ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ «ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

«Αξιοποίηση ομικών δεδομένων με σκοπό τη μελέτη και την ερμηνεία διαφορετικών αναπαραστάσεων ολοκληρωμένων δικτύων (Integrated Networks)»

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΧΡΙΣΤΙΝΑΣ ΒΑΣΙΛΟΠΟΥΛΟΥ

Πτυχιούχος Βιολογίας, Πανεπιστημίου Πατρών

Αθήνα, Σεπτέμβριος 2018

ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ «ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ»

Διπλωματική Εργασία

«Αξιοποίηση ομικών δεδομένων με σκοπό τη μελέτη και την ερμηνεία διαφορετικών αναπαραστάσεων ολοκληρωμένων δικτύων (Integrated Networks)»

Τριμελής εξεταστική επιτροπή

Καθηγητής Κωνσταντίνος Ε. Βοργιάς (Επιβλέπων)
Τομέας Βιοχημείας και Μοριακής Βιολογίας,
Τμήμα Βιολογίας, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Δρ. Ερευνητής Α' Ιωάννης Αλμυράντης,
ΕΚΕΦΕ «Δημόκριτος»

Επίκουρος Καθηγητής Ηρακλής Βαρλάμης
Τμήμα Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

Ευχαριστίες

Έστω και μέσα από αυτές τις λίγες γραμμές, θα ήθελα να ευχαριστήσω τους ανθρώπους που συνέβαλαν στην πραγματοποίηση και στην ολοκλήρωση αυτής της ερευνητικής εργασίας.

Αρχικά, θα ήθελα να ευχαριστήσω τον Δρ. Γιώργο Γιαννακόπουλο συνεργαζόμενο ερευνητή στο Εργαστήριο Μηχανικής Γνώσης και Λογισμικού (SKEL) του ΕΚΕΦΕ "Δημόκριτος", υπό την καθοδήγηση του οποίου πραγματοποιήθηκε η παρούσα διπλωματική εργασία, για την εμπιστοσύνη που μου έδειξε καλωσορίζοντάς με στην ομάδα του. Μου έμαθε να ασκώ νέους τρόπους σκέψης και κριτικής καθώς και να καθοδηγούμαι μόνη μου μέσα από την βιβλιογραφία, δίνοντάς μου την ευκαιρία να προσανατολίσω την διπλωματική μου εργασία στον τομέα που με ενδιέφερε περισσότερο. Τέλος, τον ευχαριστώ εξαιρετικά για τον πολύτιμο χρόνο, την βοήθεια, τη σωστή καθοδήγηση και την αμέριστη στήριξη που μου προσέφερε κάθε φορά που τη χρειαζόμουν.

Επιπρόσθετα, θα ήθελα να ευχαριστήσω ιδιαίτερα τον Καθηγητή Κωνσταντίνο Βοργιά, του Τμήματος Βιολογίας του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών, ο οποίος δέχθηκε να είναι επιβλέπων της παρούσας διπλωματικής εργασίας.

Ειλικρινείς ευχαριστίες θα ήθελα επίσης να εκφράσω στον Δρ. Ιωάννη Αλμυράντη Ερευνητή Α' του ΕΚΕΦΕ "Δημόκριτος" για την συνεχή στήριξη και τις πολύτιμες συμβουλές που προσέφερε στον αρχικό σχεδιασμό της παρούσας εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω τον Επίκουρο Καθηγητή Ηρακλή Βαρλάμη, του Τμήματος Πληροφορικής και Τηλεματικής του Χαροκόπειου Πανεπιστημίου που δέχθηκε να είναι μέλος της τριμελούς εξεταστικής μου επιτροπής.

Περίληψη

Στους τομείς της Συστημικής Βιολογίας και της Βιοπληροφορικής, μία από τα μεγαλύτερες σύγχρονες προκλήσεις στην ανάπτυξη υπολογιστικών εξατομικευμένων προσεγγίσεων σύνθετων ασθενειών είναι ο τρόπος με τον οποίο μπορούν να εξαχθούν, να συνδυαστούν και να αναπαρασταθούν με τον βέλτιστο τρόπο σύνολα multi- omics και κλινικών δεδομένων από δείγματα ασθενών.

Απώτερος σκοπός της παρούσας εργασίας ήταν η ανάδειξη της βελτιωμένης απόδοσης που μπορεί να προσφέρουν σε ένα μοντέλο πρόγνωσης αναπαραστάσεις υψηλότερης πολυπλοκότητας όπως είναι η τοπολογία ενός ολοκληρωμένου/ διασυνδεδεμένου δικτύου το οποίο συνδυάζει -omics δεδομένα πολλαπλών κυτταρικών επιπέδων σε σύγκριση με χαμηλότερης πολυπλοκότητας αναπαραστάσεις όπως είναι το διάνυσμα χαρακτηριστικών.

Συγκεκριμένα, στην παρούσα εργασία πραγματοποιήθηκε εξόρυξη -omics δεδομένων από την τράπεζα δεδομένων Genomic Data Commons από 412 ασθενείς που πάσχουν από μυοδιηθητικό ουροθηλιακό καρκίνωμα της ουροδόχου κύστης. Τα δεδομένα που συγκεντρώθηκαν αφορούν επίπεδα έκφρασης γονιδίων και miRNA (microRNA) καθώς και εκτίμηση επιπέδων μεθυλίωσης περιοχών του γονιδιώματος από δείγματα τα οποία έχουν ληφθεί τόσο από τον πρωτογενή όγκο όσο και από φυσιολογικό ιστό του ιδίου οργάνου.

Σε αυτό το πλαίσιο, καταφέραμε να κατασκευάσουμε ένα βελτιστοποιημένο αλγόριθμο εξόρυξης και επεξεργασίας της επιθυμητής βιολογικής πληροφορίας από 1250 αρχεία και 435 δείγματα, με απώτερο σκοπό τη πολλαπλή αναπαράσταση των δεδομένων με τη μορφή διανύσματος χαρακτηριστικών, δικτύου και τοπολογικού διανύσματος χαρακτηριστικών. Για την εκτίμηση της βελτίωσης της απόδοσης που μπορεί να προσφέρει κάθε αναπαράσταση πραγματοποιήθηκαν διερευνητικές αναλύσεις και πειράματα in silico που πραγματεύονται προβλήματα ταξινόμησης. Επιλέγοντας τα πιο διαφοροποιημένα χαρακτηριστικά πραγματοποιήθηκε κατασκευή του ολοκληρωμένου δικτύου για το μυοδιηθητικό καρκίνωμα της ουροδόχου κύστης. Στη συνέχεια, σχεδιάσαμε και κατασκευάσαμε την εξατομικευμένη αναπαράσταση ενός ασθενούς αυτή τη φορά με τη μορφή δικτύου. Συμπερασματικά, η διανυσματική αναπαράσταση προσφέρει βέλτιστα αποτελέσματα στο πρόβλημα ταξινόμησης κατηγορίας δειγμάτων, χρησιμοποιώντας τον αλγόριθμο του k-κοντινότερου γείτονα και των δέντρων απόφασης. Η τοπολογική αναπαράσταση προσφέρει καλύτερη απόδοση στο μοντέλο ταξινόμησης των σταδίων

καρκίνου (tumor stage) σε σχέση με την διανυσματική, χρησιμοποιώντας τον αλγόριθμο του k- κοντινότερου γείτονα.

Abstract

One of the greatest modern challenges in the fields of Systemic Biology and Bioinformatics is the development of personalized computational approaches in complex diseases that are able to extract, combine and represent multi-omics and clinical data from samples of patients, in an optimal way.

The ultimate goal of this work was to highlight the potentially improved performance of a prediction model that employ higher complexity representations such as the topology vector of a multi-omics integrated network, compared to lower complexity representations such as a feature vector.

Specifically, in the present study, we performed data mining from the Genomic Data Commons database of 412 patients that suffer from muscle-invasive bladder urothelial carcinoma. The final dataset comprises of gene and miRNA (microRNA) expression data as well as DNA methylation data, extracted from both primary tumor and normal tissue samples.

In this context, we succeeded in constructing an optimized algorithm for mining and processing the desired biological information from a total of 1250 files and 435 samples, as well as constructing multiple representations of the data in the form of a feature vector, network and topology vector. In addition, the reconstruction of the integrated network for the bladder urothelial carcinoma was carried out, selecting only the most differentiated features of all patients. We then designed and built a personalized network representation of each patient. Exploratory analyses and in silico experiments dealing with classification problems have been performed in order to investigate the performance of each representation. In conclusion, the feature vector representation provides optimal results for the sample class classification problem, using both the k-nearest neighbor and the decision tree algorithms. Topological representation provides better performance than the feature vector in the tumor stage classification task using the k-nearest neighbor algorithm.

Περιεχόμενα

1 Εισαγωγή

Τα τελευταία χρόνια, η έκρηξη της μεγάλης πληροφορίας και η επίδραση της συστημικής βιολογίας στην επιστημονική κοινότητα οδήγησε στην δημοσίευση αρκετών μελετών που αντιμετωπίζουν την ασθένεια ως το αποτέλεσμα διατάραξης πολύπλοκων βιολογικών δικτύων αλληλεπίδρασης, λαμβάνοντας υπόψη την επίδραση περιβαλλοντικών παραγόντων. Σε αυτό το πλαίσιο, η ενσωμάτωση βιολογικής πληροφορίας διαφορετικών επιπέδων και η μελέτη διαφορετικών αναπαραστάσεων των δεδομένων αποτελούν δύο από της μεγαλύτερες προκλήσεις της Συστημικής Βιολογίας και της Εξατομικευμένης Ιατρικής, σήμερα.

1.1 Συστημική βιολογία

Η χαρτογράφηση των γονιδιωμάτων πολλών οργανισμών και ειδικότερα του ανθρώπου (Human Genome Project), σε συνδυασμό με τη ραγδαία ανάπτυξη της τεχνολογίας, οδήγησαν στην επανάσταση των υψηλής απόδοσης τεχνολογιών (high-throughput) ή αλλιώς «-ομικών» (-omics) τεχνολογιών. Η πρόσθεση της κατάληξης -omics σε ένα μοριακό όρο υποδηλώνει την ολιστική εκτίμηση ενός συνόλου βιομορίων σε ένα συγκεκριμένο μοριακό επίπεδο.

Η συνολική αντιμετώπιση ενός βιολογικού συστήματος δημιούργησε τον όρο Βιολογία Συστημάτων ή αλλιώς Συστημική Βιολογία (Systems Biology). Η Συστημική Βιολογία αφορά σήμερα έναν κυρίαρχο διεπιστημονικό τομέα που στοχεύει στην αποσαφήνιση των μηχανισμών της ζωής μέσω της ολιστικής θεώρησης πολύπλοκων δυναμικών βιολογικών συστημάτων (Kitano 2002).

Η Συστημική Βιολογία ενσωματώνει πολλούς επιστημονικούς κλάδους που μεταξύ άλλων περιλαμβάνουν τη Βιολογία, την Επιστήμη των Υπολογιστών, τη Μηχανική, τη Βιοπληροφορική και τη Φυσική. Η μαθηματική μοντελοποίηση και η θεωρία γράφων σε συνδυασμό με τις ομικές τεχνολογίες έδωσαν τη δυνατότητα ολιστικής προσέγγισης των βιολογικών συστημάτων και θεώρησής τους ως βιομοριακά δίκτυα στα διάφορα επίπεδα κυτταρικής λειτουργίας (Barabási, Gulbahce, and Loscalzo 2011).

1.2 Η εποχή των multi-omics

Πρόσφατες μελέτες αναδεικνύουν όλο και περισσότερο την συμπληρωματική φύση των διαφορετικών -omics επιπέδων καθώς και την αναγκαία - σε ορισμένες περιπτώσεις- ενσωμάτωση πολλαπλών επιπέδων -omics προκειμένου ένα σύστημα να μελετηθεί ολιστικά.

Στην εποχή των μεγάλων δεδομένων, η Συστημική Βιολογία και ο τομέας της Βιοπληροφορικής έρχονται αντιμέτωποι με την πρόκληση της αποδοτικής ενσωμάτωσης πληροφορίας διαφορετικών μοριακών επιπέδων με σκοπό

την κατανόηση πολύπλοκων βιολογικών συστημάτων και την επαρκή αποσαφήνιση των βιολογικών λειτουργιών ενός οργανισμού. Μία από αυτές τις προκλήσεις αφορά την κατανόηση το πως αυτά τα διαφορετικά μοριακά επίπεδα σχετίζονται ποιοτικά μεταξύ τους αλλά και με τον φαινότυπο ενός οργανισμού (βλ. Σχήμα 1).

Ένα μειονέκτημα των αναλύσεων που αξιοποιούν ένα μόνο είδος ομικών δεδομένων είναι η περιορισμένη πραγματοποίηση συσχετίσεων. Ενσωματώνοντας διαφορετικά ομικά δεδομένα σε μία ανάλυση δίνεται η δυνατότητα ανάδυσης αιτιωδών σχέσεων. Για παράδειγμα, ο συνδυασμός πληροφορίας από διαφορετικά μοριακά επίπεδα στην μελέτη μίας ασθένειας μπορεί να παρέχει ευρύτερη κατανόηση σχετικά με την ροή της πληροφορίας ξεκινώντας από αρχικά γεγονότα στο γονιδίωμα, επιγενετικές τροποποιήσεις του γονιδιώματος, ή την επίδραση του περιβάλλοντος και καταλήγοντας πιθανώς σε αλλαγές στον φαινότυπο, στην αλληλεπίδραση διαφορετικών βιομορίων και στην αποφυγή ή προτίμηση διαφορετικών μοριακών/μεταβολικών μονοπατιών

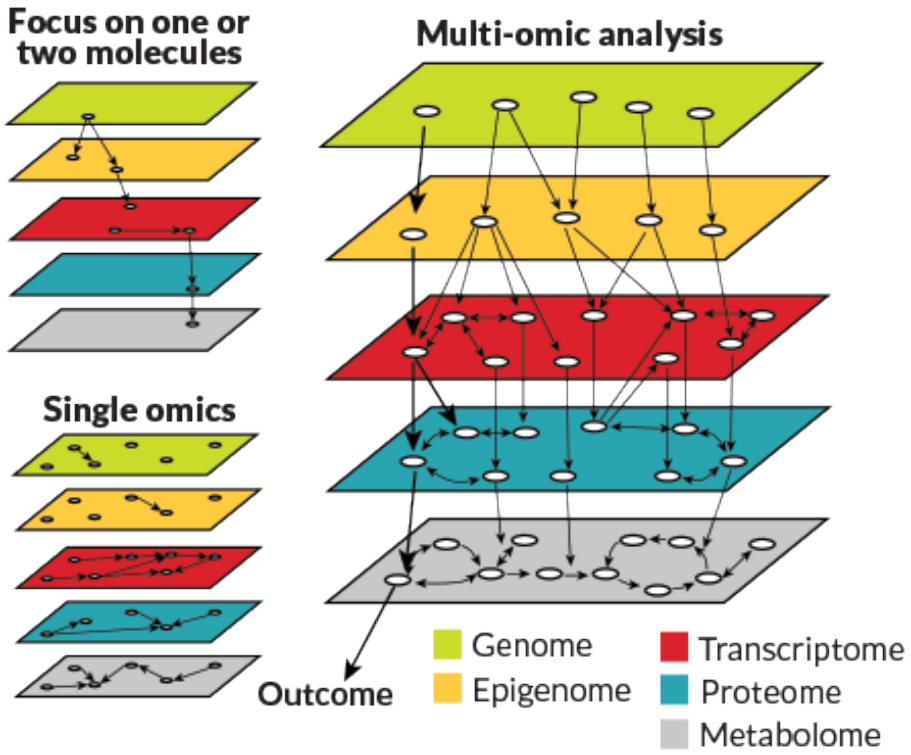
Τη πιο συχνή αναπαράσταση multi-omics δεδομένων αποτελούν τα δίκτυα. Τέτοια δίκτυα στην ελληνική βιβλιογραφία συχνά αναφέρονται ως «ολοκληρωμένα», «διασυνδεδεμένα» και «ενσωματωμένα» ενώ στην ξένη βιβλιογραφία ο πιό συχνός όρος είναι ο "Integrated". Ένα multi-omic δίκτυο αποτελείται από διαφορετικούς τύπους κόμβων που αναπαριστούν τα βιομόρια κάθε ομικού επιπέδου. Ανάλογα με το είδος της σχέσης που συνδέει ένα ζεύγος κόμβων ορίζεται ένα διαφορετικό είδος ακμής. Συχνά, παρατηρείται σε ένα multi- omic δίκτυο να συνδυάζονται πολλαπλών ειδών ακμές που να συνδέουν βιομόρια διαφορετικών επιπέδων.

1.3 Καρκίνος της ουροδόχου κύστης

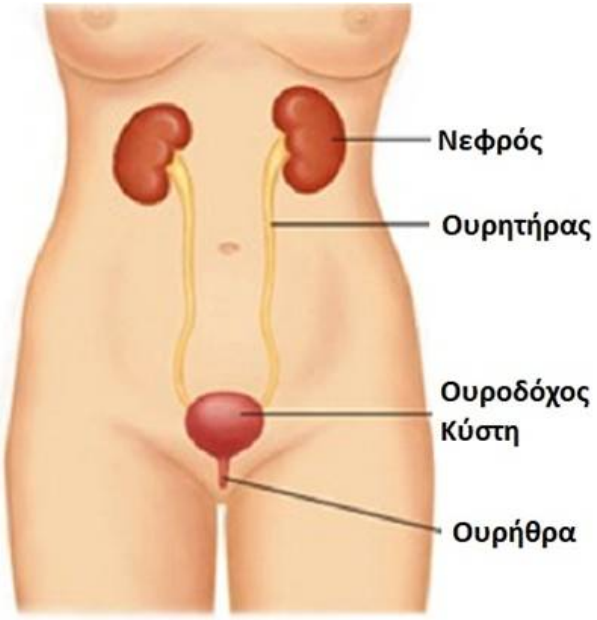
1.3.1 Γενικά

Η ουροδόχος κύστη είναι ένα κοίλο όργανο με λεπτό τοίχωμα το οποίο αποτελείται από λείες μυϊκές ίνες. Η ουροδόχος κύστη συλλέγει τα ούρα από τα νεφρά και τα αποθηκεύει πριν την απέκκρισή τους μέσω της ούρησης. Τα ούρα ρέουν μέσω του ουρητήρα, από τους νεφρούς στην ουροδόχο κύστη, από την οποία αποβάλλονται μέσω της ουρήθρας (Vander, Sherman, and Luciano 2001).

Ο καρκίνος της ουροδόχου κύστης (Bladder Cancer, BC) είναι ένας ιδιαίτερα ετερογενής και επιθετικός τύπος καρκίνου και αποτελεί τη πέμπτη πιο συχνή κακοήθεια που αναπτύσσεται παγκοσμίως στους άνδρες και την όγδοη στις γυναίκες σύμφωνα με το Εθνικό Ινστιτούτο Καρκίνου [National Cancer



Σχήμα 1: Τα πολλαπλά επίπεδα αναπαριστούν διαφορετική πληροφορία - omics.



Σχήμα 2: Το ουροποιητικό σύστημα μίας γυναίκας. Οι δύο νεφροί βρίσκονται στο πίσω μέρος του κοιλιακού τοιχώματος. Τα ούρα ρέουν μέσω του ουρητήρα, από τους νεφρούς στην ουροδόχο κύστη, από την οποία αποβάλλονται μέσω της ουρήθρας (Vander, Sherman, and Luciano 2001).

Institute (NCI)] (Siegel, Miller, and Jemal 2018).

Μερικά από τα πιο συνηθισμένα συμπτώματα του καρκίνου της ουροδόχου

κύστης είναι η ανώδυνη αιματουρία και η δυσκολία στην ούρηση. Επειδή τα συμπτώματα του καρκίνου της ουροδόχου κύστης δεν είναι εξειδικευμένα για αυτή την ασθένεια, η έγκαιρη και η σωστή διάγνωση αποτελεί μία πρόκληση. Σε αυτές τις περιπτώσεις, η πιο συχνή μέθοδος αρχικής εκτίμησης καρκίνου πραγματοποιείται μέσω υπερηχογραφήματος των νεφρών και της ουροδόχου κύστης (Lerner et al. 2016; Xylinas et al. 2014).

Ο καρκίνος της ουροδόχου κύστης προκαλείται τόσο από εξωγενείς όσο και από ενδογενείς παράγοντες. Ο κύριος εξωγενής παράγοντας που τοποθετεί συγκεκριμένες ομάδες ατόμων στην κατηγορία υψηλού κινδύνου είναι το κάπνισμα. Επίσης, πολυάριθμες μελέτες έχουν αποδώσει μεγαλύτερη πιθανότητα εμφάνισης καρκίνου της ουροδόχου κύστης σε άτομα λευκής φυλής σε σχέση με άτομα της έγχρωμης φυλής. Ένας άλλος παράγοντας, είναι η αυξημένη ηλικία του ατόμου, οι πιο συνηθισμένες περιπτώσεις ασθενών αφορούν άτομα ηλικίας άνω των 60 χρονών. Τέλος, η έκθεση ατόμων που εργάζονται σε χώρους επεξεργασίας δέρματος, με βιομηχανικά ελαστικά, βαφές και οδηγοί φορτηγών διατρέχουν μεγαλύτερο κίνδυνο.

Σύμφωνα με το Εθνικό Ινστιτούτο Καρκίνου [National Cancer Institute (NCI)], το συχνότερο μέτρο κακοήθειας ουροθηλιακών καρκινικών κυττάρων είναι το Στάδιο Κακοήθειας (Tumor Stage). Ο βαθμός κακοήθειας καθορίζεται από την μορφολογία των κυττάρων του ιστού κάτω από το μικροσκόπιο καθώς και από την πιθανότητα μεταστατικότητας του όγκου. Ο βαθμός κακοήθειας ακολουθεί διαφορετικές ταξινομικές βαθμίδες ανάλογα με τον τύπο καρκίνου, ωστόσο συνήθως κυμαίνεται μεταξύ των I έως IV. Κύτταρα με βαθμό κακοήθειας I χαρακτηρίζονται ως "καλώς διαφοροποιημένα" γιατί μοιάζουν περισσότερο με τα φυσιολογικά κύτταρα της κύστης και έχουν χαμηλή μεταστατικότητα, αντίθετα όγκος με τάδιο Κακοήθειας (Tumor Stage) III και IV δεν μοιάζουν μορφολογικά με φυσιολογικά κύτταρα, δίνουν συχνότερα μεταστάσεις και είναι περισσότερο επιθετικά και διηθητικά.

Συχνά, χαρακτηρίζουμε έναν όγκο ως υψηλού βαθμού κακοήθειας (όγκος με υψηλή πιθανότητα εξάπλωσης και γρήγορης ανάπτυξης) σε μία ή περισσότερες από τις παρακάτω περιπτώσεις:

- Ο όγκος έχει εισβάλλει στο τοίχωμα της ουροδόχου κύστης
- Υπάρχουν πολλαπλοί όγκοι
- Υπάρχουν όγκοι μεγάλου μεγέθους
- Παρουσιάζουν αυτό που είναι γνωστό ως επιτόπιο καρκίνωμα (carcinoma in situ) το οποίο είναι ένας διηθητικός τύπος της ασθένειας που περιορίζεται στο εσωτερικό τοίχωμα της ουροδόχου κύστης.

Υπάρχουν 4 τύποι όγκων της ουροδόχου κύστης που μπορούν να ξεχωρίσουν με βάση την μορφολογίας των κυττάρων στο μικροσκόπιο, οι οποίοι αφορούν το ουροθηλιακό καρκίνωμα, το καρκίνωμα από πλακώδη κύτταρα (αντιπροσωπεύει το 4% μόνο όλων των όγκων της κύστης) , το αδενοκαρκίνωμα (εξαιρετικά σπάνιο, λιγότερο από το 1% μορφή καρκίνου της ουροδόχου κύστης) και το μικροκυτταρικό καρκίνωμα (εξαιρετικά σπάνιο) (Knez et al. 2014). Το ουροθηλιακό καρκίνωμα (Urothelial Carcinoma) το οποίο και αποτέλεσε αντικείμενο αυτής της εργασίας είναι μία κακοήθεια του επιθηλίου και αποτελεί το 90% των πιο συχνών εμφανίσεων καρκίνου σε αυτή τη κατηγορία.

Το ουροθηλιακό καρκίνωμα ταξινομείται σε δύο κατηγορίες, σε μη- μυοδιηθητικό καρκίνο της ουροδόχου κύστης (Non-Muscle Invasive Bladder Cancer, NMIBC) (αποτελεί το 75% των περιπτώσεων πρώτης διάγνωσης) και σε μυοδιηθητικό καρκίνο της ουροδόχου κύστης (Muscle Invasive Bladder Cancer, MIBC) (αποτελεί το 25% των περιπτώσεων πρώτης διάγνωσης) (Robertson et al. 2017). Ο μη- μυοδιηθητικός καρκίνος της ουροδόχου κύστης δεν αποτελεί συχνή αιτία θανάτου, αποτελείται από χαμηλού βαθμού κακοήθειας όγκους και χαρακτηρίζεται από καλή πρόγνωση. Αντίθετα, ο μυοδιηθητικός καρκίνος της ουροδόχου κύστης παρουσιάζει όγκους υψηλού σταδίου που δείχνουν χαμηλή ικανότητα πρόγνωσης. Οι ασθενείς με μη-μυοδιηθητικό καρκίνο της ουροδόχου κύστης θεραπεύονται με διαουρηθρική εκτομή (transurethral resection, TUR), ενώ ασθενείς με μυοδιηθητικό καρκίνο της ουροδόχου κύστης υφίστανται ριζική κυστεκτομή και εκτροπή ούρων, ή συνδυασμένη χημειοθεραπεία. Ο μη- μυοδιηθητικός καρκίνος της ουροδόχου κύστης συχνά επανεμφανίζεται, και σε περίπου 10 30% των ασθενών μετατρέπεται σε μυοδιηθητικό καρκίνο (Xylinas et al. 2014).

1.3.2 Μοριακό υπόβαθρο

Ένας μεγάλος αριθμός μελετών αποδεικνύουν πως η έκφραση μορίων miRNA μεταβάλλεται σε ουρολογικές κακοήθειες, ρυθμίζοντας πολλαπλά μοριακά μονοπάτια που προάγουν τον καρκίνο. Συγκεκριμένα, μόρια miRNA που είναι γνωστό ότι απορρυθμίζονται στον καρκίνο της ουροδόχου κύστης στοχεύουν σε οδούς μεταγωγής σήματος που εμπλέκονται περισσότερο στην παθογένεση του καρκίνου της ουροδόχου κύστης, συγκεκριμένα του υποδοχέα 3 του αυξητικού παράγοντα των ινοβλαστών (fibroblast growth factor receptor 3, FGFR3) και της πρωτεΐνης p53 (Garzon Ramiro 2010).

Επιπλέον, μόρια miRNA των οποίων η έκφραση είναι τροποποιημένη σε κακοήθειες (”oncomirs”), συνήθως στοχεύουν είτε σε ογκογονίδια είτε σε ογκοκατασταλτικά γονίδια ανάλογα με το κυτταρικό πλαίσιο. Τα γονίδια FGFR3,

HRAS, ERBB2, CCND1, MDM2 και E2F3 αποτελούν ισχυρά εδραιωμένα ογκογονίδια, ενώ τα CDKN2A, TP53, RB1, PTEN και PTCH αντιπροσωπεύουν ογκοκατασταλτικά γονίδια που φαίνεται να εμπλέκονται στον καρκίνο της ουροδόχου κύστης (McConkey and Choi1 Woonyoung 2018). Με τη σειρά τους, τα miRNA μπορούν να λειτουργήσουν ως ογκογονίδια ή ογκοκατασταλτικά ανάλογα με τη σχετική τους έκφραση και στόχο. Η απώλεια της έκφρασης ενός ογκοκατασταλτικού miRNA είτε μέσω κάποιας διαταραγμένης επεξεργασίας μορίων miRNA είτε μέσω επιγενετικής αλλοιώσεως είτε γονιδιακής μεταλλάξεως ή διαγραφής μπορεί να προάγει την έκφραση ενός στοχευμένου ογκογονιδίου. Αντιθέτως, η αυξημένη έκφραση ενός ογκογόνου miRNA μέσω ενίσχυσης ή μετατόπισης μπορεί να οδηγήσει σε μειωμένη έκφραση ενός ογκοκατασταλτικού.

Επιπλέον, μεταλλάξεις μορίων mRNA σε περιοχές στόχων ενός miRNA μπορεί να μεταβάλλουν τη συγγένεια miR για μεταγραφή στόχους ή να δεσμεύσουν εντελώς τη δέσμευση. Συχνά, miRNA βρίσκονται κοντά σε CpG νησίδες ή ακτές που μέσω μεθυλίωσης μπορεί να οδηγηθούν σε σίγαση, ένα επιγενετικό φαινόμενο που έχει τεκμηριωθεί στον καρκίνο της ουροδόχου κύστης.

1.3.3 Θεραπευτικές επιλογές

Περιπτώσεις καρκίνων χαμηλότερου βαθμού κακοήθειας συχνά αντιμετωπίζονται μέσω της χειρουργικής αφαίρεσης του νεοπλάσματος και του ιστού που το περιβάλλει (Transurethral Resections, TUR). Εάν ο καρκίνος επανεμφανισθεί, αντιμετωπίζεται με χημειοθεραπεία ή ανοσοκατασταλτικές μεθόδους θεραπείας. Ένας άλλος τρόπος αντιμετώπισης του μυοδιηθητικού καρκινώματος είναι η ενδοκυστική έγχυση Bacille Calmette-Guerin (BCG), το οποίο είναι ένας αδρανοποιημένος βάκιλος του είδους *Mycobacterium bovis* που χρησιμοποιείται διεθνώς ως τύπο εμβολίου κατά της φυματίωσης. Ωστόσο, το BCG αποτελεί μία ισχυρή ανοσοθεραπεία και ακολουθείται συχνά για τη ρύθμιση του καρκίνου της ουροδόχου κύστης (Knez et al. 2014).

Όπως αναφέρθηκε και σε παραπάνω ενότητα, η πλειονότητα των καρκίνων της ουροδόχου κύστης είναι μη- μυοδιηθητικοί, πράγμα που σημαίνει ότι δεν έχουν προχωρήσει πέραν του μυός της ουροδόχου κύστης. Ωστόσο, σε αρκετές περιπτώσεις μη- μυοδιηθητικοί καρκίνοι μετατρέπονται σε μυοδιηθητικοί. Σε προχωρημένα στάδια καρκίνου, όταν ο όγκος έχει εξαπλωθεί βαθύτερα εντός του μυός της κύστης τότε ένα συνηθισμένο βήμα είναι αυτό της χειρουργικής επέμβασης προκειμένου να αφαιρεθεί η κύστη, ωστόσο σε κάποιες περιπτώσεις χρησιμοποιείται ένας συνδυασμός χημειοθεραπείας και

χειρουργικής επέμβασης εφόσον το νεόπλασμα είναι εκτεταμένο (Xylinas et al. 2014).

Η συχνά αναφερόμενη επιβίωση ασθενών που πάσχουν από μυοδιηθητικό ουροθυλιακό καρκίνωμα ανέρχεται στα 5 έτη και έχει παραμείνει σχετικά αμετάβλητη τα τελευταία 20 χρόνια. Ενώ η αντιμετώπιση και η θεραπεία άλλων καρκίνων έχει προχωρήσει γρήγορα, η αντιμετώπιση του καρκίνου της ουροδόχου κύστης παραμένει σχετικά στάσιμη (Chandrasekar and Evans 2016).

1.4 Η τράπεζα δεδομένων Genomic Data Commons (GDC)

Μία από τις μεγαλύτερες και πληρέστερες βάσεις καρκινικών δεδομένων ήταν η The Cancer Genome Atlas (TCGA) στεγάζοντας κλινικά, ομικά και ιστολογικά δεδομένα για 33 τύπους καρκίνων, η οποία έκλεισε την πύλη των δεδομένων της τον Ιούλιο του 2016. Πρόσφατα, οι ίδιοι θεμελιωτές της TCGA δημιούργησαν υπό την αιγίδα του Αμερικάνικου Εθνικού Ινστιτούτου Καρκίνου (U.S. National Cancer Institute (NCI)) το ερευνητικό πρόγραμμα και ταυτόχρονα πύλη δεδομένων Genomic Data Commons (GDC), διαθέσιμη στην ιστοσελίδα https://gdc.cancer.gov/.

Η τράπεζα δεδομένων GDC παρέχει ένα ολοκληρωμένο καταθετηρίο καρκινικών δεδομένων προς την επιστημονική κοινότητα με απώτερο στόχο την στήριξη της εξατομικευμένης και στοχευμένης ιατρικής. Για την καλύτερη επίτευξη του στόχου αυτού, παρέχεται η δυνατότητα στην ευρύτερη επιστημονική κοινότητα της υποβολής και του σχολιασμού δεδομένων.

Η GDC φιλοξενεί δεδομένα από τα ευρύτερα ερευνητικά προγράμματα TCGA και TARGET (Therapeutically Applicable Research to Generate Effective Therapies). Το πρόγραμμα TARGET συλλέγει δεδομένα γονιδιωματικών πειραμάτων από παιδιά που πάσχουν από διαφορετικούς τύπους καρκίνων.

Ωστόσο, η GDC δεν αποθηκεύει απλά δεδομένα υψηλής απόδοσης (-omic), κλινικά δεδομένα και βιοψίες βιολογικών δειγμάτων από ερευνητικά προγράμματα καρκίνου, αλλά έχει αναπτύξει βιοπληροφορικές μεθόδους ανάλυσης, επεξεργασίας, κανονικοποίησης, φιλτραρίσματος και σχολιασμού οι οποίες εφαρμόζονται σε όλα τα ομικά δεδομένα προκειμένου να καταστώνται άμεσα συγκρίσιμα. Τα δεδομένα είναι διαθέσιμα σε διάφορα επίπεδα επεξεργασίας, κανονικοποίησης και φιλτραρίσματος στην σελίδα της Genomic Data Commons.

Σύμφωνα με την τελευταία ενημέρωση της πύλης δεδομένων που πραγματοποιήθηκε στις 13 Ιουνίου το 2018, η GDC φιλοξενεί συνολικά δεδομένα που αφορούν 40 διαφορετικά προγράμματα, 61 πρωτογενείς ιστούς, 32.555 ασθενείς, 356.382 αρχεία, σχολιασμό 22.147 γονιδίων και 3.142.246 μεταλλα-

γών.

1.4.1 Περιήγηση στην πύλη GDC Data Portal

Η περιήγηση και η πρόσβαση στα γονιδιωματικά και κλινικά δεδομένα της βάσης δεδομένων GDC πραγματοποιείται μέσω της πύλης GDC Data Portal. Η αρχική σελίδα της GDC Data Portal μας προσφέρει πληροφορίες σχετικές με τα διαθέσιμα σύνολα δεδομένων όπως ο συνολικός αριθμός των τύπων καρκίνου και των ασθενών από όλα τα ερευνητικά προγράμματα που στεγάζει, ο συνολικός αριθμός των σχολιασμένων γονιδίων και μεταλλαγών (βλ. Σχήμα 3).

Επί προσθέτως, η αρχική σελίδα της GDC Data Portal προσφέρει τέσσερις διεπαφές (interfaces) για την αποτελεσματική πλοήγηση των δεδομένων(βλ. Σχήμα 3).

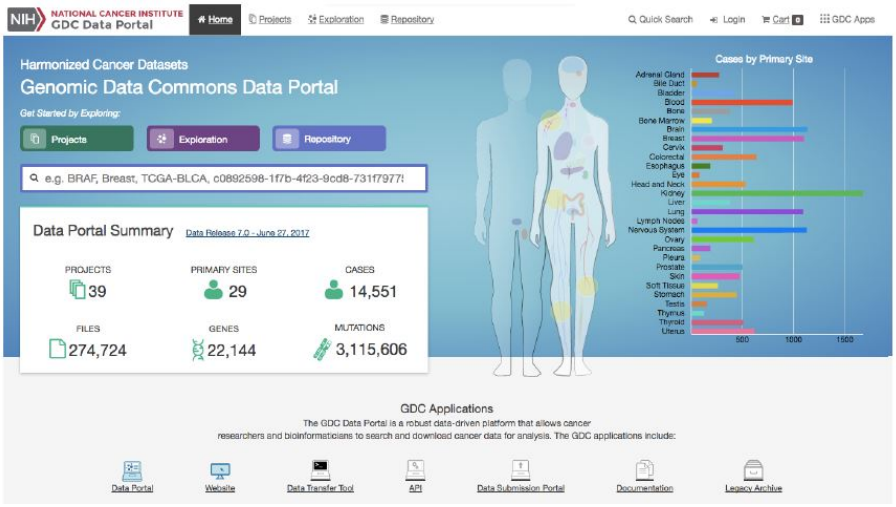
- Projects: Η συγκεκριμένη διεπαφή κατευθύνει τον χρήστη στην σελίδα των Προγραμμάτων (Projects Page), η οποία προσφέρει μία συνολική σύνοψη των καρκινικών δεδομένων, κατηγοριοποιημένη ανά το εκάστοτε πρόγραμμα. Στην GDC ως πρόγραμμα (Project) ορίζεται μία εξειδικευμένη προσπάθεια μελέτης ενός συγκεκριμένου τύπου καρκίνου η οποία αποτελεί κομμάτι ενός μεγαλύτερου καρκινικού ερευνητικού προγράμματος.

- Exploration: Μετάβαση του χρήστη στην σελίδα Περιήγησης (Exploration Page), η οποία δίνει την δυνατότητα εξερεύνησης των διαθέσιμων δεδομένων χρησιμοποιώντας πληθώρα φίλτρων που αφορούν τους ασθενείς, τα γονίδια και μεταλλάξεις σε αυτά.

- Repository: Η παρούσα διεπαφή κατευθύνει τον χρήστη στην σελίδα του Καταθετηρίου δεδομένων (Repository Page), στην οποία ο χρήστης μπορεί να περιηγηθεί στα διαθέσιμα δεδομένα και να εφαρμόσει πληθώρα επιθυμητών φίλτρων που αφορούν τόσο τα αρχεία όσο και τους ασθενείς.

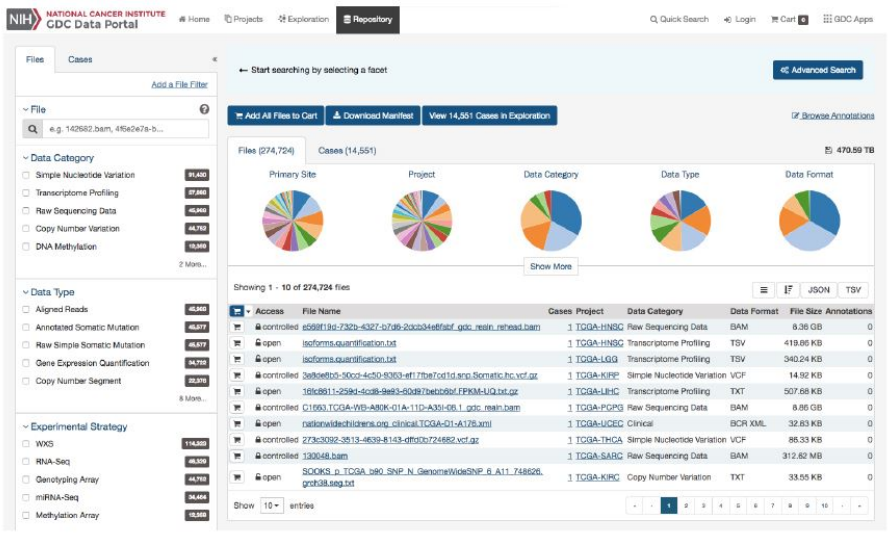
- Human Outline: Η αρχική σελίδα της GDC Data Protal, όπως φαίνεται και στην Εικόνα 1, απεικονίζει την ανατομία ενός ανθρώπινου περιγράμματος. Ο χρήστης επιλέγοντας το επιθυμητό όργανο κατευθύνεται σε μία λίστα όλων των προγραμμάτων και των ασθενών (cases) που σχετίζονται με τον καρκίνο του επιλεγμένου πρωτογενούς ιστού.

Η χρήση της διεπαφής Repository είναι ο κυρίαρχος τρόπος προκειμένου ο χρήστης να αποκτήσει πρόσβαση στα δεδομένα της πύλης GDC Data Portal.



Σχήμα 3: Η αρχική σελίδα της GDC Data Portal μας προσφέρει μία σύνοψη σχετικά με τα διαθέσιμα σύνολα δεδομένων, (https://portal.gdc.cancer.gov/, 9/8/2018).

Επιπλέον, η σελίδα Repository προσφέρει μία σύνοψη όλων των αρχείων και όλων των ασθενών τα οποία είναι διαθέσιμα δίνοντας την δυνατότητα εφαρμογής πολλαπλών φίλτρων. Οι χρήστες μπορούν να έχουν πρόσβαση στην σελίδα Repository είτε μέσω της αρχικής σελίδας της GDC Data Portal είτε κατευθείαν στην ιστοσελίδα https://portal.gdc.cancer.gov/repository.



Σχήμα 4: σελίδα της διεπαφής του Καταθετηρίου δεδομένων (Repository Page), στην οποία ο χρήστης μπορεί να περιηγηθεί στα διαθέσιμα δεδομένα και να εφαρμόσει πληθώρα επιθυμητών φίλτρων που αφορούν τόσο τα αρχεία όσο και τους ασθενείς (https://portal.gdc.cancer.gov/repository, 9/8/2018).

1.4.2 Το μοντέλο οργάνωσης της βάσης δεδομένων GDC

Το μοντέλο οργάνωσης των δεδομένων της βάσης GDC αναπαριστάται μέσω ενός κατευθυνόμενου ακυκλικού γράφου (Directed Acyclic Graph) ο οποίος αποτελείται από διασυνδεδεμένες οντότητες (entities). Κάθε οντότητα στην GDC έχει ένα σύνολο από ιδιότητες (properties) και συνδέσμους (links). Οι ιδιότητες αποτελούν ζευγάρια κλειδιών- τιμών (key- value pairs) τα οποία σχετίζονται με κάποια οντότητα. Οι τιμές των ιδιοτήτων μπορούν να είναι αριθμητικές (numeric), συμβολοσειρές (string) και λογικές μεταβλητές (boolean). Τέλος, οι σύνδεσμοι ορίζουν τις σχέσεις μεταξύ των οντοτήτων , για παράδειγμα μία προς μία, μία προς πολλές, πολλές προς πολλές.

Οι οντότητες (entities) στη βάση δεδομένων Genomic Data Commons αφορούν φακέλους, αρχεία, ασθενείς, δείγματα και επαναλήψεις δειγμάτων (aliquots) ασθενών. Κάθε οντότητα στη βάση δεδομένων Genomic Data Commons περιέχει ένα μοναδικό κωδικό UUID (Universal Unique Identifier) ο οποίος αφορά ένα 128-bit αριθμό. Οι μοναδικοί κωδικοί UUID είναι ιδιότητες που ανατίθενται σε κάθε οντότητα στην GDC και με αυτό τον τρόπο κάθε οντότητα μπορεί να ταυτοποιηθεί μοναδικά. Επιπλέον, σε κάποιες οντότητες ανατίθεται από τα ερευνητικά κέντρα υποβολής των δεδομένων ένας submitter-id κωδικός, ο οποίος είναι μοναδικός για την κάθε οντότητα.

Τα προγράμματα (programs) είναι η υψηλότερη βαθμίδα οργάνωσης των συνόλων δεδομένων της GDC. Σε κάθε πρόγραμμα γίνεται ανάθεση μία μοναδική ιδιότητα τύπου program.name. Τα σύνολα δεδομένων εντός ενός προγράμματος οργανώνονται σε επιμέρους ερευνητικές εργασίες (Projects) στις οποίες γίνεται ανάθεση μοναδικής ιδιότητας τύπου project.code . Για παράδειγμα, το όνομα της ερευνητικής εργασίας για τον καρκίνο στην ουροδόχο κύστη είναι ο μοναδικός κωδικός TCGA- BLCA του τύπου program.name-project.code. Η θεμελιώδης μονάδα οργάνωσης των δεδομένων στην Genomic Data Commons είναι η επανάληψη δείγματος (aliquots) ασθενούς και αφορά ένα μοναδικό πείραμα για ένα δείγμα ασθενούς.

1.5 Δίκτυα και τα χαρακτηριστικά τους

1.5.1 Επιστήμη Δικτύων

Τα δίκτυα βρίσκονται στην καρδιά των πολύπλοκων συστημάτων. Κάποιος μπορεί να καταλάβει την σημασία μελέτης των πολύπλοκων συστημάτων, όταν αναρωτηθεί το πόσο συχνά τα συναντάμε στην καθημερινή ζωή. Ξεκινώντας από τα μεγάλα και πολύπλοκα κοινωνικά δίκτυα, τα δίκτυα τηλεπικοινωνιών, τα παγκόσμια οδικά δίκτυα και καταλήγοντας στα πολύπλοκα

δίκτυα αλληλεπιδράσεων χιλιάδων γονιδίων με άλλα δεκάδες χιλιάδες βιομόρια, όπως μεταβολίτες, πρωτεΐνες και miRNA, που συνθέτουν την ίδια τη ζωή. Δεδομένης της σημασίας των πολύπλοκων συστημάτων και του ρόλου που έχουν τα δίκτυα σε αυτά, η κατανόηση, η μαθηματική τους μοντελοποίηση και περιγραφή, καθώς και η ανάπτυξη μοντέλων πρόγνωσης αποτελούν κάποιες από τις πιο βασικές προκλήσεις της επιστημονικής κοινότητας του 21ου αιώνα.

Η επιστήμη δικτύων (network science) είναι ένας νέος διεπιστημονικός τομέας. Η ανάδυση και η αυξανόμενη δημοτικότητα των δικτύων ως πλέον ξεχωριστός τομέας στην επιστημονική κοινότητα ξεκίνησε στην αρχές του 21ου αιώνα. Δύο επιστημονικές μελέτες του 20ού αιώνα σηματοδότησαν την αυξανόμενη ανάδυση της δικτυακής βιολογίας (Barabási and Pósfai 2016), η πρώτη αφορούσε την μελέτη των τυχαίων δικτύων (Random networks) στα πλαίσια της θεωρίας γράφων (Erdös and Rényi 1959) και η δεύτερη αφορά την επιστημονική μελέτη με τις περισσότερες αναφορές πάνω στα κοινωνικά δίκτυα (Granovetter 1973).

1.5.2 Θεωρία Γράφων

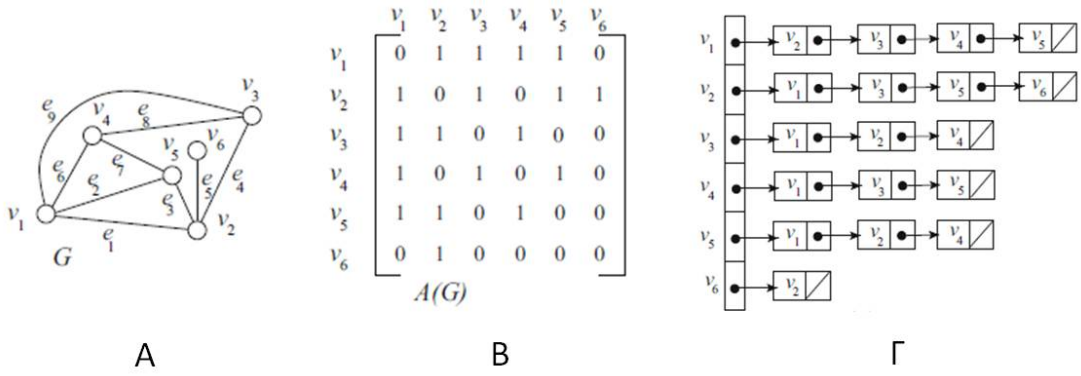
Η δικτυακή βιολογία είναι άρρηκτα συνδεδεμένη με την θεωρία γράφων. Η θεωρία γράφων (Graph Theory) ένας αναπτυσσόμενος τομέας των μαθηματικών που ο σκοπός του είναι η μελέτη και η ανάλυση δικτύων. Η θεωρία των γράφων θεωρείται ότι ξεκίνησε από τον Euler στις αρχές του 18ου αιώνα (1736).

1.5.2.1 Τρόποι αναπαράστασης ενός δικτύου

Το πρώτο βήμα για την κατανόηση ενός πολύπλοκου συστήματος είναι η μελέτη της αλληλεπίδρασης των συστατικών του δικτύου του. Ένας πρώτος απλός τρόπος να αναπαραστήσουμε τις σχέσεις ενός δικτύου είναι μέσω μίας λίστας γειτνίασης (adjacency list) που περιέχει όλα τα συστατικά του συστήματος και τις σχέσεις μεταξύ τους σε γειτονικές στήλες. Η λίστα γειτνίασης αναπαριστάται ως ένας πίνακας που κάθε στοιχείο του είναι μία λίστα (βλ. Σχήμα 5(Γ)). Η λίστα αυτή θα έχει μέγεθος ίσο με το άθροισμα των ακμών και των μη συνδεδεμένων κόμβων. Δύο βασικά μειονεκτήματα των λιστών γειτνίασης είναι ότι αρχικά δεν προτιμώνται για πυκνά δίκτυα και τέλος, χρειάζονται αρκετό υπολογιστικό χρόνο να βρουν αν μία ακμή υπάρχει ή όχι (Diestel 2016).

Στην περίπτωση πυκνών δικτύων δηλαδή δικτύων που περιέχουν πολλές ακμές, επιλέγεται ένας άλλος τρόπος αναπαράστασης που είναι οι πίνακες γειτνίασης (adjacency matrix). Σε αυτή την περίπτωση, ένα δίκτυο που πε-

ριέχει N αριθμό κόμβων θα αναπαριστάται μέσω ενός δισδιάστατου πίνακα μεγέθους N x N (βλ. Σχήμα 5(Β)). Κάθε στοιχείο ij του πίνακα έχει τιμή 1 αν υπάρχει σύνδεση μεταξύ των κόμβων i και j, και τιμή 0 αν δεν υπάρχει σύνδεση μεταξύ τους. Βέβαια, όπως θα αναφερθούμε και παρακάτω, στην περίπτωση δικτύου που περιέχει βάρη στις ακμές του, οι τιμές 1 και 0 αντικαθίστανται από τα σταθμισμένα βάρη που περιέχουν οι ακμές του δικτύου. Οι πίνακες γειτνίασης σε αντίθεση με τις λίστες, χρειάζονται μικρό υπολογιστικό χρόνο να εντοπίσουν μία ακμή. Ωστόσο, χρειάζονται αρκετό χώρο αποθήκευσης της τάξης του N^2, όπου N είναι ο αριθμός των κόμβων ενός γράφου G (Barabási and Pósfai 2016; Diestel 2016).



Σχήμα 5: (Α) Μή κατευθυνόμενος γράφος G χωρίς βάρη. Με e_x αναπαριστώνται οι ακμές και με v_x οι κόμβοι (Β) Αναπάρασταση πίνακα γειτνίασης μεγέθους ^{2} , όπου N το άθροισμα των κόμβων του G (Γ) Αναπάρασταση λίστας γειτνίασης μεγέθους N, όπου N το άθροισμα των κόμβων του G (Rahman 2017)

Ο τελευταίος και ο πιο δημοφιλής τρόπος αναπαράστασης ενός δικτύου είναι οι γράφοι, οι οποίοι προσφέρουν καλύτερη και πληρέστερη γνώση και κατανόηση των πολύπλοκων σχέσεων ενός συστήματος (βλ. Σχήμα 5(Α)). Στους γράφους τα συστατικά του δικτύου ονομάζονται κόμβοι (nodes/vertices) και οι σχέσεις μεταξύ τους, ονομάζονται ακμές (edges) (βλ. Σχήμα 8). Μια ακμή είναι ένα ζεύγος (u,v) από κόμβους. Τυπικά, ένας γράφος G ορίζεται από το σύνολο των κόμβων V και των ακμών του E, G (V, E).

1.5.2.2 Είδη δικτύων

Οι σχέσεις των στοιχείων ενός δικτύου μπορούν να περιγραφούν τόσο ποιοτικά όσο και ποσοτικά.

Ξεκινώντας με τα ποιοτικά χαρακτηριστικά ενός δικτύου, οι ακμές ενός γράφου μπορούν να έχουν κατεύθυνση (directed edges) ή να μην έχουν κατεύθυνση (undirected edges). Όπως απεικονίζεται στο Σχήμα 6, ένας γράφος ονο-

μάζεται *κατευθυνόμενος (directed graph)* αν κάθε μια από τις ακμές του είναι προσανατολισμένη προς μία κατεύθυνση και *μη κατευθυνόμενος (undirected graph)* αν οι ακμές του δεν έχουν κατεύθυνση. Τα μη κατευθυνόμενα δίκτυα αποτελούνται από σχέσεις οι οποίες είναι ταυτόσημες, δηλαδή δεν έχουν ποιοτικά χαρακτηριστικά. Αντίθετα, τα κατευθυνόμενα δίκτυα αποτελούνται από σχέσεις μεταξύ των κόμβων οι οποίες έχουν ποιοτικό χαρακτήρα, υποδηλώνοντας την προέλευση μίας λειτουργίας μεταξύ των κόμβων (Rahman 2017).

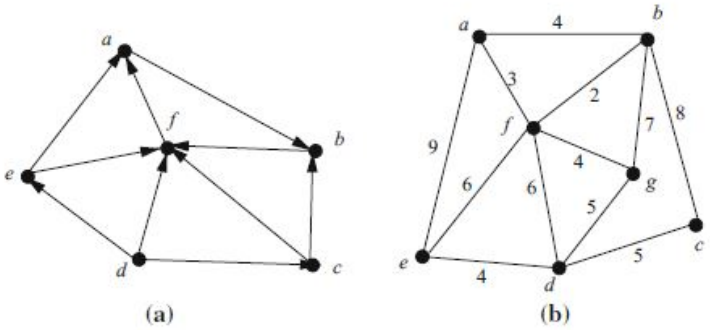
Ωστόσο, υπάρχουν γράφοι που μπορούν να συνδυάζουν κατευθυνόμενες και μη κατευθυνόμενες ακμές (Rahman 2017). Ένα χαρακτηριστικό παράδειγμα τέτοιου γράφου είναι ένα ρυθμιστικό δίκτυο (regulatory network) στο οποίο ορίζονται οι λειτουργικές αλληλεπιδράσεις της ρύθμισης της έκφρασης μεταξύ γονιδίων. Σε ένα τέτοιο δίκτυο αν το γονίδιο A ρυθμίζεται από το γονίδιο B, τότε αυτή η σχέση θα αναπαριστάται από μία κατευθυνόμενη ακμή που θα συνδέει το B προς το A.

Σχήμα 6: Στα αριστερά απεικονίζεται ένας μη-κατευθυνόμενος γράφος (Undirected Graph) που αποτελείται από 3 κόμβους 3 ακμές και στα δεξιά κατευθυνόμενος γράφος (Directed Graph) που αποτελείται από 3 κόμβους 3 ακμές.

Συχνά, οι σχέσεις μεταξύ των στοιχείων ενός δικτύου πέρα από ποιοτικά έχουν και ποσοτικά χαρακτηριστικά. Συγκεκριμένα, μία ακμή που περιγράφει τη σχέση μεταξύ δύο κόμβων σε ένα δίκτυο μπορεί πέρα από κατεύθυνση να έχει και ένα βάρος (Weight) που να εκφράζει την σταθμισμένη ποσότητα ενός συγκεκριμένου χαρακτηριστικού, όπως φαίνεται και στο Σχήμα 7.

1.5.2.3 Μέγεθος και τάξη

Στη συνέχεια, θα παραθέσουμε κάποιες βασικές ιδιότητες των γράφων, εστιάζοντας στις πιο σημαντικές.

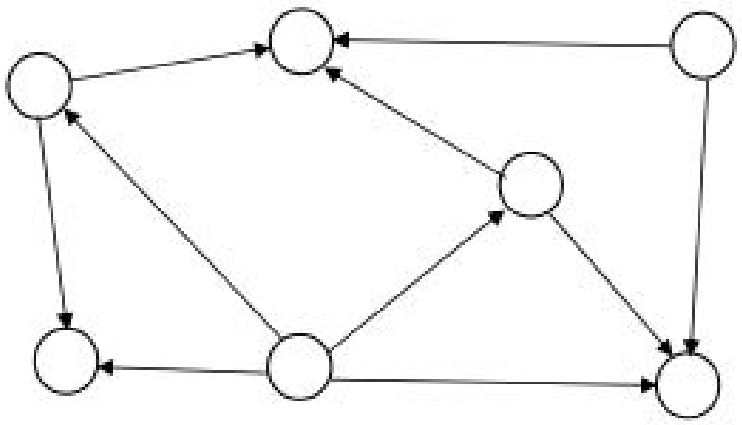


Σχήμα 7: Στα αριστερά απεικονίζεται ένας κατευθυνόμενος γράφος (Directed Graph) και στα δεξιά ένας μη κατευθυνόμενος γράφος με βάρη (Weighted Graph) (Rahman 2017).

Ξεκινώντας με δύο από τις πιο βασικές ιδιότητες των γράφων που είναι *ο αριθμός των ακμών* και *ο αριθμός των κόμβων*. Συγκεκριμένα,

- *Αριθμός των κόμβων- Τάξη*: αφορά ένα μέγεθος N το οποίο ισούται με τον συνολικό αριθμό των συστατικών του δικτύου (κόμβων) και ορίζει την τάξη (order) του γράφου
- *Αριθμός των ακμών - Μέγεθος*: αφορά ένα μέγεθος L το οποίο ισούται με τον συνολικό αριθμό των αλληλεπιδράσεων/ σχέσεων των συστατικών του δικτύου (κόμβων) και ορίζει το μέγεθος (size) του γράφου

Στο Σχήμα 8 απεικονίζεται ένα παράδειγμα ενός κατευθυνόμενου γράφου, που έχει Τάξη N= 7 και Μέγεθος L= 9.



Σχήμα 8: παράδειγμα ενός κατευθυνόμενου γράφου, που έχει συνολικό αριθμό κόμβων N= 7, Τάξη = 7 και συνολικό αριθμό ακμών L= 9, Μέγεθος = 9.

1.5.2.4 Βαθμός, μέσος βαθμός και κατανομή βαθμού

Ένα δίκτυο μεγάλου μεγέθους μπορεί να έχει ελάχιστες συνδέσεις μεταξύ των κόμβων του, ενώ παράλληλα, ένα πολύ μικρότερο δίκτυο μπορεί να είναι πολύ πιο έντονα συνδεδεμένο (πυκνότερο). Έτσι λοιπόν, για να περιγράψουμε ένα δίκτυο πέρα από το συνολικό του μέγεθος μεγάλη σημασία έχει και ο βαθμός συνδεσιμότητάς του. Για να περιγράψουμε την συνδεσιμότητα ενός γράφου χρησιμοποιούνται οι ιδιότητες του βαθμού του δικτύου.

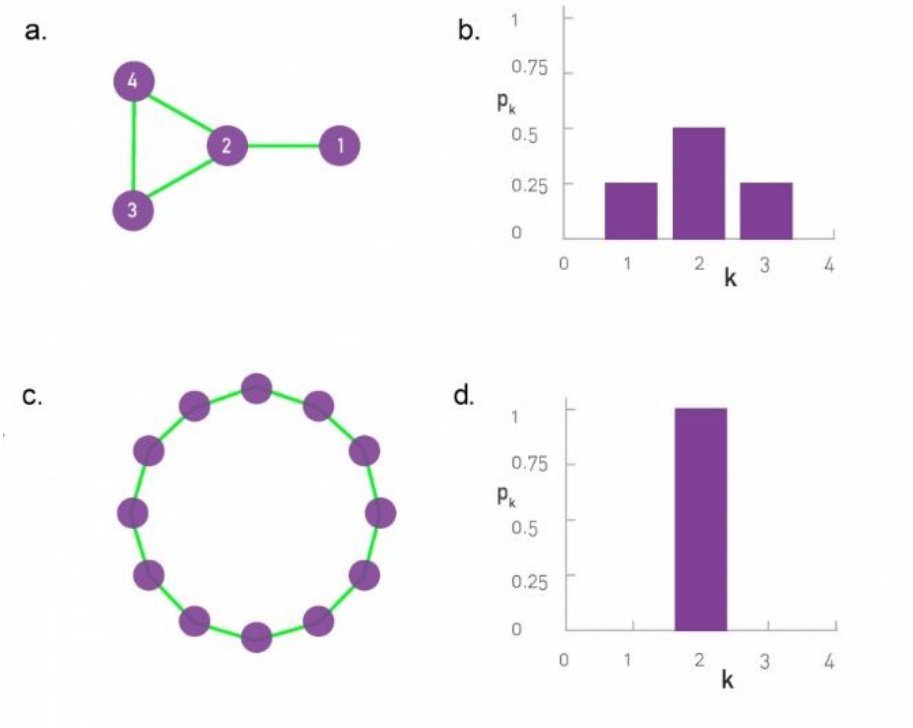
Ο βαθμός (degree) ενός κόμβου i ενός γράφου G, που ορίζεται ως k, αναπαριστά τον αριθμό των ακμών που συνδέουν τον κόμβο i με άλλους κόμβους. Για παράδειγμα, ο βαθμός k ενός κόμβου i σε ένα δίκτυο πρωτεϊνικών αλληλεπιδράσεων αναπαριστά τον αριθμό των πρωτεϊνών με τις οποίες αλληλεπιδρά η πρωτεΐνη που βρίσκεται στον κόμβο i. Κόμβοι με πολύ μεγάλο βαθμό (hubs) θεωρούνται πολύ σημαντικοί στα βιολογικά δίκτυα και η ύπαρξή τους είναι συχνή σε αυτά.

Σε ένα μη κατευθυνόμενο γράφο ο συνολικός αριθμός των ακμών, L, ισούται με το άθροισμα των βαθμών των κόμβων διαιρεμένο με το 2. Η διαίρεση του αθροίσματος των βαθμών των κόμβων διασφαλίζει πως κάθε ακμή θα μετρηθεί μία φορά. Σε κατευθυνόμενους γράφους, υπάρχουν δύο ειδών βαθμοί, ο βαθμός εισόδου (in-degree) και ο βαθμός εξόδου (out-degree) ενός κόμβου. Ο βαθμός εισόδου ενός κόμβου είναι ο αριθμός των ακμών που καταλήγουν στον κόμβο και βαθμός εξόδου ο αριθμός των ακμών που απομακρύνονται από τον κόμβο.

Επιπλέον, μέσω του βαθμού αναδύονται δύο επιπλέον σημαντικές ιδιότητες που είναι ο *μέγιστος βαθμός (maximum degree)* και ο *ελάχιστος βαθμός (minimum degree)* ενός γράφου. Ο μέγιστος βαθμός ενός γράφου G, ορίζεται ως Δ(G) και είναι η μέγιστη τιμή βαθμού από το σύνολο των βαθμών όλων των κόμβων του G. Ομοίως, ως ελάχιστος βαθμός ενός γράφου G ορίζεται ως δ(G) και είναι η ελάχιστη τιμή βαθμού από το σύνολο των βαθμών όλων των κόμβων του G.

Μία άλλη σημαντική ιδιότητα ενός δικτύου είναι ο *μέσος βαθμός (average degree)*. Για έναν κατευθυνόμενο γράφο ο μέσος βαθμός ενός κόμβου ισούται με το άθροισμα των ακμών που καταλήγουν στον κόμβο και των ακμών που απομακρύνονται από τον κόμβο.

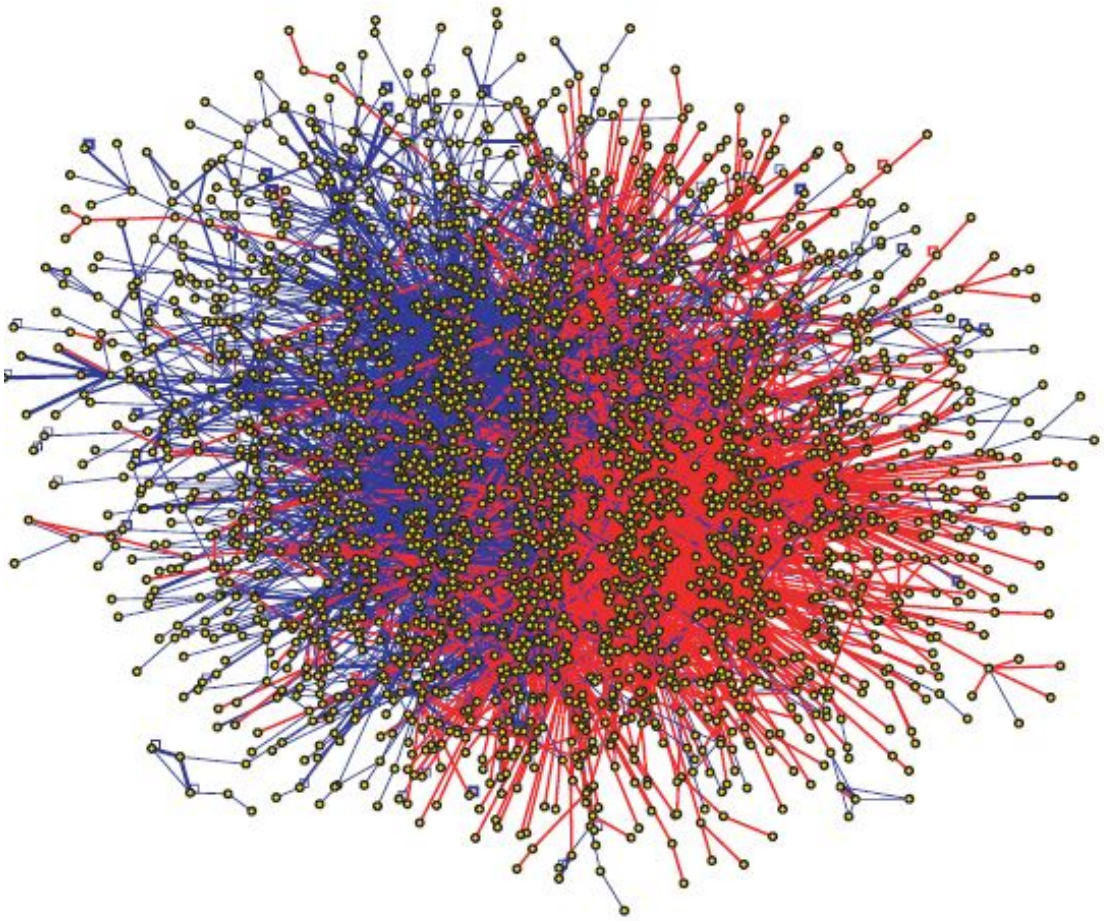
Τέλος, η κατανομή βαθμού (Degree Distribution), p, είναι μία επιπλέον ιδιότητα των γράφων που παρέχει πληροφορία για την πιθανότητα ένας τυχαίος κόμβος του δικτύου να έχει βαθμό k. Για ένα δίκτυο με N κόμβους η κατανομή βαθμού είναι ένα κανονικοποιημένο ιστόγραμμα , όπως φαίνεται και στο Σχήμα 9.



Σχήμα 9: Η κατανομή βαθμού του δικτύου που απεικονίζεται στο (a), φαίνεται στο (b). Συγκεκριμένα, ο κόμβος 1 έχει βαθμό k = 1 και p1 = 1/4 (διότι ένας από τους 4 κόμβους έχει βαθμό 1) , οι κόμβοι 3 και 4 έχουν βαθμό k = 2 και p2 = 1/2 (διότι 2 από τους 4 κόμβους έχουν βαθμό 2) και τέλος ο κόμβος 2 έχει βαθμό k = 3 και p3 = 1/4 (διότι ένας από τους 4 κόμβους έχει βαθμό 3). Ένα δίκτυο που όλοι του οι κόμβοι έχουν τον ίδιο βαθμό απεικονίζεται στο (c) και ακολουθεί το ιστόγραμμα της κατανομής του φαίνεται στο (d), (Barabási and Pósfai 2016).

Η κατανομή του βαθμού είναι καθοριστική σε πολλά φαινόμενα των δικτύων, όπως είναι η ευρωστία (robustness) που εκφράζει το πόσο ευάλωτο είναι το δίκτυο σε αλλαγές και επιθέσεις(Wagner 2000). Αντίθετα με το Σχήμα 9, στα βιολογικά δίκτυα οι βαθμοί των κόμβων διαφέρουν σε μεγάλο βαθμό. Ένα χαρακτηριστικό παράδειγμα είναι το δίκτυο πρωτεϊνικών αλληλεπιδράσεων του ανθρώπου που δημοσιεύτηκε το 2005 στο περιοδικό *Nature* και απεικονίζεται στο Σχήμα 10. Συγκεκριμένα, το ανθρώπινο δίκτυο πρωτεϊνικών αλληλεπιδράσεων είναι ένα πυκνό δίκτυο που περιέχει κόμβους με μεγάλο βαθμό. Επιπλέον, εμφανίζει την τάση αλληλεπίδρασης κόμβων με μεγάλο βαθμό με άλλους κόμβους που έχουν πολύ μικρότερο βαθμό. Τέλος, είναι σημαντικό να αναφερθεί πως ακολουθεί μία κατανομή νόμου- δύναμης (Power-law degree distribution), (Rual et al. 2005). Τα δομικά χαρακτηριστικά δικτύων που ακολουθούν την κατανομή νόμου- δύναμης είναι εξαιρετικά αποτελεσματικά στη μετάδοση πληροφορίας, την ταχεία επικοινωνία μεταξύ των μελών τους αλλά κυρίως τους προσδίδουν τη βασική ιδιότητα της ευρωστίας

(robustness) (Wagner 2000).



Σχήμα 10: Το δίκτυο πρωτεϊνικών αλληλεπιδράσεων του ανθρώπου, (Human Interactome) (Rual et al. 2005). Με κίτρινο χρώμα απεικονίζονται οι πρωτεΐνες που αποτελούν τους κόμβους του δικτύου. Ακμές με κόκκινο και μπλε χρώμα συμβολίζονται οι επιβεβαιωμένες πρωτεϊνικές αλλελπιδράσεις μέσω πειράματος υψηλής απόδοσης και μέσω βιβλιογραφικής αναζήτησης, αντίστοιχα.

1.5.2.5 Μονοπάτια και αποστάσεις

Στον μικρόκοσμο, οι αποστάσεις έχουν καθοριστικό ρόλο στο τρόπο αλληλεπίδρασης μεταξύ των βιομορίων. Για παράδειγμα, η απόσταση δύο πρωτεϊνών στο κυτταρόπλασμα ενός κυττάρου καθορίζει αν αυτές οι δύο πρωτεΐνες θα αλληλεπιδράσουν.

Συχνά στην ανάλυση δικτύων μας ενδιαφέρει να μελετήσουμε την σχέση μίας πληθώρας κόμβων οι οποίοι δεν συνδέονται άμεσα μεταξύ τους. Ένας τρόπος για να περιγράψουμε την σχέση μεταξύ τέτοιων κόμβων είναι η δικτυακή απόσταση που εκφράζεται μέσω της *απόστασης μονοπατιού (path length)*. Ως απόσταση μονοπατιού ορίζουμε τον αριθμό των ακμών που περιέχει το εκάστοτε μονοπάτι.

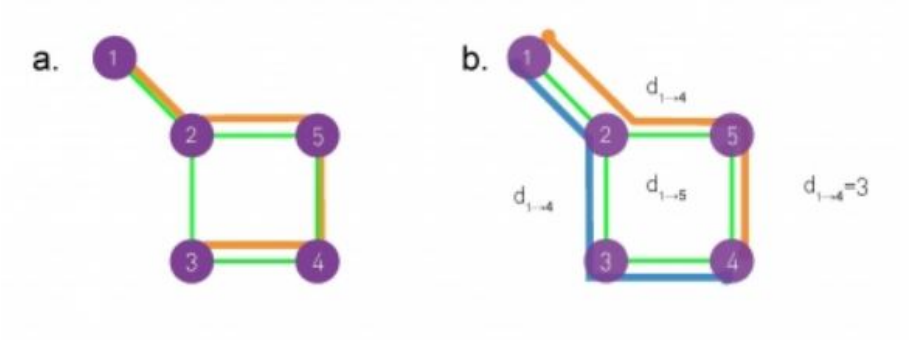
Ωστόσο, ο υπολογισμός της δικτυακής απόστασης δύο κόμβων, ειδικότερα στην περίπτωση πυκνών δικτύων, μπορεί να γίνει με πολλούς τρόπους, αφού προκύπτουν πολλά διαφορετικά μονοπάτια διαφορετικής απόστασης που μπορούν να συνδέουν τους δύο κόμβους (βλ. Σχήμα 11). Για αυτό το λόγο, στην ανάλυση δικτύων χρησιμοποιείται συχνά μία σημαντική ιδιότητα των γράφων που είναι το *ελάχιστο μονοπάτι (Shortest Path).* Ως ελάχιστο μονοπάτι μεταξύ των κόμβων ενός δικτύου, ορίζεται η ελάχιστη απόσταση ακμών που μπορεί κάποιος να διατρέξει για να βρεθεί από τον ένα κόμβο στον άλλο. Όταν δύο κόμβοι χαρακτηρίζονται από μικρή απόσταση ελάχιστου μονοπατιού τότε μπορούμε να πούμε πως αυτό το ζεύγος κόμβων έχει στενότερη σχέση σε σύγκριση με άλλους κόμβους που έχουν μεγαλύτερη δικτυακή απόσταση ελάχιστου μονοπατιού μεταξύ τους. Σε μικρά δίκτυα ο υπολογισμός του ελάχιστου μονοπατιού είναι σχετικά εύκολος (βλ. Σχήμα 11), ωστόσο, όταν η ανάλυση αφορά ένα πραγματικό βιολογικό δίκτυο όπως η ανάλυση των ανθρώπινων πρωτεϊνικών αλληλεπιδράσεων (βλ.Σχήμα 10), τότε είναι φανερό πως η ανάλυση αντιμετωπίζει πολλές προκλήσεις. Για αυτό το σκοπό, έχει αναπτυχθεί πληθώρα αλγορίθμων για τον υπολογισμό του ελάχιστου μονοπατιού, με πιο δημοφιλή να θεωρείται ο Breadth First Search (BFS), ο οποίος υπολογίζει την απόσταση μεταξύ δύο κόμβων i και j, ξεκινώντας από τον κόμβο i και προχωρώντας μέσω διαδοχικών ελέγχων των άμεσων γειτόνων μέχρι να φτάσει στο κόμβο j (Zhou and Hansen 2006).

Επιπλέον, η ιδιότητα του ελάχιστου μονοπατιού χρησιμοποιείται συχνά για τον υπολογισμό μίας επιπλέον ιδιότητας των γράφων που είναι η *διάμετρος του δικτύου*. Η διάμετρος του δικτύου, ορίζεται ως d_{max} και ισούται με τη μέγιστη απόσταση ελάχιστου μονοπατιού μεταξύ οποιοδήποτε κόμβων στο δίκτυο.

1.5.2.6 Συνδεσιμότητα

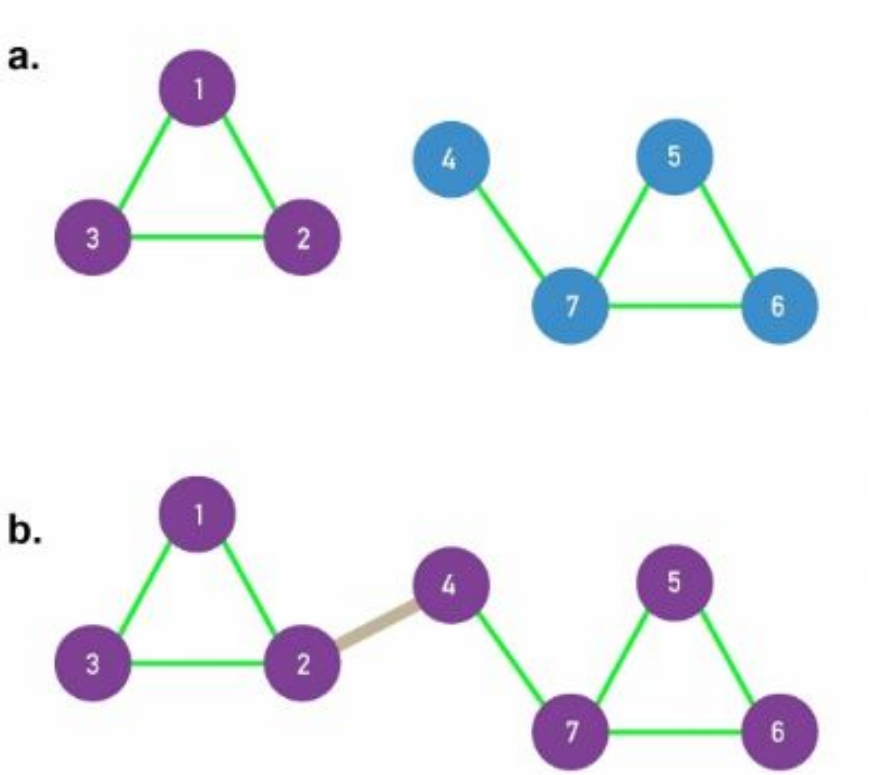
Ονομάζουμε ένα δίκτυο συνδεδεμένο (connected network) όταν όλα τα ζευγάρια κόμβων του είναι συνδεδεμένα μεταξύ τους. Αντίθετα, ένα δίκτυο ονομάζεται μη συνδεδεμένο όταν τουλάχιστον ένα ζευγάρι κόμβων δεν είναι συνδεδεμένο (βλ. Σχήμα 12).

Η συνδεσιμότητα (Connectivity/Connectedness) αποτελεί κεντρική ιδιότητα των περισσότερων δικτύων. Ως συνδεσιμότητα κ(G) ενός συνδεδεμένου γράφου G ορίζεται ως ο ελάχιστος αριθμός κόμβων που μπορούν να αφαιρεθούν ώστε να καταλήξουμε σε έναν μη συνδεδεμένο γράφο. Επιπλέον, ως συνδεσιμότητα ακμών κ'(G) ενός συνδεδεμένου γράφου G ορίζεται ως ο ελάχιστος αριθμός ακμών που μπορούν να αφαιρεθούν ώστε να καταλήξουμε σε



Σχήμα 11: Ως μονοπάτι ορίζεται μία σειριακή αλληλουχία κόμβων που αποτελείται από N+1 κόμβους και N ακμές (Barabási and Pósfai 2016). Με πράσινο χρώμα αναπαριστώνται οι ακμές του δικτύου και με μωβ χρώμα οι αριθμημένοι κόμβοι. Στο (b) παρατηρούμε με μπλε και πορτοκαλί χρώμα τα δύο εναλλακτικά μονοπάτια από τον κόμβο 1 στον κόμβο 4. Συγκεκριμένα, με μπλε ορίζεται το μονοπάτι 1 - 2 - 3 - 4 και με πορτοκαλί το μονοπάτι 1 - 2 -5 - 5. Και τα δύο μονοπάτια, έχουν μήκος 3.

έναν μη συνδεδεμένο γράφο (Rahman 2017).



Σχήμα 12: a. Παράδειγμα μη συνδεδεμένου δικτύου. Τόσο οι κόμβοι 1,2,3 όσο και οι 4,5,6,7 συνδέονται μεταξύ τους. Ωστόσο δεν υπάρχει κάποια σύνδεση/μονοπάτι μεταξύ των δύο υπο-ομάδων του δικτύου. b. Παράδειγμα συνδεδεμένου δικτύου. Παρατηρούμε πως υπάρχει μονοπάτι μεταξύ όλων των ζευγαριών κόμβων του δικτύου. Με γκρι αναπαριστάται η ακμή που ενώνει τις δύο υπο-ομάδες που ήταν πριν ξεχωριστές. Η γκρί ακμή ονομάζεται ως *γέφυρα* (Barabási and Pósfai 2016).

1.6 Μηχανική μάθηση

1.6.1 Εισαγωγή στην Μηχανική Μάθηση

Κάποιες από τις μεγαλύτερες προκλήσεις στην μελέτη και στην ανάλυση πολύπλοκων συστημάτων είναι η κατανόηση της φύσης των δεδομένων και η μοντελοποίηση της δομής τους. Βρισκόμενοι στην εποχή των μεγάλων δεδομένων (big data) αυτές οι προκλήσεις αποκτούν όλο και περισσότερη σημασία, καθώς οι επιστήμονες είναι πλέον αντιμέτωποι με τεράστια ποσά πληροφορίας η οποία χαρακτηρίζεται από πολυδιάστατα δεδομένα, τα οποία συχνά περιέχουν μεγάλα ποσά ”θορύβου”. Έτσι λοιπόν, τις τελευταίες δεκαετίες αποτέλεσε το κέντρο των ενδιαφερόντων πληθώρας επιστημονικών πεδίων η εύρεση και η μελέτη αποτελεσματικών αλγορίθμων η οποίοι θα είναι ικανοί να πραγματοποιούν ανάλυση των δεδομένων, μοντελοποίηση του προς

μελέτη συστήματος καθώς και πρόγνωση στο μέλλον. Αυτή η πληθώρα επιστημονικών πεδίων που συνέβαλλε σε αυτόν τον στόχο όπως η Επιστήμη των Υπολογιστών (Computer Science), η Εξόρυξη Δεδομένων (Data Mining), η Στατιστική (Statistics), η Αναγνώριση Προτύπων (Pattern Recognition), η Επεξεργασία και Ανάλυση Σήματος και Ήχου (Signal and Image Processing and Analysis), και η Βιοπληροφορική (Bioinformatics), σήμερα στεγάζονται κάτω από την ίδια ομπρέλα που ονομάζεται *Μηχανική Μάθηση*.

Ο αρχικός ορισμός της Μηχανικής Μάθησης (Machine Learning) δόθηκε το 1959 από τον Arthur Samuel ο οποίος θεωρείται πλέον πρωτοπόρος στην τεχνική νοημοσύνη. Συγκεκριμένα, ο Samuel οριοθέτησε την μηχανική μάθηση ως το *πεδίο μελέτης το οποίο δίνει την ικανότητα στους υπολογιστές να μαθαίνουν μία εργασία χωρίς να έχουν προγραμματιστεί εντελώς για αυτήν* (Samuel 1959). Η ιστορία πίσω από την σύλληψη του ορισμού της Μηχανικής μάθησης ξεκινάει το 1950, όταν ο Arthur Samuel δημιούργησε ένα πρόγραμμα για σκάκι. Το πρόγραμμα χρησιμοποιούσε την δομή των Δέντρων Αναζήτησης (Search Tree) και μία συνάρτηση υπολογισμού σκορ, προκειμένου ο αλγόριθμος να βρει την καλύτερη κίνηση στο σκάκι και τελικά να κερδίσει την παρτίδα σκακιού. Το σημαντικό με αυτό το πρόγραμμα ήταν πως σε κάθε νέο παιχνίδι "θυμόταν" όλες τις κινήσεις που είχε ήδη δει και επιπλέον, είχε υποβληθεί σε δεκάδες χιλιάδες παιχνίδια με τον εαυτό του. Έτσι, το πρόγραμμά του Samuel ενώ ξεκίνησε από ένα αρχάριο επίπεδο κατάφερε με τον καιρό να αποκτήσει ένα υψηλό επίπεδο εμπειρίας ικανό να μπορεί να αντιμετωπίζει κάποιους από τους καλύτερους παίκτες σκακιού.

Ωστόσο, το 1997 αποδόθηκε από τον Tom Mitchell ένας πιο μοντέρνος ορισμός της Μηχανικής μάθησης, ο οποίος είναι ο εξής: *"Ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από μία εμπειρία E ως προς μια κλάση εργασιών T και ένα μέτρο απόδοσης P, αν η απόδοσή του σε εργασίες της κλάσης T, όπως εκτιμάται από το μέτρο απόδοσης P, βελτιώνεται με την εμπειρία E*". Σύμφωνα με αυτόν τον ορισμό, το παράδειγμα με το πρόγραμμα σκακιού μεταφράζεται ως εξής:

- E = η εμπειρία που κερδίζει μέσω των πολλαπλών παιχνιδιών σκακιού
- T = η εργασία του υπολογιστή να παίξει σκάκι
- P = η πιθανότητα νίκης του υπολογιστή σε μία παρτίδα σκάκι.

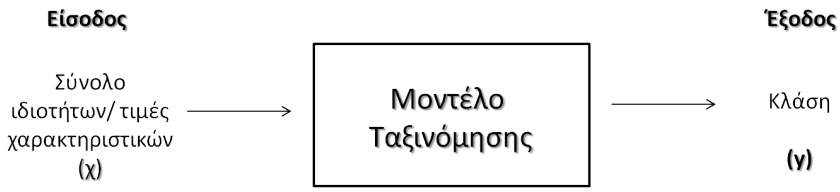
1.6.2 Ταξινόμηση

Ένας από τους πιο ώριμους και πιο δημοφιλείς τύπους μηχανικής μάθησης είναι η *Ταξινόμηση (Classification)*. Η ταξινόμηση αποτελεί μία εργασία

ανάθεσης ενός αγνώστου μοτίβου σε μία ή περισσότερες γνωστές κατηγορίες. Αυτή η διαδικασία είναι γνωστή σε πάρα πολλά προβλήματα και προκλήσεις τόσο της καθημερινής ζωής όσο και του τομέα της έρευνας. Τέτοια προβλήματα περιλαμβάνουν την κατηγοριοποίηση των ηλεκτρονικών μηνυμάτων σε ανεπιθύμητων και μη, ηλεκτρονικών μηνυμάτων (spam- non spam), λαμβάνοντας πληροφορία από τον τίτλο τους και το περιεχόμενό τους. Επίσης, στους τομείς της βιοϊατρικής και της βιοπληροφορικής συχνή είναι η προσπάθεια κατηγοριοποίησης πολλαπλών προφιλ ασθενών σε επιμέρους υποτύπους της προς μελέτη ασθένειας.

Ως *ταξινομητής (classifier)* ή *μοντέλο ταξινόμησης* ορίζεται ένα σύστημα το οποίο έχει ως είσοδο ένα διάνυσμα (vector) από διακριτές και/ή συνεχείς *τιμές χαρακτηριστικών (feature values)* και έχει ως έξοδο μία μοναδική διακριτή τιμή , *την κλάση (class).* (Domingos 2012) (βλ. Σχήμα 13). Παραδείγματα ταξινομητών αποτελούν μεταξύ άλλων τα νευρωνικά δίκτυα, μηχανές διανυσματικής υποστήριξης και τα δέντρα αποφάσεων.

Σε ένα μοντέλο ταξινόμησης, η είσοδος των δεδομένων (Input) αποτελεί στην ουσία ένα σύνολο από εγγραφές. Κάθε εγγραφή ονομάζεται και *στιγμιότυπο (instance)* και χαρακτηρίζεται από πολλά ζευγάρια τιμών (χ,y), αριθμού ίσου με το σύνολο των χαρακτηριστικών. Το χ αντιστοιχεί σε μία τιμή ενός χαρακτηριστικού από το σύνολο των χαρακτηριστικών (features) που ονομάζεται και ως *γνώρισμα (attribute)* και το y αφορά μία ετικέτα κλάσης (Class Label) (Tan, Steinbach, and Kumar 2006). Για παράδειγμα αν επιθυμούσε κάποιος να αναπτύξει ένα μοντέλο το οποίο να ταξινομεί ασθενείς με καρκίνο του μαστού και υγιείς ανθρώπους, τότε θα έπρεπε να συλλέξει δεδομένα από ασθενείς και από υγιείς, οι οποίες θα ήταν οι δύο ετικέτες κατηγοριών (τιμές y για κάθε εγγραφή) και να τα οργανώσει σε τέτοια μορφή ώστε η τιμή χ κάθε εγγραφής να αφορά ενδεχομένως κάποιο μοριακό χαρακτηριστικό όπως η έκφραση ενός από το σύνολο των γονιδίων ή/και τη συγκέντρωση μίας από το σύνολο των πρωτεϊνών που περιέχει το σύνολο των δεδομένων.



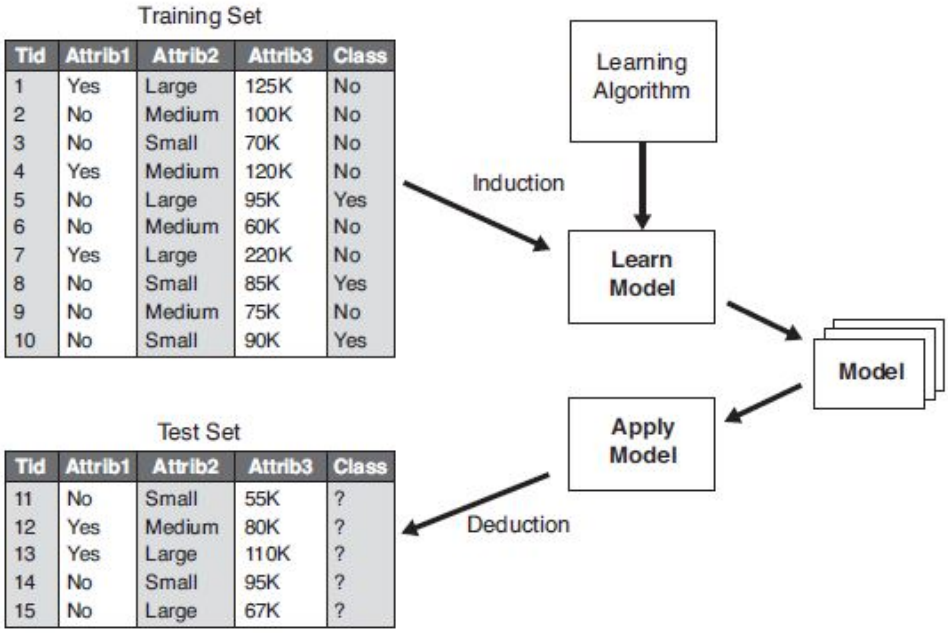
Σχήμα 13: Σχεδιαγραμματική αναπαράσταση του μοντέλου ταξινόμησης. Ως είσοδο (x) του μοντέλου είναι ένα σύνολο χαρακτηριστικών και ως έξοδο (y) η αντιστοίχιση του συνόλου σε μία γνωστή κλάση

Μετά από την εξόρυξη των δεδομένων, το πρώτο και σημαντικό βήμα

στην ανάπτυξη ενός μοντέλου ταξινόμησης είναι η λήψη αποφάσεων σχετικά με την επιλογή των κατάλληλων αναπαραστάσεων που θα πρέπει να έχουν τα δεδομένα. Συχνά, επιλέγεται η αναπαράσταση ενός διανύσματος, το οποίο ονομάζεται *διάνυσμα χαρακτηριστικών (feature vector)*. Αυτό το διάνυσμα στην ουσία μπορεί να περιγραφεί μέσω ενός πίνακα που έχει μήκος ίσο με το σύνολο των χαρακτηριστικών. Κάθε στιγμιότυπο χαρακτηρίζεται από F ζευγάρια τιμών (χ,y), όπου F είναι ο αριθμός των χαρακτηριστικών, έτσι ώστε κάθε στιγμιότυπο να έχει μία συγκεκριμένη θέση σε έναν F-διάστατο χώρο που ονομάζεται *χώρος των χαρακτηριστικών (feature space)*. Ωστόσο, σπάνια σε μελέτες μηχανικής μάθησης το μοντέλο ταξινόμησης χρησιμοποιεί το αρχικό σύνολο των χαρακτηριστικών καθώς πολλές φορές τα χαρακτηριστικά είναι περισσότερα από τα ίδια τα στιγμιότυπα, κάνοντας την ταξινόμηση ακόμη πιο δύσκολη. Για αυτό το σκοπό, συχνά πραγματοποιείται ένα επιπλέον βήμα, αυτό της *επιλογής των πιο σημαντικών χαρακτηριστικών (feature selection)* (Duda, Hart, and Stork 2001).

1.6.2.1 Κατασκευή του μοντέλου ταξινόμησης

Επόμενο βήμα σε ένα πρόβλημα ταξινόμησης αποτελεί η κατασκευή του μοντέλου (βλ. Σχήμα 14). Ειδικότερα, πρώτο στάδιο της κατασκευής του μοντέλου είναι η εκπαίδευση του ταξινομητή. Αυτό πραγματοποιείται μέσω της επιλογής ενός *συνόλου εκπαίδευσης (training set)* το οποίο περιέχει δεδομένα από εγγραφές που έχουν γνωστή κλάση. Μετά το στάδιο εκπαίδευσης, πραγματοποιείται ο σχεδιασμός μία συνάρτησης *f*, η οποία στην ουσία αποτελεί τον ίδιο τον ταξινομητή, και σκοπό έχει η κατασκευή ενός μοντέλου το οποίο θα μπορεί να ενσωματώνει με τον καλύτερο τρόπο τη σχέση μεταξύ των ιδιοτήτων του συνόλου εισόδου και των ετικετών των κλάσεων. Αφού, επιλεχθεί η κατάλληλη συνάρτηση και έχει κατασκευαστεί το μοντέλο ταξινόμησης τότε το μοντέλο θα πρέπει να είναι ικανό να πραγματοποιεί όσο το δυνατόν πιο σωστές προγνώσεις ταξινόμησης σε εγγραφές που δεν έχει ξαναδεί. Αυτή η διαδικασία εφαρμογής του μοντέλου πραγματοποιείται σε ένα *σύνολο ελέγχου (testing set)* το οποίο αποτελείται από εγγραφές που δεν περιέχουν πληροφορία σχετική με την κλάση τους (Breiman et al. 1984). Για παράδειγμα, έστω ότι οι κλάσεις του προβλήματος ταξινόμησης είναι δύο, τότε εγγραφές από το σύνολο ελέγχου θα ταξινομηθούν είτε στη πρώτη κλάση είτε στη δεύτερη, όπως φαίνεται και στο Σχήμα 16. Στην πιο απλή περίπτωση, η συνάρτηση f του ταξινομητή είναι γραμμική χωρίζοντας το επίπεδο σε δύο μέρη.

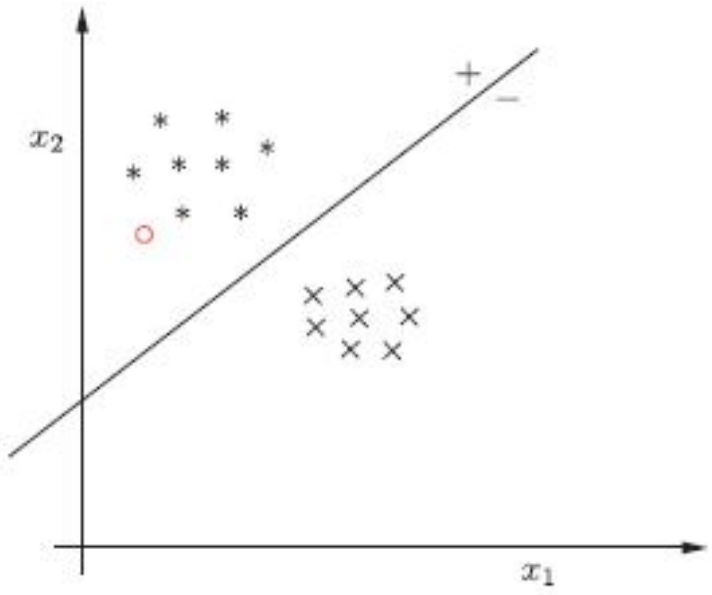


Σχήμα 14: Σχεδιαγραμματική αναπαράσταση της κατασκευής ενός μοντέλου ταξινόμησης. Πρώτο στάδιο είναι η εκπαίδευση του μοντέλου, μέσω του συνόλου εκπαίδευσης (training set) και στη συνέχεια η σχεδίαση ενός κατάλληλου αλγορίθμου εκμάθησης. Αφού έχει κατασκευαστεί το μοντέλο, τελικά εφαρμόζεται σε ένα σύνολο ελέγχου (testing set) (Tan, Steinbach, and Kumar 2006).

1.6.3 Πρόβλεψη τιμής (regression)

Πέρα από τα προβλήματα ταξινόμησης στα οποία επιθυμούμε να κατατάξουμε ένα προφίλ σε κάποια κλάση, συχνά είναι και τα προβλήματα πρόγνωσης μίας τιμής. Τέτοια προβλήματα αντιμετωπίζονται μέσω *της παλινδρόμησης (regression)* στόχος της οποίας είναι η πρόβλεψη της τιμής της εξόδου μίας μεταβλητής όταν είναι γνωστές οι τιμές εισόδου.

Τυπικά, η παλινδρόμηση *είναι μία διαδικασία προσδιορισμού της σχέσης μιας μεταβλητής y (εξαρτημένη μεταβλητή ή έξοδος) με μια ή περισσότερες άλλες μεταβλητές* $x_{[n]}$ *(ανεξάρτητες μεταβλητές ή είσοδοι)* (Breiman et al. 1984). Ειδικότερα, σε πειραματικές μελέτες η ανεξάρτητη μεταβλητή X αποτελεί την παράμετρο της οποίας μπορούμε να καθορίσουμε τις τιμές της (όπως η εκτίμηση γονιδιακής έκφρασης), ενώ εξαρτημένη μεταβλητή Y είναι εκείνη η παράμετρος στην οποία γίνονται αντανάκλαση οι τιμές των ανεξάρτητων μεταβλητών X (όπως η βιωσιμότητα ασθενούς). Όταν οι δύο αυτές μεταβλητές συνδέονται μέσω μίας γραμμικής σχέσης του τύπου $Y = f(X)$ και μπορούμε ακριβώς να προβλέψουμε την τιμή της μεταβλητής Ψ βάση της τιμής X, τότε αυτές οι δύο μεταβλητές συνδέονται μέσω *συναρτησιακής - προσδιοριστικής (deterministic) σχέσης*. Αντίθετα, στην περίπτωση που η σχέση μεταξύ εξαρτημένης και ανεξάρτητων μεταβλητών δεν είναι προσδιοριστικές, και το



Σχήμα 15: Ο ταξινομητής (σε αυτή την απλή περίπτωση) έχει σχεδιαστεί για να χωρίζει τα δεδομένα εκπαίδευσης σε δύο κλάσεις, έχοντας στην θετική του πλευρά τα σημεία που προέρχονται από μια τάξη και στην αρνητική πλευρά του τα άλλα. Το "κόκκινο" σημείο, του οποίου η κλάση είναι άγνωστη, κατατάσσεται στην ίδια κλάση με τα "αστέρια", αφού βρίσκεται στην θετική πλευρά του ταξινομητή (Theodoridis 2015).

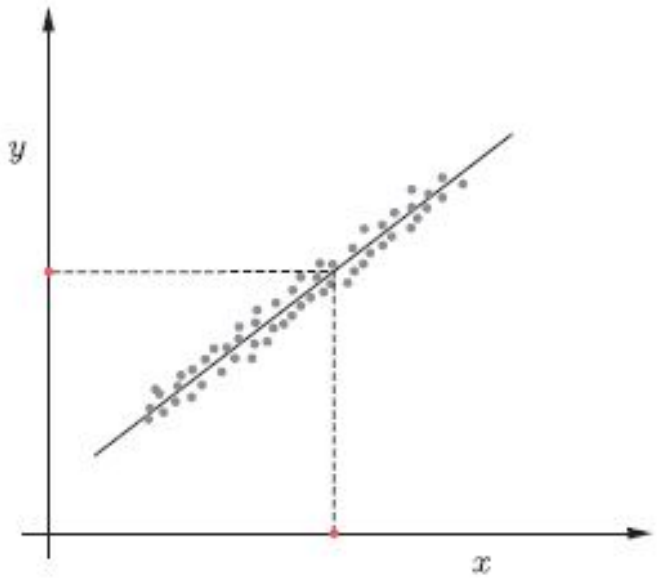
διάγραμμα διασποράς τους αφορά ένα νέφος σημείων γύρω από μια καμπύλη τότε λέμε πως αυτές οι σχέσεις είναι *στοχαστικές – στατιστικές (stochastic, probabilistic) σχέσεις*.

Μια μέθοδος που χρησιμοποιείται για την περιγραφή της στοχαστικής εξάρτησης μεταβλητών είναι η *μέθοδος των ελαχίστων τετραγώνων*. Στην πιο απλή περίπτωση όπου το μελετούμενο σύστημα αποτελείται από δύο μεταβλητές, τότε η προσέγγιση της μεθόδου έχει την μορφή του τύπου (1)

$$(y_j) = \alpha + \beta \times (x_j) + (\epsilon_j) \tag{1}$$

όπου y_j είναι η πραγματική τιμή, (α + β χ x_j) είναι η θεωρητική τιμή και ϵ_j είναι οι αποκλίσεις από την πραγματική τιμή y_j. Επομένως, σκοπός της προσέγγισης είναι να εκτιμηθούν οι α και β έτσι ώστε να ελαχιστοποιείται η τιμή του ε.

Μία βασική διαφορά μεταξύ της ταξινόμησης και της παλινδρόμησης είναι, πως στην τελευταία, η μεταβλητή εξόδου y δεν είναι διακριτή αλλά παίρνει συνεχείς τιμές σε ένα διάστημα στον πραγματικό άξονα. Η διαδικασία της παλινδρόμησης αποτελεί ουσιαστικά ένα πρόβλημα προσαρμογής καμπύλης.



Σχήμα 16: Απεικονίζεται το γράφημα μιας γραμμικής συνάρτησης f η οποία έχει σχεδιαστεί ώστε να ταιριάζει με το σύνολο εκπαίδευσης (x) (γκρι σημεία). Δεδομένης μίας τιμής χ, (κόκκινο σημείο) η πρόβλεψη της αντίστοιχης τιμής εξόδου δίνεται από το y = f (x) (Theodoridis 2015).

1.6.4 Αξιολόγηση επίδοσης στη μηχανική μάθηση

Μετά την κατασκευή του μοντέλου πρόγνωσης ή ταξινόμησης, το τελευταίο και πιο σημαντικό βήμα είναι αυτό της αξιολόγησης της επίδοσης του μοντέλου. Η αξιολόγηση βασίζεται στον αριθμό των σωστών και των λάθος προγνώσεων που έγιναν από το μοντέλο στο σύνολο ελέγχου.

Στη δυαδική ταξινόμηση (ταξινόμηση των δεδομένων σε δύο κατηγορίες) χρησιμοποιούνται συχνά οι έννοιες των True Positive, True Negative, False Positive και False Negative. Το πιο βασικό μέτρο αξιολόγησης και πιο συχνά χρησιμοποιούμενο σε μελέτες ταξινόμησης είναι η *ακρίβεια (accuracy)*. Η ακρίβεια υπολογίζεται από τον τύπο (2) και είναι ο λόγος του αριθμού των σωστών προγνώσεων προς τον συνολικό των προγνώσεων που πραγματοποιήθηκαν. Με άλλα λόγια, η ακρίβεια εκφράζει το πόσο κοντά βρίσκεται μία πρόγνωση στο σωστό αποτέλεσμα.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

όπου,

TP (True Positive) = τιμές που προγνώστηκαν ως θετικές και ήταν όντως θετικές

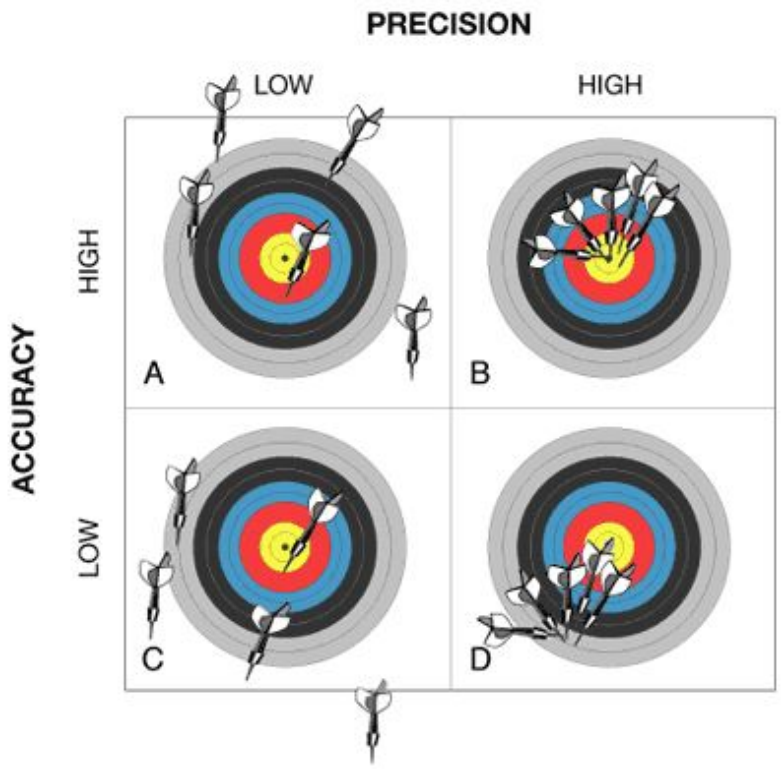
TN (True Negative) = τιμές που προγνώστηκαν ως αρνητικές και ήταν όντως

αρνητικές
FP (False Positive) = τιμές που προγνώστηκαν ως θετικές και ήταν αρνητικές
FN (False Negative) = τιμές που προγνώστηκαν ως αρνητικές και ήταν θετικές

Μία άλλη μετρική για την εκτίμηση της απόδοσης ενός αλγορίθμου ταξινόμησης είναι η *πιστότητα (precision)*. Τυπικά, η πιστότητα υπολογίζεται από τον τύπο (3) και είναι ο λόγος του αριθμού των σωστών προγνώσεων προς τον συνολικό των προγνώσεων που πραγματοποιήθηκαν. Η πιστότητα εκφράζει το πόσο συνεπής είναι το μοντέλο πρόγνωσης σε σωστές προβλέψεις.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Όπως παρουσιάζεται και στο Σχήμα 17, ένα μοντέλο ταξινόμησης δεν αρκεί απλά να χαρακτηρίζεται μόνο από καλή ακρίβεια ή μόνο από καλή πιστότητα, έτσι ώστε να αποτελεί ένα αξιόπιστο μοντέλο.



Σχήμα 17: Αναπαράσταση των ανεξάρτητων κριτηρίων της ακρίβειας (accuracy) και της πιστότητας (precision). Κάθε στήλη έχει τον ίδιο βαθμό πιστότητας και κάθε γραμμή τον ίδιο βαθμό ακρίβειας. Ο συνδυασμός υψηλής ακρίβειας και υψηλής πιστότητας όπως φαίνεται στο (B) χαρακτηρίζουν ένα αξιόπιστο μοντέλο πρόγνωσης. (http://starasp.uit.yorku.ca/psycho/en/postscript.asp, 27/8/2018)

Άλλη χρήσιμη μετρική είναι η *Ανάκληση (Recall)* ή *Αληθής Θετικός Ρυθμός (True Positive Rate)* ή *Ευαισθησία (Sensitivity)* η οποία υπολογίζεται από τον

τύπο (4) και εκφράζει το σύνολο των θετικών αποτελεσμάτων που όντως ήταν θετικά.

$$Precision = \frac{TP}{TP + FN} \tag{4}$$

Επιπλέον, ένα άλλο μέτρο το οποίο χρησιμοποιείται συχνά ως μέτρο αξιολόγησης ενός μοντέλου πρόγνωσης είναι το F-measure ή F1- score, το οποίο συνδυάζει τις μετρικές της ανάκλησης και της πιστότητας και υπολογίζεται σύμφωνα με τον τύπο (5).

$$F - measure = \frac{2 \times precision \times recall}{precision + recall} \tag{5}$$

Επιπρόσθετα, μία μετρική που χρησιμοποιείται για την εκτίμηση μοντέλων ταξινόμησης είναι το *Standard Error of the mean*, το οποίο είναι ένα μέτρο που υπολογίζει την τυπική απόκλιση της κατανομής των δειγμάτων και ισούται με τον μαθηματικό τύπο 6:

$$\text{Standard Error of the mean} = \frac{\text{Τυπική απόκλιση}}{\sqrt[2]{\text{Αριθμός δειγμάτων}}} \tag{6}$$

Το Mean Squared Error (MSE) είναι ένα μέγεθος εκτίμησης της προγνωστικότητας ενός μοντέλου πρόγνωσης τιμής. Ειδικότερα το MSE υπολογίζει τον μέσο όρο του τετραγώνου της διαφοράς μεταξύ των πραγματικών τιμών και των προβλεπόμενων τιμών. Το MSE υπολογίζεται από τον τύπο (7).Το πλεονέκτημα του MSE, όπως φαίνεται και από τον τύπο είναι πως η επίδραση των μεγαλύτερων σφαλμάτων γίνεται πιο έντονη σε σχέση με μικρότερα σφάλματα, και έτσι, το μοντέλο μπορεί πλέον να επικεντρωθεί περισσότερο στα μεγαλύτερα σφάλματα.

$$\text{Mean Squared Error} MSE = \frac{1}{N}\sum_{j=1}^{N}((y_j) - (y'_j))^2 \tag{7}$$

όπου y'_j είναι η εκτιμώμενη τιμή και y_j η αντίστοιχη πραγματική τιμή για το δείγμα j.

1.7 Σχετική βιβλιογραφία

Η συστημική βιολογία είναι ένα ευρύ διεπιστημονικό πεδίο που στοχεύει στην αποσαφήνιση των μηχανισμών της ζωής μέσα από την ολιστική προσέγγιση των σύνθετων δυναμικών βιολογικών συστημάτων (Vidal 2009). Η μεγάλη ανάπτυξη και πρόοδος των υψηλής απόδοσης αναλύσεων (-omics) κατέστησαν δυνατή την ανίχνευση και την ποσοτικοποίηση εκατοντάδων ή

ακόμη και χιλιάδων βιομορίων με υψηλή ευαισθησία και ακρίβεια, επιτρέποντάς μας να ανιχνεύσουμε ακόμη και μικρές αλλαγές στη φυσιολογία ενός βιολογικού συστήματος που προηγουμένως θεωρούνταν ασήμαντη. Η εφαρμογή της Συστημικής Βιολογίας στους τομείς της υγείας και της βιομηχανίας μπορεί να εμβαθύνει τις γνώσεις και την κατανόηση μας στην ανάπτυξη των ασθενειών καθώς και να προσφέρει νέα δυναμική στον σχεδιασμό νέων θεραπειών και φαρμάκων (Gov, Kori, and Arga 2017; Higdon et al. 2015; Tang et al. 2012; Zhang et al. 2016; Zhang, Burdette, and Wang 2014).

Στο πλαίσιο της συστημικής βιολογίας, ένας κοινός φαινότυπος ασθένειας αντικατοπτρίζει τη διαταραχή σύνθετων ενδοκυττάριων δικτύων που επηρεάζονται από πολλούς περιβαλλοντικούς παράγοντες (Barabási, Gulbahce, and Loscalzo 2011; Loscalzo and Barabasi 2011). Τα βιολογικά δίκτυα διαδραματίζουν βασικό ρόλο στη μοντελοποίηση πολλών σύνθετων ασθενειών, δημιουργώντας γραφήματα που αποτελούνται από κόμβους και ακμές, αντιπροσωπεύοντας τα επιμέρους στοιχεία του συστήματος και τις σχέσεις τους αντίστοιχα (Gov, Kori, and Arga 2017). Η ταυτότητα αυτών των στοιχείων ορίζει τον τύπο του δικτύου. Αυτοί οι τύποι μπορεί να περιλαμβάνουν, μεταξύ άλλων, δίκτυα πρωτεϊνικών αλληλεπιδράσεων, μεταβολικά και γονιδιακής συν-έκφρασης δίκτυα (Erdoğan, Kurt, and Diri 2017).

Ένα μοναδικό omics επίπεδο μπορεί να παρέχει περιορισμένη κατανόηση και γνώση για την αιτιολογία μιας νόσου (Zhang et al. 2017). Πρόσφατες μελέτες, αποδεικνύουν πως η ενσωμάτωση των διαφορετικών omics επιπέδων μπορεί να παρέχει πολλές και χρήσιμες πληροφορίες για αναδυόμενες ιδιότητες του συστήματος, όπως είναι οι μηχανισμοί γονιδιακής ρύθμισης και αποκρίσεις μονοπατιών σηματοδότησης σε μια ασθένεια (Geschwind and Konopka 2009; Gov, Kori, and Arga 2017; Higdon et al. 2015; Hu et al. 2017; Sun and Hu 2016). Ωστόσο,η μεθοδολογία για τον αποτελεσματικό συνδυασμό διαφορετικών μοριακών επιπέδων είναι ακόμη ανεπαρκής. Επιπλέον, οι προκλήσεις που προκύπτουν στην ανάλυση και την αναπαράσταση των δεδομένων, λαμβάνοντας υπόψη τον εκθετικά αυξανόμενο όγκο δεδομένων είναι πολλές. Μία από αυτές τις προκλήσεις περιλαμβάνουν τον τρόπο εντοπισμού και κατανόησης των περίπλοκων σχέσεων που αναδύονται από την ενσωμάτωση διαφορετικών δεδομένων υψηλής απόδοσης (Hu et al. 2017). Για το σκοπό αυτό, την τελευταία δεκαετία έχει δημιουργηθεί ένας μεγάλος αριθμός "κατευθυνόμενων από τα δεδομένα" εργαλείων (data-driven tools) που στόχος τους είναι η ενσωμάτωση και η ανάλυση multi-omics δεδομένων (Gundersen et al. 2017; Ramos et al. 2017; Silva et al. 2016; Yue et al. 2018).

Η συλλογή δεδομένων omics και λειτουργικών συσχετισμών από ένα μεγάλο αριθμό ασθενών είναι ένα πολύ σημαντικό βήμα για την κατασκευή

αξιόπιστων μαθηματικών μοντέλων. Υπάρχει ένας μεγάλος αυξανόμενος αριθμός αποθετηρίων δεδομένων που καλύπτουν ένα μεγάλο φάσμα ασθενειών και περιέχουν πολύτιμες πληροφορίες περιλαμβάνοντας κλινικά, λειτουργικά και omics δεδομένα που λαμβάνονται τόσο από ασθενείς όσο και από τις ομάδες ελέγχου , μεταξύ άλλων τέτοια δεδομένα αφορούν την Ποικιλομορφία αριθμού αντιγράφων (Copy number variation, CNV), την DNA μεθυλίωση, τους Σημειακούς πολυμορφισμούς (Single nucleotide polymorphisms, SNP), την Μεταγραφομική (Transcriptomics), την Πρωτεομική (Proteomics) και την Μεταβολομική (Metabolomics) (Yue et al. 2018). Έτσι, ένας πειραματικός σχεδιασμός ο οποίος συνδυάζει πολλές διαφορετικές πηγές πληροφορίας και επομένως διασφαλίζει την πρόσβαση σε ένα μεγάλο αριθμό ασθενών και δεδομένων ενισχύει την εγκυρότητα και την ανάπτυξη της προσέγγισης.

Σε δικτυακές μελέτες συνήθως ακολουθούνται κάποια βασικά βήματα τα οποία αφορούν αρχικά την κατάλληλη συλλογή και ανάλυση δεδομένων. Πρώτο και σημαντικό βήμα είναι η κατασκευή του δικτύου, η οποία πραγματοποιείται κύρια μέσω της αποσαφήνισης των συσχετιστικών σχέσεων των συστατικών του δικτύου (Morgun, Dong, and Yambartsev 2014). Αυτές οι συσχετιστικές σχέσεις μπορούν να ληφθούν μεταξύ άλλων μέσω συσχετιστικών μετρικών,όπως είναι ο συντελεστής Pearson (Pearson Correlation Coefficient), η αμοιβαία πληροφορία (Mutual Information) και η άμεση πληροφορία (Direct Information) (Erdoğan, Kurt, and Diri 2017; Jiang, Martinez-Ledesma, and Morcos 2017). Στη συνέχεια, μία συνήθης πρακτική των δικτυακών μελετών είναι ο προσδιορισμός των λειτουργικών υπο-ενοτήτων (functional modules) και των μοριακών υποτύπων της ασθένειας (Tornow and Mewes 2003). Μία προσέγγιση σε αυτό το βήμα είναι η συσταδοποίηση τόσο του δικτύου όσο και των δεδομένων (network clustering approaches) που μπορεί να πραγματοποιηθεί μέσω πολλαπλών αλγορίθμων (k-means, Ιεραρχική συσταδοποίηση), παρέχοντας έτσι πολύτιμη εικόνα για τη διάκριση μεταξύ επιμέρους υποομάδων στο δίκτυο και μοριακών υποτύπων (Morgun, Dong, and Yambartsev 2014; Zhang et al. 2016). Τέλος, αυτές οι ομάδες συνήθως μπορούν να βαθμολογηθούν προκειμένου να εκτιμηθεί η έκταση στην οποία επηρεάζονται σε μοριακό επίπεδο από τη συγκεκριμένη ασθένεια.

Επί προσθέτως, η τοπολογία του βιολογικού δικτύου μπορεί να παρέχει ζωτικές πληροφορίες για το βιολογικό σύστημα (Lei and Ruan 2013). Για παράδειγμα, η προδιαγραφή των δημοφιλών κόμβων (hub nodes) υποδεικνύει εκείνα τα ιδιαίτερα συστατικά του δικτύου που έχουν τον μεγαλύτερο αριθμό σχέσεων στο δίκτυο και ως εκ τούτου διαδραματίζουν σημαντικό ρόλο σε αυτή τη συγκεκριμένη βιολογική κατάσταση. Επιπλέον, έχει αποδειχθεί ότι τα γειτονικά συστατικά του δικτύου τείνουν να συμμετέχουν στην ίδια βιο-

λογική διαδικασία ή ακόμα και στην ίδια ασθένεια (Barabási, Gulbahce, and Loscalzo 2011; Lei and Ruan 2013), επομένως, η διάκριση των λειτουργικών υπο-ενοτήτων των δικτύου (functional modules) δίνει το πλεονέκτημα της αναγνώρισης υποομάδων που αποτελούνται από λειτουργικά συναφή βιομόρια (Somvanshi and Venkatesh 2014; Tornow and Mewes 2003).

Οι προσεγγίσεις μηχανικής μάθησης σε συνδυασμό με την ανάλυση δικτύων αποτελεί δημοφιλής στρατηγική μεταξύ των μελετών που κινούνται στην σφαίρα της εξατομικευμένης ιατρικής (Ding, Zu, and Gu 2016; Pineda et al. 2015). Συγκεκριμένα, τέτοιες προσεγγίσεις επιτρέπουν την εξαγωγή σημαντικών χαρακτηριστικών που μπορούν να προσφέρουν ένα βιολογικό αποτύπωμα του συστήματος ενδιαφέροντος (Zhang, Burdette, and Wang 2014). Είναι κοινή στρατηγική η χρήση τόσο των δεδομένων που έχουν επαληθευτεί από τη βιβλιογραφία όσο και των πειραματικών δεδομένων να χρησιμοποιούνται ως δεδομένα εισόδου, προκειμένου να επιλεχθούν καλύτερα χαρακτηριστικά που σχετίζονται με τον φαινότυπο (Peng, Li, and Wang 2017). Αυτά τα χαρακτηριστικά μπορούν να χρησιμοποιηθούν για να διακρίνουν τις υποομάδες των ασθενών και να προβλέψουν αποτελεσματικά γονίδια που σχετίζονται με την ασθένεια (Zhang et al. 2016). Αυτές οι πληροφορίες μπορούν να διερευνηθούν περαιτέρω προκειμένου να ανακαλυφθούν πιθανοί βιοδείκτες που μπορούν να οδηγήσουν σε νέες στρατηγικές για έγκαιρη ανίχνευση, πρόληψη και θεραπεία ασθενειών (Ding, Zu, and Gu 2016; Gov, Kori, and Arga 2017; Zhang et al. 2017).

Αντικείμενο μελέτης στην παρούσα εργασία, αποτέλεσε το μυοδιηθητικό ουροθηλιακό καρκίνωμα της ουροδόχου κύστης, ο οποίος χαρακτηρίζεται από μεγάλη ετερογένεια και μικρή προγνωστικότητα. Πολυάριθμες μελέτες έχουν δημοσιευθεί τα τελευταία χρόνια, χρησιμοποιώντας μοναδικά επίπεδα -omics αλλά και ενσωματώνοντας multi- omics δεδομένα από ασθενείς με καρκίνο τους ουροδόχου κύστης για να προβλέψουν την βιωσιμότητα ασθενών, να αποσαφηνίσουν την ανάπτυξη της ασθένειας, την αλληλεπίδραση διαφορετικών βιομορίων από διαφορετικά μοριακά επίπεδα σε μεγάλη κλίμακα, αλλά και για την βελτιστοποίηση μεθόδων ενσωμάτωσης multi- omic δεδομένων (Pineda et al. 2015; Platts et al. 2011; Zhu et al. 2011) .Σε αυτό το πλαίσιο, ο τομέας της εξατομικευμένης ιατρικής έχει ανάγκη από την ανάπτυξη αλγορίθμων με βελτιωμένη προγνωστική ικανότητα. Μία πρώτη συνεισφορά της παρούσας μελέτης σε αυτήν την ευρύτερη προσπάθεια, είναι η μελέτη και η σύγκριση διαφορετικών αναπαραστάσεων multi- omic δεδομένων και ο τρόπος που χαρακτηριστικά υψηλότερου επιπέδου όπως είναι η τοπολογία του δικτύου, μπορούν να βελτιώσουν την προβλεπτική ικανότητα ενός μοντέλου, που στην δική μας γνώση στο χώρο του καρκίνου της ουροδόχου κύστης

μελετάται πρώτη φορά.

Μία κοίνη στρατηγική στις- omics μελέτες είναι ο περιορισμός του πειραματικού σχεδιασμού στην μοντελοποίηση του φαινομένου, ενσωματώνοντας πληροφορία από το σύνολο των δειγμάτων/ ασθενών. Σε αυτό το πλαίσιο, μία επιπλέον συνεισφορά της παρούσας εργασίας είναι η επέκταση της διαδικασίας ανάλυσης από την μοντελοποίηση του φαινομένου του ουροθηλιακού καρκινώματος της ουροδόχου κύστης στην εξατομικευμένη αναπαράσταση ενός δείγματος ασθενούς. Για αυτό το σκοπό, πραγματοποίηθηκε συλλογή multi- omics και κλινικών δεδομένων από δείγματα ανθρώπων που πάσχουν από καρκίνο της ουροδόχου κύστης. Στη συνέχεια, πραγματοποιήθηκε αναπαράσταση των δεδομένων σε διανυσματική μορφή και αξιολόγηση σε συνδυασμό με μηχανική μάθηση . Σε επόμενο βήμα, κατασκευάστηκε η δεύτερη αναπαράσταση δεδομένων σε μορφή multi- omic δικτύου και αξιοποιήθηκαν τοπολογικά χαρακτηριστικά του δικτύου για τη βελτίωση της προβλεπτικής ικανότητας.

2 Στόχος

Απώτερος σκοπός αυτής της εργασίας ήταν η ανάδειξη της βελτιωμένης απόδοσης που μπορεί να προσφέρουν σε ένα μοντέλο πρόγνωσης αναπαραστάσεις υψηλότερης πολυπλοκότητας όπως είναι τα τοπολογικά χαρακτηριστικά ενός διασυνδεδεμένου δικτύου το οποίο συνδυάζει -omics δεδομένα διαφορετικών κυτταρικών επιπέδων σε σύγκριση με την αναπαράσταση ενός διανύσματος.

Ειδικότερα, η παρούσα εργασία εστιάζει στην μελέτη του ουροθηλιακού καρκινώματος της ουροδόχου κύστης, ο οποίος χαρακτηρίζεται από μεγάλη ετερογένεια και μικρή προγνωστικότητα. Για αυτό το σκοπό πραγματοποιήθηκε εξόρυξη, μελέτη και ανάλυση δειγμάτων που προέρχονται από ασθενείς που πάσχουν από μυοδιηθητικό ουροθηλιακό καρκίνωμα της ουροδόχου κύστης. Τα δεδομένα τα οποία συλλέχθηκαν αφορούσαν:

- επίπεδα γονιδιακής έκφρασης
- επίπεδα έκφρασης μορίων miRNA
- εκτίμησης επιπέδων μεθυλίωσης του γονιδιώματος
- κλινικά δεδομένα των ασθενών.

3 Μέθοδοι και εργαλεία

3.1 Εξόρυξη δεδομένων

Σημαντικό βήμα στη παρούσα διπλωματική εργασία αποτέλεσε η αποτελεσματική εξόρυξη -ομικών και κλινικών δεδομένων από την τράπεζα δεδομένων Genomic Data Commons. Η βάση δεδομένων Genomic Data Commons πραγματοποιεί υποβολή, επεξεργασία και διανομή γονιδιωματικών, κλινικών και ιστολογικών δεδομένων από ερευνητικά προγράμματα καρκίνου. Περισσότερες πληροφορίες για τα δεδομένα της GDC είναι διαθέσιμες στην επίσημη ιστοσελίδα της βάσης (https://gdc.cancer.gov/about-data).

Τα δεδομένα τα οποία χρησιμοποιήθηκαν στην παρούσα μελέτη είναι αποτέλεσμα επιστημονικών μελετών για τον χαρακτηρισμό του μυοδιηθητικού ουροθηλιακού καρκινώματος της ουροδόχου κύστης. Οι επιστημονικές αυτές μελέτες αποτέλεσαν κομμάτι της ευρύτερης προσπάθειας χαρακτηρισμού τύπων καρκίνου του προγράμματος The Cancer Genome Atlas (TCGA) (Robertson et al. 2017; Zhang and Wang 2015).Επιπλέον, οι τύποι δεδομένων που επιλέχθηκε να μελετηθούν, αφορούν επίπεδα έκφρασης miRNA, mRNA, επίπεδα μεθυλίωσης του DNA, καθώς και κλινικά δεδομένα των ασθενών.

3.1.1 Πρόσβαση στα δεδομένα της πύλης GDC Data Portal

Πρώτο βήμα στην εξόρυξη των δεδομένων ήταν η περιήγηση και η κατανόηση της δομής της προς μελέτη πύλης δεδομένων GDC. Η περιήγηση και η πρόσβαση στα γονιδιωματικά και κλινικά δεδομένα της βάσης δεδομένων GDC πραγματοποιείται μέσω της GDC Data Portal. Η μετάβαση στην πύλη δεδομένων πραγματοποιείται μέσω του Tab "Launch Data Portal" στην αρχική σελίδα της βάσης δεδομένων GDC.

Η χρήση της διεπαφής Repository είναι ο κυρίαρχος τρόπος προκειμένου ο χρήστης να αποκτήσει πρόσβαση στα δεδομένα της πύλης GDC Data Portal. Όπως αναφέρθηκε σε παραπάνω υποενότητα, η σελίδα Repository προσφέρει μία σύνοψη όλων των αρχείων και όλων των ασθενών τα οποία είναι διαθέσιμα δίνοντας την δυνατότητα εφαρμογής πολλαπλών φίλτρων. Οι χρήστες μπορούν να έχουν πρόσβαση στην σελίδα Repository είτε μέσω της αρχικής σελίδας της GDC Data Portal είτε κατευθείαν στην ιστοσελίδα https://portal.gdc.cancer.gov/repository.

Στη σελίδα Repository μπορούμε να εισάγουμε πολλαπλά φίλτρα σχετικά τόσο με τους ασθενείς όσο και με το είδος των προς λήψη δεδομένων/αρχείων. Συγκεκριμένα, οι ιδιότητες των αρχείων στις οποίες μπορούν να εφαρμοστούν τα επιθυμητά φίλτρα είναι οι παρακάτω:

- File: Εισαγωγή ειδικού κωδικού (UUID) ή όνομα αρχείου

- Data Category: Ορισμός κατηγορίας δεδομένων υψηλού επιπέδου όπως ”Transcriptome Profiling”
- Data Type: Ορισμός είδους επεξεργασίας δεδομένων/τύπου δεδομένων χαμηλότερου επιπέδου σε σχέση με το Data Category, όπως “Aligned Reads” ή “Gene Expression Quantification”
- Experimental Strategy: Πειραματική στρατηγική που ακολουθήθηκε για τον μοριακό χαρακτηρισμό του καρκίνου
- Workflow Type: Ορισμός του τύπου της βιοπληροφορικής εργασίας ροής που ακολουθήθηκε για την παραγωγή και επεξεργασία των δεδομένων
- Data Format: Ο τύπος του αρχείου
- Platform: Η τεχνολογική πλατφόρμα που χρησιμοποιήθηκε για την παραγωγή των δεδομένων
- Access Level: Κατηγοριοποίηση των αρχείων σε ευρέως διαθέσιμα (Open access) και σε ελεγχόμενης πρόσβασης (Controlled) προς τους χρήστες. Τα αρχεία Open access είναι ευρέως διαθέσιμα προς όλους τους επισκέπτες της βάσης χωρίς να απαιτείται ειδική σύνδεση(log in) του χρήστη στην βάση δεδομένων GDC. Αντίθετα, αν ο χρήστης επιλέξει να πραγματοποιήσει λήψη ενός ελεγχόμενου (Controlled) αρχείου τότε η βάση προειδοποιεί τον χρήστη πως δεν μπορεί να πραγματοποιηθεί η λήψη του αρχείου χωρίς ο χρήστης να έχει συνδεθεί με ατομικό λογαριασμό στην βάση δεδομένων. Αξίζει να σημειωθεί πως όλα τα αρχεία τα οποία επιλέχθηκαν προς λήψη ήταν ανοιχτά διαθέσιμα προς τους χρήστες (”Open Access”).

Επιπλέον, προσφέρεται και ένα δεύτερο επίπεδο εισαγωγής φίλτρων που αφορά τους ασθενείς το οποίο περιλαμβάνει τα παρακάτω πεδία:

- Case: Εισαγωγή ειδικού κωδικού ασθενούς (UUID)
- Primary Site: Ιστός του προς μελέτη πρωτογενούς όγκου
- Cancer Program: Επιλογή του ευρύτερου ερευνητικού προγράμματος το οποίο μπορεί να αποτελείται από διαφορετικά εστιασμένες καρκινικές μελέτες
- Project: Μία ερευνητική προσπάθεια η οποία εστιάζει σε ένα τύπο καρκίνου

- Disease Type: Τύπος του προς μελέτη καρκίνου
- Gender: Ορισμός φύλου ασθενούς
- Age at Diagnosis: Ηλικία στην οποία διαγνώσθηκε με καρκίνο ο ασθενής
- Vital Status: Κατηγοριοποίηση των ασθενών ως προς την βιωσιμότητα τους από την τελευταία φορά επαφής μαζί τους
- Days to Death: Αριθμός των ημερών που έχει ζήσει ο ασθενής από την ημερομηνία διάγνωσής του
- Race: Φυλή του ασθενούς
- Ethnicity: Εθνικότητα του ασθενούς.

Όσον αφορά το επίπεδο των ασθενών, τα φίλτρα τα οποία τροποποιήθηκαν για ασθενείς που πάσχουν από ουροθηλιακό καρκίνωμα της ουροδόχου κύστης και για δείγματα μη καρκινικών ιστών από ασθενείς που πάσχουν από ουροθηλιακό καρκίνωμα της ουροδόχου κύστης, τα οποία θεωρήθηκαν ως δείγματα ελέγχου (Control) εμφανίζονται στον Πίνακα 1. Ειδικότερα, όπως φαίνεται και στον Πίνακα 1, για την ομάδα ασθενών επιλέχθηκε τα δείγματα να προέρχονται αυστηρά από ιστό πρωτογενούς όγκου ενώ αντίστοιχα για την ομάδα ελέγχου από φυσιολογικό ιστό του ιδίου οργάνου. Τα φίλτρα τα οποία εφαρμόστηκαν στο επίπεδο των αρχείων αναλύονται ξεχωριστά σε παρακάτω κεφάλαια.

Φίλτρα	Ασθενείς	Ομάδα ελέγχου (Control)
Sample Type	Primary Tumor	Solid Tissue Normal
Primary Site	Bladder	Bladder
Program	TCGA	TCGA
Project	TCGA- BLCA	TCGA- BLCA
Disease Type	Bladder Urothelial Carcinoma	Bladder Urothelial Carcinoma

Πίνακας 1: Τα φίλτρα τα οποία εφαρμόστηκαν για την επιλογή των ασθενών, όπως ακριβώς εμφανίζονται στην πύλη δεδομένων της GDC. Όλα τα δείγματα λήφθηκαν από ασθενείς με ουροθηλιακό καρκίνωμα της ουροδόχου κύστης στα πλαίσια του προγράμματος The Cancer Genome Atlas, TCGA. Η πρώτη και η δεύτερη στήλη αφορούν δείγματα από ασθενείς με ουροθηλιακό καρκίνωμα της ουροδόχου κύστης (Bladder Urothelial Carcinoma). Ωστόσο, η πρώτη στήλη αφορά δείγματα που προέρχονται αυστηρά από ιστό πρωτογενούς όγκου (Primary Tumor) της ουροδόχου κύστης (Bladder) ενώ αντίστοιχα για την ομάδα ελέγχου τα δείγματα προέρχονται από φυσιολογικό ιστό του ιδίου οργάνου (Solid Tissue Normal, Bladder).

3.1.2 Δεδομένα γονιδιακής έκφρασης

Το γονιδίωμα ενός οργανισμού ορίζεται ως το σύνολο DNA που διαθέτει, περιλαμβάνοντας όλη την πληροφορία για την φυσιολογία, ανάπτυξη, εξέλιξη και απόκριση του οργανισμού. Το 2003 ολοκληρώθηκε η παγκόσμια ερευνητική προσπάθεια του Human Genome Project για την ολιστική ανάλυση του ανθρώπινου γονιδιώματος, παρέχοντας την πρώτη επίσημη καταγραφή των αποτελεσμάτων της αλληλούχισης (Lander et al. 2001). Το ανθρώπινο γονιδίωμα αποτελείται παραπάνω από 3 δισεκατομμύρια ζεύγη βάσεων, 23 ζεύγη χρωμοσωμάτων και συνολικά από περίπου 22,000 γονίδια. Σε κάθε κύτταρο ενός οργανισμού δεν είναι ενεργοποιημένα ανά πάσα στιγμή όλα τα γονίδια που διαθέτει. Αντίθετα κάθε γονίδιο εκφράζεται με διαφορετική ένταση στον χρόνο εξαρτώμενο από τις λειτουργίες που επιτελεί το κύτταρο, την δράση των ρυθμιστικών μηχανισμών και από την απόκρισή του στο περιβάλλον.

Η μεταγραφομική (transcriptomics) συνιστά την ολιστική μελέτη του προτύπου συγκέντρωσης των μορίων mRNA ενός οργανισμού, δηλαδή την έκφραση των γονιδίων του. Η ποσοτικοποίηση της έκφρασης μορίων mRNA βρίσκεται στο επίκεντρο του ερευνητικού ενδιαφέροντος στο χώρο της βιοπληροφορικής και παράλληλα αποτελεί τον πιο διαδεδομένο, άμεσο και συστηματικό τρόπο μελέτης της γονιδιακής έκφρασης. Ενώ παράλληλα η υψηλή απόδοση, ταχύτητα και το χαμηλό κόστος των εφαρμογών νέας γενιάς επιτρέπουν στους ερευνητές να μελετήσουν τα βιολογικά συστήματα σε νέα επίπεδα.

Φίλτρα στο επίπεδο των αρχείων	Δεδομένα γονιδιακής έκφρασης
Data Category	Transciptome Profiling
Data type	Gene Expression Quantification
Experimental strategy	RNA-Seq
Workflow Type	HTSeq - FPKM - UQ
Access	open

Πίνακας 2: Φίλτρα τα οποία εφαρμόστηκαν για την επιλογή αρχείων από δεδομένα γονιδιακής έκφρασης, όπως ακριβώς εμφανίζονται στην πύλη δεδομένων της GDC.

Τα δεδομένα γονιδιακής έκφρασης που αξιοποιήθηκαν σε αυτή τη μελέτη είναι αποτέλεσμα αλληλούχισης υψηλής απόδοσης (ή αλληλούχισης επόμενης γενιάς, next generation sequencing, NGS) (βλ. Πίνακα 2). Η πορεία ανάλυσης ποσοτικοποίησης μορίων mRNA που ακολουθεί η GDC, υπολογίζει γονιδιακά επίπεδα έκφρασης τα οποία είναι διαθέσιμα προς λήψη σε διαφορετικά επίπεδα επεξεργασίας και κανονικοποίησης. Τα επίπεδα αυτά αφορούν την εκτίμηση των παρακάτω τιμών:

- τα HT-Seq αριθμός αναγνώσεων ανά μετάγραφο (raw read count),
- HT-Seq - Fragments per Kilobase of transcript per Million mapped reads (FPKM), και
- HTSeq - FPKM UQ (upper quartile normalization).

Αυτές οι τιμές παράγονται μέσω μίας προτυποποιημένης πορείας ανάλυσης που εφαρμόζεται σε όλα τα δεδομένα. Τα HTSeq - counts αφορούν τον χαρτογραφημένο αριθμό αναγνώσεων για κάθε γονίδιο ο οποίος υπολογίζεται από το πακέτο της γλώσσας προγραμματισμού Python HTSeq. Τα αρχεία HTSeq-Count είναι διαθέσιμα σε μία tab-delimited μορφή με την πρώτη στήλη να αντιστοιχεί στα Ensembl gene ID και μία δεύτερη στο χαρτογραφημένο αριθμό αναγνώσεων ανά γονίδιο. Αυτά τα αρχεία στη συνέχεια επεξεργάζονται για να παραχθούν οι διορθωμένες τιμές FPKM και FPKM-UQ . Συγκεκριμένα, οι τιμές FPKM δηλαδή τα "θραύσματα ανά 1000 βάσεις εξωνίου ανά εκατομμύριο διαβασμάτων που έχουν χαρτογραφηθεί" [Fragments Per Kilobase of exon per Million fragments mapped (FPKM)] είναι κανονικοποιημένες τιμές που αφορούν την εκτίμηση του βαθμού έκφρασης γονιδίων. Η κανονικοποίηση των τιμών FPKM είναι διπλή, αρχικά ως προς το μήκος του (ανά χιλιάδα βάσεων) κι έπειτα ως προς το σύνολο των παραχθεισών αλληλουχιών (ανά εκατομμύριο αλληλουχιών). Με αυτόν τον τρόπο, η τιμή FPKM είναι ανεξάρτητη από το μήκος του γονιδίου όσο και με του "βάθους" της αλληλούχισης.

Στην παρούσα μελέτη για την εκτίμηση του βαθμού έκφρασης των γονιδίων επιλέξαμε να χρησιμοποιήσουμε τις κανονικοποιημένες τιμές FPKM - UQ (κανονικοποίηση στο άνω τεταρτημόριο) (βλ. Πίνακα 2). Η μορφή τέτοιων αρχείων απεικονίζεται στο Σχήμα 18, όπου η πρώτη στήλη αντιστοιχεί στο Ensembl ID του εκάστοτε γονιδίου (με το σύμβολο "." να ορίζεται η εκάστοτη έκδοση) και η δεύτερη στήλη σε κανονικοποιημένες τιμές FPKM- UQ. Είναι σημαντικό να υπογραμμιστεί πως δεν επιλέχθηκε να μελετηθούν οι ισομορφές των γονιδίων επομένως δεν περιλαμβάνονται στην παρούσα μελέτη φαινόμενα συναρμογής (alternative splicing).

Για λόγους καλύτερης αναπαράστασης και κατανόησης των δεδομένων, επιλέξαμε να αναπαραστήσουμε με ένα κοινό τρόπο τα σύμβολα των γονιδίων στο επίπεδο της γονιδιακής έκφρασης και στο επίπεδο της μεθυλίωσης του DNA. Ειδικότερα, επιλέξαμε ως κοινή αναφορά των γονιδίων το HUGO/HGNC Gene Symbol. Δεδομένου πως ο γονιδιακός σχολιασμός των αρχείων γονιδιακής έκφρασης έγινε με το Ensembl gene ID, μία πρώτη πρόκληση που αντιμετωπίστηκε στα δεδομένα γονιδιακής έκφρασης ήταν η μετατροπή των Ensembl gene IDs σε Gene Symbols με βάση την σωστή εκδοχή του ανθρώ-

```
ENSG00000242268.2   0.0
ENSG00000270112.3   0.0
ENSG00000167578.15  110195.448157
ENSG00000273842.1   0.0
ENSG00000078237.5   82239.1281105
ENSG00000146083.10  304200.857881
ENSG00000225275.4   0.0
ENSG00000158486.12  1033.91180278
ENSG00000198242.12  5521714.81747
ENSG00000259883.1   1467.90553
ENSG00000231981.3   0.0
ENSG00000269475.2   0.0
ENSG00000201788.1   0.0
ENSG00000134108.11  985327.482134
ENSG00000263089.1   264.367008557
ENSG00000172137.17  1993.69670634
ENSG00000167700.7   788648.551583
ENSG00000234943.2   0.0
ENSG00000240423.1   1790.75184119
ENSG00000060642.9   167989.678339
```

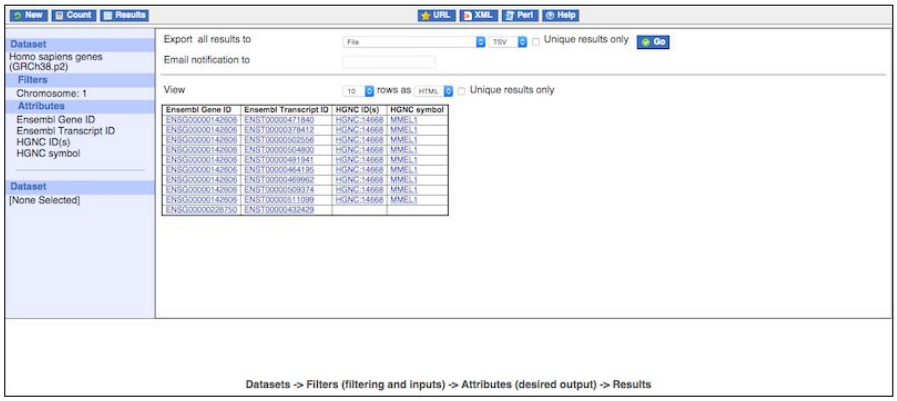
Σχήμα 18: Παράδειγμα δομής αρχείου εκτίμησης της γονιδιακής έκφρασης της τράπεζας δεδομένων GDC. Η πρώτη στήλη αντιστοιχεί στο μοναδικό Ensembl ID του εκάστοτε γονιδίου και η δεύτερη στήλη σε FPKM - UQ (αριθμός Fragments per Kilobase of transcript per Million mapped reads (FPKM), κανονικοποιημένος στο άνω τεταρτημόριο)

.

πινου γονιδιώματος αναφοράς που χρησιμοποιήθηκε για την χαρτογράφηση των δεδομένων, που ήταν το GRCh38. Για αυτό το σκοπό, χρησιμοποιήθηκε το διαδικτυακό εργαλείο BioMart το οποίο στεγάζεται στην ιστοσελίδα της Ensembl (http://www.ensembl.org/biomart/martview). Το BioMart είναι ένα εύχρηστο εργαλείο στο οποίο μέσω της εφαρμογής πολλαπλών φίλτρων ο χρήστης μπορεί να δημιουργήσει σύνολα δεδομένων που αντιστοιχίζουν βιολογική πληροφορία πολλαπλών επιπέδων προερχόμενη από διαφορετικές βάσεις δεδομένων (πχ. Ensembl Gene ID - HGNC Gene Symbol - Ensembl Transcript ID κ.ο.κ.) (βλ. Σχήμα 19). Επιπλέον, το BioMart δίνει την δυνατότητα απευθείας λήψης των αποτελεσμάτων στην μορφή που επιθυμεί ο χρήστης (TSV, CSV, HTML και XLS) αλλά και την αποστολή τους στην ηλεκτρονική διεύθυνση του χρήστη.

3.1.3 Δεδομένα εκτίμησης επιπέδων μεθυλίωσης του DNA

Η μεθυλίωση του DNA (DNA methylation) είναι μία επιγενετική διαδικασία η οποία συναντάται συχνά στα ευκαρυωτικά κύτταρα και έχει συσχετιστεί με μεταγραφική ανενεργότητα όταν εντοπίζεται σε περιοχές υποκινητών. Ειδικότερα, η μεθυλίωση του DNA αναφέρεται στην ομοιοπολική τροποποίηση των βάσεων κυτοσίνης στη θέση C-5, σε νησίδες CpG του γονιδιώματος, και



Σχήμα 19: Το BioMart είναι ένα εύχρηστο εργαλείο που επιτρέπει την εξόρυξη βιολογικής πληροφορίας (http://www.ensembl.org/biomart/martview). Στην αριστερή στήλη ο χρήστης μπορεί να εισάγει πολλαπλά φίλτρα (αριστερή στήλη) και να δημιουργήσει σύνολα δεδομένων στα οποία αντιστοιχίζεται βιολογική πληροφορία πολλαπλών επιπέδων (πχ. Ensembl Gene ID - HGNC Gene Symbol - Ensembl Transcript ID). Στα δεξιά εμφανίζονται τα τελικά σύνολα δεδομένων, όπως διαμορφώθηκαν από την εφαρμογή φίλτρων. Το εργαλείο δίνει την δυνατότητα απευθείας λήψης των αποτελεσμάτων αλλά και την αποστολή τους στην ηλεκτρονική διεύθυνση του χρήστη.

.

καταλύεται in vivo από τις DNA μεθυλοτρανσφεράσες. Τόσο τα δεδομένα έκφρασης, όσο κι αυτά της μεθυλίωσης προέρχονται, όπως αναφέραμε και παραπάνω, από μεθοδολογίες αλληλούχισης νέας γενιάς και συγκεκριμένα από την αναλυτική πλατφόρμα Illumina Human Methylation 450.

Στην παρούσα μελέτη για την εκτίμηση του βαθμού μεθυλίωσης των γονιδίων χρησιμοποιήσαμε το μέγεθος της B-τιμής (Methylation Beta-value) (βλ. Πίνακα 5). Η B-τιμή αναπαριστά τον λόγο έντασης μεθυλίωσης σε μία συστοιχία (array) προς την συνολική ένταση της συστοιχίας (ισούται με το άθροισμα των επιπέδων μεθυλίωσης ενός μεθυλιωμένου και ενός μη μεθυλιωμένου ανιχνευτή) και παίρνει τιμές μεταξύ του 0 και 1, με 0 να αναπαριστά μία μη μεθυλιωμένη περιοχή ενώ 1 μία υπερμεθυλιωμένη περιοχή του γονιδιώματος (Zhou, Laird, and Shen 2017).

Η μορφή των δεδομένων εκτίμησης επιπέδων μεθυλίωσης του DNA απεικονίζεται στο Σχήμα 20 και αφορά αρχεία εξαιρετικά μεγάλου μεγέθους (150MB). Ειδικότερα, τέτοια αρχεία αποτελούνται από 9 στήλες οι οποίες παραθέτονται στον Πίνακα 3. Οι στήλες που χρησιμοποιήθηκαν προς εξόρυξη και επεξεργασία από τον παραπάνω είδος αρχείου ήταν η Beta value και το Gene Symbol. Η στήλη Beta value αναπαριστά τιμές του μεγέθους της B-τιμής (Beta-value) η οποία εκφράζει την εκτίμηση του βαθμού μεθυλίωσης των γονιδίων. Το Gene Symbol είναι το σύμβολο του γονιδίου που σχετίζεται με την CpG νησίδα, συμπεριλαμβάνοντας μόνο γονίδια που βρίσκονται εντός

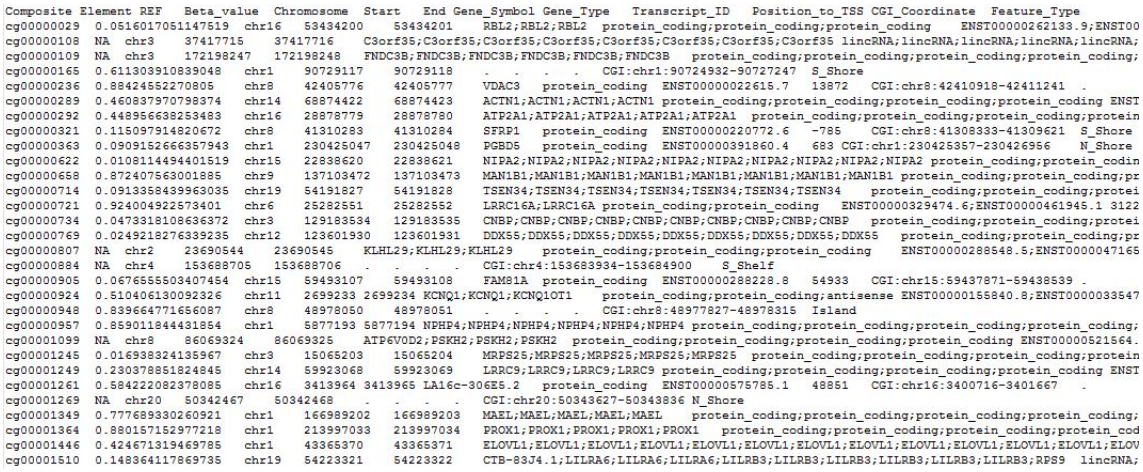
1.500 bp ανοδικά του σημείου έναρξης μεταγραφής (TSS) μέχρι το τέλος του σώματος του γονιδίου.

Τύπος αρχείου	Στήλες αρχείου
Tab-delimited ASCII Text	• Composite Element REF • Beta value • Chromosome • Start • End • Gene Symbol • Gene Type • Transcript ID • Position to TSS • CGI Coordinate • Feature Type

Πίνακας 3: Δεδομένα εκτίμησης επιπέδων μεθυλίωσης του DNA. Η στήλη Composite Element REF αντιστοιχεί σε ένα μοναδικό κωδικό για ένα ανιχνευτή συστοιχίας (array probe) που αντιστοιχεί σε μία CpG νησίδα στο γονιδίωμα. Η στήλη Beta value αναπαριστά τιμές του μεγέθους της Β-τιμής (Beta-value) η οποία εκφράζει την εκτίμηση του βαθμού μεθυλίωσης των γονιδίων. Οι στήλες Chromosome, Start και End, περιέχουν πληροφορία για το χρωμόσωμα που βρίσκεται το σημείο πρόσδεσης του ανιχνευτή, η αρχή της CpG νησίδας στο χρωμόσωμα και το τέλος της CpG νησίδας στο χρωμόσωμα, αντίστοιχα. Το Gene Symbol είναι το HUGO σύμβολο του γονιδίου που σχετίζεται με την CpG νησίδα, συμπεριλαμβάνοντας μόνο γονίδια που βρίσκονται εντός 1.500 bp ανοδικά του σημείου έναρξης μεταγραφής (TSS) μέχρι το τέλος του σώματος του γονιδίου. Η στήλη Gene Type αποτελεί μία γενικότερη κατηγοριοποίηση των γονιδίων (π.χ.ψευδογονίδια, miRNA). Η Transcript ID αντιστοιχεί στο Ensembl transcript IDs κάθε μεταγράφου που σχετίζεται με τα γονίδια της στήλης Gene Symbol. Η Position to TSS είναι μία στήλη που μας πληροφορεί για την απόσταση σε bp από την CpG νησίδα μέχρι το σημείο έναρξης της μεταγραφής του συσχετιζόμενου γονιδίου. Τέλος, η CGI Coordinate αφορά τις συντεταγμένες έναρξης και λήξης της CpG νησίδας.

Φίλτρα στο επίπεδο των αρχείων	Δεδομένα μεθυλίωσης του DNA
Data Category	DNA methylation
Data type	Methylation Beta Value
Experimental strategy	Methylation Array
Workflow Type	Liftover
Platform	Illumina Human Methylation 450
Access	open

Πίνακας 4: Φίλτρα τα οποία εφαρμόστηκαν για την επιλογή αρχείων από δεδομένα μεθυλίωσης του DNA, όπως ακριβώς εμφανίζονται στην πύλη δεδομένων της GDC.



```
Composite Element REF   Beta_value  Chromosome  Start   End Gene_Symbol Gene_Type   Transcript_ID   Position_to_TSS CGI_Coordinate  Feature_Type
cg00000029  0.0516017051147519  chr16   53434200    53434201    RBL2;RBL2;RBL2  protein_coding;protein_coding;protein_coding    ENST00000262133.9;ENST00
cg00000108  NA  chr3    37417715    37417716    C3orf35;C3orf35;C3orf35;C3orf35;C3orf35;C3orf35;C3orf35;C3orf35;C3orf35 lincRNA;lincRNA;lincRNA;lincRNA;lincRNA;
cg00000109  NA  chr3    172198247   172198248   FNDC3B;FNDC3B;FNDC3B;FNDC3B;FNDC3B;FNDC3B   protein_coding;protein_coding;protein_coding;protein_coding;
cg00000165  0.611303910839048   chr1    90729117    90729118    .   .   .   CGI:chr1:90724932-90727247  S_Shore
cg00000236  0.884246552270805   chr8    42405776    42405777    VDAC3   protein_coding  ENST00000022615.7   13872   CGI:chr8:42410918-42411241  .
cg00000289  0.4608379707983874  chr14   68874422    68874423    ACTN1;ACTN1;ACTN1;ACTN1 protein_coding;protein_coding;protein_coding;protein_coding ENST
cg00000292  0.448956638253483   chr16   28878779    28878780    ATP2A1;ATP2A1;ATP2A1;ATP2A1;ATP2A1  protein_coding;protein_coding;protein_coding;protein
cg00000321  0.115097914820672   chr8    41310283    41310284    SFRP1   protein_coding  ENST00000220772.6   -785    CGI:chr8:41308333-41309621  S_Shore
cg00000363  0.0909152666357943  chr1    230425047   230425048   PGBD5   protein_coding  ENST00000391860.4   683 CGI:chr1:230425357-230426956    N_Shore
cg00000622  0.0108114494401519  chr15   22838620    22838621    NIPA2;NIPA2;NIPA2;NIPA2;NIPA2;NIPA2;NIPA2;NIPA2;NIPA2 protein_coding;protein_codin
cg00000658  0.872407563001885   chr9    137103472   137103473   MAN1B1;MAN1B1;MAN1B1;MAN1B1;MAN1B1;MAN1B1;MAN1B1;MAN1B1 protein_coding;protein_coding;pr
cg00000714  0.0913358439963035  chr19   54191827    54191828    TSEN34;TSEN34;TSEN34;TSEN34;TSEN34;TSEN34;TSEN34    protein_coding;protein_coding;protei
cg00000721  0.924004922573401   chr6    25282551    25282552    LRRC16A;LRRC16A protein_coding;protein_coding   ENST00000329474.6;ENST00000461945.1 3122
cg00000734  0.0473318108636372  chr3    129183534   129183535   CNBP;CNBP;CNBP;CNBP;CNBP;CNBP;CNBP;CNBP;CNBP   protein_coding;protein_coding;protei
cg00000769  0.0249218276339235  chr12   123601930   123601931   DDX55;DDX55;DDX55;DDX55;DDX55;DDX55;DDX55;DDX55;DDX55   protein_coding;protein_coding;pr
cg00000807  NA  chr2    23690544    23690545    KLHL29;KLHL29;KLHL29    protein_coding;protein_coding;protein_coding    ENST00000288548.5;ENST0000047165
cg00000884  NA  chr4    153688705   153688706   .   .   .   CGI:chr4:153683934-153684900    S_Shelf
cg00000905  0.0676555503407454  chr15   59493107    59493108    FAM81A  protein_coding  ENST00000288228.8   54933   CGI:chr15:59437871-59438539 .
cg00000924  0.510406130092326   chr11   2699233 2699234 KCNQ1;KCNQ1;KCNQ1OT1   protein_coding;protein_coding;antisense ENST00000155840.8;ENST0000033547
cg00000948  0.839664771656087   chr8    48978050    48978051    .   .   .   CGI:chr8:48977827-48978315  Island
cg00000957  0.859011844431854   chr1    5877193 5877194 NPHP4;NPHP4;NPHP4;NPHP4;NPHP4;NPHP4 protein_coding;protein_coding;protein_coding;protein_coding;
cg00001099  NA  chr8    86069324    86069325    ATP6V0D2;PSKH2;PSKH2;PSKH2  protein_coding;protein_coding;protein_coding;protein_coding ENST00000521564.
cg00001245  0.016938324135967   chr3    15065203    15065204    MRPS25;MRPS25;MRPS25;MRPS25;MRPS25  protein_coding;protein_coding;protein_coding;protein
cg00001249  0.230378851824845   chr14   59923068    59923069    LRRC9;LRRC9;LRRC9;LRRC9 protein_coding;protein_coding;protein_coding;protein_coding ENST
cg00001261  0.584222082378085   chr16   3413964 3413965 LA16c-306E5.2   protein_coding  ENST00000575785.1   48851   CGI:chr16:3400716-3401667   .
cg00001269  NA  chr20   50342467    50342468    .   .   .   CGI:chr20:50343627-50343836 N_Shore
cg00001349  0.777689330260921   chr1    166989202   166989203   MAEL;MAEL;MAEL;MAEL;MAEL    protein_coding;protein_coding;protein_coding;protein_coding;
cg00001364  0.880157152977218   chr1    213997033   213997034   PROX1;PROX1;PROX1;PROX1;PROX1    protein_coding;protein_coding;protein_coding;protein_cod
cg00001446  0.424671319469785   chr1    43365370    43365371    ELOVL1;ELOVL1;ELOVL1;ELOVL1;ELOVL1;ELOVL1;ELOVL1;ELOVL1;ELOVL1;ELOVL1;ELOVL1;ELOVL1;ELOV
cg00001510  0.148364117869735   chr19   54223321    54223322    CTB-83J4.1;LILRA6;LILRA6;LILRA6;LILRB3;LILRB3;LILRB3;LILRB3;LILRB3;LILRB3;RPS9  lincRNA;
```

Σχήμα 20: Παράδειγμα δομής αρχείου εκτίμησης της μεθυλίωσης του DNA.

.

3.1.4 Δεδομένα miRNA

Μία άλλη πτυχή της μεταγραφομικής αφορά την μελέτη του προτύπου συγκέντρωσης μορίων miRNA. Τα microRNA (miRNA) αποτελούν μικρά, συντηρημένα μόρια RNA μήκους 21–23 νουκλεοτιδίων. Κύριος ρόλος των miRNA είναι η ρύθμιση της γονιδιακής έκφρασης στα ευκαρυωτικά κύτταρα μέσω της πρόσδεσης τους στη 3'-αμετάφραστη περιοχή (3'-UTR) μορίων mRNA. Επιπλέον, ένα μόριο mRNA μπορεί να είναι ο στόχος πολλαπλών miRNAs, και κάθε μεμονωμένο miRNA έχει την ικανότητα να στοχεύει εκατοντάδες γονίδια.

Η έξοδος της προτυποποιημένης πορείας ανάλυσης της έκφρασης miRNA που ακολουθεί η GDC περιλαμβάνει δύο ειδών μετρήσεις. Τα δύο είδη εκτίμησης αφορούν ακατέργαστες μετρήσεις μικρο- αναγνώσεων αλληλουχιών (raw reads counts) καθώς και κανονικοποιημένες μετρήσεις σε ανάγνωσεις ανά εκατομμύριο αντιστοιχισμένες αναγνώσεις (reads per million mapped reads, RPM). Το πρώτο περιέχει αθροιστική έκφραση για όλες ευθυγραμμισμένες αναγνώσεις με γνωστά miRNAs της βάσης miRBase (http://www.mirbase.org/). Εάν υπάρχουν πολλαπλές ευθυγραμμίσεις σε διαφορετικά miRNA ή διαφορετικές περιοχές του ίδιου miRNA, η ανάγνωση επισημαίνεται ως "cross-mapped" . Τέλος, στην παρούσα εργασία δεν μελετήθηκαν ισομορφές των μορίων miRNA.

Η μορφή των GDC αρχείων που περιλαμβάνουν δεδομένα εκτίμησης επιπέδων έκφρασης μορίων miRNA απεικονίζεται στο Σχήμα 21. Ειδικότερα, τέτοια αρχεία αποτελούνται από 5 στήλες οι οποίες παραθέτονται στον Πίνακα 5. Οι στήλες που χρησιμοποιήθηκαν προς εξόρυξη και επεξεργασία από το παραπάνω είδος αρχείου ήταν η miRNA ID η οποία περιλαμβάνει το μοναδικό κωδικό κάθε μορίου miRNA όπως ορίζεται από την βάση miRBase και ο αριθμός αναγνώσεων ανά εκατομμύριο διαβασμάτων miRNA που έχουν

χαρτογραφηθεί (Reads per million miRNA mapped).

```
miRNA_ID    read_count  reads_per_million_miRNA_mapped  cross-mapped
hsa-let-7a-1    57613   7139.317499 N
hsa-let-7a-2    56732   7030.145286 N
hsa-let-7a-3    57254   7094.830751 N
hsa-let-7b  63372   7852.964236 N
hsa-let-7c  10788   1336.832957 N
hsa-let-7d  2224    275.594781  N
hsa-let-7e  12856   1593.096450 N
hsa-let-7f-1    24963   3093.377931 N
hsa-let-7f-2    25203   3123.118375 N
hsa-let-7g  3046    377.455802  N
hsa-let-7i  2558    316.983566  N
hsa-mir-1-1 345 42.751888   N
hsa-mir-1-2 371 45.973770   N
hsa-mir-100 38123   4724.145610 N
hsa-mir-101-1   37506   4647.687885 N
hsa-mir-101-2   37475   4643.846411 N
hsa-mir-103a-1  61752   7652.216239 Y
hsa-mir-103a-2  62063   7690.754898 Y
hsa-mir-103b-1  0   0.000000    N
hsa-mir-103b-2  0   0.000000    N
hsa-mir-105-1   1   0.123919    N
```

Σχήμα 21: Παράδειγμα δομής αρχείου εκτίμησης επιπέδων έκφρασης μορίων miRNA.

.

Τύπος αρχείου	Στήλες αρχείου
Tab-delimited ASCII Text	• miRNA ID • read count • reads per million miRNA mapped • cross-mapped

Πίνακας 5: Δεδομένα εκτίμησης επιπέδων έκφρασης μορίων miRNA. Η στήλη miRNA ID αφορά το μοναδικό κωδικό κάθε μορίου miRNA όπως ορίζεται από την βάση miRBase. Η στήλη read count αφορά ακατέργαστες μετρήσεις μικρο- αναγνώσεων αλληλουχιών. Η στήλη reads per million miRNA mapped είναι μία κανονικοποιημένη μορφή των μετρήσεων read count και αντιστοιχεί στον αριθμό αναγνώσεων ανά εκατομμύριο διαβασμάτων miRNA που έχουν χαρτογραφηθεί. Τέλος, στη στήλη cross-mapped ορίζεται εάν υπάρχουν πολλαπλές ευθυγραμμίσεις σε διαφορετικά miRNA ή διαφορετικές περιοχές του ίδιου miRNA.

Φίλτρα	Δεδομένα έκφρασης miRNA
Data Category	Transciptome Profiling
Data type	miRNA Expression Quantification
Experimental strategy	miRNA-Seq
Workflow Type	miRNA Profiling
Access	open

Πίνακας 6: Φίλτρα τα οποία εφαρμόστηκαν για την επιλογή αρχείων από δεδομένα έκφρασης miRNA, όπως ακριβώς εμφανίζονται στην πύλη δεδομένων της GDC.

3.1.5 Κλινικά δεδομένα

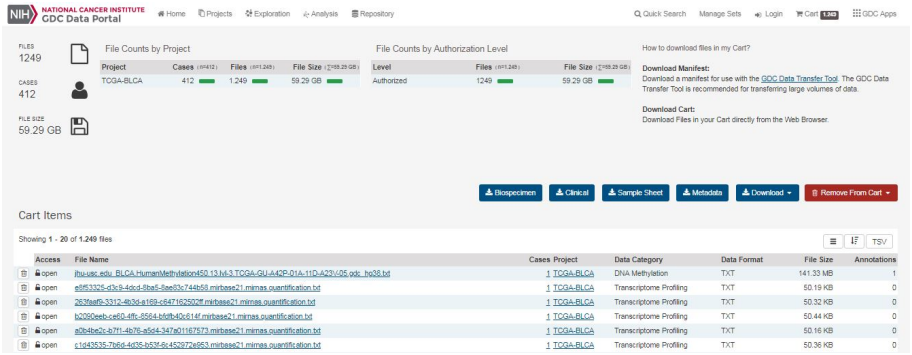
Το αρχείο των κλινικών δεδομένων που παρέχει η GDC περιλαμβάνει δημογραφική και παθολογική πληροφορία τόσο για τους ίδιους τους ασθενείς από του οποίους λήφθηκε το εκάστοτε δείγμα όσο και για τα δείγματα αυτά καθεαυτά. Ειδικότερα, η πληροφορία που περιλαμβάνει ένα κλινικό αρχείο της GDC είναι εξαιρετικά πλούσια. Ενδεικτικά, κάποια από τα πεδία που περιλαμβάνει ένα κλινικό αρχείο είναι ο μοναδικός κωδικός κάθε ασθενούς (Case ID), το φύλο, η ηλικία, η ημερομηνία γέννησης, εθνικότητα, φυλή που ανήκει ο ασθενής, επιπλέον περιλαμβάνονται το ερευνητικό πρόγραμμα καρκίνου, κατηγοριοποίηση του σταδίου του όγκου, αν ο ασθενής πάσχει από κάποια άλλη ασθένεια, αν είναι ή ήταν καπνιστής, κατάσταση βιωσιμότητας (νεκρός/ζωντανός) και ο αριθμός ημερών από την ημερομηνία διάγνωσης του ασθενούς μέχρι την ημερομηνία θανάτου. Είναι σημαντικό να σημειωθεί πως η GDC δίνει την δυνατότητα πρόσθεσης επιπλέον πληροφορίας πεδίων στα κλινικά αρχεία που πραγματοποιείται στην διεπαφή του Καταθετηρίου (Repository Page).

Από το παραπάνω είδος αρχείων, αξιοποιήθηκαν οι στήλες του μοναδικού κωδικού κάθε ασθενούς (Στήλη: Case ID) και η κατηγοριοποίηση του σταδίου του όγκου (Στήλη: Tumor stage). Η στήλη Tumor stage αποτελεί μία γενικότερη κατηγοριοποίηση του σταδίου του καρκίνου με βάση συγκεκριμένων κριτηρίων που αφορούν το μέγεθος του όγκου, αν οι λεμφαδένες του ασθενούς έχουν καρκινικά κύτταρα και αν έχει γίνει μετάσταση του πρωτογενούς όγκου σε άλλα σημεία του σώματος. Επιπλέον, οι αποδεκτές τιμές για το στάδιο του όγκου εξαρτώνται από τον ιστό του πρωτογενούς όγκου, τον τύπο και το αποδεκτό σύστημα σταδιοποίησης.

case_id	gender	year_of_birth	race	ethnicity	year_of_death	classification_of_tumor	last_known_disease_status	
db84[C:\perltest\nikiforos\test\clinical.tsv]f90	TCGA-BT-A42B	TCGA-BLCA	male	1952	white	not hispanic or latino	--	not reported	not reported	
0f4fb205-b13c-4a85-97b3-779429e6cddd	TCGA-BL-A3JM	TCGA-BLCA	male	1948	white	not hispanic or latino	2010	not reported	not repo
22424a05-0546-44c2-8ab7-abb8e49c3363	TCGA-G2-AA3B	TCGA-BLCA	female	1934	white	not hispanic or latino	--	not reported	not reported	
104c8e20-1b29-4d77-9633-359a35be48fb	TCGA-DK-AA6U	TCGA-BLCA	male	1949	white	not hispanic or latino	--	not reported	not reported	
10422fe9-8d70-40a9-9875-3fe9fab68823	TCGA-GV-A40G	TCGA-BLCA	male	1935	white	not hispanic or latino	--	not reported	not reported	
1029514b-a32c-4c43-a440-a5ba2709f717	TCGA-ZF-AA51	TCGA-BLCA	female	1940	white	not hispanic or latino	--	not reported	not reported	
f5c187ef-dc08-4466-932a-43aa33285673	TCGA-GV-A6ZA	TCGA-BLCA	male	1958	white	not hispanic or latino	--	not reported	not reported	
045fe498-6a07-41ff-9956-dc0fb0483e7f	TCGA-E7-A7PW	TCGA-BLCA	male	1950	asian	not hispanic or latino	--	not reported	not reported	
286cf147-b7f7-4a05-8e41-7fbd3717ad71	TCGA-E7-A7XN	TCGA-BLCA	male	1947	asian	not hispanic or latino	--	not reported	not reported	
24d56331-4bca-4ab8-9fb7-a95a30b4c8fe	TCGA-ZF-AA5P	TCGA-BLCA	male	1948	white	not hispanic or latino	--	not reported	not reported	
41e8c196-ef93-47c6-ad06-0031f23d4347	TCGA-FD-A43P	TCGA-BLCA	male	1938	black or african american	not hispanic or latino	--	not repo
cd50c69f-0f45-4c57-b61a-8c6b41ab4f7e	TCGA-G2-A3VY	TCGA-BLCA	male	1946	asian	not hispanic or latino	--	not reported	not reported	
0198d8e5-a14b-44ee-b1c4-d248994845ad	TCGA-DK-AA74	TCGA-BLCA	male	1935	white	not hispanic or latino	--	not reported	not reported	
8da76432-3004-47e5-a474-bd94bd4c0b33	TCGA-FD-A5BY	TCGA-BLCA	female	1949	white	not hispanic or latino	--	not reported	not reported	
d46f7976-6061-4ca9-8294-6c215703662b	TCGA-UY-A78L	TCGA-BLCA	male	--	white	not hispanic or latino	--	not reported	not reported	C
ada95115-275f-4229-aa00-f298a5ff931c	TCGA-XF-AAMT	TCGA-BLCA	female	1931	white	not hispanic or latino	2006	not reported	not repo
f027d254-ec01-4882-a8a7-964ae90b7d49	TCGA-DK-A1AB	TCGA-BLCA	female	1936	white	not hispanic or latino	--	not reported	not reported	
841b4582-a268-4a55-a9a2-47c7e5c3b69f	TCGA-XF-A8HE	TCGA-BLCA	male	1957	white	not hispanic or latino	--	not reported	not reported	
036b9681-ab58-4734-a8fd-e0d5d078bf5a	TCGA-DK-A2I1	TCGA-BLCA	female	1937	white	not hispanic or latino	--	not reported	not reported	
63443198-4fcf-40bd-befb-4a6920d29a4	TCGA-ZF-AA4W	TCGA-BLCA	male	1953	white	not hispanic or latino	--	not reported	not reported	
faf4c71c-e4f3-4513-8c44-9114a5d3729a	TCGA-E7-A41J	TCGA-BLCA	male	1956	asian	not hispanic or latino	--	not reported	not reported	
7d198ee2-d80a-4759-804f-a7ef85971843	TCGA-SY-A9G5	TCGA-BLCA	male	1945	white	not hispanic or latino	--	not reported	not reported	
9743e684-25dd-4dd9-9b84-6ae9518d1818	TCGA-5N-A9KM	TCGA-BLCA	female	1939	white	not hispanic or latino	2013	not reported	not repo
aaa55de1-8f53-4b56-ba0a-67cff909cf81	TCGA-ZF-A9R5	TCGA-BLCA	male	1952	white	not hispanic or latino	--	not reported	not reported	
e6f7bf6c-92b1-4795-94fc-b189d0c7877f	TCGA-ZF-AA4X	TCGA-BLCA	male	1953	white	not hispanic or latino	--	not reported	not reported	
c5b20ccc-523d-491c-8536-008e05c57212	TCGA-DK-A1A3	TCGA-BLCA	male	1950	white	not hispanic or latino	--	not reported	not reported	
28e25ffb-8b0c-4839-a735-ed7b78e3213a	TCGA-G2-A2EC	TCGA-BLCA	female	1952	black or african american	not hispanic or latino	--	not repo
ebedab7c-dd7c-448e-8324-78c86f13be8b	TCGA-BT-A20N	TCGA-BLCA	male	1931	not reported	not reported	2005	not reported	not repo
bb38805e-2ccc-4169-ba01-34b98adf5bd0	TCGA-FJ-A871	TCGA-BLCA	male	1964	white	not hispanic or latino	2013	not reported	not repo
1ecfdf16-f5e9-4a08-8834-37f4b474179d	TCGA-4Z-AA7M	TCGA-BLCA	male	1948	white	not reported	--	not reported	not reported	C67.
610811d5-e45e-4a4f-bdc1-8d700efd6e43	TCGA-K4-A6FZ	TCGA-BLCA	female	1937	white	not hispanic or latino	--	not reported	not reported	

Σχήμα 22: Παράδειγμα δομής αρχείου κλινικών δεδομένων.

.



Σχήμα 23: Μετά την επιλογή των επιθυμητών αρχείων, πραγματοποιείται η προσθήκη τους στην "Κάρτα" (Card) όπου εμφανίζονται όλα τα προς λήψη αρχεία και το σύνολο των ασθενών που έχουμε επιλέξει.

3.1.6 Λήψη των δεδομένων

Μετά την επιλογή των επιθυμητών αρχείων τα οποία ικανοποιούν τους περιορισμούς που θέσαμε, πραγματοποιήθηκε η προσθήκη τους στην "Κάρτα" (Card) όπου εμφανίζονται όλα τα προς λήψη αρχεία και το σύνολο των ασθενών που έχουμε επιλέξει (βλ. Σχήμα). Ανάλογα με το κάθε είδος δεδομένων και το σύνολο ασθενών που επιλέχθηκε, η GDC δίνει την δυνατότητα λήψης πληθώρας αρχείων που περιέχουν πληροφορίες που σχετίζονται με τους ασθενείς που επιλέξαμε στην Κάρτα. Αυτά τα αρχεία περιλαμβάνουν, ένα αρχείο με όλα τα κλινικά δεδομένα των ασθενών, τόσο ενός manifest.txt αρχείου το οποίο χρησιμοποιείται από το Data Transfer Tool εργαλείο όσο και ενός αρχείου αντιστοίχισης ενός μοναδικού κωδικού κάθε αρχείου με το ένα μοναδικό κωδικό κάθε ασθενή (UUID).

Η GDC προσφέρει δύο πιθανούς τρόπους λήψης των αρχείων που έχουν προστεθεί στην Κάρτα. Ο πρώτος τρόπος είναι χειροκίνητος και σχετικά χρονοβόρος και επομένως συνιστάται για μικρό όγκο δεδομένων. Ο δεύτερος τρόπος λήψης είναι αυτοματοποιημένος, γρήγορος και πραγματοποιείται μέσω του εργαλείου Data Transfer Tool. Το Data Transfer Tool είναι ένα εργαλείο γραμμής εντολών το οποίο προσφέρεται δωρεάν προς τον χρήστη από την GDC. Το Data Transfer Tool δέχεται ως είσοδο ένα manifest.txt αρχείο το οποίο δημιουργείται από την GDC Data Portal όταν έχει γίνει η τελική επιλογή των επιθυμητών αρχείων/ασθενών. Το αποτέλεσμα του Data Transfer Tool είναι η τοπική λήψη του συνόλου των αρχείων ανά ασθενή που έχουμε επιλέξει, ταξινομημένα σε φακέλους που περιέχουν ένα μοναδικό UUID. Λόγω του μεγάλου όγκου δεδομένων, στην παρούσα εργασία ακολουθήθηκε ο δεύτερος τρόπος λήψης αρχείων. Με αντίστοιχο τρόπο συλλέχθηκαν και τα κλινικά δεδομένα, με χρήση της Κάρτας.

3.2 Δημιουργία γενικού πινάκα με ολοκληρωμένη πληροφορία για κάθε ασθενή

Το στιγμιότυπο (instance) που επιλέχθηκε να μελετηθεί ήταν το δείγμα ενός ασθενούς. Επομένως, σημαντικό στάδιο της ανάλυσης ήταν η ευθυγράμμιση όλων των ασθενών/δειγμάτων για κάθε ομικό επίπεδο καθώς και η αποτελεσματική προσθήκη των επιθυμητών στηλών από τα κλινικά δεδομένα σε έναν ενιαίο πίνακα. Συγκεκριμένα, σε αυτόν τον ενιαίο πίνακα κάθε στήλη θα πρέπει να αντιπροσωπεύει ένα χαρακτηριστικό το οποίο μπορεί να περιλαμβάνει ή την έκφραση ενός γονιδίου ή την έκφραση ενός μορίου miRNA ή την τιμή εκτίμησης της μεθυλίωσης ενός γονιδίου και κάθε γραμμή το προφίλ ενός δείγματος ασθενούς. Για αυτό το σκοπό, σχεδιάστηκε και υλοποιήθηκε κατάλληλο πρόγραμμα σε γλώσσα προγραμματισμού Perl, και σε αυτή την ενότητα, θα αναλυθεί ο γενικός αλγόριθμος του προγράμματος αυτού.

Μετά την λήψη όλων των κατάλληλων αρχείων από το σύνολο των ασθενών, κάθε αρχείο δεδομένων ενός ασθενούς αποθηκεύεται αυτόματα σε ξεχωριστό φάκελο. Το πρόγραμμα που σχεδιάσαμε δέχεται ως πρώτη είσοδο ένα αρχείο (Sample Sheet) το οποίο περιλαμβάνει πληροφορία αντιστοίχισης ειδικών κωδικών μεταξύ των φακέλων, των ονομάτων των αρχείων, των ειδικών κωδικών κάθε ασθενούς, καθώς και το είδος ανάλυσης που έχει υποβληθεί κάθε δείγμα. Με βάση αυτό το αρχείο δημιουργούνται 3 ειδών πίνακες κατακερματισμού (hashes) (ομαδοποιημένοι με το είδος των δεδομένων) οι οποίοι αντιστοιχίζουν κάθε μοναδικό ID δείγματος (Case ID) με ένα μοναδικό μονοπάτι στα αρχεία του υπολογιστή (μοναδικός κωδικός φακέλου/ όνομα αρχείου) που αντιστοιχεί στο μονοπάτι των αρχείων που επιθυμούμε να επεξεργαστούμε.

Στη συνέχεια, κατασκευάζεται επιπλέον πίνακας κατακερματισμού στον οποίο αντιστοιχίζεται κάθε δείγμα με μία συμβολοσειρά η οποία αποτελείται από επιμέρους υποσυμβολοσειρές που συνδέονται με ";" οι οποίες αποτελούν το είδος των αναλύσεων που έχουν πραγματοποιηθεί στο δείγμα. Στην GDC είναι δυνατό ένας ασθενής που αντιστοιχεί σε μοναδικό Case ID να αντιστοιχεί σε παραπάνω από ένα δείγματα (Sample) με μοναδικά Sample IDs, και κάθε δείγμα να έχει πολλαπλές αναλύσεις (aliquots) τα οποία έχουν μοναδικά Aliquot IDs. Επιπλέον, αξίζει να σημειωθεί πως δεν είναι απαραίτητο ένα δείγμα ασθενούς να έχει αναλυθεί με όλες τις μεθόδους υψηλής απόδοσης. Επομένως, για να καταστεί ισοδύναμη και συγκρίσιμη η πληροφορία κάθε ασθενούς, ήταν απαραίτητο να εξασφαλίσουμε πως κάθε ασθενής θα αναπαριστάται από ένα προφίλ που αποτελείται από το σύνολο των ομικών δεδομένων.

Στην συνέχεια, κατασκευάστηκαν τρεις διαφορετικές υπορουτίνες για το κάθε είδος δεδομένων προς επεξεργασία. Οι υπορουτίνες αυτές δέχονται ως είσοδο τον αντίστοιχο πίνακα κατακερματισμού που αντιστοιχίζει κάθε μοναδικό ID δείγματος (Case ID) με ένα μοναδικό μονοπάτι στα αρχεία του υπολογιστή που οδηγεί στο αρχείο που περιέχει τα αντίστοιχο είδος δεδομένων. Ο σκοπός και των τριών υπορουτίνων είναι η κατασκευή λιστών (πραγματοποιήθηκε χρήση του πακέτου Set::Object) που θα περιέχουν το σύνολο των μοναδικών γονιδίων/miRNAs που περιέχει το κάθε αρχείο (η δομή δεδομένων Set απαγορεύει τις διπλότυπες εγγραφές).

Η πρώτη υπορουτίνα σχεδιάστηκε για την κατασκευή λίστας γονιδίων από αρχεία γονιδιακής έκφρασης καθώς και για την μετατροπή των Ensembl gene IDs σε HUGO/ HGNC Gene Symbols. Για τον δεύτερο σκοπό, το πρόγραμμα δέχεται ως είσοδο το κατάλληλο αρχείο που αντιστοιχίζει πληροφορία Ensembl gene IDs με HUGO/HGNC Gene Symbols, έτσι όπως σχεδιάστηκε από εμάς χρησιμοποιώντας το εργαλείο Biomart. Μέσα στην υπορουτίνα γίνεται κλήση μίας άλλης υπορουτίνας (exractmRNAVector) που σχεδιάστηκε για την βέλτιστη απομόνωση πληροφορίας από αρχεία γονιδιακής έκφρασης και την κατασκευή κατάλληλου πίνακα κατακερματισμού που θα έχει ως κλειδί ένα HUGO/HGNC Gene Symbol και ως τιμή την εκτίμηση FPKM-UQ. Αξίζει να σημειωθεί πως για την διάκριση των γονιδίων που αντιστοιχούν σε έκφραση μορίων mRNA από τα γονίδια που έχουν συσχετιστεί με επίπεδα μεθυλίωσης, προσθέσαμε ειδικό πρόθεμα `"m_"` μπροστά από κάθε συμβολοσειρά HUGO Gene Symbol των δεδομένων γονιδιακής έκφρασης.

Η δεύτερη υπορουτίνα σχεδιάστηκε για την κατασκευή λίστας γονιδίων από αρχεία εκτίμησης επιπέδων έκφρασης μορίων miRNA. Η υπορουτίνα δέχεται ως είσοδο το αντίστοιχο πίνακα κατακερματισμού που αντιστοιχίζει κάθε μοναδικό ID δείγματος (Case ID) με ένα μοναδικό μονοπάτι στα αρχεία του υπολογιστή που οδηγεί στο αρχείο που περιέχει τα αρχεία έκφρασης μορίων miRNA. Έπειτα, κατασκευάζει ένα σετ από μοναδικά miRNA IDs και καλεί μία επιπλέον υπορουτίνα (exractmiRNAVector) που σχεδιάστηκε για την κατάλληλη απομόνωση πληροφορίας από αρχεία έκφρασης μορίων miRNA και την κατασκευή κατάλληλου πίνακα κατακερματισμού που θα έχει ως κλειδί ένα miRNA ID και ως τιμή την εκτίμηση RPM.

Επί προσθέτως, σχεδιάστηκε η τρίτη υπορουτίνα η οποία όπως και οι δύο προηγούμενες αποθηκεύει σε λίστα το σύνολο των γονιδίων που συσχετίζονται με ένα ανιχνευτή συστοιχίας (array probe) μίας CpG νησίδας. Στην συνέχεια, πραγματοποιείται κλήση επιπλέον υπορουτίνας (exractMethylationVector) που σχεδιάστηκε για την κατάλληλη απομόνωση πληροφορίας από αρχεία μεθυλίωσης του DNA και την κατασκευή κατάλληλου πίνακα κατακερματισμού

που θα έχει ως κλειδί ένα HUGO/ HGNC Gene Symbol και ως τιμή την εκτίμηση beta- value. Ωστόσο, σε αυτό το είδος δεδομένων ήταν πιο απαιτητική η σχεδίαση καθώς μία επιπλέον πρόκληση που χρειάστηκε να αντιμετωπίσουμε ήταν πως ένας ανιχνευτής CpG νησίδας είναι δυνατό να συσχετίζεται με παραπάνω από ένα γονίδια και αντίστοιχα ένα γονίδιο να συσχετίζεται με παραπάνω από μία CpG Νησίδα. Επομένως, για την επίλυση αυτού του "υπολογιστικού" προβλήματος, σχεδιάστηκε και ενσωματώθηκε στην ανάλυση κατάλληλη υπορουτίνα (averageBetaValuePerGene) η οποία υπολογίζει τον μέσο όρο των Β-τιμών όλων των CpG νησίδων του κάθε αρχείου που είναι συσχετισμένες για το εκάστοτε γονίδιο. Επιπλέον, επειδή κάποιοι ανιχνευτές CpG νησίδων δεν είχαν τιμή ενσωματώθηκε στην υπορουτίνα ο έλεγχος "N/A" τιμών (Not Acquired Values) καθώς σε αυτό το είδος αρχείων η τιμή 0 έχει βιολογική σημασία.

Τέλος, αφού κατασκευαστούν οι τρεις πίνακες κατακερματισμού χρησιμοποιούνται ως είσοδο σε μία τελευταία υπορουτίνα (writeOutput) που κατασκευάστηκε για την κατάλληλη καταγραφή των δεδομένων σε ένα TSV αρχείο αλλά και για την ενσωμάτωση των κλινικών δεδομένων. Το πρώτο βήμα της υπορουτίνας είναι η καταγραφή της πρώτης γραμμής που συνιστά το σύνολο όλων των μοναδικών γονιδίων / miRNAs ανά είδος δεδομένων καθώς και επιθυμητών στηλών από τα κλινικά δεδομένα. Όσον αφορά τα κλινικά δεδομένα, αξίζει να υπογραμμιστεί το πεδίο του Days to Death που αντιστοιχεί στον αριθμό ημερών από την ημέρα διάγνωσης έως την ημέρα θανάτου ήταν κενό για τους ασθενείς που είναι ζωντανοί σύμφωνα με την τελευταία καταγραφή. Για την αντιμετώπιση αυτού του προβλήματος, αποφασίσαμε στις κενές τιμές να εισάγουμε ως τιμή το αποτέλεσμα της μέγιστης τιμής σε αυτό το πεδίο πολλαπλασιασμένη επί δύο. Στη συνέχεια, καταγράφονται τα μοναδικά ID δειγμάτων (Case ID) και οι τιμές ανά είδος δεδομένων ανά ασθενή όπως έχουν αποθηκευτεί στους πίνακες κατακερματισμού που κατασκευάστηκαν. Για τα δείγματα που δεν έχουν αναλυθεί σε ένα από τους τρεις τύπους δεδομένων προστίθενται "N/A" (Not Acquired) τιμές.

3.3 Κανονικοποίηση και διόρθωση δεδομένων

Η κανονικοποίηση αποτελεί ένα από τα πιο σημαντικά στάδια σε κάθε ανάλυση. Σκοπός της κανονικοποίησης είναι να καταστήσει το σύνολο των δεδομένων συγκρίσιμο καθώς και να πιστοποιήσει την σταθερότητα και την επαναληψιμότητα τους. Τα multi- omic δεδομένα των οποίων πραγματοποιήθηκε λήψη από την τράπεζα δεδομένων GDC ήταν ήδη σε μία κανονικοποιημένη μορφή αποτρέποντας συστηματικά λάθη που συνήθως προέρχονται

από την αναλυτική πλατφόρμα.

Στην παρούσα εργασία, πρώτο βήμα κανονικοποίησης αποτέλεσε η ανάδειξη σχετικών αναλογικών σχέσεων μεταξύ των καρκινικών δειγμάτων και των δειγμάτων ελέγχου. Για αυτό το σκοπό, αρχικά υπολογίστηκε ξεχωριστά ο μέσος όρος κάθε τιμής γονιδίου/miRNA όλων των δειγμάτων ελέγχου. Στην συνέχεια, για τον υπολογισμό της σχετικής αλλαγής (relative change) όπως φαίνεται στην Σχέση (8), κάθε τιμή έκφρασης γονιδίου/miRNA και εκτίμηση μεθυλίωσης γονιδίου καρκινικού δείγματος αφαιρέθηκε από την αντίστοιχη μέση τιμή γονιδίου/miRNA των δειγμάτων ελέγχου και το υπόλοιπο της αφαίρεσης διαιρέθηκε με την αντίστοιχη μέση τιμή γονιδίου/miRNA των δειγμάτων ελέγχου.

$$\texttt{Relative Change}_{(Y_t - Y_c)} = \frac{Y_i - Mean(Y_c)}{Mean(Y_c)} \tag{8}$$

όπου, Y_t είναι η τιμή του στοιχείου Y του καρκινικού δείγματος t, Y_c είναι η τιμή του στοιχείου Y του δείγματος ελέγχου c, Y_i είναι η τιμή του στοιχείου Y οποιοδήποτε δείγματος, Mean (Y_c) αντιστοιχεί στο μέσο όρο του Y_c.

Στην συνέχεια, για τον περιορισμό των δεδομένων σε ένα μικρότερο εύρος αλλά και για την καλύτερη ανάδειξη αναλογικών σχέσεων υπολογίσαμε τον λογάριθμο με βάση το 2 το πηλίκο της μορφής που αναφέραμε παραπάνω για κάθε τιμή για όλα τα δείγματα. Έτσι, για παράδειγμα, μία διπλάσια αύξηση στην έκφραση οδηγεί σε μία κανονικοποιημένη τιμή ίση με 1 έχοντας ίση απόλυτη τιμή με μία μείωση στο μισό στην έκφραση η οποία είναι ίση με -1. Ωστόσο, καθώς δεν ορίζεται αρνητικός λογάριθμος, για να πιστοποιήσουμε μόνο θετικές τιμές πηλίκου της σχέσης (8), αθροίσαμε το αποτέλεσμα του πηλίκου της σχέσης (8) με το 2. Έτσι, μία διπλάσια αύξηση στην έκφραση του καρκινικού δείγματος σε σχέση με το δείγμα ελέγχου οδηγεί σε μία κανονικοποιημένη τιμή ίση με log_2(2 + 1) = 1,58, αντίθετα ένας υποδιπλασιασμός στην έκφραση σε σχέση με το δείγμα ελέγχου οδηγεί σε μία κανονικοποιημένη τιμή ίση με log_2(2 + (-0.5)) = 0,58, για μηδενικές αλλαγές σε τιμές του καρκινικού δείγματος σε σχέση με το δείγμα ελέγχου ισχύει log_2(2 + (-0)) = 1, ενώ για την μικρότερη δυνατή τιμή (μηδενική τιμή) του καρκινικού δείγματος σε σχέση με το δείγμα ελέγχου ισχύει log_2(2 + (-1)) = 0.

Επιπλέον, μία επιπλέον πρόκληση των δεδομένων που κληθήκαμε να αντιμετωπίσουμε ήταν η ύπαρξη "N/A"(Not Acquired) τιμών. Για λόγους διατήρησης της ευθυγράμμισης των δεδομένων καθώς και διατήρησης του αριθμού των δειγμάτων/ χαρακτηριστικών, επιλέξαμε να αντικαταστήσουμε τις "N/A" τιμές με τον μέσο όρο αυτού του χαρακτηριστικού από όλα τα δείγματα της εκάστοτε ομάδας. Ωστόσο, χαρακτηριστικά που αποτελούνταν μόνο από

”N/A” τιμές αγνοήθηκαν από την ανάλυση.

3.4 Διαφορετικές αναπαραστάσεις για μηχανική μάθηση

3.4.1 Διανυσματική αναπαράσταση

Η πρώτη αναπαράσταση που κατασκευάσαμε και μελετήσαμε ήταν η διανυσματική. Το διάνυσμα αποτελεί μία ευρεία έννοια με πολλές εφαρμογές σε πολλούς τομείς. Ένα διάνυσμα στα πλαίσια της μηχανικής μάθησης, ορίζεται ως ένας πίνακας ή μία λίστα μεγέθους N, όπου N είναι το σύνολο των χαρακτηριστικών του μοντέλου. Η δομή δεδομένων του πίνακα ή της λίστας αποτελείται από 1 διάσταση, καθώς απλώς αποθηκεύει δεδομένα στη μνήμη πχ (3.65, 7.8, 100.5). Αντίθετα, η αναπαράσταση ενός πίνακα που αποτελείται από 3 χαρακτηριστικά (3.65, 7.8, 100.5) στη μηχανική μάθηση, πραγματοποιείται μέσω της τοποθέτησης ενός σημείου (που αναπαριστά ένα στιγμιότυπο) σε ένα τρισδιάστατο χώρο. Κάθε χαρακτηριστικό προσθέτει μία επιπλέον διάσταση στην οποία κάθε χαρακτηριστικό έχει συγκεκριμένη τιμή (attribute).

Στο δικό μας βιολογικό πρόβλημα, όπως αναφέραμε σε παραπάνω ενότητα, ορίσαμε ως στιγμιότυπο το ένα δείγμα το οποίο αναπαριστάται από δύο κατηγορίες (Καρκινικό Δείγμα- Δείγμα Ελέγχου). Επιπλέον, το σύνολο των γονιδίων, των miRNA και της κλινικής πληροφορίας που καταφέραμε να απομονώσουμε από το σύνολο των δεδομένων αποτελούν τα χαρακτηριστικά μας. Επομένως, ένα διάνυσμα χαρακτηριστικών (feature vector) αναπαριστά ένα δείγμα και κάθε χαρακτηριστικό έχει μία τιμή που αντιστοιχεί στα παρακάτω:

- εκτίμηση επιπέδων έκφρασης γονιδίων
- εκτίμηση επιπέδων έκφρασης miRNA
- εκτίμηση επιπέδων μεθυλίωσης του DNA
- κατηγορική τιμή για το είδος του δείγματος
- κατηγορική τιμή για το στάδιο καρκίνου του ασθενούς

3.4.2 Δίκτυα

Ο δεύτερος τρόπος αναπαράστασης που θέλαμε να μελετήσουμε το βιολογικό μας σύστημα ήταν το δίκτυο. Όπως και στην διανυσματική αναπαράσταση, επιλέξαμε ως στιγμιότυπο να είναι το δείγμα ενός ασθενούς.

3.4.2.1 Κατασκευή Δικτύων

Το πρώτο βήμα στην αναπαράσταση των δεδομένων με τη μορφή δικτύου ήταν η λήψη αποφάσεων σχετικά με το τι θα αναπαριστούν οι κόμβοι καθώς και την φύση της σχέσης δύο κόμβων που συνδέονται με μία ακμή. Λόγω της ετερογένειας των δεδομένων μας ακολουθήσαμε την κοινή στρατηγική που ακολουθείται σε multi- omics μελέτες, η οποία αφορά την ενσωμάτωση όλων των δεδομένων σε ένα κοινό δίκτυο όπου υπάρχουν 3 ειδών κόμβοι, οι οποίοι είναι:

- γονίδια τα οποία έχουν συσχετιστεί με CpG νησίδες στις οποίες έχουν εκτιμηθεί τα επίπεδα μεθυλίωσης
- γονίδια των οποίων έχουν εκτιμηθεί τα επίπεδα έκφρασής τους
- miRNA των οποίων έχουν εκτιμηθεί τα επίπεδα έκφρασής τους

Επομένως, κάθε κόμβος αποτελεί ένα χαρακτηριστικό από τον αρχικό διάνυσμα χαρακτηριστικών που κατασκευάσαμε σε προηγούμενο βήμα.

Επόμενο βήμα στην αναπαράσταση δικτύων ήταν η κατασκευή του δικτύου για τον καρκίνο της ουροδόχου κύστης, λαμβάνοντας υπόψη πληροφορία από όλους τους ασθενείς. Σε αυτό το δίκτυο, κάθε χαρακτηριστικό αναπαριστάται μέσω της κατασκευής ενός πίνακα που περιέχει το σύνολο των τιμών αυτού του χαρακτηριστικού στο σύνολο όλων των δειγμάτων. Δύο κόμβοι θα συνδέονται με ακμή εφόσον συσχετίζονται αρνητικά ή θετικά. Η μετρική συσχέτισης που χρησιμοποιήσαμε είναι ο συντελεστής συσχέτισης Pearson (Pearson Correlation Coefficient). Ο συντελεστής Pearson παίρνει τιμές από το -1 έως το 1. Μία συσχέτιση μεταξύ δύο μεταβλητών που έχει τιμή κοντά στο -1 υποδεικνύει πως αυτές οι δύο μεταβλητές έχουν μία τέλεια αρνητική συσχέτιση. Αντίθετα, αν για δύο μεταβλητές ο συντελεστής συσχέτισης Pearson έχει τιμή ίση με 1, τότε αυτές οι δύο μεταβλητές έχουν μία τέλεια θετική συσχέτιση. Το κατώφλι που ορίσαμε για την συσχέτιση δύο κόμβων ήταν το 0,3 (απόλυτη τιμή). Επομένως, δύο κόμβοι με αυξημένη συσχέτιση ενώνονται μέσω μίας ακμής και εισάγουμε στην ακμή τους ένα βάρος που ισούται με την τιμή του συντελεστή συσχέτισης Pearson.

Μετά την κατασκευή του multi- omics δικτύου για όλα τα δείγματα καρκίνου, ακολουθήσαμε διαφορετικές στρατηγικές προκειμένου να αναπαραστήσουμε με τη μορφή δικτύου αυτή τη φορά το προφίλ ενός ασθενή. Το γενικότερο multi- omics δίκτυο που κατασκευάστηκε σε προηγούμενο βήμα, χρησιμοποιήθηκε ως το βασικό δίκτυο πάνω στο οποίο χαρτογραφούνται οι τιμές σχετικής αλλαγής κάθε ασθενούς. Με άλλα λόγια, στο δίκτυο αυτό κάθε

κόμβος είναι ένα χαρακτηριστικό το οποίο έχει ως βάρος μία κανονικοποιημένη τιμή του χαρακτηριστικού του συγκεκριμένου ασθενούς.

Λόγω του εξαιρετικά μεγάλου αριθμού χαρακτηριστικών που αναπαριστάται κάθε στιγμιότυπο (ο οποίος είναι ίσος με 73667), συμπεριλήφθηκαν στην δικτυακή αναπαράσταση μόνο γονίδια και miRNA τα οποία είχαν σχετική αλλαγή πάνω από 65. Αυτό το κατώφλι ορίστηκε για λόγους υπολογιστικής ισχύος.

3.4.3

Τοπολογική Αναπαράσταση

Μετά την κατασκευή των δικτύων, επόμενο βήμα ήταν να καταλήξουμε από ένα μεγάλο αριθμό κόμβων σε ένα μικρότερο σύνολο "σημαντικών κόμβων", όπου θα μπορούσαν πιο εύκολα να αναδυθούν ιδιότητες και χαρακτηριστικά του δικτύου.

Η ανάδειξη σημαντικών κόμβων είναι μία διαδικασία με πολλές προκλήσεις και με μία μεγάλη ποικιλία στρατηγικών. Στην παρούσα εργασία, ακολουθήσαμε την στρατηγική της ενεργοποίησης διάδοσης (Spreading activation) για την ανάδειξη υψηλά συσχετιζόμενων υποδικτύων. Η ενεργοποίηση διάδοσης είναι μια μέθοδος για την αναζήτηση υψηλά συσχετισμένων υποδικτύων που έχει εφαρμοστεί ευρέως σε νευρωνικά και σε σημασιολογικά δίκτυα (Tsatsaronis, Vazirgiannis, and Androutsopoulos 2007). Η συνήθης διαδικασία ενεργοποίησης ξεκινά με την επισήμανση με βάρη ενός συνόλου κόμβων προέλευσης (source nodes). Η ενεργοποίηση αυτή διαδίδεται μέσω του δικτύου, σταδιακά εξασθενεί και τελικά τερματίζεται όταν δύο εναλλακτικές διαδρομές φτάνουν στον ίδιο κόμβο. Η ιδέα πίσω από το μοντέλο ενεργοποίησης διάδοσης είναι πως κόμβοι που μοιράζονται σε μεγάλο βαθμό κοινά χαρακτηριστικά, τείνουν να διαθέτουν αντίστοιχες ενεργοποιήσεις (Collins and Loftus 1975), με άλλα λόγια, κόμβοι που έχουν υψηλή συνδεσιμότητα "ενεργοποιούν" σε μεγαλύτερο βαθμό τους γειτονικούς του κόμβους σε σχέση με άλλους κόμβους που δεν είναι τόσο "δημοφιλείς".

Στην παρούσα εργασία, η ενεργοποίηση διάδοσης επαναλαμβάνεται 100 φορές και σε κάθε επανάληψη όλοι οι κόμβοι είναι κόμβοι προέλευσης. Τα βάρη των κόμβων προέλευσης είναι ίσα με την κανονικοποιημένη τιμή του χαρακτηριστικού που αναπαριστά ο κόμβος του εκάστοτε ασθενούς. Κάθε κόμβος μοιράζει το μισό του βάρος αναλογικά προς τους γειτονικούς του κόμβους. Για παράδειγμα, έστω ότι ένας κόμβος προέλευσης έχει βάρος ίσο με 1 και έχει δύο γειτονικούς κόμβους, όπου με τον πρώτο έχουν συσχέτιση ίση με 0,3 και με τον δεύτερο ίση με 0,8 τότε θα μοιράσει το 50% του βάρους

του που είναι ίσο με 0,5 στους δύο γειτονικούς του κόμβους, δίνοντας 0,4 στον κόμβο που έχουν συσχέτιση 0,8 και μόνο 0,1 στον άλλο κόμβο. Αξίζει να σημειωθεί πως αρνητικές και θετικές συσχετίσεις τις θεωρούμε ισοδύναμα σημαντικές. Με αυτόν τον τρόπο, όχι μόνο ευνοούνται οι κόμβοι με υψηλή συνδεσιμότητα αλλά και με μεγάλη συσχέτιση. Μετά την ενεργοποίηση διάδοσης που εφαρμόστηκε σε όλα τα δίκτυα ασθενών, ορίσαμε ένα κατώφλι στο οποίο κρατήσαμε το 25% των "σημαντικότερων" κόμβων με τα υψηλότερα βάρη όπως αυτά τροποιήθηκαν.

Τελευταίο βήμα στην δικτυακή αναπαράσταση ήταν η εξόρυξη τοπολογικών χαρακτηριστικών από τους τροποποιημένους γράφους. Τα χαρακτηριστικά τα οποία επιλέχθηκαν για την αναπαράσταση της τοπολογίας ενός ολοκληρωμένου δικτύου είναι ο αριθμός κόμβων N, ο αριθμός των ακμών E, ο μέσος όρος ελάχιστου μονοπατιού, ο μέσος όρος συνδεσιμότητας των κόμβων, μέση κεντρικότητα βαθμού (Mean degree centrality), αριθμός κλικών γράφου (graph number of cliques) και ο αριθμός συνδεδεμένων στοιχείων (number of connected components).

3.5 Διερευνητική στατιστική ανάλυση των δεδομένων

Για την ανακάλυψη μίας πιθανής ενδιαφέρουσας δομής αλλά και σημαντικών χαρακτηριστικών που αναδύονται από τα δεδομένα που συγκεντρώσαμε, πραγματοποιήσαμε κάποιες πρώτες διερευνητικές αναλύσεις των δεδομένων που αφορούν την Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis-PCA) χρησιμοποιώντας την γλώσσα προγραμματισμού Python.

3.5.1 Ανάλυση Κύριων Συνιστωσών (Principal Component Analysis-PCA)

Η Ανάλυση Κύριων Συνιστωσών γνωστή και ως PCA (Principal component analysis) είναι ένα εργαλείο μείωσης διαστάσεων το οποίο χρησιμοποιείται ευρέως σε διαφορετικές αναλύσεις μεγάλων δεδομένων διότι είναι μία πολυπαραμετρική μέθοδος εξόρυξης σχετικής πληροφορίας από περίπλοκα σύνολα δεδομένων.

Η PCA βρίσκει τον πιο αντιπροσωπευτικό χώρο χαμηλότερων διαστάσεων που εμπεριέχει το μεγαλύτερο ποσοστό πληροφορίας του αρχικού πίνακα δεδομένων N x K. Πιο συγκεκριμένα, έστω ότι το σύνολο των δεδομένων μας αποτελείται από N χαρακτηριστικά/μεταβλητές που στην παρούσα εργασία είναι το σύνολο των γονιδίων και των miRNA και που αντιστοιχούν σε K δείγματα ασθενών, προκύπτει ένας πίνακας δεδομένων N x K, με N σειρές

και Κ στήλες. Όλα τα στοιχεία Κ/δείγματα ασθενών του πίνακα δεδομένων μπορούν να αντιπροσωπευθούν ως σημεία στον Ν-διάστατο χώρο.

Ειδικότερα, ο αλγόριθμος της PCA εντοπίζει ένα γραμμικό συνδυασμό των μεταβλητών, ο οποίος επιτρέπει την εξόρυξη μέγιστης διακύμανσης από τα δεδομένα. Στη συνέχεια, αφαιρείται αυτή η μέγιστη διακύμανση και πραγματοποιείται εύρεση δεύτερου γραμμικού συνδυασμού στον οποίο μεγιστοποιείται η διακύμανση. Αυτή η διαδικασία επαναλαμβάνεται για πολλές φορές και ονομάζεται μέθοδος των κυρίων συνιστωσών. Οι κύριες συνιστώσες είναι γραμμικοί συνδυασμοί των αρχικών μεταβλητών με ενσωματωμένο βάρος την συνεισφορά κάθε μεταβλητής στην συνολική διακύμανση σε μία ορθογώνια διάσταση. Σε κάθε άξονα αντιστοιχεί μία ιδιοτιμή, ενώ το άθροισμα όλων των ιδιοτιμών που αντιστοιχούν σε κάθε άξονα ισούται με το σύνολο της διακύμανσης των τιμών. Η τιμή της διακύμανσης κάθε άξονα αντιπροσωπεύει το ποσό της πληροφορίας που αντιστοιχεί στον αρχικό χώρο δεδομένων (κύρια συνιστώσα/άξονας 1 > κύρια συνιστώσα/άξονας 2 > κύρια συνιστώσα/άξονας 3). σε

Σε κάθε εικόνα ανάλυσης PCA αναγράφεται ένα ποσοστό σε κάθε άξονα το οποίο αντιστοιχεί στο ποσοστό πληροφορίας που αυτός περιέχει. Σε ένα γράφημα PCA είναι σημαντικό να λαμβάνεται υπόψη το ποσοστό της πληροφορίας που αντικατοπτρίζεται σε αυτόν το τρισδιάστατο χώρο συνολικά αλλά και από κάθε άξονα ξεχωριστά.

3.6 Πρόβλημα Ταξινόμησης

Μετά την κατασκευή των τοπολογικών πινάκων χαρακτηριστικών για κάθε δείγμα, επόμενο βήμα ήταν να ελέγξουμε αν η τοπολογική αναπαράσταση προσέφερε καλύτερη και πιο στοχευμένη ικανότητα ταξινόμησης από την ευθύγραμμη αναπαράσταση.

Ειδικότερα, το πρόβλημα ταξινόμησης που επιλέξαμε, αφορούσε την ικανότητα διαχωρισμού καρκινικών δειγμάτων από δείγματα ελέγχου. Για αυτό το σκοπό, χρησιμοποιήθηκε ο επιβλεπόμενος αλγόριθμος μηχανικής μάθησης που είναι τα δέντρα αποφάσεων (decision trees). Τα δέντρα αποφάσεων χρησιμοποιούνται ευρέως τόσο για προβλήματα ταξινόμησης όσο και για προβλήματα εκτίμησης τιμής. Η βασική ιδέα πίσω από τον αλγόριθμο των δέντρων αποφάσεων είναι πως προσπαθούν να βρουν αυτά τα χαρακτηριστικά τα οποία περιέχουν την περισσότερη πληροφορία σχετικά με το χαρακτηριστικό στόχου (target feature) (όπου στο παρόν πρόβλημα ταξινόμησης είναι το είδος δείγματος) και στη συνέχεια να χωρίσουν το σύνολο δεδομένων με βάση τις τιμές αυτών των χαρακτηριστικών έτσι ώστε κάθε υποσύνολο να

ανήκει σε μία κατηγορία. Για την υλοποίηση της ταξινόμησης χρησιμοποιήθηκε ο αλγόριθμος των δέντρων απόφασης (decision trees), χρησιμοποιώντας την κλάση DecisionTreeClassifier της γλώσσας προγραμματισμού python. Η DecisionTreeClassifier δέχεται ως είσοδο ένα πίνακα X x Y μεγέθους [αριθμός δειγμάτων, αριθμός χαρακτηριστικών] του συνόλου εκπαίδευσης και έναν πίνακα Y μεγέθους [αριθμός δειγμάτων] του συνόλου εκπαίδευσης κρατώντας τις ετικέτες κάθε ομάδας, δηλαδή είτε "καρκινικό δείγμα" είτε "δείγμα ελέγχου".

Ένας άλλος τρόπος με τον οποίο επιλέξαμε να ταξινομήσουμε τα δεδομένα μας ήταν ο k-nearest neighbor (K-NN). Ο k-nearest neighbor είναι μία μη- παραμετρική μέθοδος που χρησιμοποιείται τόσο για προβλήματα ταξινόμησης όσο και για προβλήματα εκτίμησης τιμής. Η ιδέα πίσω από αυτόν τον αλγόριθμο είναι πως ένα στιγμιότυπο ταξινομείται σε μία κατηγορία με βάση την πιο συχνή κλάση των πιο κοντινών γειτόνων του. Στα δικά μας πειράματα θέσαμε k=3.

Για την εκτίμηση της απόδοσης του κάθε μοντέλου αλλά και για την αποτελεσματική σύγκριση των επιδόσεων όλων των μοντέλων, χρησιμοποιήθηκε ένα σύνολο από μετρικές που χρησιμοποιούνται ευρέως στη βιβλιογραφία σε προβλήματα ταξινόμησης. Ξεκινώντας, με το F1-score το οποίο μπορεί να ερμηνευτεί ως ο σταθμισμένος μέσος όρος ακρίβειας και ανάκλησης, με την βέλτιστη τιμή του να φθάνει στο 1 και τη χειρότερη στο 0. Η σχετική συμβολή της ακρίβειας και της ανάκλησης στο F1-score είναι ίση (βλ. μαθηματικό τύπο 5). Λόγω της μεγάλης ανισσοροπίας που υπήρχε στον αριθμό των δύο κλάσεων (412 καρκινικά δείγματα προς 23 δείγματα ελέγχου), χρησιμοποιήσαμε δύο μετρικές προκειμένου να αντιμετωπίσουμε αυτή την πρόκληση. Αυτές οι μετρικές αφορούν το μέσο f1- macro και το μέσο f1-micro. Ειδικότερα, η μετρική micro είναι ευαίσθητη στην ύπαρξη ανισσοροπίας στον αριθμό των κλάσεων (class imbalance). Αντίθετα, η μετρική micro υπολογίζει τις μετρικές για κάθε ετικέτα κλάσης και στο τέλος κατασκευάζει τον μέσο όρο χωρίς να λαμβάνει υπόψη την ανισσοροπία στον αριθμό των κλάσεων. Για κάθε τιμή macro/micro υπολογίζεται η τυπική απόκλιση αυτής της κάθε τιμής micro/macro για το σύνολο των προγνώσεων. Επιπλέον, ένα άλλο μέτρο επίδοσης που χρησιμοποιήθηκε σε όλα τα πειράματά ταξινόμησης είναι η Πιστότητα (Accuracy). Τέλος, για κάθε μοντέλο ταξινόμησης υπολογίζεται το Standard Error of the mean, το οποίο είναι ένα μέτρο που υπολογίζει την τυπική απόκλιση της κατανομής των δειγμάτων και ισούται με τον μαθηματικό τύπο 6.

4 Αποτελέσματα

4.1 Κλινικά και παθολογικά χαρακτηριστικά των ασθενών

Το σύνολο των δεδομένων τα οποία συγκεντρώθηκαν στην παρούσα εργασία, αφορούν 412 ασθενείς από τους οποίους πραγματοποιήθηκε λήψη δειγμάτων όγκου υψηλού βαθμού (high grade) και πάσχουν από ουροθηλιακό καρκίνωμα της ουροδόχου κύστης. Είναι σημαντικό να αναφερθεί πως μέχρι την ώρα της δειγματοληψίας δεν είχαν υποβληθεί σε χημειοθεραπεία. Επιπλέον, η λήψη δειγμάτων όγκου πραγματοποιήθηκε από 32 διαφορετικά σημεία ιστού που χαρακτηρίστηκαν από ειδικούς παθολογοανατόμους ως καθαρά ουροθηλιακά ή μικτής ιστολογίας.

Από το σύνολο των ασθενών, οι 35 είχαν λάβει προηγουμένως ενδοαγγειακή ανοσοθεραπεία με Bacille Calmette-Guerin (BCG), και 12 είχαν λάβει προεγχειρητική χημειοθεραπεία (Neoadjuvant chemotherapy, NAC) μετά την απόκτηση του όγκου. Επιπλέον, συνολικά 120 ασθενείς ήταν ζωντανοί, 163 είχαν επανεμφανίσει όγκο και 182 είχαν πεθάνει. Η ενδιάμεση παρακολούθηση των ασθενών ήταν 20,9 μήνες για όσους ζούσαν στην τελευταία παρακολούθηση. Τουλάχιστον 122 (67%) θάνατοι ασθενών σχετίζονταν με τον καρκίνο. Τα δείγματα συνολικά χαρακτηρίστηκαν από κλινικά δεδομένα και από 6 αναλυτικές πλατφόρμες υψηλής απόδοσης.

4.2 Γενικές πληροφορίες για το σύνολο δεδομένων

Πραγματοποιήθηκε επιτυχώς η λήψη αρχείων τα οποία περιέχουν multi-omics και κλινικά δεδομένα 412 ασθενών που πάσχουν από ουροθηλιακό καρκίνωμα της ουροδόχου κύστης χρησιμοποιώντας την τράπεζα δεδομένων Genomic Data Commons.

Ύστερα από εφαρμογή των επιθυμητών κριτηρίων για την λήψη αρχείων που αφορούν δείγματα πρωτογενούς όγκου το αποτέλεσμα αφορούσε συνολικά 412 ασθενείς και 1250 αρχεία (1249 αρχεία για τα 3 -omics δεδομένα και 1 αρχείο κλινικών δεδομένων). Όσον αφορά την επιλογή και λήψη των δειγμάτων ελέγχου που αφορούσαν δείγματα μη καρκινικού ιστού από την ουροδόχο κύστη ασθενών, συνολικά καταλήξαμε με 23 δείγματα τα οποία συνολικά αντιστοιχούσαν σε 60 αρχεία (59 αρχεία για τα 3 -omics δεδομένα και 1 αρχείο κλινικών δεδομένων) όπως φαίνεται στους Πίνακες 7 και 8, αντίστοιχα. Αξίζει να σημειωθεί πως δεν είναι απαραίτητο ένα δείγμα ασθενούς να έχει αναλυθεί με όλες τις μεθόδους υψηλής απόδοσης και αντίστοιχα είναι δυνατό ένα δείγμα ασθενούς να έχει αναλυθεί παραπάνω από μία φορές. Αυτό εξηγεί το γεγονός πως ο αριθμός των αρχείων δεν είναι ίσος πάντα με τον αριθμό των ασθενών, όπως παρατηρούμε στους Πίνακες 7 και 8.

Τύπος δεδομένων	Αριθμός ασθενών	Αριθμός Αρχείων
Δεδομένα γονιδιακής έκφρασης	19	19
Δεδομένα έκφρασης miRNA	19	19
Δεδομένα μεθυλίωσης του DNA	21	21
Κλινικά δεδομένα	23	1
Συνολικά	23	60

Πίνακας 7: Αριθμός αρχείων που πραγματοποιήθηκε λήψη για το κάθε είδος δεδομένων από μη καρκινικό ιστό ουροδόχου κύστης ασθενών που πάσχουν από μυοδιηθητικό ουροθηλιακό καρκίνωμα της ουροδόχου κύστης (Δείγματα ελέγχου).

Τύπος δεδομένων	Αριθμός ασθενών	Αριθμός Αρχείων
Δεδομένα γονιδιακής έκφρασης	408	414
Δεδομένα έκφρασης miRNA	409	417
Δεδομένα μεθυλίωσης του DNA	412	418
Κλινικά δεδομένα	412	1
Συνολικά	412	1250

Πίνακας 8: Αριθμός αρχείων που πραγματοποιήθηκε λήψη για το κάθε είδος δεδομένων ασθενών που πάσχουν από ουροθηλιακό καρκίνωμα της ουροδόχου κύστης

4.3 Διερευνητική ανάλυση δεδομένων

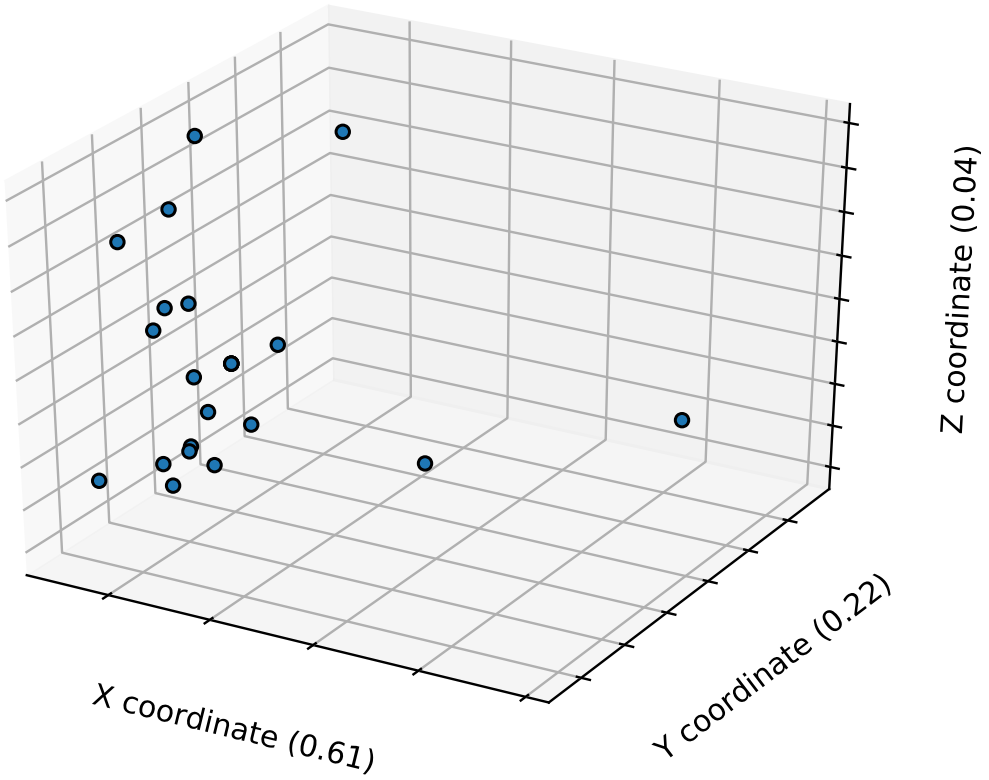
Τα πρώτα ερωτήματα που θέσαμε στα δεδομένα μας ήταν διερευνητικής φύσεως με απώτερο σκοπό την καλύτερη κατανόηση της δομής τους καθώς και την ανακάλυψη ιδιαίτερων χαρακτηριστικών τους. Στη συνέχεια, χρησιμοποιώντας τις δύο αναπαραστάσεις multi- omics δεδομένων που κατασκευάσαμε (την διανυσματική και την τοπολογική αναπαράσταση) αξιολογήθηκε συγκριτικά η απόδοσή τους τόσο σε προβλήματα ταξινόμησης όσο και σε προβλήματα εκτίμησης τιμής. Είναι σημαντικό να τονιστεί πως πριν από την υλοποίηση κάθε πειράματος προηγήθηκε κατάλληλη κανονικοποίηση στο σύνολο των δεδομένων κάθε δείγματος ως προς το μέσο πρότυπο των δειγμάτων ελέγχου.

4.3.1 Διερευνητική ανάλυση των διανυσματικών αναπαραστάσεων

Η πρώτη αναπαράσταση που κατασκευάστηκε και μελετήθηκε ήταν η διανυσματική. Για αυτό το σκοπό, όπως αναφέρθηκε σε παραπάνω ενότητα, κατασκευάστηκαν δύο διανύσματα χαρακτηριστικών, ένα για τα καρκινικά δείγματα (Tumor samples) και ένα για τα δείγματα ελέγχου (Control Samples/ Normal Tissue Samples). Συνολικά, ο αριθμός των χαρακτηριστικών για την διανυσματική αναπαράσταση είναι ίσος με 73667. Ενώ, ο συνολικός αριθμός των στιγμιοτύπων που κατασκευάσαμε ίσος με 435.

Ένα πρώτο ερώτημα που θέσαμε ήταν ο έλεγχος ύπαρξης συνοχής στη διανυσματική αναπαράσταση των δεδομένων ελέγχου. Για την υλοποίηση αυτού του ερωτήματος χρησιμοποιήθηκε η πολυπαραμετρική μέθοδος της Ανάλυσης Κυρίων Συνιστωσών (Principal Component Analysis, PCA) στους πίνακες χαρακτηριστικών όλων των δειγμάτων ελέγχου. Στο σχήμα 24 εμφανίζεται το γράφημα της PCA, όπου κάθε κουκκίδα αντιπροσωπεύει το multi- omic προφίλ ενός δείγματος. Όσο πιο κοντά βρίσκονται δύο κουκκίδες/δείγματα τόσο μεγαλύτερη είναι η ομοιότητα των multi-omic προφίλ τους, με τη βαρύτητα της ομοιότητας να είναι διαφορετική ανάλογα με την κύρια συνιστώσα. Η πρώτη κύρια συνιστώσα αποτελείται από το 61.0% της διακύμανσης του αρχικού πειραματικού χώρου, η δεύτερη αποτελείται από το 22.0% και η τρίτη από το 4%. Συγκεκριμένα, από τα αποτελέσματα της Ανάλυσης Κυρίων Συνιστωσών για τα δείγματα ελέγχου παρατηρείται πως η πλειοψηφία των δειγμάτων ομαδοποιείται ως προς την πρώτη κύρια συνιστώσα που έχει το μεγαλύτερο ποσοστό πληροφορίας. Ωστόσο παρατηρούμε 2 δείγματα τα οποία διαφοροποιούνται αρκετά από το σύνολο των δειγμάτων ελέγχου.

Επόμενο βήμα στην διερευνητική ανάλυση ήταν ο έλεγχος διαφοροποίησης των προφίλ δειγμάτων ελέγχου από τα προφίλ των καρκινικών δειγμάτων χρησιμοποιώντας την διανυσματική τους αναπαράσταση. Τα αποτελέσματα της Ανάλυσης Κυρίων Συνιστωσών φαίνονται στο Σχήμα 25. Το συνολικό ποσοστό διακύμανσης από τις τρεις κύριες συνιστώσες είναι αρκετά υψηλό και επομένως αντικατοπτρίζει μεγάλο ποσοστό του πραγματικού πειραματικού χώρου. Με μία πρώτη ματιά, δεν παρατηρούμε σημαντική διαφοροποίηση των δύο σχετικών ομάδων. Ωστόσο, παρατηρούμε πως τα δείγματα ελέγχου ομαδοποιούνται μεταξύ τους σχηματίζοντας μία συμπαγή μάζα, ενώ τα καρκινικά δείγματα διασπείρονται περικυκλικά των δειγμάτων ελέγχου. Σε αυτό το σημείο, υπενθυμίζεται πως τα δείγματα ελέγχου έχουν ληφθεί από μη καρκινικό ιστό ουροδόχου κύστης από ασθενείς που πάσχουν από ουροθηλιακό καρκίνωμα της ουροδόχου κύστης, επομένως δεν αναμέναμε τόση διαφοροποίηση όσο αν χρησιμοποιούσαμε ως δείγμα ελέγχου ιστό ουροδόχου κύστης από έναν υγιή άνθρωπο. Επομένως, αυτό το μοτίβο ομαδοποίησης των δύο ομάδων δειγμάτων δικαιολογεί την σχετική ομοιότητά τους ως προς το μοριακό τους προφίλ και παράλληλα επιδεικνύει τον βαθμό κακοήθειας και αποκλίνουσας συμπεριφοράς των καρκινικών δειγμάτων που ως προς την πρώτη κύρια συνιστώσα που έχει τη μεγαλύτερη βαρύτητα απομακρύνονται από τα δείγματα ελέγχου.

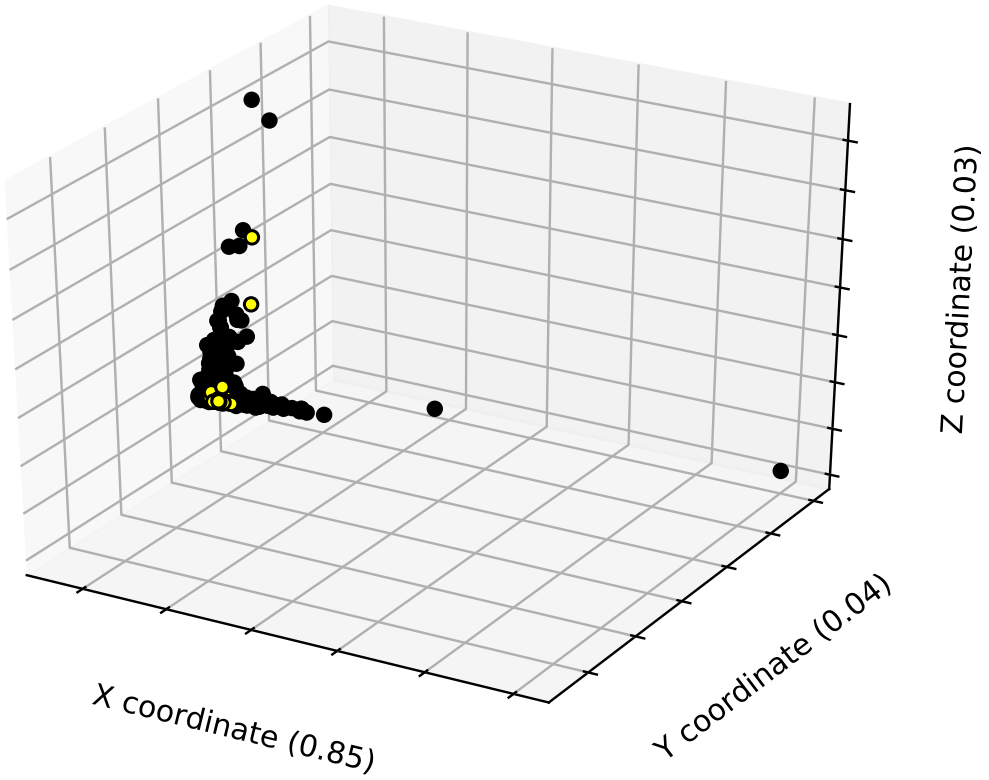


Σχήμα 24: Ανάλυση κύριων συνιστωσών (PCA) στην διανυσματική αναπαράσταση των multi- omic προφίλ όλων των δειγμάτων ελέγχου. Τα X, Y και Z coordinates αναφέρονται στις τρεις διαστάσεις των κύριων συσιστωσών και κάθε συνιστώσα χαρακτηρίζεται από ένα νούμερο που αφορά την % διακύμανση των προτύπων από τον κανονικό πειραματικό χώρο που μεταφέρονται στις κύριες συνιστώσες.

4.4 Διερευνητική ανάλυση της τοπολογικής αναπαράστασης

Μετά την επιτυχή κατασκευή των εξατομικευμένων δικτύων για κάθε ασθενή, πραγματοποιήθηκε εξόρυξη πληροφορίας για την τοπολογία κάθε δικτύου. Αυτή η πληροφορία αναπαριστάται μέσω ενός τοπολογικού διανύσματος χαρακτηριστικών, το οποίο περιέχει της τιμές 5 τοπολογικών μέτρων. Τα χαρακτηριστικά τα οποία επιλέχθηκαν για την αναπαράσταση της τοπολογίας ενός γράφου είναι ο αριθμός κόμβων N, ο αριθμός των ακμών E, ο μέσος όρος ελάχιστου μονοπατιού, ο μέσος όρος συνδεσιμότητας των κόμβων, μέση κεντρικότητα βαθμού (Mean degree centrality), αριθμός κλικών γράφου (graph number of cliques) και ο αριθμός συνδεδεμένων στοιχείων (number of connected components).

Στο Σχήμα 26, παρουσιάζεται το γράφημα της PCA για την τοπολογική

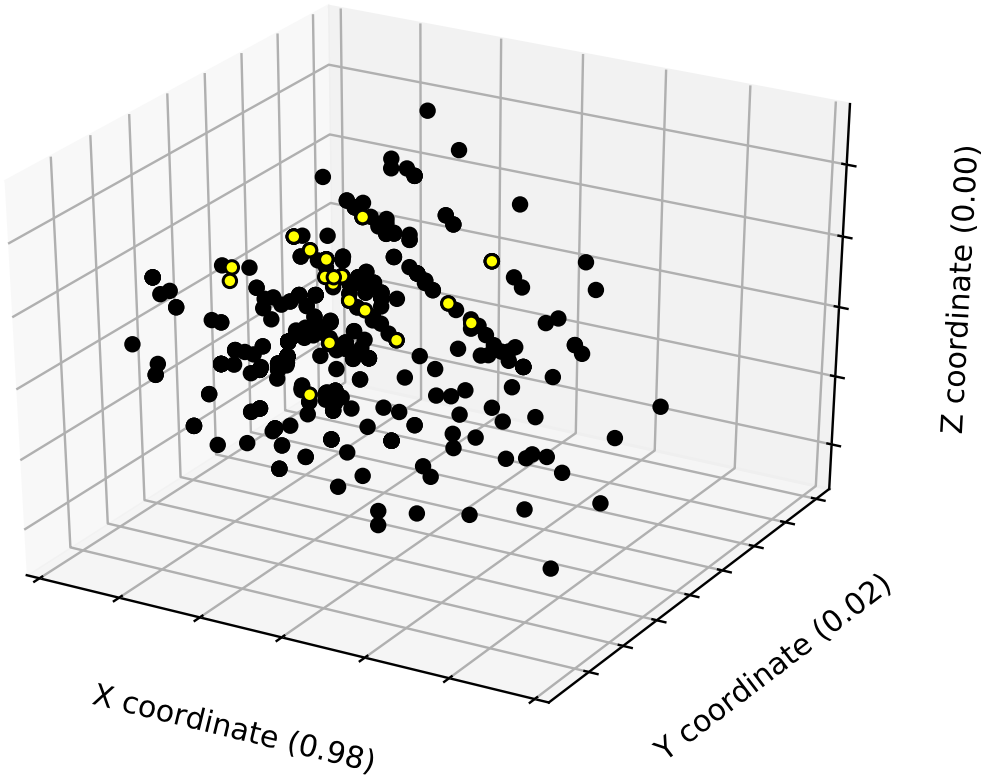


Σχήμα 25: Ανάλυση κύριων συνιστωσών (PCA) στην διανυσματική αναπαράσταση των multi- omic προφίλ στο σύνολο των δειγμάτων της ανάλυσης. Συγκεκριμένα, με κίτρινο χρώμα αναπαριστώνται τα δείγματα ελέγχου ένω με μαύρο τα καρκινικά δείγματα. Τα X, Y και Z coordinates αναφέρονται στις τρεις διαστάσεις των κύριων συσιστωσών και κάθε συνιστώσα χαρακτηρίζεται από ένα νούμερο που αφορά την % διακύμανση των προτύπων από τον κανονικό πειραματικό χώρο που μεταφέρεται στις κύριες συνιστώσες.

αναπαράσταση των multi- omic προφίλ στο σύνολο των δειγμάτων της ανάλυσης. Συγκεκριμένα, με κίτρινο χρώμα αναπαριστώνται τα δείγματα ελέγχου ένω με μαύρο τα καρκινικά δείγματα. Παρατηρούμε πως και σε αυτή την περίπτωση αναπαράστασης αναδεικνύεται η αποκλίνουσα συμπεριφορά των καρκινικών δειγμάτων έχοντας μεγαλύτερη διασπορά. Αντίθετα, τα δείγματα ελέγχου είναι πιο συμπυκνωμένα και βρίσκονται στο πυρήνα του καρκινικού "νέφους".

4.5 Κατασκευή δικτύων

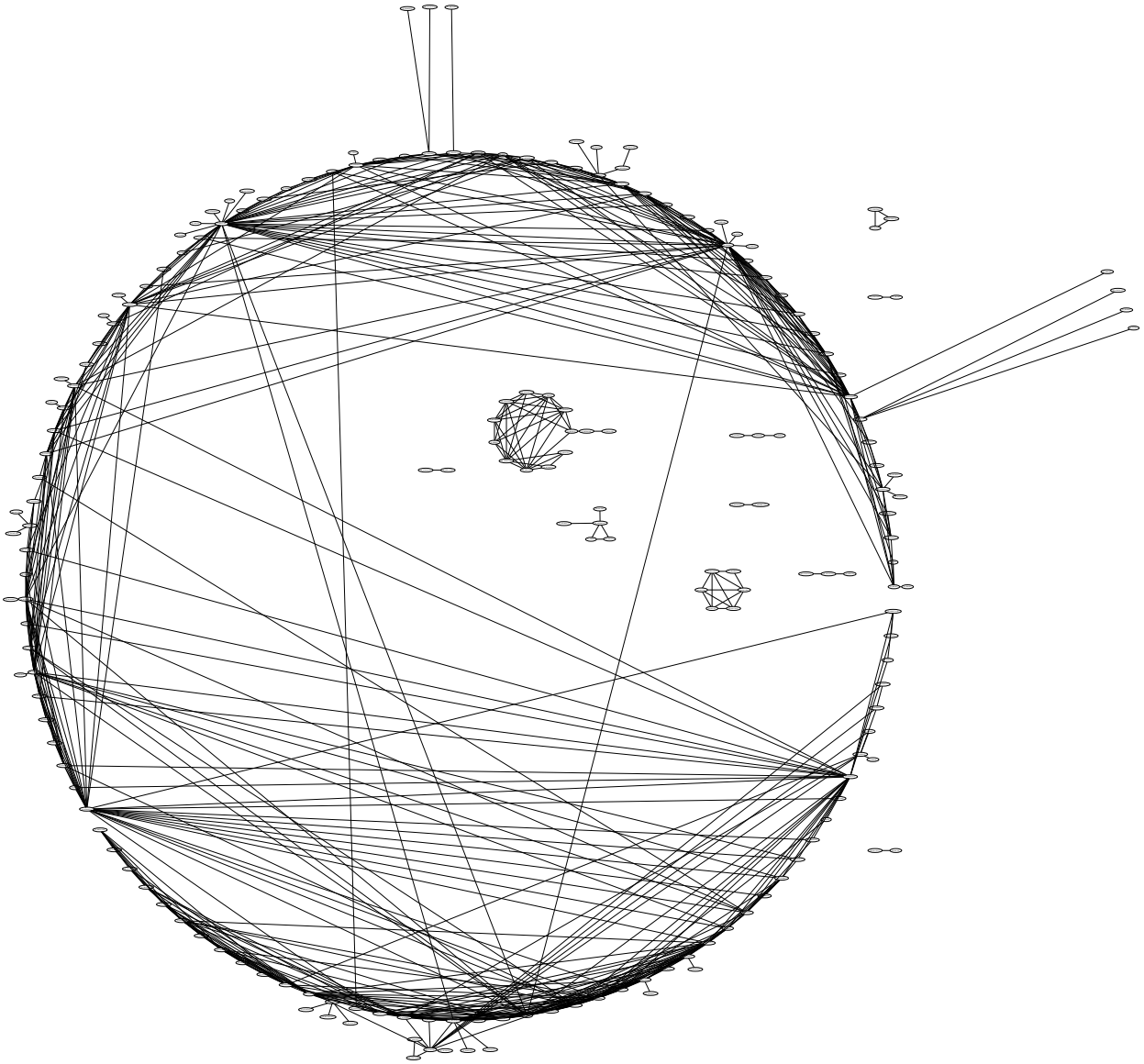
Όπως αναφέρθηκε σε παραπάνω ενότητα για την δικτυακή και τοπολογική αναπαράσταση ενός ασθενούς, πραγματοποιήθηκε ως πρώτο βήμα η κατασκευή ενός γενικού multi- omic διασυνδεδεμένου δικτύου για το ουροθηλιακό καρκίνωμα της ουροδόχου κύστης, λαμβάνοντας πληροφορία από



Σχήμα 26: Ανάλυση κύριων συνιστωσών (PCA) στην τοπολογική αναπαράσταση των multi- omic προφίλ στο σύνολο των δειγμάτων της ανάλυσης. Συγκεκριμένα, με κίτρινο χρώμα αναπαριστώνται τα δείγματα ελέγχου ένω με μαύρο τα καρκινικά δείγματα. Τα X, Y και Z coordinates αναφέρονται στις τρεις διαστάσεις των κύριων συσιστωσών και κάθε συνιστώσα χαρακτηρίζεται από ένα νούμερο που αφορά την % διακύμανση των προτύπων από τον κανονικό πειραματικό χώρο που μεταφέρονται στις κύριες συνιστώσες.

όλους τους ασθενείς.

Στο Σχήμα 27 παρουσιάζεται το ανακατασκευασμένο ολοκληρωμένο δίκτυο που αναπαριστά το multi-omic προφίλ κανονικοποιημένων δειγμάτων από ασθενείς που πάσχουν από μυοδιηθητικό ουροθηλιακό καρκίνωμα της ουροδόχου κύστης. Για την κατασκευή του δικτύου λήφθηκε πληροφορία από όλα τα δείγματα των ασθενών προκειμένου το ανακατασκευασμένο γενικό δίκτυο να αντικατοπτρίζει ολιστικά την φυσιολογία του ασθενούς. Συνολικά, το ολοκληρωμένο δίκτυο αποτελείται από 195 κόμβους και 555 ακμές. Παρατηρούμε πως πρόκειται για ένα εξαιρετικά πυκνό δίκτυο, με υψηλή συνδεσιμότητα και με αρκετούς δημοφιλείς κόμβους. Επιπλέον, το είδος των ”σημαντικών” κόμβων έτσι όπως επιλέχθηκε από την ανάλυση αφορούν μόνο μόρια miRNA και γονίδια των οποίων έχει εκτιμηθεί η συνολική τους έκφραση.



Σχήμα 27: Το γενικό ολοκληρωμένο δίκτυο που αναπαριστά τα multi-omic προφίλ κανονικοποιημένων δειγμάτων από ασθενείς που πάσχουν από μυοδιηθητικό ουροθηλιακό καρκίνωμα της ουροδόχου κύστης.

Λόγω του εξαιρετικά μεγάλου αριθμού χαρακτηριστικών που αναπαριστάται κάθε στιγμιότυπο (ο οποίος είναι ίσος με 73667 χαρακτηριστικά), συμπεριλήφθηκαν στην δικτυακή αναπαράσταση μόνο γονίδια και miRNA τα οποία είχαν σχετική αλλαγή πάνω από 65. Αυτό το κατώφλι ορίστηκε για λόγους περιορισμένης υπολογιστικής ισχύος.

Το δεύτερο είδος δικτύου που κατασκευάστηκε αφορούσε την αναπαράσταση ενός αυτή τη φορά ασθενούς. Για αυτό το σκοπό, χρησιμοποιήθηκε το γενικό ολοκληρωμένο δίκτυο ως χάρτης πάνω στον οποίο τοποθετούνται τα βάρη και οι τιμές των χαρακτηριστικών κάθε ασθενούς. Μετά την εφαρμογή της ενεργοποίησης διάδοσης αναδεικνύονται οι κόμβοι με τη μεγαλύτερη

συνδεσιμότητα και ταυτόχρονα οι σχέσεις μεταξύ των μορίων που χαρακτηρίζονται από υψηλή συσχέτιση (τόσο αρνητική όσο και θετική). Στη συνέχεια, εφαρμόσαμε ένα ιδιαίτερα αυστηρό κατώφλι το οποίο συγκρατεί το 25% των πιο σημαντικών κόμβων (δηλαδή των κόμβων με τα μεγαλύτερα σκορ μετά την εφαρμογή της ενεργοποίησης διάδοσης) σε κάθε δίκτυο ασθενούς. Με αυτόν τον τρόπο, αναδεικνύεται εξατομικευμένα πλέον, η τοπολογία του δικτύου κάθε ασθενούς.

Στα Σχήματα 28 και 29, εμφανίζονται ενδεικτικά δύο εξατομικευμένα δίκτυα δύο δειγμάτων ασθενών όπως έχουν τροποποιηθεί μετά την εφαρμογή της ενεργοποίησης διάδοσης και του κατωφλίου συγκρατώντας μόνο τους "σημαντικούς κόμβους". Το δίκτυο του σχήματος 29 αναπαριστά το δίκτυο δείγματος που έχει ληφθεί από τον πρωτογενή όγκο ασθενούς και το σχήμα 28 αναπαριστά το δίκτυο δείγματος που έχει ληφθεί από φυσιολογικό ιστό του ιδίου οργάνου (ουροδόχος κύστη) που βρίσκεται ο πρωτογενής όγκος, αντίστοιχα. Με μία γρήγορη ματιά, παρατηρούμε τις αλλαγές στα μη- συνδεδεμένα υποδίκτυα που περιέχουν και οι δύο γράφοι στο εσωτερικό τους. Και τα δύο δίκτυα χαρακτηρίζονται από υψηλή συνδεσιμότητα και αποτελούνται από 49 κόμβους.

4.6 Αποτελέσματα πειραμάτων ταξινόμησης

Προκειμένου να μελετηθεί και να συγκριθεί η απόδοση που προσφέρει κάθε τύπος αναπαράστασης σε ένα πρόβλημα ταξινόμησης, εφαρμόσαμε τους αλγορίθμους του k-κοντινότερου γείτονα (k-nearest neighbor) και των δέντρων απόφασης (decision trees).

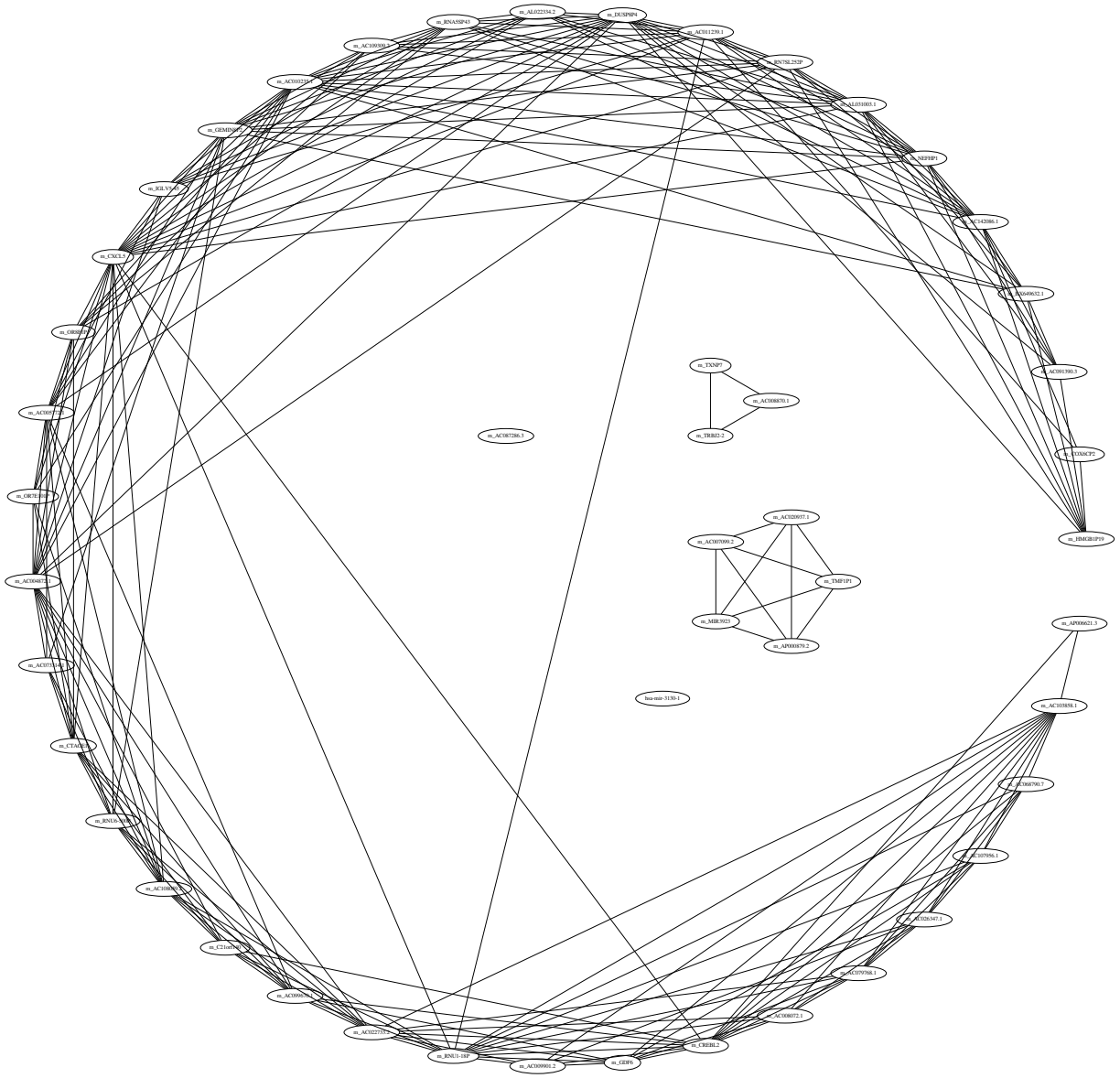
Σε όλα τα πειράματα ταξινόμησης, λόγω του εξαιρετικά μεγάλου αριθμού χαρακτηριστικών που έχει η διανυσματική αναπαράσταση, χρησιμοποιήθηκαν οι τρεις διαστάσεις συνδυασμένων χαρακτηριστικών που επέλεξε η PCA.

4.6.1 Αποτελέσματα πειραμάτων ταξινόμησης κατηγορίας δείγματος

Το πρώτο πείραμα ταξινόμησης που εκτελέσαμε αφορούσε την κατασκευή ενός μοντέλου που θα παράγει προγνώσεις σχετικά με την κατηγορία κάθε δείγματος (Καρκινικό δείγμα- δείγμα φυσιολογικού ιστού).

4.6.1.1 Αποτελέσματα πειραμάτων ταξινόμησης διανυσματικής αναπαράστασης

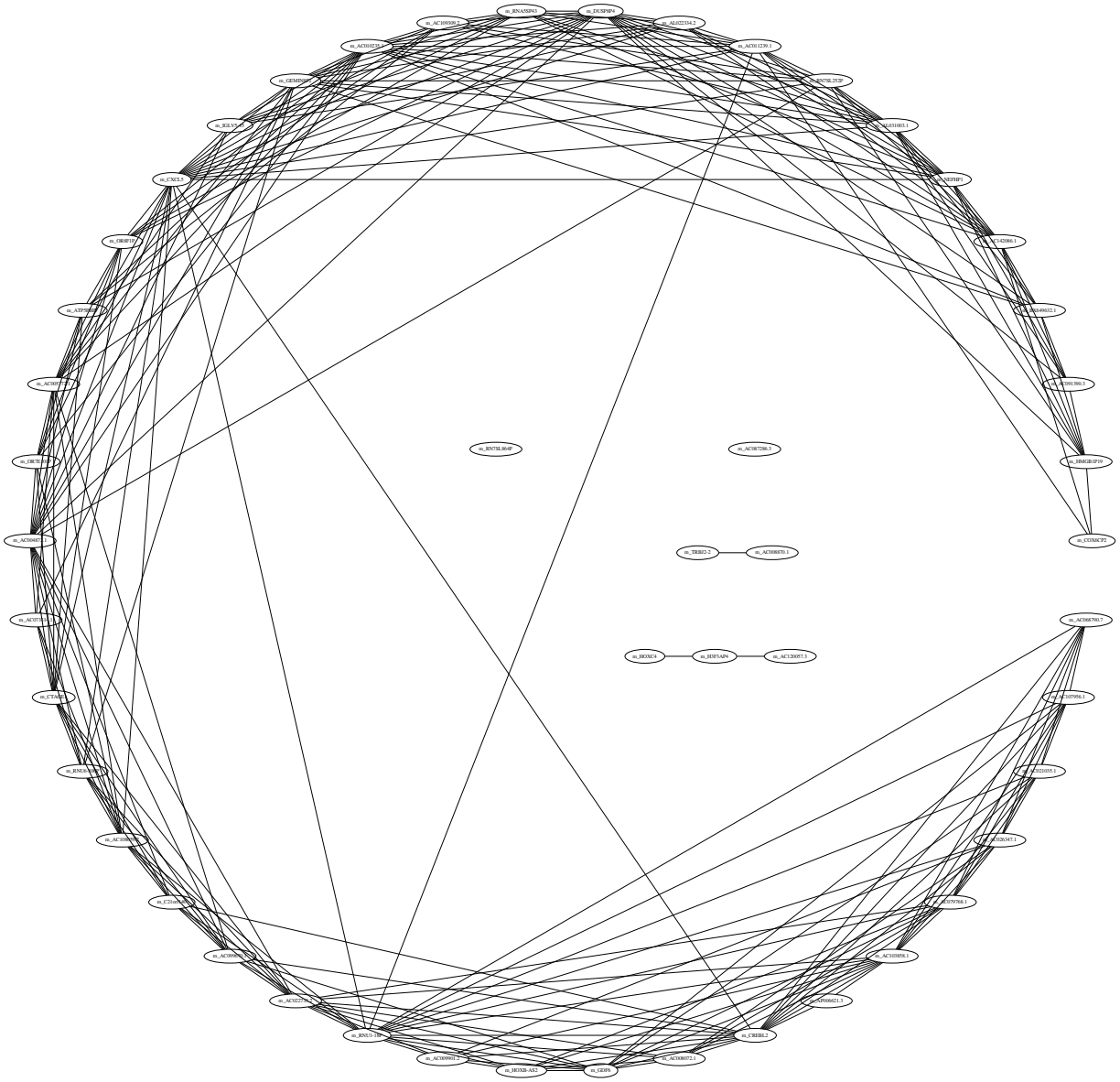
Σε αυτή την υποενότητα παρουσιάζονται τα αποτελέσματα της απόδοσης του μοντέλου ταξινόμησης που κατασκευάστηκε χρησιμοποιώντας τον αλγό-



Σχήμα 28: Το ολοκληρωμένο/διασυνδεδεμένο δίκτυο που αναπαριστά το multi-omic προφίλ ενός κανονικοποιημένου δείγματος υγιούς ιστού ουροδόχου κύστης από ασθενή που πάσχει από μυοδιηθητικό ουροθηλιακό καρκίνωμα της ουροδόχου κύστης.

ριθμο του k-κοντινότερου γείτονα (k-nearest neighbor) για την διανυσματική αναπαράσταση θέτοντας k=3. Οι βέλτιστες παράμετροι του μοντέλου υπολογίστηκαν μέσω του Stratified Shuffle Split με 20 επαναλήψεις. Στους Πίνακες 9, 10 και 11 παρουσιάζονται οι μετρικές εκτίμησης της απόδοσης F1 macro, F1 micro και Accuracy.

Ίδια αποτελέσματα απόδοσης προσέφερε το μοντέλο ταξινόμησης χρησιμοποιώντας τον αλγόριθμο των δέντρων απόφασης για την διανυσματική αναπαράσταση.



Σχήμα 29: Το ολοκληρωμένο/διαδυνδεδεμένο δίκτυο που αναπαριστά το multi-omic προφίλ ενός κανονικοποιημένου δείγματος πρωτογενούς όγκου από ιστό ουροδόχου κύστης από ασθενή που πάσχει από μυοδιηθητικό ουροθηλιακό καρκίνωμα της ουροδόχου κύστης.

F1 macro	st. dev.	SE
1	0	± 0

Πίνακας 9: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική F1 macro. Τα αποτελέσματα της απόδοσης του μοντέλου για κάθε επανάληψη του Stratified Shuffle Split [1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.1. 1. 1. 1. 1. 1. 1. 1. 1.]

F1 micro	st. dev.	SE
1	0	±0

Πίνακας 10: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική F1 micro. Τα αποτελέσματα της απόδοσης του μοντέλου για κάθε επανάληψη του Stratified Shuffle Split [1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.1. 1. 1. 1. 1. 1. 1. 1. 1.]

Accuracy	st. dev.	SE
1	0	±0

Πίνακας 11: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική Accuracy.

4.6.1.2 Αποτελέσματα πειραμάτων ταξινόμησης τοπολογικής αναπαράστασης

Σε αυτή την υποενότητα παρουσιάζονται τα αποτελέσματα της απόδοσης του μοντέλου ταξινόμησης που κατασκευάστηκε χρησιμοποιώντας τον αλγόριθμο του k-κοντινότερου γείτονα (k-nearest neighbor) για την τοπολογική αναπαράσταση, με k=3. Στους Πίνακες 12, 13 και 14 παρουσιάζονται οι μετρικές εκτίμησης της απόδοσης F1 macro, F1 micro και Accuracy.

F1 macro	st. dev.	SE
0.56	0.17	±0.5

Πίνακας 12: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική F1 macro. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.50 (st. dev. 0.12). Το Baseline performance εκφράζει ένα κατώφλι στο οποίο αντιστοιχούν τυχαίες εκτιμήσεις. Τα αποτελέσματα της απόδοσης του μοντέλου για κάθε επανάληψη του Stratified Shuffle Split [0.72619048 0.47727273 0.82170543 0.48837209 0.48837209 0.47619048 0.48837209 0.48837209 0.48837209 1. 0.48837209 0.47619048 0.48780488 0.475 0.48780488 1. 0.48780488 0.475 0.48780488 0.48780488].

F1 micro	st. dev.	SE
0.94	0.03	±0,01

Πίνακας 13: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική F1 micro. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.89 (st. dev. 0.05).Τα αποτελέσματα της απόδοσης του μοντέλου για κάθε επανάληψη του Stratified Shuffle Split [0.91304348 0.91304348 0.95652174 0.95454545 0.95454545 0.90909091 0.95454545 0.95454545 0.95454545 1. 0.95454545 0.90909091 0.95238095 0.9047619 0.95238095 1. 0.95238095 0.9047619 0.95238095 0.95238095].

Παρόμοια αποτελέσματα απόδοσης προσέφερε το μοντέλο ταξινόμησης χρησιμοποιώντας τον αλγόριθμο των δέντρων απόφασης για την τοπολογική

Accuracy	st. dev.	SE
0.94	0.03	±0,01

Πίνακας 14: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική Accuracy. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.91 (st. dev. 0.03).

αναπαράσταση.

4.6.2 Αποτελέσματα πειραμάτων ταξινόμησης σταδίου κακοήθειας (tumor stage)

Το δεύτερο πείραμα ταξινόμησης που εκτελέσαμε αφορούσε την κατασκευή ενός μοντέλου που θα παράγει προγνώσεις σχετικά με το στάδιο κακοήθειας (tumor stage) που έχει κατηγοριοποιηθεί κάθε πρωτογενής όγκος κάθε ασθενούς, σύμφωνα με τα κλινικά δεδομένα που συγκεντρώσαμε (Τύπου I - IV). Η κατανομή κάθε κατηγορίας ανά αριθμό δειγμάτων εμφανίζεται στον Πίνακα 15.

Βαθμός κακοήθειας	I	II	III	IV	Not Reported
Αριθμός ασθενών	2	131	141	136	2

Πίνακας 15: Αποτελέσματα αριθμού δειγμάτων ασθενών ανά στάδιο κακοήθειας (tumor stage) .

4.6.2.1 Αποτελέσματα πειραμάτων ταξινόμησης διανυσματικής αναπαράστασης

Σε αυτή την υποενότητα παρουσιάζονται τα αποτελέσματα της απόδοσης του μοντέλου ταξινόμησης που κατασκευάστηκε χρησιμοποιώντας τον αλγόριθμο των δέντρων άποφασης (Decision Tree) και το κ-κοντινότερο γείτονα (k-nearest neighbor) για την διανυσματική αναπαράσταση. Στους Πίνακες 16, 17 και 18 παρουσιάζονται οι μετρικές εκτίμησης της απόδοσης F1 macro, F1 micro και Accuracy.

F1 macro	st. dev.	SE
0.31	0.08	±0.07

Πίνακας 16: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική F1 macro. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.31 (st. dev. 0.12).

F1 micro	st. dev.	SE
0.35	0.08	±0.07

Πίνακας 17: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική F1 micro. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.35 (st. dev. 0.12).

Accuracy	st. dev.	SE
0.35	0.08	±0.07

Πίνακας 18: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική Accuracy. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.33 (st. dev. 0.09).

4.6.2.2 Αποτελέσματα πειραμάτων ταξινόμησης τοπολογικής αναπαράστασης

Παρουσιάζονται τα αποτελέσματα της απόδοσης των μοντέλων ταξινόμησης που κατασκευάστηκαν χρησιμοποιώντας τον αλγόριθμο του k-κοντινότερου γείτονα (k-nearest neighbor) και τον αλγόριθμο των δέντρων απόφασης για την τοπολογική αναπαράσταση. Ξεκινώντας με τα αποτελέσματα του μοντέλου που κατασκευάστηκε χρησιμοποιώντας τον αλγόριθμο του k-κοντινότερου γείτονα (k-nearest neighbor), k=3 . Στους Πίνακες 19, 20 και 21 παρουσιάζονται οι μετρικές εκτίμησης της απόδοσης F1 macro, F1 micro και Accuracy.

F1 macro	st. dev.	SE
0.35	0.13	±0.12

Πίνακας 19: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική F1 macro. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.32 (st. dev. 0.15). Το Baseline performance εκφράζει ένα κατώφλι στο οποίο αντιστοιχούν τυχαίες εκτιμήσεις.

F1 micro	st. dev.	SE
0.38	0.12	±0.11

Πίνακας 20: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική F1 micro. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.32 (st. dev. 0.07).

Accuracy	st. dev.	SE
0.38	0.12	±0.11

Πίνακας 21: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική Accuracy. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.34 (st. dev. 0.09).

Ακολούθως παρουσιάζονται τα αποτελέσματα για την εκτίμηση της απόδοσης του μοντέλου που κατασκευάστηκε χρησιμοποιώντας τον αλγόριθμο των δέντρων απόφασης (decision tree) παρουσιάζονται στους Πίνακες 22, 23 και 24.

F1 macro	st. dev.	SE
0.29	0.08	±0.07

Πίνακας 22: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική F1 macro. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.31 (st. dev. 0.01). Το Baseline performance εκφράζει ένα κατώφλι στο οποίο αντιστοιχούν τυχαίες εκτιμήσεις.

F1 micro	st. dev.	SE
0.34	0.07	±0,07

Πίνακας 23: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική F1 micro. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.34 (st. dev. 0.11).

Accuracy	st. dev.	SE
0.35	0.07	±0,07

Πίνακας 24: Αποτελέσματα απόδοσης του μοντέλου χρησιμοποιώντας τη μετρική Accuracy. Το Baseline performance σε αυτή τη περίπτωση είναι ίσο με 0.36 (st. dev. 0.10).

5 Ευρήματα και συζήτηση

Η υπόθεση που κάναμε στην παρούσα εργασία ήταν πως η τοπολογία ενός δικτύου από μόνη της μπορεί να μας προσφέρει στοχευμένη, εξατομικευμένη και επαρκή πληροφορία για ένα πρόβλημα ταξινόμησης ή πρόβλεψης τιμής, αγνοώντας εντελώς τα αρχικά μοριακά και κλινικά χαρακτηριστικά ενός ασθενούς.

Στην πορεία της προσπάθειας αποσαφήνισης λεπτομερειών υλοποίησης των ερευνητικών μας προσπαθειών υπήρξαν αρκετά εμπόδια τα οποία αντιμετωπίσαμε. Αρχικά, μελετήθηκε μία νέα και περίπλοκη βάση καρκινικών δεδομένων, η οποία ενσωματώνει τεράστια και ετερογενή πληροφορία καρκινικών δειγμάτων. Σε αυτό το πλαίσιο, σχεδιάστηκε επιτυχώς ο αλγόριθμος εξόρυξης και επεξεργασίας της επιθυμητής πληροφορίας από ένα σύνολο 1250 αρχείων από 435 δειγμάτων, παράλληλα ενσωματώνοντας την διαφορετική βιολογική πληροφορία και τις διαφορετικές προκλήσεις που διέθετε κάθε είδος δεδομένων.

Καταφέραμε να κατασκευάσουμε ένα βελτιστοποιημένο αλγόριθμο για την πολλαπλή αναπαράσταση, ενσωμάτωση καθώς και κανονικοποίηση multi-omics και κλινικών δεδομένων από δείγματα ασθενών που πάσχουν από μυοδιηθητικό ουροθηλιακό καρκίνωμα της ουροδόχου κύστης. Στον αλγόριθμο αυτό ενσωματώθηκε κατάλληλο κομμάτι υλοποίησης διερευνητικών στατιστικών αναλύσεων καθως΄και πειραμάτων για προβλήματα ταξινόμησης και πρόβλεψης τιμής.

Στη συνέχεια, επιλέξαμε να ανακατασκευάσουμε το ολοκληρωμένο δίκτυο για το μυοδιηθητικό καρκίνωμα της ουροδόχου κύστης το οποίο ενσωματώνει πληροφορία από το σύνολο των ασθενών. Καθώς ο συνολικός αριθμός των χαρακτηριστικών (73667) ήταν απαγορευτικός από την πλευρά της υπολογιστής ισχύος . Για αυτό το σκοπό, ορίσαμε ένα αρκετά αυστηρό (από βιολογικής άποψης) κατώφλι για την επιλογή των πιο σημαντικών κόμβων. Ειδικότερα, το κατώφλι συμπεριέλαβε μόνο τα γονίδια και microRNA που είχαν πάνω από 67 φορές διαφοροποιημένη σχετική αλλαγή (σε σχέση με το μέσο όρο προφίλ των δειγμάτων ελέγχου). Το αποτέλεσμα αφορούσε ένα εξαιρετικά πυκνό και μεγάλου μεγέθους δίκτυο. Συνολικά, αποτελείται από δύο από τα τρία είδη δεδομένων που αποτελούνται τα αρχικά διανύσματα των ασθενών, (δεν επιλέχθηκαν από την εφαρμογή του κατωφλίου γονίδια των οποίων έχουν εκτιμηθεί τα επίπεδα μεθυλίωσής τους).

Μετά την γενική κατασκευή του δικτύου για το ουροθηλιακό καρκίνωμα της ουροδόχου κύστης, σχεδιάσαμε και κατασκευάσαμε την εξατομικευμένη αναπαράσταση ενός ασθενούς αυτή τη φορά με τη μορφή δικτύου. Το αποτέλεσμα (Βλ. Σχήμα 28 και 29) τοπολογικά είναι διαφορετικό τόσο μεταξύ των δύο εξατομικευμένων γράφων όσο και μεταξύ των δύο εξατομικευμένων και

του γενικού ολοκληρωμένου γράφου.

Τα πειράματα ταξινόμησης κατηγορίας δείγματος (διαχωρισμός καρκινικού δείγματος από δείγμα ελέγχου) και σταδίων καρκίνου αποτέλεσαν προβλήματα με αρκετές προκλήσεις. Μία πρώτη πρόκληση ήταν πως και οι δύο κατηγορίες δειγμάτων είναι πιθανό να ομοιάζουν σε μεγάλο βαθμό στην φυσιολογία τους και επομένως στα multi-omic προφίλ τους, καθώς προέρχονται από τους ίδιους ασθενείς (η ομάδα ελέγχου είναι υποσύνολο του συνολικού αριθμού ασθενών), από το ίδιο όργανο και σε αρκετές περιπτώσεις σύμφωνα με το πρωτόκολλο της διαδικασίας συλλογής των δειγμάτων που ακολουθήθηκε στο πρόγραμμα TCGA, μπορεί να απέχουν στον ιστό μόνο 2cm. Αυτό το γεγονός φαίνεται και από το γράφημα της PCA (βλ.25) στο οποίο τα δείγματα ελέγχου ομαδοποιούνται μεταξύ τους σε ένα συμπαγή πυρήνα, ωστόσο τα καρκινικά δείγματα είναι αρκετά κοντά τους επιδεικνύοντας μία αποκλίνουσα συμπεριφορά από αυτά. Όπως αναφέρεται και από την βιβλιογραφία κύτταρα αρχικών σταδίων κακοήθειας (στάδιο I, II) καρκίνου έχουν αρκετές ομοιότητες με τα φυσιολογικά κύτταρα, επομένως είναι πιθανό ο πυρήνας που σχηματίζεται γύρω από τα δείγματα ελέγχου να αποτελούν δείγματα ασθενών αρχικών σταδίων. Επί προσθέτως, μία επιπλέον δυσκολία ήταν η έντονη ανισορροπία του μεγάλου αριθμού καρκινικών δειγμάτων (412) προς τον αριθμό των δειγμάτων ελέγχου (23). Όμοια, τα πειράματα ταξινόμησης για το στάδιο του καρκίνου εμφάνισαν προκλήσεις οι οποίες είναι σημαντικό να ληφθούν υπόψη στην ερμηνεία των αποτελεσμάτων. Όπως φαίνεται στον Πίνακα 15, υπάρχει εξαιρετική ανισορροπία στον αριθμό δειγμάτων κάθε κλάσης κάτι που κάνει ιδιαίτερα δύσκολο το πρόβλημα της ταξινόμησης. Για να αντιμετωπίσουμε αυτή τη δυσκολία, εφαρμόσαμε τόσο μετρικές οι οποίες είναι ευαίσθητες στο class imbalance όσο και μετρικές που δεν είναι, προκειμένου να είμαστε αντικειμενικοί. Επιπλέον, για αυτό το σκοπό χρησιμοποιήθηκε η μέθοδος "Stratified Shuffle Split" για την εκτίμηση του μοντέλου στην οποία εφαρμόζεται αναλογική αντιπροσώπευση κάθε κατηγορίας.

Αυτές οι προκλήσεις προσδίδουν μία επιπλέον σημασία στην απόδοση της διανυσματικής αναπαράστασης στο πρόβλημα ταξινόμησης. Η διανυσματική αναπαράσταση είχε την βέλτιστη απόδοση στο πρόβλημα ταξινόμησης κατηγορίας δείγματος (διαχωρισμός καρκινικού δείγματος από δείγμα ελέγχου), χρησιμοποιώντας τον αλγόριθμο του k-κοντινότερου γείτονα. Η σημασία αυτής της απόδοσης αναδεικνύεται καλύτερα από το γεγονός πως χρησιμοποιήθηκαν μόνο οι τρεις πιο αντιπροσωπευτικές διαστάσεις που επέλεξε η PCA προκειμένου το μοντέλο να μπορεί να διαχωρίζει τέλεια ένα δείγμα αν προέρχεται από πρωτογενή όγκο ενός ιστού ή από υγιή ιστό του ιδίου οργάνου. Επομένως, η ενσωμάτωση τριών ομικών δεδομένων σε αυτό το πρόβλημα πα-

ρέχει ολοκληρωμένη, στοχευμένη και επαρκή πληροφορία στη διανυσματική μορφή.

Οι τοπολογικοί πίνακες χαρακτηριστικών προσφέρουν μία καλή διαφοροποίηση των δύο ομάδων δειγμάτων σε σύγκριση με την διανυσματική αναπαράσταση, όπως φαίνεται από το γράφημα PCA του Σχήματος 26. Και σε αυτή την περίπτωση αναδεικνύεται η αποκλίνουσα συμπεριφορά των καρκινικών δειγμάτων, όπως επιβεβαιώνει το γράφημα PCA της διανυσματικής αναπαράστασης. Ωστόσο, ένας καλύτερος δείκτης για να συμπεράνουμε τα ιδιαίτερα χαρακτηριστικά που προσφέρουν οι τοπολογικοί πίνακες χαρακτηριστικών ήταν τα πειράματα ταξινόμησης. Παρατηρούμε πως και στα δύο πειράματα ταξινόμησης τα μοντέλα που κατασκευάστηκαν με βάση την τοπολογική αναπαράσταση τα πάνε καλύτερα από την τυχαιότητα όπως την ορίζει το baseline performance. Ένα σημαντικό αρχικό εύρημα που επιβεβαιώνει σε ένα βαθμό την αρχική μας υπόθεση είναι η καλύτερη απόδοση που προσφέρει το μοντέλο ταξινόμησης των σταδίων καρκίνου της τοπολογικής αναπαράστασης σε σχέση με την διανυσματική, χρησιμοποιώντας τον αλγόριθμο του k- κοντινότερου γείτονα.

Σε αυτό το σημείο αξίζει να αναφέρουμε πως στα περισσότερα πειράματα in silico ο αλγόριθμος των δέντρων απόφασης είχε παρόμοια απόδοση με τον αλγόριθμο του k-κοντινότερου γείτονα (k- nearest neighbor). Ωστόσο, εξαίρεση αποτελούν τα πειράματα ταξινόμησης του σταδίου καρκίνου χρησιμοποιώντας την τοπολογική αναπαράσταση, στα οποία ο αλγόριθμος του k-κοντινότερου γείτονα (k- nearest neighbor) είχε καλύτερη απόδοση από τα δέντρα απόφασης.

6 Ανακεφαλαίωση και μελλοντικοί στόχοι

Καταφέραμε να κατασκευάσουμε ένα βελτιστοποιημένο αλγόριθμο για την πολλαπλή αναπαράσταση multi- omics και κλινικών δεδομένων από δείγματα ασθενών που πάσχουν από ουροθηλιακό καρκίνωμα της ουροδόχου κύστης.

Ένας σημαντικός μελλοντικός στόχος είναι η μεγαλύτερη ανάδειξη της σημασίας των τοπολογικών χαρακτηριστικών. Για την επίτευξη αυτού του σκοπού, θα προσθέσουμε επιπλέον χαρακτηριστικά στα δίκτυα που κατασκευάσαμε, χαμηλώνοντας το πολύ αυστηρό κατώφλι που θέσαμε, προκειμένου να αναδειχθούν σχέσεις που τώρα δεν συμπεριλαμβάνονται και ενδεχομένως έχουν σημαντικό ρόλο στην τοπολογία. Ένας άλλος τρόπος είναι να αναδείξουμε την καθολική σημασία των τοπολογικών χαρακτηριστικών εφαρμόζοντας τον βελτιστοποιημένο αλγόριθμο σε άλλους τύπους καρκίνου, ή άλλους τύπους ασθενειών.

Επιπλέον, άμεσος στόχος πέρα από τα πειράματα in silico προβλημάτων ταξινόμησης που διενεργήθηκαν είναι η προσθήκη προβλημάτων εκτίμησης τιμής. Ειδικότερα, ένα πρώτο πρόβλημα εκτίμησης τιμής που επιθυμούμε να αντιμετωπίσουμε είναι η πρόγνωση του κλινικού πεδίου "Days to death", το οποίο έχει ήδη ενσωματωθεί στους πίνακες χαρακτηριστικών κάθε ασθενούς και ορίζει το χρονικό διάστημα ζωής ενός ασθενούς από την μέρα διάγνωσης καρκίνου μέχρι την ημέρα θανάτου του. Τα πειράματα εκτίμησης τιμής θα πραγματοποιηθούν χρησιμοποιώντας τη γραμμική παλινδρόμηση Lasso ("Least Absolute Shrinkage and Selection Operator"), μία μέθοδος με εξαιρετική δυναμική, η οποία συνήθως χρησιμοποιείται σε ιδιαίτερα μεγάλα σύνολα δεδομένων τα οποία εμφανίζουν πολυσυγγραμικότητα.

Ένας επιπλέον μελλοντικός στόχος της παρούσας εργασίας αποτελεί η ανάλυση μονοπατιών (pathway analysis) των διαφορικά εκφρασμένων μορίων mRNA, miRNA καθώς και των διαφορικά μεθυλιωμένων γονιδίων, προκειμένου να αναδειχθούν σημαντικά κυτταρικά μονοπάτια τα οποία επηρεάζονται στον καρκίνο της ουροδόχου κύστης.

Πρόσθετο σημαντικό βήμα είναι η βιβλιογραφική αναζήτηση των κεντρικών κόμβων και των στενά συνδεδεμένων υποδικτύων τόσο του γενικού ολοκληρωμένου δικτύου για το ουροθηλιακό καρκίνωμα της ουροδόχου κύστης, για την βαθύτερη κατανόηση της βιολογίας του συστήματος μελέτης μας.

Επί προσθέτως, η προσθήκη μίας επιπλέον ομάδας ελέγχου η οποία θα αφορά δείγματα ιστού ουροδόχου κύστης από υγιείς ανθρώπους μπορεί να προσφέρει μία νέα δυναμική στον πειραματικό σχεδιασμό μας. Ειδικότερα, θα δοθεί η ικανότητα μελέτης υπό νέο πρίσμα των επιδράσεων του καρκίνου στην φυσιολογία των κυττάρων της ουροδόχου κύστης, καθώς και να αποσαφηνιστούν οι διαφορές μεταξύ των δύο πλέον ομάδων ελέγχου.

Τέλος, ένας άλλος στόχος αποτελεί η εξόρυξη και η ενσωμάτωση επιπλέον

επιπέδων omics, όπως είναι η Ποικιλομορφία αριθμού αντιγράφων (Copy number variation, CNV) και οι Σημειακοί πολυμορφισμοί (Single nucleotide polymorphisms, SNP) για την περαιτέρω ανάδειξη της γενετικής ποικιλομορφίας σε ασθενείς που πάσχουν από ουροθηλιακό καρκίνωμα της ουροδόχου κύστης σε σύγκριση με δείγματα από υγιείς ανθρώπους.

7 Βιβλιογραφία

Barabási, Albert-László and Márton Pósfai (2016). "Network science". In: *Cambridge University Press*. URL: http://barabasi.com/networksciencebook/.

Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo (2011). "Network Medicine: A Network-based Approach to Human Disease". In: *Nat Rev Genet* 12.1, pp. 56–68. DOI: 10.1038/nrg2918.Network.

Breiman, Leo et al. (1984). *Classification And Regression Trees*. Vol. 1. 1, p. 366. ISBN: 0412048418. DOI: 10.1002/widm.8.

Chandrasekar, Thenappan and Christopher P Evans (2016). "Autophagy and urothelial carcinoma of the bladder: A review." In: *Investigative and clinical urology* 57 Suppl 1, pp. 89–97. ISSN: 2466-054X. DOI: 10.4111/icu.2016.57.S1.S89. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4910764&tool=pmcentrez&rendertype=abstract.

Collins, Allan M. and Elizabeth F. Loftus (1975). "A spreading-activation theory of semantic processing". In: *Psychological Review* 82.6, pp. 407–428. ISSN: 0033295X. DOI: 10.1037/0033-295X.82.6.407.

Diestel, Reinhard (2016). *Graph Theory*, p. 447. ISBN: 978-3-662-53621-6. DOI: 10.1007/978-3-662-53622-3.

Ding, Zijian, Songpeng Zu, and Jin Gu (2016). "Systems biology Evaluating the molecule-based prediction of clinical drug responses in cancer". In: *Systems Biology* 32.June, pp. 2891–2895. DOI: 10.1093/bioinformatics/btw344.

Domingos, Pedro (2012). "A few useful things to know about machine learning". In: *Communications of the ACM* 55.10, p. 78. ISSN: 00010782. DOI: 10.1145/2347736.2347755. URL: http://dl.acm.org/citation.cfm?doid=2347736.2347755.

Duda, Richard O., Peter E. Hart, and David G. Stork (2001). *Pattern Classification*. pattern1. DOI: 10.1007/BF01237942.

Erdoğan, Cihat, Zeyneb Kurt, and Banu Diri (2017). "Estimation of the proteomic cancer coexpression sub networks by using association estimators". In: *PLoS ONE* 12.11, pp. 1–19. ISSN: 19326203. DOI: 10.1371/journal.pone.0188016.

Erdös, P and A Rényi (1959). "On random graphs". In: *Publicationes Mathematicae* 6, pp. 290–297. ISSN: 00029947. DOI: 10.2307/1999405.

Garzon Ramiro Marcucci Guido, Croce Carlo M. (2010). "Targeting MicroRNAs in Cancer: Rationale, Strategies and Challenges". In: *Nat Rev Drug Discov.* 8.9, pp. 1385–1395. ISSN: 08966273. DOI: 10.2217/nnm.12.167.Gene.

Geschwind, D H and G Konopka (2009). "Neuroscience in the era of functional genomics and systems biology". In: *Nature* 461.7266, pp. 908–915. ISSN: 0028-0836. DOI: nature08537[pii]{\textbackslash}r10.1038/nature08537. URL: http://www.ncbi.nlm.nih.gov/pubmed/19829370.

Gov, Esra, Medi Kori, and Kazim Yalcin Arga (2017). "Multiomics Analysis of Tumor Microenvironment Reveals Gata2 and miRNA-124-3p as Potential Novel Biomarkers in Ovarian Cancer". In: *OMICS: A Journal of Integrative Biology* 21.10, omi.2017.0115. ISSN: 1557-8100. DOI: 10.1089/omi.2017.0115. URL: http://online.liebertpub.com/doi/10.1089/omi.2017.0115.

Granovetter, Mark (1973). "The Strength of Weak Ties". In: *American journal of sociology* 78.6, pp. 1360–1380. ISSN: 0002-9602. DOI: 10.1086/225469. URL: http://sfx.unimelb.edu.au/sfxlcl3?sid=google&auinit=MS&aulast=Granovetter&atitle=Thestrengthofweakties&title=Americanjournalofsociology&volume=78&issue=6&date=1973&spage=1360&issn=0002-9602%0Ahttps://sociology.stanford.edu/sites/default/files/publi.

Gundersen, Sveinung et al. (2017). "GSuite HyperBrowser : integrative analysis of dataset collections across the genome and epigenome". In: *GigaScience* April, pp. 1–12. DOI: 10.1093/gigascience/gix032.

Higdon, Roger et al. (2015). "The Promise of Multi-Omics and Clinical Data Integration to Identify and Target Personalized Healthcare Approaches in Autism Spectrum Disorders". In: *OMICS* 19.4, pp. 197–208. DOI: 10.1089/omi.2015.0020.

Hu, Bangli et al. (2017). "Identification of novel therapeutic target genes and pathway in pancreatic cancer by integrative analysis." In: *Medicine* 96.42, e8261. ISSN: 1536-5964 (Electronic). DOI: 10.1097/MD.0000000000008261.

Jiang, Xian Li, Emmanuel Martinez-Ledesma, and Faruck Morcos (2017). "Revealing protein networks and gene-drug connectivity in cancer from direct information". In: *Scientific Reports* 7.1, pp. 1–13. ISSN: 20452322. DOI: 10.1038/s41598-017-04001-3. URL: http://dx.doi.org/10.1038/s41598-017-04001-3.

Kitano, Hiroaki (2002). "Computational systems biology". In: *Nature* 420.6912, pp. 206–210. ISSN: 0028-0836. DOI: 10.1038/nature01254. URL: http://www.nature.com/nature/journal/v420/n6912/full/nature01254.html%5Cnhttp://www.nature.com/doifinder/10.1038/nature01254.

Knez, Virginia M. et al. (2014). "Clear cell urothelial carcinoma of the urinary bladder: A case report and review of the literature". In: *Journal of Medical Case Reports* 8.1, pp. 1–7. ISSN: 17521947. DOI: 10.1186/1752-1947-8-275.

Lander, E S et al. (2001). "Initial sequencing and analysis of the human genome". In: *Nature* 409.6822, pp. 860–921. ISSN: 0028-0836. DOI: 10.1038/35057062. URL: http://www.ncbi.nlm.nih.gov/pubmed/11237011%5Cnhttp://www.nature.com/nature/journal/v409/n6822/pdf/409860a0.pdf.

Lei, Chengwei and Jianhua Ruan (2013). "A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity".

In: *Bioinformatics* 29.3, pp. 355–364. ISSN: 13674803. DOI: 10.1093/bioinformatics/bts688.

Lerner, Seth P. et al. (2016). "Bladder cancer molecular taxonomy: Summary from a consensus meeting". In: *Bladder Cancer* 2.1, pp. 37–47. ISSN: 23523735. DOI: 10.3233/BLC-150037.

Loscalzo, Joseph and Albert-Laszlo Laszlo Barabasi (2011). "Systems Biology and the Future of Medicine". In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 3.6, pp. 619–627. ISSN: 19395094. DOI: 10.1002/wsbm.144.Systems.

McConkey, David J. and Choi1 Woonyoung (2018). "Molecular Subtypes of Bladder Cancer". In: *Oncology Reports*, pp. 1–7. ISSN: 15346269. DOI: 10.1007/s11912-018-0727-5. URL: https://doi.org/10.1007/s11912-018-0727-5.

Morgun, Andrey, Xiaoxi Dong, and Anatoly Yambartsev (2014). "Reverse enGENEering of regulatory networks from Big Data : a guide for a biologist Title : Reverse enGENEering of regulatory networks from Big Data : a guide for a". In: *Bioinformatics and Biology insights*, pp. 0–27. DOI: 10.4137/BBI.S12467.RECEIVED.

Peng, Chen, Ao Li, and Minghui Wang (2017). "Discovery of Bladder Cancer-related Genes Using Integrative Heterogeneous Network Modeling of Multi-omics Data". In: *Scientific Reports* November, pp. 1–11. DOI: 10.1038/s41598-017-15890-9.

Pineda, Silvia et al. (2015). "Integration Analysis of Three Omics Data Using Penalized Regression Methods: An Application to Bladder Cancer". In: *PLoS Genetics* 11.12, pp. 1–22. ISSN: 15537404. DOI: 10.1371/journal.pgen.1005689.

Platts, Adrian E et al. (2011). "Disease Progression And Solid Tumor Survival: A Transcriptome Decoherence Model". In: *Mol Cell Probes* 24.1, pp. 1–18. DOI: 10.1016/j.mcp.2009.09.005.Disease.

Rahman, Md. Saidur (2017). *Basic Graph Theory*. ISBN: 978-3-319-49474-6. DOI: 10.1007/978-3-319-49475-3. URL: http://link.springer.com/10.1007/978-3-319-49475-3.

Ramos, Marcel et al. (2017). "Software for the Integration of Multiomics Experiments in Bioconductor". In: *Cancer Research*, pp. 39–43. DOI: 10.1158/0008-5472.CAN-17-0344.

Robertson, A. Gordon et al. (2017). "Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer". In: *Cell* 171.3, pp. 540–556. ISSN: 10974172. DOI: 10.1016/j.cell.2017.09.007.

Rual, Jean François et al. (2005). "Towards a proteome-scale map of the human protein-protein interaction network". In: *Nature* 437.7062, pp. 1173–1178. ISSN: 00280836. DOI: 10.1038/nature04209.

Samuel, A L (1959). "Machine Learning Using the Game of Checkers". In: *IBM Journal of Research and Development* 44.1.2, pp. 206 –226. ISSN: 0018-8646. DOI: 10.1147/rd.441.0206. URL: https://pdfs.semanticscholar.org/e9e6/bb5f2a04ae30d8ecc9287f8b702eedd7b772.pdf%0Ahttps://pdfs.semanticscholar.org/e9e6/bb5f2a04ae30d8ecc9287f8b702eedd7b772.pdf%0Ahttp://ieeexplore.ieee.org/abstract/document/5389202/.

Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal (2018). "Cancer statistics, 2018". In: *CA: A Cancer Journal for Clinicians* 68.1, pp. 7–30. ISSN: 00079235. DOI: 10.3322/caac.21442. URL: http://doi.wiley.com/10.3322/caac.21442.

Silva, Tiago C. et al. (2016). "TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages". In: *F1000Research* 5.0, p. 1542. ISSN: 2046-1402. DOI: 10.12688/f1000research.8923.2. URL: https://f1000research.com/articles/5-1542/v2.

Somvanshi, Pramod Rajaram and K. V. Venkatesh (2014). "A conceptual review on systems biology in health and diseases: From biological networks to modern therapeutics". In: *Systems and Synthetic Biology* 8.1, pp. 99–116. ISSN: 18725325. DOI: 10.1007/s11693-013-9125-3.

Sun, Yan V. and Yi Juan Hu (2016). *Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases*. Vol. 93. Elsevier Ltd, pp. 148–190. ISBN: 9780128048016. DOI: 10.1016/bs.adgen.2015.11.004. URL: http://dx.doi.org/10.1016/bs.adgen.2015.11.004.

Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar (2006). "Classification : Basic Concepts , Decision Trees , and". In: *Introduction to Data Mining* 67.17, pp. 145–205. ISSN: 00224405. DOI: 10.1016/0022-4405(81)90007-8. URL: http://www-users.cs.umn.edu/~kumar/dmbook/index.php.

Tang, Binhua et al. (2012). "Cancer omics : From regulatory networks to clinical outcomes". In: *CANCER LETTERS*. ISSN: 0304-3835. DOI: 10.1016/j.canlet.2012.11.033. URL: http://dx.doi.org/10.1016/j.canlet.2012.11.033.

Theodoridis, Sergios (2015). *Machine Learning: A Bayesian and Optimization Perspective*, p. 1075. ISBN: 978-0-12-801722-7. DOI: 10.1017/CBO9781107415324.004. URL: https://books.google.com.sg/books?id=NHODBAAAQBAJ.

Tornow, Sabine and H. W. Mewes (2003). "Functional modules by relating protein interaction networks and gene expression". In: *Nucleic Acids Research* 31.21, pp. 6283–6289. ISSN: 03051048. DOI: 10.1093/nar/gkg838.

Tsatsaronis, George, Michalis Vazirgiannis, and Ion Androutsopoulos (2007). "Word sense disambiguation with spreading activation networks generated from thesauri". In: *IJCAI International Joint Conference on Artificial Intelligence*, pp. 1725–1730. ISSN: 10450823. DOI: 10.1145/1459352.1459355.

Vander, Arthur, James Sherman, and Dorothy Luciano (2001). "Human Physiology: The Mechanisms of Body Function -8th ed." In: *McGraw- Hill*.

Vidal, Marc (2009). "A unifying view of 21st century systems biology". In: *FEBS Letters* 583.24, pp. 3891–3894. ISSN: 00145793. DOI: 10.1016/j.febslet.2009.11.024. URL: http://dx.doi.org/10.1016/j.febslet.2009.11.024.

Wagner, Andreas (2000). "Robustness against mutations in genetic networks of yeast". In: 24.april, pp. 355–361.

Xylinas, Evanguelos et al. (2014). "Urine markers for detection and surveillance of bladder cancer". In: *Urologic Oncology: Seminars and Original Investigations* 32.3, pp. 222–229. ISSN: 18732496. DOI: 10.1016/j.urolonc.2013.06.001. URL: http://dx.doi.org/10.1016/j.urolonc.2013.06.001.

Yue, Zongliang et al. (2018). "PAGER 2.0: An update to the pathway, annotated-list and gene-signature electronic repository for Human Network Biology". In: *Nucleic Acids Research* 46.D1, pp. D668–D676. ISSN: 13624962. DOI: 10.1093/nar/gkx1040.

Zhang, Di et al. (2016). "Identification of ovarian cancer subtype-specific network modules and candidate drivers through an integrative genomics approach". In: *Oncotarget* 7.4, pp. 4298–4309. ISSN: 1949-2553. DOI: 10.18632/oncotarget.6774. URL: http://www.oncotarget.com/fulltext/6774.

Zhang, Hui et al. (2017). "Biomarker MicroRNAs for Diagnosis of Oral Squamous Cell Carcinoma Identified Based on Gene Expression Data and MicroRNA-mRNA Network Analysis". In: *Computational and mathematical methods in medicine* 2017, p. 9803018. ISSN: 17486718. DOI: 10.1155/2017/9803018.

Zhang, Qingyang, Joanna E Burdette, and Ji-ping Wang (2014). "Integrative network analysis of TCGA data for ovarian cancer". In: *BMC systems biology*, pp. 1–18. DOI: 10.1186/s12918-014-0136-9.

Zhang, Qingyang and Ji-Ping Wang (2015). "A Bayesian network approach for modeling mixed features in TCGA ovarian cancer data". In: *bioRxiv*, p. 033332. DOI: 10.1101/033332. URL: http://biorxiv.org/content/early/2016/03/26/033332.abstract.

Zhou, Rong and Eric A. Hansen (2006). "Breadth-first heuristic search". In: *Artificial Intelligence* 170.4-5, pp. 385–408. ISSN: 00043702. DOI: 10.1016/j.artint.2005.12.002.

Zhou, Wanding, Peter W. Laird, and Hui Shen (2017). "Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip

probes". In: *Nucleic acids research* 45.4, e22. ISSN: 13624962. DOI: 10.1093/nar/gkw967.

Zhu, Jialou et al. (2011). "A systematic analysis on DNA methylation and the expression of both mRNA and microRNA in bladder cancer". In: *PLoS ONE* 6.11. ISSN: 19326203. DOI: 10.1371/journal.pone.0028223.