-101b, microRNA101b, etc.) είτε σε αυτά περιλαμβάνεται το πρόθεμα του είδους είτε όχι (37).

Οι πρωτεΐνες αλλά και τα διάφορα γονίδια εντοπίζονται στα κείμενα με την τεχνική του string matching κάνοντας χρήση του εργαλείου syngrep. Το εργαλείο syngrep, κάνει χρήση

του αλγορίθμου Aho – Corasick, ο οποίος χρησιμοποιεί τεχνικές ανάλυσης περιεχομένου ώστε να αυξήσει την ακρίβειά του (37). Έτσι λοιπόν, με βάση τον αλγόριθμό αυτό, η σάρωση για συγκεκριμένες λίστες microRNA, γονιδίου και πρωτεΐνης σε ολόκληρη τη βάση του Pubmedd, απαιτεί περίπου 30 λεπτά σε ένα υπολογιστή με 4 CPU-πυρήνες (37).

Η διεπαφή χρήστη του MirSel παρέχει τη δυνατότητα υποβολής ερωτημάτων κάνοντας χρήση των παρακάτω κριτηρίων φιλτραρίσματος (37):

- Το φίλτρο «αυστηρότητας» το οποίο προβαίνει σε αυστηρή αντιστοίχιση του string που εντοπίζεται σε σχέση με αυτό που υπάρχει ήδη καταχωρημένο στις οντότητες του λεξικού
- Το φίλτρο της μοναδικής πρότασης επιστρέφει τα ζεύγη microRNA – γονιδίων που εμφανίζονται σε μεμονωμένες προτάσεις και όχι σε ολόκληρα τα abstract
- Το φίλτρο συσχέτισης αφορά στην αντιστοίχισης μίας συγκεκριμένης συσχέτισης microRNA – γονιδίου
- Το φίλτρο ταξινόμησης, έχεις ως στόχο να ενισχύσει τις αντιστοιχίσεις που αφορούν σε συγκεκριμένους οργανισμούς
- Το φίλτρο συνωνύμων γονιδίων αποκλείει τα συνώνυμα πρωτεΐνης ή γονιδίων που αναφέροντα σε πολλαπλά γονίδια ή πρωτεΐνες που βρίσκονται καταχωρημένες στις βάσεις δεδομένων
- Το φίλτρο της βάσης δεδομένων εμφανίζει τα ζεύγη εάν και μόνο αυτά περιέχονται σε άλλες βάσεις δεδομένων ή αφορούν σε υπολογιστικές προβλέψεις στόχων microRNA – γονιδίων

Η αξιολόγηση της απόδοσης του αλγορίθμου παρουσιάζεται στον πίνακα που ακολουθεί. Όπως μπορεί κανείς να παρατηρήσει, ο αλγόριθμος παρουσιάζεται να είναι σχετικά αποδοτικός, χωρίς όμως να είναι ο βέλτιστος καθώς η αναζήτηση που πραγματοποιεί εντοπίζεται μονάχα στο χώρο του abstract του κειμένου και όχι στο σύνολο αυτού.

Performance evaluation	abstracts	sentences	cases	Recall	precision	f-meas
(a) miRNA occurrences	50	89	79	0.96	1.00	0.98
(b) miRNA-gene associations	50	89	181	0.90	0.65	0.76
(c) like b, after disambiguation	50	89	181	0.88	0.78	0.83
(d) like b, with keywords	20	29	103	0.89	0.70	0.78
(e) like b, association types	20	29	103	0.87	0.62	0.73

Εικόνα 13 : Αξιολόγηση του MirSel (37)

4.2.2 Ο MirCancer

Ο αλγόριθμος mirCancer λειτουργεί με μία παρόμοια λογική με τους παραπάνω αλγορίθμους ακολουθώντας τα βήματα που περιγράφονται στη συνέχεια. Οι συσχετίσεις που βρίσκονται καταχωρημένες στη βάση δεδομένων, εξάγονται με βάση ένα σύστημα εξόρυξης κειμένου που βασίζεται σε συγκεκριμένους κανόνες και συνοδεύεται από έλεγχο και επιβεβαίωση που διεξάγεται από curators. Για να γίνει η συλλογή όλων των σχετικών δημοσιεύσεων γίνεται αναζήτηση στο PubMed η οποία βασίζεται στο ερώτημα ((((mir) OR mirna) OR microrna) OR micro-rna) OR micro rna (38). Τα αποτελέσματα παρέχονται από το PubMed σε ένα αρχείο XLM το οποίο περιέχει όλες τις λεπτομέρειες του άρθρου σε συνδυασμό πάντα με το abstract αυτού.

Η δυνατότητα εξαγωγής πληροφορίας του συστήματος βασίζεται αποκλειστικά στα αναγνωριστικά των ονομάτων των οντοτήτων, δηλαδή των ονομάτων των microRNAs αλλά και των ονομάτων των διαφόρων τύπων των καρκίνων (38). Και στην περίπτωση του αλγορίθμου αυτού, όπως και σε αυτών που περιγράφηκαν παραπάνω, η αναγνώριση των microRNAs μέσα στο κείμενο είναι μία σχετικά εύκολη διαδικασία καθώς συνήθως ο τρόπος με τον οποίο αναφέρονται και περιγράφονται αυτά αποτελεί μία πολύ απλή διαδικασία. Έτσι

λοιπόν, ο αλγόριθμος αυτός κάνει χρήση των κανονικών εκφράσεων ώστε να εντοπίσει τα ονόματα των microRNAs.

Η αναγνώριση του ονόματος του καρκίνου είναι μία διαφορετική διαδικασία και συντελείται διαμέσου της άμεσης σύγκρισης του κειμένου με ονόματα που βρίσκονται καταχωρημένα σε συγκεκριμένη βιβλιοθήκη. Η βιβλιοθήκη αυτή ονομάζεται ICD-O (International Classification of Diseases for Oncology) και χρησιμοποιείται τα τελευταία 25 χρόνια (38). Η ταξινόμηση των ονομάτων στη βιβλιοθήκη αυτή γίνεται με βάση δύο άξονες, τον άξονα της μορφολογίας ο οποίος περιγράφει τη συμπεριφορά του όγκου και τον άξονα της τοπολογίας ο οποίος περιγράφει το σημείο εκκίνησης του καρκίνου. Έτσι λοιπόν κάθε εγγραφή που αφορά στη μορφολογία συνοδεύεται από το γράμμα Μ ενώ κάθε εγγραφή που αφορά στη θέση από το γράμμα C (38). Παράλληλα, υπάρχουν αντίστοιχες αριθμητικές εγγραφές που χαρακτηρίζουν το όνομα και τη συμπεριφορά του καρκίνου. Για παράδειγμα, ένας μέτριος καρκίνος του πνεύμονα μπορεί να έχει το χαρακτηριστικό C34.9, M8046/3 όπου το μέρος C34.9 θα καταδεικνύει ότι ο καρκίνος εντοπίζεται στον πνεύμονα ενώ το υπόλοιπο M8046 θα αντιστοιχεί στον κωδικό ενός μέτριού μεγέθους όγκου καρκίνου με το τελευταίο ψηφίο, το 3 να περιγράφει τη συμπεριφορά του κακοήθους όγκου (38).

Η εξόρυξη του κειμένου πραγματοποιείται από τον αλγόριθμο όχι μόνο στο abstract του κειμένου αλλά και στον τίτλο αυτού και εάν στα σημεία αυτά εντοπιστούν όροι όπως το microRNA, το όνομα του καρκίνου και ο τρόπος έκφρασης τότε το άρθρο αυτό παραπέμπεται για περαιτέρω επεξεργασία.

Τα αποτελέσματα που αφορούν στον αλγόριθμο αυτό παρουσιάζονται στον πίνακα που ακολουθεί :

Database	Recall	Precision	F-measure
miRCancer	78.50%	100.00%	88.00%
Mir2Disease	28.50%	100.00%	44.40%
Mir2Disease (Before 2010)	77.00%	100.00%	87.00%

Εικόνα 14 : Αξιολόγηση του MiRCancer (38)

Όπως μπορεί κανείς να παρατηρήσει ο αλγόριθμος είναι ιδιαίτερα αποτελεσματικός σε ότι αφορά στην ακρίβειά του παρόλα αυτά όμως παρουσιάζει σχετικά χαμηλά ποσοστά σε ότι αφορά στη δυνατότητα της ανάκλησης. Κύριο αίτιο για το οποίο μπορεί να συμβαίνει αυτό είναι το γεγονός ότι η συσχέτιση συγκεκριμένου microRNA με τύπο καρκίνου ενδεχόμενα να μην αναφέρεται σε μία και μεμονωμένη πρόταση (38). Ακόμη, σε πολλά άρθρα παρατηρούνται ορθογραφικά λάθη, τυπογραφικά λάθη ή ακόμη και περίεργοι συλλαβισμοί που δυσχεραίνουν σημαντικά τον αλγόριθμο στο να συλλάβει την αλληλεπίδραση και να την καταγράψει.

Κεφάλαιο 5

5.1 Το σετ δεδομένων που δημιουργήθηκε με στόχο την εκπαίδευση του TarMiner

Όπως σημειώθηκε πολλές φορές παραπάνω, η εκπαίδευση ενός αλγορίθμου είναι μία ιδιαίτερα σημαντική διαδικασία καθώς τον βελτιώνει σημαντικά σε ότι αφορά στην αποτελεσματικότητα του ενώ την ίδια στιγμή διευκολύνει σημαντικά τους curators σε ότι αφορά στην εργασία τους.

Με στόχο λοιπόν τη βελτίωση του ήδη ιδιαίτερα αξιόπιστου TarMiner, δημιουργήθηκε ένα αρκετά μεγάλο σετ δεδομένων που προήλθε από την ανάγνωση 150 επιστημονικών άρθρων. Η πληροφορία που συλλέχθηκε είναι ιδιαίτερα σημαντική για την εξέλιξη του αλγορίθμου καθώς αυτή πλέον προέρχεται από τη μεγαλύτερη και καλύτερη βάση που έχει καταγραφεί στον τομέα αυτό, από την TarBase. Εάν λοιπόν ο αλγόριθμος αυτός εκπαιδευτεί βασιζόμενος στην πληρέστερη πληροφορία θα παράγει ακόμη καλύτερα αποτελέσματα, θα εξοικονομήσει σημαντικό χρόνο και θα βοηθήσει σημαντικά στην περαιτέρω επεξεργασία της πληροφορίας.

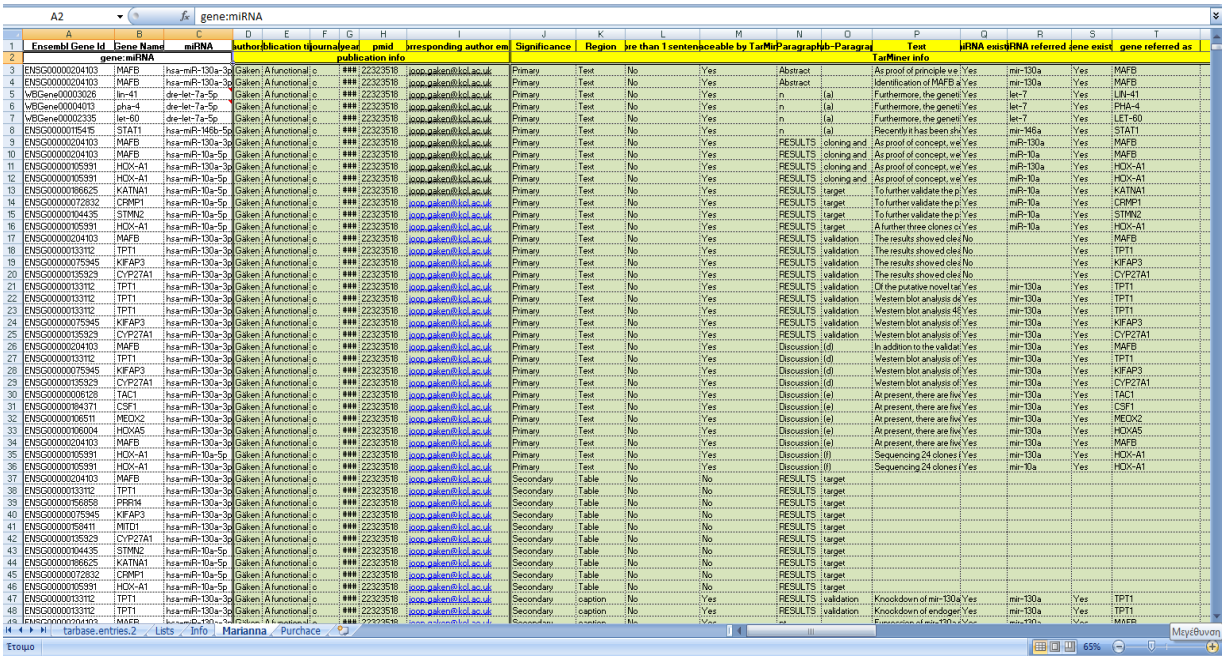
Ο αλγόριθμος που χρησιμοποιούσε έως σήμερα το TarBase έκανε αναζήτηση μονάχα στα abstract των επιστημονικών άρθρων ενώ ο εντοπισμός και η καταγραφή των υπόλοιπων πληροφοριών έπρεπε να γίνει με τη συμβολή των curators. Ο TarMiner έχει υπερπηδήσει την δυσκολία αυτή και έχει στη διάθεσή του την πληρέστερη πληροφορία από τη μεγαλύτερη βάση αλληλεπιδράσεων ανάμεσα σε microRNAs και γονίδια.

Όπως είναι αντιληπτό, ο αριθμός των επιστημονικών άρθρων που συγγράφονται καθημερινά είναι τεράστιος, και έτσι σε καμία περίπτωση δεν είναι ούτε εύκολο αλλά ούτε και εφικτό η πληροφορία αυτή να ανανεώνεται καθημερινά από ανθρώπους. Για το λόγο αυτό είναι απαραίτητοι οι αλγόριθμοι εξόρυξης κειμένου και ένας ιδιαίτερα βελτιωμένος αλγόριθμος όπως αναμένεται να είναι ο TarMiner στη νέα έκδοσή του θα βοηθήσει σημαντικά την

πρόοδο της επιστήμης και θα εξοικονομήσει χρόνο και κόπο από ένα σημαντικά μεγάλο αριθμό ανθρώπων.

Η βαρύτητα λοιπόν της εκπαίδευσης τέτοιων αλγορίθμων είναι πολύ μεγάλη και πρέπει να γίνεται με τρόπο τέτοιο που θα βοηθά τον αλγόριθμο να λειτουργήσει αποτελεσματικά. Έτσι δημιουργούνται τα διάφορα σετ δεδομένων που τον εκπαιδεύουν στην ανάγνωση της πληροφορίας και εξετάζουν την αποτελεσματικότητά του. Για όλους τους παραπάνω λόγους, δημιουργήθηκε ένα αρκετά μεγάλο σετ δεδομένων με στόχο να εκπαιδεύσει τον TarMiner. Στην παραπάνω κατεύθυνση βοήθησε σημαντικά η σχετική πληροφορία που αντλήθηκε από τη βάση TarBase που είναι η πληρέστερη και πλέον ενημερωμένη βάση σε ότι αφορά την πληροφορία που σχετίζεται με τις αλληλεπιδράσεις ανάμεσα στα microRNAs και τα γονίδια. Ιδιαίτερα χρήσιμη αποδείχθηκε και η βάση Ensmbl της Biomart από την οποία εκμαιεύθηκε η πληροφορία σχετικά με το μονοσήμαντο κωδικό που ανατίθεται σε κάθε γονίδιο. Ειδικότερα, η βάση αυτή, εμφανίστηκε για πρώτη φορά το 1999, και μάλιστα πριν χαρτογραφηθεί πλήρως το ανθρώπινο γονίδιο. Στόχος δημιουργίας της εν λόγω βάσης ήταν η αυτοματοποίηση της διαδικασίας της υποσημείωσης των γονιδίων αλλά και η οργάνωσης διαφόρων πληροφοριών που σχετίζονται άμεσα με της αλληλουχίες των γωνιωμάτων. Η βάση αυτή, είναι διαθέσιμη με τη μορφή ιστοσελίδας και αποτελεί ένα φορητό σύστημα λογισμικού ανοικτού κώδικα για τη διαχείριση των γονιδιωμάτων. Τα αναγνωριστικά των γονιδίων που συμπεριλαμβάνονται στη βάση αυτή ξεκινούν με ENSG όταν αφορούν στον άνθρωπο, με WBC όταν αφορούν στο Caenorhabditis elegans, με ENSMUSG όταν αναφέρονται στο όταν αναφέρονται στο musmuscolis, και τέλος με ENSRNOG όταν αναφέρονται στο rat.

Το σετ δεδομένων που δημιουργήθηκε με στόχο την εκπαίδευση του TarMiner είχε τη μορφή που παρουσιάζεται στην παρακάτω εικόνα :



Εικόνα 15 : Μορφή του σετ δεδομένων

Και περιλάμβανε τα πεδία όπως αυτά αναφέρονται στον πίνακα που ακολουθεί :

Πεδίο	Επεξήγηση	Παράδειγμα
Ensemble Gene Id	Μονοσήμαντος Κωδικός που δίνεται από τη βάση Ensemble	ENSG00000204103
Gene Name	Όνομα του γονιδίου	MAFB
microRNA	Όνομα του microRNA	hsa-miR-130a-3p
Authors	Ονόματα των συγγραφέων	Gäken J, Mohamedali AM, Jiang J, Malik F, Stangl D, Smith AE, Chronis C, Kulasekararaj AG, Thomas NS, Farzaneh F, Tavassoli

		M, Mufti GJ
Publication	Τίτλος της δημοσίευσης	A functional assay for microRNA target identification and validation
Journal	Περιοδικό στο οποίο δημοσιεύθηκε το άρθρο	Nucleic Acids Research
Year	Έτος δημοσίευσης	2012
PMID	Μονοσήμαντος κωδικός με τον οποίο το άρθρο καταχωρήθηκε στο PubMed	22323518
Corresponding Author Email	E – mail επικοινωνίας με τους συγγραφείς	joop.gaken@kcl.ac.uk
Significance	Σημαντικότητα της αλληλεπίδρασης	Primary
Region	Περιοχή στην οποία εντοπίστηκε η αλληλεπίδραση	Text
More than one sentence	Περισσότερες από μία προτάσεις που χρειάστηκαν για την εξαγωγή πληροφορίας	No
Traceable by TarMiner	Ο αλγόριθμος έχει τη δυνατότητα ή όχι να	Yes

	εντοπίσει την πληροφορία	
Paragraph	Σε ποια ενότητα βρίσκεται η πληροφορία	Introduction
Sub – Paragraph	Σε ποια παράγραφο συγκεκριμένης ενότητας βρίσκεται η πληροφορία	(b)
Text	Κείμενο στο οποίο εντοπίζεται η πληροφορία	Identification of MAFB and five additional targets and their subsequent confirmation as mir-130a targets by western blot analysis and knockdown experiments validates this strategy for the functional identification of miRNA targets
microRNA exists	Υπάρχει ή δεν υπάρχει το MicroRNA στην πρόταση	Yes
microRNA is referred as	Ο τρόπος με τον οποίο αναφέρεται το MicroRNA στην πρόταση	Mir-130a
Gene exists	Υπάρχει ή δεν υπάρχει το γονίδιο στην πρόταση	Yes

Gene is referred as	Ο τρόπος με τον οποίο MAFB αναφέρεται το γονίδιο στην πρόταση

Πίνακας 2 : Πεδία που αναφέρονται στο σετ δεδομένων

Για να είναι καλύτερη η εκπαίδευση του αλγορίθμου αλλά και για να είναι σε θέση οι δημιουργοί του να προσμετρήσουν την αποτελεσματικότητα αυτού κατέστη αναγκαία η συλλογή τόσο ενός positive set όσο και ενός negative set από τα διαθέσιμα προς ανάγνωση άρθρα. Στο positive set λοιπόν συγκαταλέχθηκε οποιαδήποτε πληροφορία εντοπίστηκε μέσα σε κείμενο, σε λεζάντα εικόνας, σε κείμενο στο supplement ή σε λεζάντα εικόνας στο supplement ενώ στο negative set οποιαδήποτε πληροφορία απεικονιζόταν απλά σε εικόνα ή σε πίνακα με μορφή εικόνας και άρα δεν ήταν δυνατό να αναγνωστούν από τον αλγόριθμο.

Στο positive set εφαρμόστηκε ένας ακόμη περιορισμός ο οποίος διαχώριζε την πληροφορία σε κύρια και δευτερεύουσα. Έτσι λοιπόν οποιαδήποτε πληροφορία εμφανίστηκε στο κείμενο θεωρήθηκε πρωτευούσης σημασίας ενώ οποιαδήποτε πληροφορία εμφανίστηκε στα υπόλοιπα μέρη του άρθρου δευτερευούσης.

Οι παρακάτω πίνακες παρουσιάζουν τους εν λόγω διαχωρισμούς

Positive Set	Negative Set
Κείμενο	Εικόνα
Λεζάντα Εικόνας	Πίνακας
Supplement	

Λεζάντα εικόνας στο Supplement	

Πίνακας 3 : Διαχωρισμός σε positive & negative set

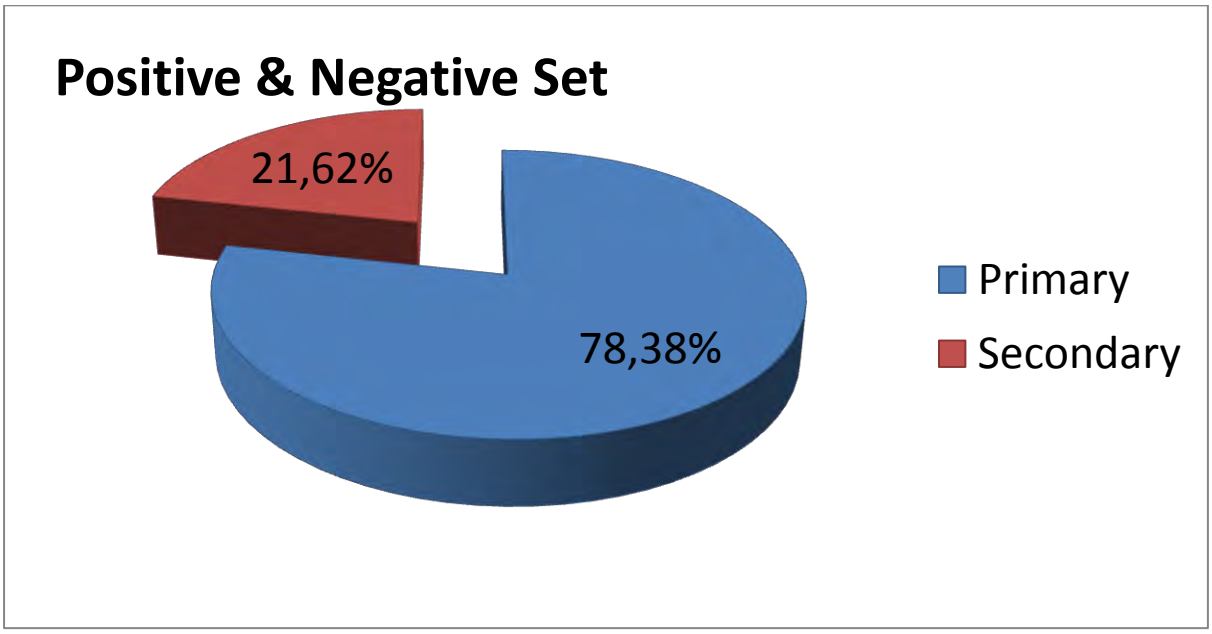
Primary Information	Secondary Information
Κείμενο	Λεζάντα Εικόνας
	Κείμενο στο Supplement
	Λεζάντα εικόνας στο Supplement
	Λεζάντα σε πίνακα
	Λεζάντα σε πίνακα που βρίσκεται στο Supplement

Πίνακας 4 : Διαχωρισμός σε primary & secondary information

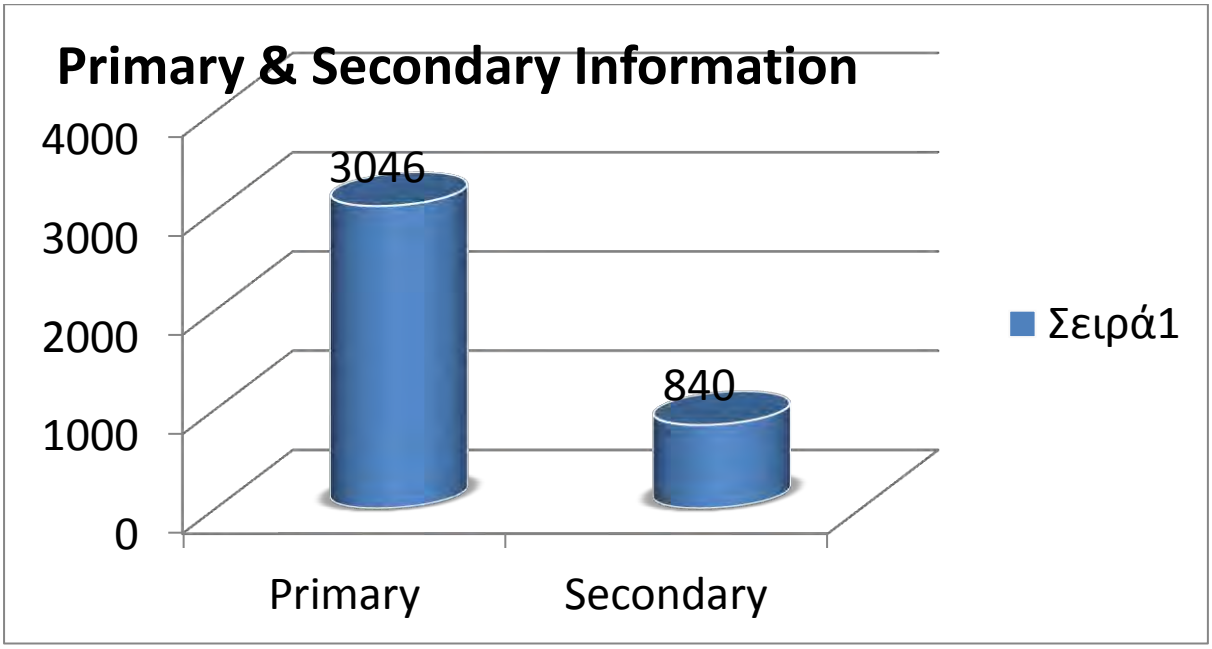
Το σύνολο των εγγραφών που καταχωρήθηκαν για τη δημιουργία του σετ δεδομένων έφτασε τις 4.000 και προήλθε όπως προαναφέρθηκε από το manual curation περίπου 150 επιστημονικών άρθρων. Η λογική πίσω από την οποία δημιουργήθηκε το σετ αυτό ήταν η ορθή συμπλήρωση των παραπάνω πεδίων με επιμέλεια και χωρίς να αναφέρεται περισσότερες των δύο ή τριών φορών η ίδια πληροφορία για μία αλληλεπίδραση. Σημαντικό ήταν να καταγραφεί η χρήσιμη πληροφορία σίγουρα στο abstract και έπειτα στην κατά τη διαδικασία της πειραματικής μεθόδου με την οποία επιβεβαιώθηκε η αλληλεπίδραση.

Στο σύνολο λοιπόν των 4.000 εγγραφών, 3.850 εγγραφές ανήκαν στο positive set και αρα είναι αναγνώσιμες από τον αλγόριθμο, ενώ 150 ανήκουν στο negative και αγνοούνται από αυτόν. Παρατηρούμε ότι η συντριπτική πληροφορία που εντοπίστηκε στα κείμενα μπορεί να αναγνωστεί από τον TarMiner, γεγονός που θεωρείται ιδιαίτερα σημαντικό για την εκπαίδευση και την εξέλιξη αυτού. Επιπλέον, από τις 3.850 αυτές εγγραφές, οι 3.046

εντοπίστηκαν στο κύριο σώμα του κειμένου ενώ οι μονάχα οι υπόλοιπες 804 σε διάφορα άλλα σημεία του άρθρου όπως οι λεζάντες ή το supplement.



Εικόνα 16 : Positive & Negative Set



Εικόνα 17 : Primary & Secondary information

Το συμπέρασμα λοιπόν το οποίο διεξάγεται στην περίπτωση αυτή είναι ότι ο αλγόριθμος TarMiner είναι σε θέση να εντοπίσει σε πολλά σημεία στο κείμενο όλη την απαραίτητη χρήσιμη πληροφορία, να συμπεριλάβει το επιστημονικό άρθρο ανάμεσα σε αυτά που ενδιαφέρουν το επιστημονικό πεδίο και να βοηθήσει σημαντικά τον curator στην επιπλέον καταχώρηση της αλληλεπίδρασης ανάμεσα στο γονίδιο και το microRNA.

Συμπεράσματα

Η επιστήμη της Βιοπληροφορικής αποτελεί έναν νέο και γρήγορα ανερχόμενο τομέα καθώς κατορθώνει να συνδυάσει τη γνώση της βιολογίας με την υπολογιστική δύναμη της πληροφορικής. Το επιστημονικό αυτό πεδίο ασχολείται έντονα με τη δράση των microRNAs αλλά και με τον τρόπο που αυτά αλληλεπιδρούν με τα διάφορα γονίδια και συντελούν σημαντικά στην εμφάνιση των ασθενειών.

Προς αυτή την κατεύθυνση, έχει δημιουργηθεί μία σειρά βάσεων δεδομένων που περιγράφει τις αλληλεπιδράσεις, τα μέλη που λαμβάνουν χώρα, την πειραματική μέθοδο με την οποία εντοπίστηκε η αλληλεπίδραση κ.α. Όπως είναι αντιληπτό, ο όγκος της πληροφορίας αυτής είναι τεράστιος και έτσι για να εντοπισθεί αυτή αλλά και να καταχωρηθεί απαιτείται σοβαρή και έντονη προσπάθεια από τους διάφορους curators.

Με στόχο να εξοικονομηθεί χρόνος αλλά και να είναι εγκυρότερη η πληροφορία που καταχωρείται, δημιουργήθηκε μία σειρά αλγορίθμων εξόρυξης κειμένου που εντοπίζει τη ζητούμενη πληροφορία μέσα σε ένα επιστημονικό άρθρο. Ένας τέτοιος αλγόριθμος είναι και ο TarMiner που έως και σήμερα αποτελεί το βέλτιστο αλγόριθμο στην κατηγορία του.

Για να εκπαιδευθεί ο παραπάνω αλγόριθμος, δημιουργήθηκε ένα τεράστιο σετ τα δεδομένα του οποίου συλλέχθηκαν από σειρά επιστημονικών άρθρων και καταχωρήθηκαν σε ένα αρχείο. Τα δεδομένα αυτά προέρχονται από την πλέον επικαιροποιημένη αλλά και μεγαλύτερη βάση σήμερα, την TarBase και αναμένεται να εκπαιδεύσουν τον αλγόριθμο με τον βέλτιστο δυνατό τρόπο ώστε να τον κάνουν ακόμη πιο αποδοτικό.

training set των άλλων kai limitations

Βιβλιογραφία

1. **Burnet, L.** *Essential Genetics A course bokk.* Cambridge : Cambrigde University Press, 1986.

2. **Cohen, N.** *Cell Structure, function and metabolism.* s.l. : Hodder & Stoughton. The Open University, 1991.

3. **Jones, M. and Jones, G.** *Advanced Biology.* Gambridge : University of Cambridge, 1997.

4. **Mader, S.S.** *Biology.* Boston : MGraw-Hill. Higher Education, 2004. 5th Edition.

5. UOA. *users.uoa.gr.* [Online] [Cited: 15 6 2016.] http://users.uoa.gr/~panagiotavl/domi3.html.

6. **Alberts, B., Bray, D. and Lewis, J.** *Molecular Biology of the Cell.* New York : Garland, 1994. 3rd Edition.

7. **Snustad, D.P and Simmons, M.J.** *Principles of Genetics.* New York : John Wiley and Sons, 2000.

8. *Βιολογία Ενιαίου Λυκείου Θετικής Κατεύθυνησς.* s.l. : ΟΕΔΒ.

9. *Protein and AminoAcid Requirements in Human Nutrition.* s.l. : World Health Organisation, 2007.

10. **Brown, T.A.** *Gene Cloning and DNA Analysis.* Manchester : University of Manchester, 2010. ISBN 978-1-4051-8173-0.

11. Boundless. *www.boundless.com.* [Online] [Cited: 17 5 2016.] https://www.boundless.com/biology/textbooks/boundless-biology-textbook/dna-structure-and-function-14/dna-structure-and-sequencing-100/the-structure-and-sequence-of-dna-433-11661/.

12. **Brandt, Mark.** www.rose-hulman.edu. *www.rose-hulman.edu.* [Online] [Cited: 17 5 2016.] https://www.rose-hulman.edu/~brandt/Chem430/RNA.pdf.

13. www.onbeyondz.net. *www.onbeyondz.net/.* [Online] [Cited: 13 4 2016.] http://www.onbeyondz.net/biology-transcription-translation-and-proteins-oh-my.html.

14. **Livak, K.J and Schmitgenn, T.D.** Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2−ΔΔCT Method. *Methods.* 2001, Vol. Volume 25, Issue 4.

15. **Travers, Andrew and Mushkelishvili, Georgi.** DNA structure and function. *FEBS Journal.* 282, 2015, 12.

16. Biology Pages. *www.biology-pages.info.* [Online] [Cited: 22 3 2016.] http://www.biology-pages.info/T/Transcription.html.

17. **Librado, P. and Rozas, J.** DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics.* 25, 2009, Vol. 11, 1451-1452.

18. **Vlachos, Ioannis, et al.** DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleid Acids.* 2015.

19. **Narri, Kim.** MicroRNA biogenesis: coordinated cropping and dicing. *Nature Reviews Molecular Cell Biology.* 2005, 376+385.

20. MirStart. *mirstart.mbc.nctu.edu.tw.* [Online] [Cited: 24 2 2016.] http://mirstart.mbc.nctu.edu.tw/about.php.

21. **MacFarlane, L.A and Murphy, P.R.** MicroRNA: Biogenesis, Function and Role in Cancer. *Current Genomics.* 11, 2010, 537-561.

22. MDPI. *www.mdpi.com.* [Online] [Cited: 25 3 2016.] http://www.mdpi.com/2311-553X/1/1/17/htm.

23. **Crick, F.H.C.** *The Structure of DNA.*

24. [Online] [Cited: 10 6 2016.] https://sites.google.com/site/bioplerophorike/home/orismos.

25. Kallipos. *repository.kallipos.gr.* [Online] [Cited: 7 1 2016.] https://repository.kallipos.gr/bitstream/11419/5017/1/02_chapter_1.pdf.

26. **Hafner, Markus, et al.** PAR-CliP - A Method to Identify Transcriptome-wide the Binding Sites of RNA Binding Proteins. *Journal of Visualized Experiments.* 2034, 2010, 41.

27. **Darnell, Robert.** HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA.* 2010, Vol. 12, 266-286.

28. **Vlachos, Ioannis, et al.** DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleid Acids Research.* 2015, 11.

29. MirTarBase. *mirtarbase.mbc.nctu.edu.tw.* [Online] http://mirtarbase.mbc.nctu.edu.tw/.

30. miRecords. [Online] http://c1.accurascience.com/miRecords/.

31. miR2Disease. [Online] http://www.mir2disease.org/.

32. *https://www.idtdna.com.* [Online] https://www.idtdna.com/pages/docs/educational-resources/a-basic-pcr-protocol.pdf?sfvrsn=5.

33. AppliedBiosystems. [Online] http://www6.appliedbiosystems.com/support/tutorials/pdf/rtpcr_vs_tradpcr.pdf.

34. [Online] https://www.thermofisher.com/gr/en/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-elisa.html.

35. **Mitchell, Tom. M.** *Machine Learning.* s.l. : McGraw-Hill Science/Engineer, 1997. ISBN:0070428077 .

36. **Tsoupidi, Rodothea, et al.** TarMiner: automatic extraction of miRNA targets from literature. New York : SSDBM '15 Proceedings of the 27th International Conference on Scientific and Statistical Database Management , 2015. 12.

37. **Naeem, Haroon, et al.** miRSel: Automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics.* 135, 2010, 11.

38. miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics.* 2013, 638-644.

39. IVF. *www.ivf.gr.* [Online] 2015. http://www.ivf.gr/developments_9.html.

40. [Online] http://www.onbeyondz.net/biology-transcription-translation-and-proteins-oh-my.html.