Ατομική Διπλωματική Εργασία

Δημιουργία Πλατφόρμας για την Σύγκριση Αλγόριθμων για την Ανάλυση Συσχετίσεων μεταξύ Σημειακών Νουκλεοτιδικών Πολυμορφισμών και Φαινοτύπων

Μαρία Κυριάκου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάιος 2014

2013-2014

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Δημιουργία Πλατφόρμας για την Σύγκριση Αλγόριθμων για την Ανάλυση Συσχετίσεων μεταξύ Σημειακών Νουκλεοτιδικών Πολυμορφισμών και Φαινοτύπων

Μαρία Κυριάκου

Επιβλέπων Καθηγητής

Δρ. Κωνσταντίνος Παττίχη

Η Ατομική Διπλωματική Εργασία υποβλήθηκε προς μερική εκπλήρωση των απαιτήσεων απόκτησης του πτυχίου Πληροφορικής του Τμήματος Πληροφορικής του Πανεπιστημίου Κύπρου

Ευχαριστίες

Θα ήθελα να ευχαριστήσω όλους όσοι συνέβαλαν με οποιονδήποτε τρόπο για την εκπόνηση της διπλωματικής μου εργασίας. Ευχαριστώ θερμά το Δρ. Παττίχη Κωνσταντίνο για την επίβλεψη της διπλωματικής καθ' όλη την διάρκεια του έτους, και την ευκαιρία που μου έδωσε να μελετήσω ένα πολύ ενδιαφέρον αντικείμενο. Επίσης, είμαι ευγνώμων για την ιδιαίτερη βοήθεια του κύριου Αριστοδήμου Αρίστου, για την καθοδήγηση του και την εξαιρετική συνεργασία που είχαμε. Σε όλη την διάρκεια του έτους ήταν πάντα διαθέσιμος να ασχοληθεί με οποιανδήποτε απορία μου, και ήταν πρόθυμος για προσεκτική ανάγνωση της εργασίας μου, καθώς επίσης και για τις πολύτιμες υποδείξεις του.

Επιπλέον ευχαριστίες σε όλους τους καθηγητές μου στο Πανεπιστήμιο Κύπρου που με εμπλούτισαν με γνώσεις και με καθοδήγησαν στο πολύ ευρύ αντικείμενο της επιστήμης της πληροφορικής.

Θα ήθελα να ευχαριστήσω τους φίλους μου που ήταν το έρεισμα μου όλα τα χρόνια των σπουδών μου. Βέβαια, το μεγαλύτερο ευχαριστώ το οφείλω στην οικογένεια μου για την όλη συμπαράσταση τους και την ολόψυχη αγάπη του πατέρα και της μητέρας μου που ήταν η κινητήρια δύναμη να συνεχίζω σε κάθε εμπόδιο που αντιμετώπιζα. Αφιερώνω αύτη την διπλωματική εργασία στα αδέλφια μου, Κυριάκο, Ρουπίνα, Χρυσταλλένη, Παντελίτσα.

Περίληψη

Ο σκοπός της διπλωματικής εργασίας ήταν η δημιουργία πλατφόρμας για την σύγκριση αλγόριθμων για την ανάλυση συσχετίσεων μεταξύ σημειακών νουκλεοτιδικών πολυμορφισμών (SNP) και φαινοτύπων. Όπως υποστηρίζεται από επιστήμονες η συσχέτιση σημειακών νουκλεοτιδικών πολυμορφισμών οφείλεται για την ύπαρξη ή όχι μιας ασθένειας. Έτσι η συγκέντρωση αλγορίθμων που μπορούν να ανιχνεύσουν τους σημειακούς νουκλεοτιδικούς πολυμορφισμούς που ευθύνονται για μια ασθένεια θεωρείται αναγκαία.

Αρχικά, γίνεται αναφορά σε βασικούς όρους της μοριακής βιολογίας και ακολούθως γίνεται μελέτη των αλγορίθμων που επρόκειτο να χρησιμοποιήσουμε για την ανάλυση συσχετίσεων μεταξύ σημειακών νουκλεοτιδικών πολυμορφισμών και φαινοτύπων. Οι αλγόριθμοι που επιλεχθήκαν χωρίζονται σε 3 κατηγορίες, σε αλγορίθμους οι οποίοι κάνουν έλεγχο 1 προς 1 SNP (Chi Square-Fisher Exact), σε αλγορίθμους οι οποίοι ελέγχουν την αλληλεπίδραση 2 προς 2 SNPs (BOOST-Regression) και σε αλγορίθμους οι οποίοι ελέγχουν περισσότερα από δύο SNPs (MDR), ούτως ώστε να μελετήσουμε όλα τα επίπεδα συσχετισμού.

Στην συνέχεια υλοποιήθηκαν τα εργαλεία για κάθε αλγόριθμο, όπου και έγινε η εισαγωγή τους στην πλατφόρμα Galaxy.

Ακολούθως, πήραμε μερικά σύνολα πλασματικών δεδομένων που υπέστηκαν προεπεξεργασία και τρέξαμε τους αλγορίθμους για να παρατηρήσουμε αν επιτυγχάνουν, και αν ναι ποιο είναι το ποσοστό επιτυχίας τους. Έτσι τρέχοντας για κάθε σύνολο δεδομένων τους αλγορίθμους μπορέσαμε να τους συγκρίνουμε και να εξάγουμε συμπεράσματα για την ορθή ανταπόκριση των αλγορίθμων.

Περιεχόμενα

Ακρώνυμα

DNA Deoxyribonucleic acid

RNA Ribonucleic acid

SNP Single Nucleotide Polymorphism

APOE Apolipoprotein E

LD Linkage Disequilibrium

A Adenine

T Thymine

C Cytosine

G Guanine

RF Random Forest

MDR Multifactor Dimensionality Reduction

CV Cross Validation

U Uniform Distribution

NU Non Uniform Distribution

Κεφάλαιο 1

Εισαγωγή

1.1 Γενική Εισαγωγή για την διπλωματική

Στις μέρες μας γίνονται πολλές μελέτες για να ελεγχθεί η συσχέτιση μεταξύ γενετικών σημείων για την ύπαρξη ή όχι ασθενειών. Με σύμμαχο την πληροφορική, σήμερα μπορούμε να μελετήσουμε το DNA και ακολούθως τους σημειακούς νουκλεοτιδικούς πολυμορφισμούς. Σημειακός νουκλεοτιδικός πολυμορφισμός (SNP) είναι μια παραλλαγή που παρατηρείται στην αλληλουχία του DNA σε ένα νουκλεοτίδιο (Αδενίνη(A), Κυτοσίνη(C), Γουανίνη(G), Θυμίνη(T)). Η μελέτη των SNPs προσπαθεί να μελετήσει κατά πόσο επηρεάζεται ο φαινότυπος του ανθρώπου σε μια αλλαγή της τιμής ενός νουκλετιδίου.

Μέχρι σήμερα έχουν ανακαλυφθεί περισσότεροι από 3 εκατομμύρια απλοί νουκλεοτιδικοί πολυμορφισμοί. Αυτή η πληροφορία είναι πολύ σημαντική για την εύρεση ακολουθιών που σχετίζονται με ασθένειες εφόσον υποστηρίζεται ότι τα SNPs συσχετίζονται άμεσα με περιπλοκές και ανίατες ασθένειες. Οι επιστήμονες υποστηρίζουν ότι είναι πιο πιθανόν να ευθύνονται πέραν του ενός SNP για την ύπαρξη μια ασθένειας. Ωστόσο, υπάρχουν περιπτώσεις όπου κάποια ασθένεια οφείλετε και σε μόνο ένα SNP (*mendelian diseases*).

Επιπλέον, υποστηρίζεται ότι οι γενετικές αυτές παραλλαγές έχουν σαν επίπτωση και την ευαισθησία μας σε ασθένειες ή ακόμα την αντίδραση του οργανισμού μας στα φάρμακα. Εύστοχο είναι το παράδειγμα ότι μια παραλλαγή στην βάση APOE σχετίζεται, με μεγάλο κίνδυνο στην νόσο του Alzheimer.

Όπως έχει ήδη αναφερθεί η Πληροφορική αποτελεί το επιστημονικό υπόβαθρο επί του οποίου στηρίζονται μεγάλες έρευνες, αφού μπορούμε να τρέξουμε σχετικά γρήγορα διάφορες εφαρμογές και αλγορίθμους. Στηριζόμενοι σε αυτό, πλάνο μας είναι να συγκεντρώσουμε ποικίλα από αλγορίθμους, οι οποίοι θα αναλύουν την συσχέτιση μεταξύ SNPs και φαινοτύπων, για να τους συγκρίνουμε.

1.2 Στόχος διπλωματικής εργασίας

Όπως είναι ευρέως γνωστό η γενετική μας ποικιλομορφία αποτελείτε μόλις στο 0,1% του γενετικού μας υλικού, εφόσον οι άνθρωποι είμαστε όμοιοι με 99,9% του γονότυπου μας [1]. Η ανακάλυψη των παραλλαγών στην ακολουθία του DNA, προσφέρουν την ευκαιρία για να βρεθούν οι αιτίες διάφορων περίπλοκων ασθενειών, εφόσον ισχύει η υπόθεση ότι τα SNPs συσχετίζονται με τον φαινότυπο. Στόχος της διπλωματικής λοιπόν είναι η συγκέντρωση και η διαμόρφωση αλγορίθμων, για τον έλεγχο συσχέτισης κάθε SNP ανεξάρτητα από τα άλλα, με τον φαινότυπο αλλά και η εύρεση αλγορίθμων που ελέγχουν κατά πόσο η συσχέτιση οφείλεται σε περισσότερα του ενός SNP που αλληλεπιδρούν μεταξύ τους. Σήμερα είναι ευρέως διαδεδομένοι αλγόριθμοι οι οποίοι συγκρίνουν την συσχέτιση ενός SNP με την ύπαρξη ή όχι μιας ασθένειας. Τα τελευταία χρόνια, νέοι αλγόριθμοι έχουν προταθεί που ελέγχουν την αλληλεπίδραση μεταξύ 2 ή περισσότερων SNPs και τη συσχέτισή τους με την ύπαρξη ή όχι μιας ασθένειας. Επιπλέον στόχος αυτής της διπλωματικής είναι η ενσωμάτωση τέτοιων αλγορίθμων σε μια πλατφόρμα όπου οι συγκεκριμένοι αλγόριθμοι θα μπορούν να συγκριθούν μεταξύ τους. Η σύγκριση θα γίνεται με πλασματικά δεδομένα που θα δοθούν από άλλο πρόγραμμα. Επιπλέον, τα εργαλεία αυτά θα διατεθούν σε on-line πλατφόρμα ούτως ώστε να μπορούν να χρησιμοποιούνται από διάφορους χρήστες. Συνεπώς στο τέλος με βάση αυτά θα παρουσιαστούν οι εισηγήσεις για την επίτευξη μιας ολοκληρωμένης πλατφόρμας όπου θα μπορεί ο κάθε

χρήστης να εισάγει τον δικό του αλγόριθμο και θα μπορεί να τον συγκρίνει με τους υπόλοιπους που προϋπάρχουν στην πλατφόρμα.

1.3 Δομή διπλωματικής εργασίας

Λόγω του ότι θα χρησιμοποιούνται στην διπλωματική όροι οι οποίοι ανήκουν στην Μοριακή Βιολογία, το δεύτερο κεφάλαιο είναι αφιερωμένο για επεξήγηση των βασικών ορισμών και διαδικασιών. Στο τρίτο κεφάλαιο θα αναφέρουμε την μεθοδολογία μας. Εδώ θα συναντήσουμε την πλατφόρμα Galaxy η οποία θα μας προσφέρει την δυνατότητα να επιτύχουμε τον πιο πάνω στόχο. Στο τέταρτο κεφάλαιο θα αναλυθούν οι αλγόριθμοι που θα εισάγουμε στην πλατφόρμα μας, καθώς επίσης θα τους κατηγοριοποιήσουμε λόγω του ότι θα έχουμε αλγόριθμούς που κάνουν τον έλεγχο 1 προς 1 SNP, θα έχουμε αλγόριθμούς που κάνουν τον έλεγχο 2 προς 2 SNPs και αλγορίθμους με περισσότερες συγκρίσεις SNPs. Στο πέμπτο κεφάλαιο θα αναλύσουμε τα αποτελέσματα καθώς θα συγκρίνουμε την αποδοτικότητα κάθε αλγορίθμου σε σχέση με τους υπόλοιπους. Ωστόσο θα εισαχθούν μερικές βελτιστοποιήσεις. Στο τελευταίο κεφάλαιο, που είναι το έκτο στην σειρά θα καταγράφουν τα συμπεράσματα και η μελλοντικές βλέψης για το έργο αυτό.

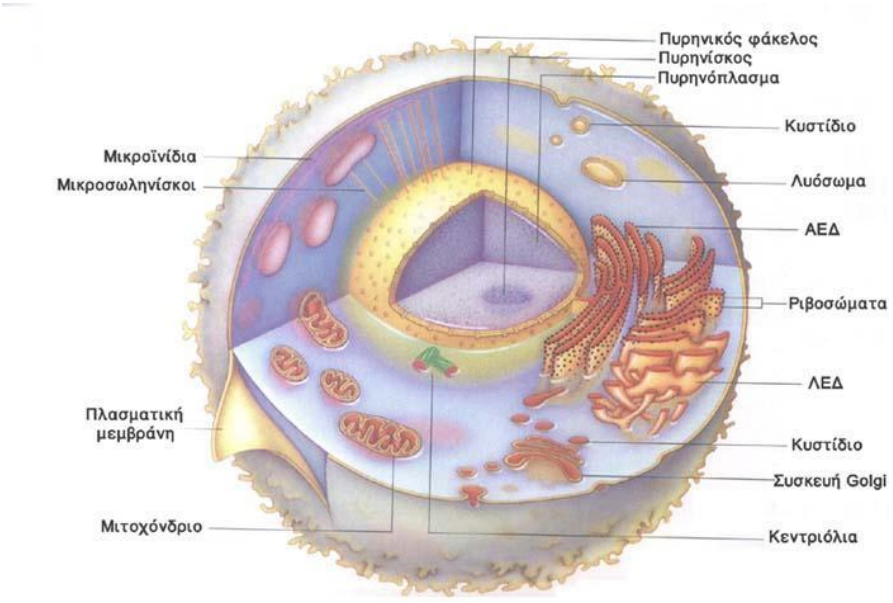
Κεφάλαιο 2

Βασικές Έννοιες Μοριακής Βιολογίας

2.1 Ανθρώπινο Κύτταρο

Όλοι οι ζωντανοί οργανισμοί αποτελούνται από κύτταρα, και συγκεκριμένα ο άνθρωπος αποτελείται από ευκαριωτικά κύτταρα. Το κύτταρο είναι μια εξαιρετικά ευμετάβλητη δομική μονάδα που μπορεί να πολλαπλασιάζεται καθώς και να προσαρμόζεται ώστε να

ανταποκρίνεται στις απαιτήσεις του οργανισμού. Περιγράφοντας την βασική κυτταρική δομή, παρατηρούμε ότι όλα τα κύτταρα περιβάλλονται από την πλασματική (κυτταρική) μεμβράνη καθώς περιέχουν τον πυρήνα και το κυτταρόπλασμα. Η πλασματική μεμβράνη αποτελείται από φωσφολιπίδια, τα οποία έχουν ως ιδιότητα μια υδρόφιλη κεφαλή και δύο υδρόφοβες ουρές. Όταν λοιπόν τα φωσφολιπίδια περιβάλλονται από νερό φτιάχνουν μια διπλοστιβάδα, και ο λόγος που το κάνουν αυτό είναι ότι αποτελούν δομικά συστατικά των μεμβρανών του κυττάρου καθώς δημιουργούν μια διαχωριστική μεμβράνη μεταξύ δύο διαφορετικών υδάτινων περιβαλλόντων. Το κυτταρόπλασμα καταλαμβάνει χώρο από την πλασματική μεμβράνη μέχρι την εξωτερική επιφάνεια του πυρήνα και αποτελείται από νερό, πρωτεΐνες, υδατάνθρακες και λίπη. Το σημαντικότερο ωστόσο οργανίδιο στο κύτταρο είναι ο πυρήνας. Περιβάλλεται από μια διπλή μεμβράνη που σχηματίζει τον πυρηνικό φάκελο. Το πυρηνόπλασμα αποτελείται κυρίως από DNA και πρωτεΐνες. Μέσα στο πυρηνόπλασμα υπάρχουν μια ή περισσότερες περιοχές, οι πυρινίσκοι, όπου εμπεριέχεται το Ριβονουκλεϊκό οξύ(RNA). Στην συνέχεια το RNA συνδυάζεται με πρωτεΐνες και σχηματίζονται τα ριβοσώματα έξω από τον πυρήνα [4].

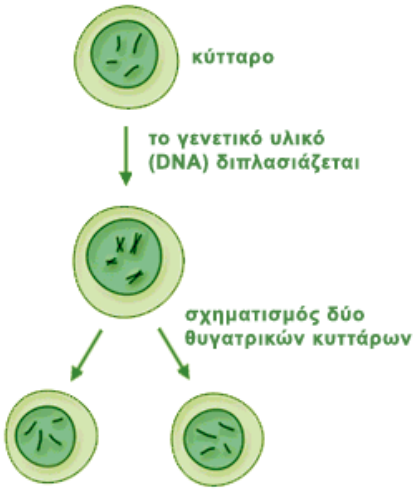


Εικόνα 2.1 Ευκαριωτικό κύτταρο παρθέν από ‘SCH’

(http://users.sch.gr/xristosvl/_private/BIOLOGIA/a_gymnasiou/eykar_kyttaro3.jpg)

2.2 Κυτταρική διαίρεση

Για την ανάπτυξη του πολυκύτταρου οργανισμού απαιτείται συνεχής αύξηση του αριθμού των κυττάρων. Τα νέα αυτά κύτταρα μπορούν να παραχθούν μέσω δύο τρόπων. Πρώτος τρόπος είναι με την διαδικασία της μίτωσης, όπου γίνεται πυρηνική διαίρεση των πατρικών κυττάρων και συνεπώς ακολουθείται από την διαίρεση του κυτταροπλάσματος. Ο δεύτερος τρόπος διαίρεσης του κυττάρου είναι η διαδικασία της μείωσης. Στον άνθρωπο η διαδικασία της μείωσης γίνεται μόνο στα γενετικά κύτταρα των όρχεων και των ωοθηκών. Σημαντικό στην διαδικασία της κυτταρικής διαίρεσης είναι η αντιγραφή και η αναδιοργάνωση του γενετικού υλικού του κυττάρου που βρίσκεται στον πυρήνα με τη μορφή χρωμοσωμάτων [4].



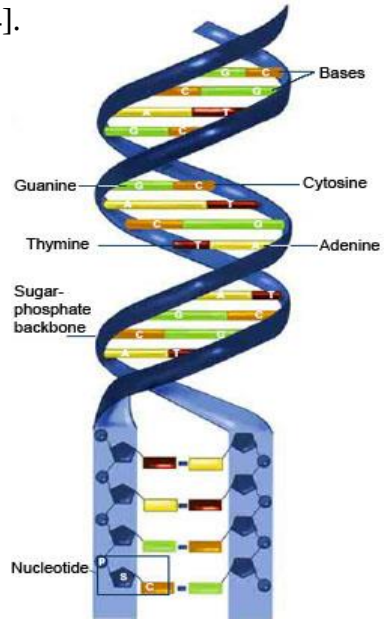
Εικόνα 2.2 Κυτταρική Διαίρεση και σχηματισμός θυγατρικών κυττάρων παρθέν από ‘ΚΠΕΚ’

(http://kpe-kastor.kas.sch.gr/leaf/photos/cell-division.gif)

2.3 Δεσοξυριβοζονουκλεϊνικό οξύ

Το δεσοξυριβοζονουκλεϊκό οξύ (DNA) ανήκει στην κατηγορία των νουκλεϊκών οξέων, και είναι το μόριο που ως συστατικό των χρωματοσωμάτων μεταφέρει τις γενετικές

πληροφορίες. Με την βοήθεια του άλλου νουκλεϊκού οξέος, που ονομάζεται ριβονουκλεϊκό οξύ (RNA), το DNA δίνει σε κάθε ζωντανό κύτταρο, οδηγίες για το ποίες πρωτεΐνες πρέπει να παράγει. Κάθε νουκλεοτιδική υπομονάδα του μορίου DNA περιλαμβάνει την πεντόζη δεσοξυριβόζη, την φωσφορική ομάδα και μια από τις αζωτούχες βάσεις, Αδενίνη, Κυτοσίνη, Θυμίνη, Γουανίνη. Ένα πολυνουκλεοτίδιο συνιστάτε από ένα σκελετό αποτελούμενο από σάκχαρα (Πεντόζες) και φωσφορικές ομάδες, πάνω στο οποίο συνδέονται και προεξέχουν οι αζωτούχες βάσεις. Αυτό που αποκαλούμε DNA αποτελείται από δύο πολυνουκλεοτιδικές αλυσίδες και δημιουργούν τη διπλή έλικα (όπως οι ίνες σχοινιού), που οι βάσεις βρίσκονται στο εσωτερικό του έλικα και ο σακχαροφωσφορικός κορμός έξω από αυτό. Τα δύο πολυνουκλεοτίδια παραμένουν ενωμένα με δεσμούς υδρογόνου μεταξύ των γειτονικών βάσεων [2]. Οι βάσεις είναι ταξινομημένες σε ζεύγη που παραμένουν σταθερές. Η Αδενίνη ενώνεται πάντα με την Θυμίνη (A-T) και η Γουανίνη με την Κυτοσίνη (G-C). Η δομή του DNA προτάθηκε από τους Watson και Crick το 1953, και μας εξηγείται πως εξυπηρετεί η δομή αυτή την αντιγραφή (τον διπλασιασμό) του DNA(2.4). Εκτός από την αντιγραφή, το μόριο του DNA έχει ακόμα ένα χαρακτηριστικό, να αποθηκεύει την πληροφορία που επιτρέπει στο κύτταρο να συνθέτει συγκεκριμένες πρωτεΐνες [4].

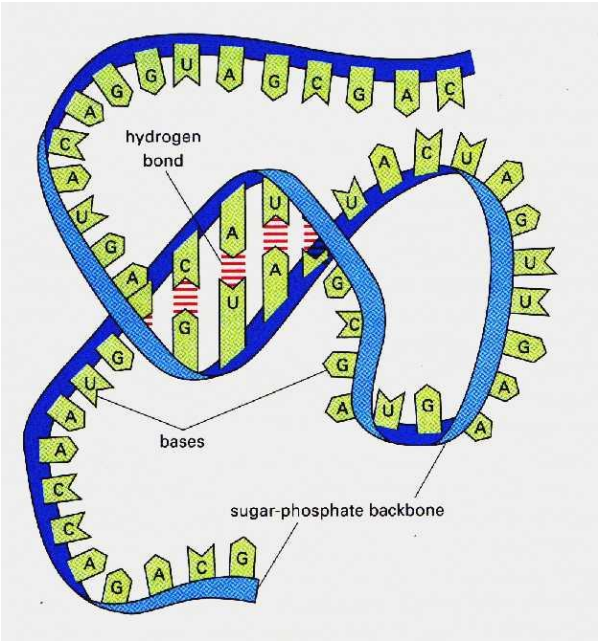


Εικόνα 2.3 Δομή DNA παρθέν από το 'NIH'

(http://publications.nigms.nih.gov/thenewgenetics/images/ch1_nucleotide.jpg)

2.4 Ριβονουκλεϊκό οξύ

Στην κατηγορία των νουκλεϊκών οξέων κατατάσσεται και το ριβονουκλεϊκό οξύ (RNA). Το μόριο του RNA αποτελείται από την πεντόζη ριβόζη και μία από τις αζωτούχες βάσης, Αδενίνη, Κυτοσίνη, Γουανίνη και Ουρακίλη. Αντίθετα από το DNA, αποτελείται από μόνο ένα κλώνο, παρόλο που μπορεί σε κάποια σημεία να σχηματίζει τμήματα διπλής έλικας, παρόμοια με αυτή του DNA. Τα τμήματα αυτά στηρίζονται από δεσμούς υδρογόνου μεταξύ των συμπληρωματικών ζευγών των βάσεων. Διάφοροι τύποι του RNA συμμετέχουν στην σύνθεση των πρωτεϊνών. [4]



Εικόνα 2.4 RNA παρθέν από 'UNIC'

(http://www.uic.edu/classes/phys/phys461/phys450/ANJUM04/RNA_sstrand.jpg)

2.5 Αντιγραφή

Όπως έχουμε ήδη προαναφερθεί η αντιγραφή είναι μια από τις ικανότητες του DNA. Το πρώτο βήμα της αντιγραφής είναι η ελίκαση δηλαδή ο διαχωρισμός των αλυσίδων του δίκλωνου DNA. Κάθε αρχική αλυσίδα μπορεί τώρα να αποτελέσει το δείγμα που θα καθορίσει την σειρά των νουκλεοτιδίων κατά μήκος της συμπληρωματικής αλυσίδας. Το

επόμενο βήμα είναι το DNA πολυμεράση, δηλαδή τα συμπληρωματικά νουκλεοτίδια στοιχίζονται και ενώνονται, σχηματίζοντας 2 νέους έλικες. Κάθε θυγατρικό DNA αποτελείται από ένα πολυνουκλεοτίδιο από τον πατρικό έλικα και από ένα πολυνουκλεοτίδιο από μια νέα αλυσίδα [4].

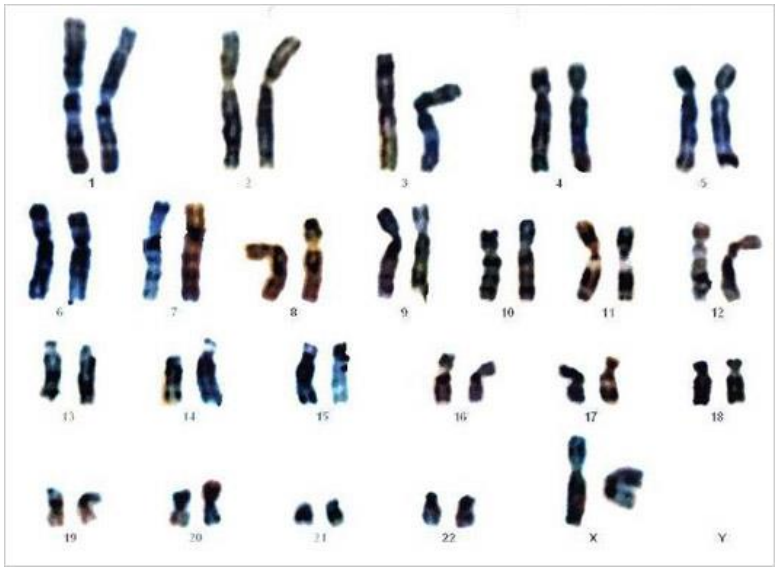
Εικόνα 2.5 Αντιγραφή DNA παρθέν από 'NIH'

(http://images.nigms.nih.gov/imageRepository/2543/DNA_Replication.jpg)

2.6 Χρωμοσώματα

Η ετυμολογία της λέξης χρωμόσωμα αποτελείται από τις λέξεις χρώμα και σώμα, και το όνομα αυτό οφείλεται λόγω της ιδιότητας του χρωμοσώματος να χρωματίζεται έντονα από τις ιδιαίτερες χρωστικές ουσίες. Τα χρωμοσώματα του κυττάρου που δεν βρίσκονται στη φάση της κυτταρικής διαίρεσης είναι πολύ λεπτές διασκορπισμένες δομές. Ακριβώς όμως πριν από την κυτταρική διαίρεση γίνονται πιο χοντρές, με σχήμα χ δηλαδή δύο βραχιόνων οι οποίοι είναι ενωμένοι σε μια περιοχή η οποία ονομάζεται κεντρομερίδιο, η οποία είναι μη χρωματισμένη. Τα χρωμοσώματα εντούτοις αποτελούνται από το DNA και διάφορες

πρωτεΐνες που είναι υπεύθυνες να κατασκευάσουν το DNA και να ελέγχουν τις λειτουργίες του. Τα σωματικά κύτταρα των οργανισμών περιέχουν μια χαρακτηριστική ομάδα χρωμοσωμάτων με αριθμό και σχήμα, ευρέως γνωστό από το Καρυόγραμμα (απεικόνιση όλων των χρωμοσωμάτων). Τα χρωμοσώματα ενός ζεύγους ονομάζονται ομόλογα, καθώς τα κύτταρα που έχουν ομόλογα χρωμοσώματα ονομάζονται διπλοειδείς. Αντιθέτως, απλοειδείς, είναι τα κύτταρα που περιλαμβάνουν το ένα χρωμόσωμα από το κάθε ζεύγος, όπως είναι οι γαμέτες (Χ κ Υ). Στον άνθρωπο υπάρχουν συνολικά 46 χρωμοσώματα διατεταγμένα σε 22 ομόλογα ζεύγη συν τα φυλετικά χρωμοσώματα Χ και Υ. Το 23^{ο} ζεύγος χρωμοσωμάτων καθορίζει το φύλο. Αυτό καθορίζεται από τον άντρα, λόγω του ότι όλα τα ωάρια περιέχουν ένα Χ χρωμόσωμα, ενώ τα σπερματοζωάρια μπορούν να περιέχουν είτε ένα Χ χρωμόσωμα είτε ένα Υ χρωμόσωμα. Στα θηλυκά το 23^{ο} ζεύγος περιέχει «ΧΧ» χρωμοσώματα, ενώ στα αρσενικά «ΧΥ». Αφού κάθε αρσενικό πρέπει να παρουσιάζει ένα Υ χρωμόσωμα, το οποίο μπορεί να κληρονομηθεί μόνο από τον πατέρα του, το αρσενικό Υ χρωμόσωμα αντιπροσωπεύει ένα αυθεντικό στοιχείο της πατρικής του κληρονομιάς [2].

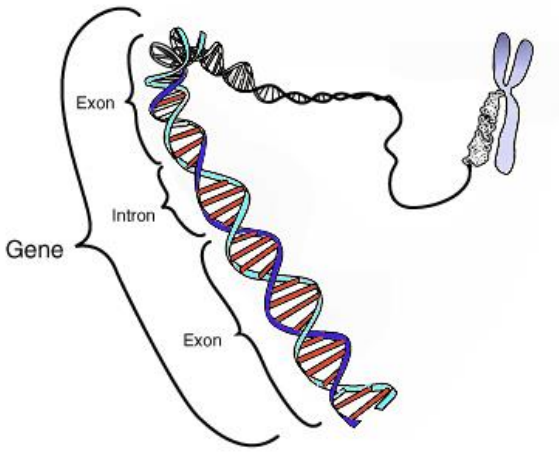


Εικόνα 2.6 Καρυόγραμμα, τα 23 ζεύγη ομόλογων χρωμοσωμάτων παρθέν από ‘FOPH’

(http://www.bag.admin.ch/php/modules/mediamanager/sendobject.php?lang=de&image=NHzLpZeg7t,lnp6I0NTU042l2Z6ln1acy4Zn4Z2qZpnO2Yuq2Z6gpJCGdn52e2ym162bpYbqjKaypeg-)

2.7 Γονίδια

Ένα γονίδιο αποτελείται από μια αλληλουχία βάσεων του DNA, βρίσκεται στα χρωμοσώματα, και αποτελεί τη βασικότερη μονάδα κληρονομικότητας. Έχει την ικανότητα να μεταβιβάζει μια πληροφορία από ένα κύτταρο και κατ' επέκταση μεταβιβάζεται από μια γενιά στην άλλη. Τα ευκαριωτικά γονίδια αποτελούνται από διάφορα μέρη και περιλαμβάνει τα ιντρόνια, τα οποία είναι περιοχές που δεν κωδικοποιούν κάτι και τα εξώνια, που είναι περιοχές οι οποίες κωδικοποιούν προϊόντα. Τα ιντρόνια βρίσκονται ανάμεσα στα εξώνια, όπως παρουσιάζεται στην Εικόνα 2.7. [5].



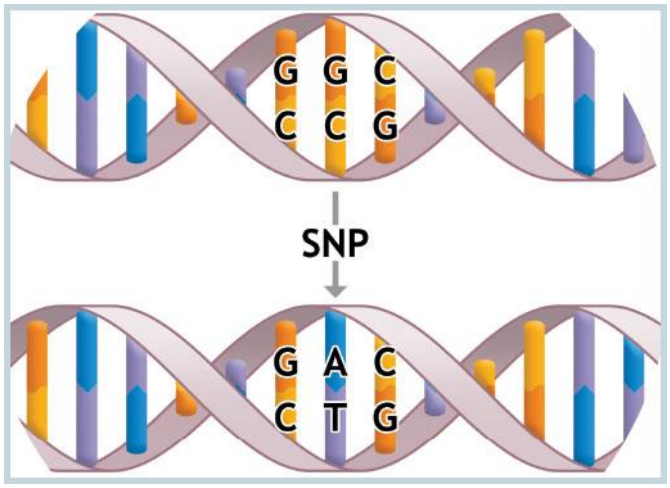
Εικόνα 2.7 Ιντρόνια ανάμεσα σε Εξώνια παρθέν από 'National Human Genome Research Institute'

(http://upload.wikimedia.org/wikipedia/commons/0/07/Gene.png)

2.8 Σημειακοί Νουκλεοτιδικοί Πολυμορφισμοί (SNP)

Όταν αναφερόμαστε στον όρο σημειακού νουκλεοτιδικού πολυμορφισμού αναφερόμαστε σε μια παραλλαγή στην αλληλουχία του DNA, όπου παρατηρείται όταν ένα νουκλεοτίδιο (A, C, T, G) διαφέρει στο γονιδίωμα με άλλα άτομα. Για παράδειγμα, παίρνουμε δυο αλληλουχίες DNA από το άτομο Β και το άτομο Γ, AAGCAA**A**TT και AAGCAA**C**TT αντίστοιχα. Οι δυο αυτές αλληλουχίες έχουν μόνη διαφορά στο $7^{ο}$ νουκλεοτίδιο, καθώς παρατηρούμε ότι η νουκλεοτιδική Αδενίνη έχει αντικατασταθεί με το νουκλεοτίδιο της

Κυτοσίνης. Μπορούμε εδώ να πούμε ότι έχουμε δύο αλληλόμορφα γονίδια. Αλληλόμορφα γονίδια είναι αυτά που ενεργούν για το ίδιο γνώρισμα αλλά με διαφορετικό τρόπο. Επιπλέον χαρακτηριστικό των SNPs είναι ότι η γονιδιωματική κατανομή των SNPs δεν είναι ομοιογενής, παρ' όλα αυτά παρατηρείται αυξημένη η εμφάνισή τους σε ιντρόνια παρά εξώνια. Σχεδόν όλα τα κοινά SNPs έχουν μόνο δύο αλληλόμορφα. Η πυκνότητα των SNPs μπορεί να προβλεφθεί από μικροδορυφόρους. Έρευνες έδειξαν ότι εμφανίζονται μια φορά σε 300 νουκλεοτίδια, και συμπερασματικά καταλήγουμε στο γεγονός ότι υπάρχουν 10 εκατομμύρια SNPs στο ανθρώπινο γονιδίωμα. Τα περισσότερα SNPs δεν έχουν επίπτωση στην ανάπτυξη του ατόμου ή στην υγεία του. Ωστόσο, σε μελέτες έχει παρατηρηθεί ότι οι γενετικές αυτές παραλλαγές διαδραματίζουν σημαντικό ρόλο για την μελέτη της ανθρώπινης υγείας. Οι ερευνητές, έχουν προσδιορίσει SNPs τα οποία μπορούν να βοηθήσουν έτσι ώστε να προβλεφθεί η ανταπόκριση του ατόμου σε ορισμένα φάρμακα και κατά κύριο λόγο την ευαισθησία σε πολυάριθμες και πολύπλοκες ασθένειες. Παράλληλα, με την μελέτη των SNPs μπορούμε να παρακολουθούμε την κληρονομικότητα των γονιδίων που μεταφέρει η νόσος [6].



Εικόνα 2.8 Σημειακός Νουκλεοτιδικός Πολυμορφισμός (SNP) παρθέν από 'IBBL'

(http://www.ibbl.lu/wp-content/uploads/2012/07/SNPs.jpg)

2.9 Linkage Disequilibrium (LD)

Ο όρος Ανισορροπίας Συνδέσμων (Linkage Disequilibrium - LD) αναφέρεται στη μη τυχαία συσχέτιση αλληλόμορφων γονιδίων σε δύο η περισσότερους γονικούς τόπους. Με άλλα λόγια LD είναι η κατάσταση στην οποία τα αλληλόμορφα διαφορετικών γονιδίων μπορεί να μην κληρονομούνται σε τυχαία ισορροπία, αλλά να κληρονομούνται πάντα (ή πολύ συχνά) μαζί. Το γεγονός αυτό, οδηγεί σε υψηλή στατιστική συσχέτιση τους [12]. Αυτό οδηγεί σε πιο συχνή κοινή κληρονόμηση κάποιων αλληλίων (allele – αλληλόμορφο γονίδιο) από εκείνη που θα αναμέναμε αν βασιζόμασταν σε κανόνα τυχαίου σχηματισμού απλότυπων. Μερικοί παράγοντες που επηρεάζουν το LD είναι η φυσική επιλογή, το ποσοστό ανασυνδιασμού, το ποσοστό μετάλλαξης, γενετικής παρέκκλισης.

Εάν όλοι οι πολυμορφισμοί ήταν ανεξάρτητοι και τυχαίοι μεταξύ τους, τότε θα χρειαζόταν να τους γονοτυπήσουμε όλους, για να βρούμε την γενική αιτία μιας νόσου. Εάν όμως βρούμε πως ένας πολυμορφισμός σχετίζεται με μια νόσο, τότε φτάνουμε σε δύο ενδεχόμενα: είτε ότι πολυμορφισμός αυτός είναι η πραγματική αιτία της νόσου, είτε πως ο πολυμορφισμός αυτός είναι σε ανισορροπία σύνδεσης με την αιτία της νόσου.[6]

2.10 Επίσταση

Η Επίσταση περιγράφει το φαινόμενο κατά το οποίο το φαινοτυπικό αποτέλεσμα ενός γονιδίου εξαρτάτε από την δράση ενός άλλου μη αλληλόμορφου γονιδίου. Δηλαδή μια παραλλαγή ή ένα αλλήλιο, εμποδίζει σε μια άλλη παραλλαγή, σε άλλη περιοχή να επιδράσει στο φαινότυπο του ατόμου [6].

Για παράδειγμα, ας αναφερθούμε στα σκυλιά Labradors, στα οποία ένα γονίδιο ευθύνεται για την εναπόθεση μελανίνης. Το κυρίαρχο αλληλόμορφο είναι το B που προκαλεί απόθεση μεγάλων ποσοτήτων μελανίνης και υπολειπόμενο είναι το b που προκαλεί λιγότερη απόθεση μελανίνης. Έτσι ένας σκύλος είναι μαύρος με BB ή Bb και καφέ bb. Έτσι το αλλήλιο B είναι επικρατέστερο του b.

Ένα άλλο γονίδιο ελέγχει κατά πόσον αποτίθεται μελανίνη ή όχι. Το κυρίαρχο είναι το E που επιτρέπει απόθεση μελανίνης και υπολειπόμενο το e. Έτσι ένας σκύλος με EE και Ee μπορεί να αποθέσει μελανίνη άρα θα είναι καφέ ή μαύρος, αλλά ένας ee σκύλος δεν μπορεί να εναποθέσει μελανίνη, και είναι κίτρινος. Το αλλήλιο E είναι επικρατέστερο του e [3].

	Genotype at locus E		
Genotype at locus B	**EE**	**Ee**	**Ee**
BB	Black	Black	Yellow
Bb	Black	Black	Yellow
bb	Brown	Brown	Yellow

Πίνακας 2.1 Παράδειγμα φαινοτύπου από διαφορετικούς γονότυπους σε δύο περιοχές που αλληλεπιδρούν επιστατικά, βάση του ορισμού της επίστασης που έδωσε ο Bateson

Παρατηρώντας τον πίνακα 2.1 βλέπουμε ότι εφόσον ο σκύλος έχει αλληλόμορφο E, τότε αν εμφανίζεται το B στον γονότυπο, το χρώμα του είναι μαύρο αλλιώς αν έχουμε bb το χρώμα του είναι καφέ. Αν στο γονότυπο του σκύλου υπάρχει το ee τότε το χρώμα του είναι κίτρινο ανεξάρτητα από την τιμή του γονότυπου B (BB ή Bb ή bb). Το ee χαρακτηρίζεται ως επιστατικό, αφού όταν υπάρχει, αναστέλλει την έκφραση του μαύρου ή καφέ σκυλιού [3].

Κεφάλαιο 3

Αλγόριθμοι Εύρεσης Συσχετίσεων μεταξύ SNPs και Φαινοτύπων

3.1 Chi-Square

Το Chi-Square τεστ είναι ένα πολύ δημοφιλές μη παραμετρικό τεστ. Αν και υπάρχουν πολλές παραλλαγές του θα αναφερθούμε στην πιο απλή μορφή του. Με προϋπόθεση το γεγονός ότι υπάρχει ένα σύνολο δεδομένων, στόχος είναι με την χρησιμοποίηση των δεδομένων να προσδιορισθεί μια αναλογία του πληθυσμού σε ποια κατηγορία ανήκει. Σημαντικό για την επίτευξη του στόχου είναι η null hypothesis, η οποία προσδιορίζει πως είτε οι αναλογίες δεν διαφέρουν σε άλλους πληθυσμούς είτε δεν υπάρχει κάποια συγκεκριμένη προτίμηση στις κατηγορίες. Συγκεκριμένα η μηδενική υπόθεση καθορίζει τον αριθμό που αναμένεται να ανήκει σε κάθε κατηγορία, με βάση τις μετρήσεις από τα δεδομένα εκπαίδευσης [7]. Για παράδειγμα, αν υπολογίζεται ο αριθμός των αρσενικών και

θηλυκών ανθρώπων που έχουν κατά πλάκα σκλήρυνση σε δύο χώρες, η null hypothesis θα είναι ότι η αναλογία των αρσενικών είναι η ίδια και στις 2 χώρες.

Ο στατιστικός δείκτης ελέγχου για την αξιολόγηση είναι το X^2. Οι μαθηματικές σχέσεις υπολογισμού είναι οι εξής:

$X^2 = \Sigma \frac{(fo-fe)^2}{fe}$, όπου το *fo* αναμενόμενη συχνότητα και *fe* *παρατηρούμενη συχνότητα*

Στα πλεονεκτήματα του αλγορίθμου, εντάσσεται το γεγονός ότι πρόκειται για ένα μη παραμετρικό τεστ, άρα δεν απαιτείται επαλήθευση της καταλληλότητας κάποιας υπόθεσης. Επιπλέον, τα δεδομένα που χρησιμοποιούνται είναι συχνότητες και όχι μετρήσεις.

Μειονέκτημα του Chi-Square τεστ είναι ότι δεν είναι αξιόπιστο για πολύ μικρές συχνότητες.

3.2 Fisher's Exact

Το Fisher's Exact τεστ έχει προταθεί από τους Diaconis and Sturmfels (1998) και έχει περίπου τις ίδιες λειτουργίες με το προαναφερθέν Chi-Square τεστ, παρά μόνο έχει την επιπλέον ιδιότητα ότι ανταποκρίνεται και σε μικρά δείγματα κάτι που το Chi-Square τεστ υστερεί. Στις περισσότερες χρήσεις του Fisher τεστ, περιλαμβάνεται ο πίνακας «contingency table», μεγέθους 2 X 2, ωστόσο στην αρχή της δοκιμής μπορούμε να έχουμε ένα πίνακα μεγέθους n x m. Για παράδειγμα, έχουμε ένα σύνολο από ανθρώπους, άντρες και γυναίκες και όπως βλέπουμε στον πιο κάτω πίνακα έχουμε τον αριθμό αυτών που πάσχουν από μια νόσο και αυτούς που είναι υγιείς. Με βάση την εξίσωση , η οποία υπολογίζει την πιθανότητα, βρίσκουμε την ακριβή πιθανότητα και του πιο κάτω πίνακα [13].

	Women	Men	**Total**
Disease	A	B	a+b
No Disease	C	d	c+d
Total	a+c	b+d	(a+b+c+d)=n

Πίνακας 3.1 Contingency Table με σύνολο θηλυκών και αρσενικών που πάσχουν από μια ασθένεια και αυτών που είναι υγιείς

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

Μια παραλλαγή του Fisher's Exact test είναι το Barnard's exact test, που είναι ένας εξίσου καλός αλγόριθμός για συσχέτιση δεδομένων [13].

3.3 Logistic Regression

Το Logistic Regression είναι ένα εργαλείο για έλεγχο (πρόβλεψη) κατά πόσο το φαινοτυπικό αποτέλεσμα ενός γονιδίου εξαρτάται από την δράση ενός άλλου μη αλληλόμορφου γονιδίου (επίσταση). Εφαρμόζεται όταν οι παράγοντες είναι διχοτόμοι, δηλαδή μπορούν να λάβουν μόνο δύο πιθανές τιμές. Το Logistic Regression χρησιμοποιείται για την εκτίμηση της πιθανότητας για επίσταση με βάση την ύπαρξη διαφόρων προσδιοριστών.

Εξίσωση του είναι:

$$\ln(p/(1p)) = \alpha + \beta xB + \gamma xC + ixBxC$$

όπου xB και xC αντιστοιχούν σε δυο SNPs στους γονότοπους B και C αντίστοιχα, i αντιπροσωπεύει ένα όρο αλληλεπίδρασης και β και γ είναι regression coefficients (συντελεστές) που αντιπροσωπεύουν την ισχύ του B και C. Στην περίπτωση που ο συντελεστής i σε ένα προσαρμοσμένο μοντέλο δεν είναι μηδέν τότε σημαίνει ότι τα δύο SNPs αλληλεπιδρούν μεταξύ τους. Στην αντίθετη περίπτωση όπου ο το i είναι μηδέν, τότε

συμπεραίνουμε ότι αν υπάρχει μια αλληλεπίδραση (μπορεί και όχι), τότε δεν μπορεί να είναι πολλαπλασιαστική.

Μειονέκτημα του Logistic Regression είναι το γεγονός πως δεν ανταποκρίνεται σε μη γραμμικά προβλήματα, και κατ' επέκταση σε μη γραμμικές αλληλεπιδράσεις γονιδίων [8].

3.4 Random Forest (RF)

Ο RF επεκτείνει την έννοια της σημασίας για ζεύγη αλληλεπιδράσεων των SNPs, καθώς προσπαθεί να συλλάβει κοινές επιδράσεις. Ο αλγόριθμος αυτός αποτελεί μια ειδική κατηγορία των συνδυαστικών μεθόδων ταξινόμησης, η οποία χρησιμοποιεί για ταξινομητές δένδρα απόφασης. Παρακάτω δίνεται μια σειρά από τα βήματα για την επεξήγηση της λειτουργίας των random forest:

1. Αρχικά, δημιουργούνται τα δένδρα απόφασης με την βοήθεια του bootstrap. Εφαρμόζοντας την τεχνική του bootstrap καταφέρνουμε να δημιουργήσουμε από το σύνολο των δεδομένων (data set), περισσότερα σύνολα δεδομένων εκπαίδευσης(training sets) όπου αυτό μας είναι απαραίτητο αφού έχουμε πολυάριθμα δένδρα απόφασης . Τα training sets καθορίζονται για κάθε δένδρο, ωστόσο υπάρχουν δεδομένα που δεν χρησιμοποιήθηκαν. Τα δεδομένα αυτά ονομάζονται out-of-bag data, όπου χρησιμοποιούνται για δεδομένα επαλήθευσης(testing data).
2. Κάθε δένδρο ταξινομεί με τυχαία επιλογή χαρακτηριστικών σε κάθε κόμβο. Παράλληλα το κάθε δένδρο ψηφίζει μια κλάση. Έτσι κάθε κλάση έχει έναν αριθμό ψήφων.
3. Η απόφαση καθορίζεται με την κλάση η οποία κέρδισε στην ψηφοφορία.

Από τα παραπάνω βήματα συμπεραίνουμε ότι παίζει σημαντικό ρόλο το πώς δημιουργούνται τα δέντρα απόφασης. Κάθε δένδρο λοιπόν, δημιουργείται βάση του αλγορίθμου:

i. Θέσε: N = αριθμός των objects ή cases του training set και

M = αριθμός των μεταβλητών-χαρακτηριστικών (input variables)

ii. Για κάθε δένδρο

επέλεξε από το data set, N cases στην τύχη με εναπόθεση (with replacement).

Αυτές οι N cases αποτελούν το training set για κάθε δένδρο.

iii. Επέλεξε τυχαία m από τις M μεταβλητές (όπου m<M) και το καλύτερο split που μπορεί να γίνει σε αυτές τις m μεταβλητές επιλέγεται για να χρησιμοποιηθεί στο node. – Η τιμή του m παραμένει σταθερή κατά τη διάρκεια κατασκευής όλου του δάσους και παίζει πολύ σημαντικό ρόλο για το classification error του δάσους.

iv. Κάθε δένδρο αναπτύσσεται στο μεγαλύτερο δυνατό βαθμό χωρίς να πραγματοποιείται

«κλάδεμα». [9]

Όπως έχει αναφερθεί, δημιουργούνται πολλά δένδρα απόφασης, με αποτέλεσμα να αποφεύγετε η περίπτωση για υπερεκπαίδευση του δικτύου. Ωστόσο κάθε δένδρο δεν επιδέχεται κλάδεμα και καλλιεργείται σε πλήρες βάθος, δηλαδή αναπτύσσεται στο μεγαλύτερο βαθμό που μπορεί να αναπτυχθεί.

Τα σημαντικά SNPs βρίσκονται πιο ψηλά στο δένδρο. Αξιοσημείωτη είναι η έννοια του permutation testing, όπου είναι ένας τρόπος για την αναγνώριση των σημαντικών SNP. Η λειτουργία του είναι ως εξής: Αντικαθιστούμε τιμές ενός SNP με τυχαίες τιμές. Αν το αποτέλεσμα αλλάζει τότε συμπεραίνουμε ότι το εν λόγω SNP είναι σημαντικό.

Τα πλεονεκτήματα του αλγορίθμου αυτού τον καθιστούν ιδιαίτερα σημαντικό για τον έλεγχο αλληλεπίδρασης των SNP. Ένα από αυτά που έχει ήδη αναφερθεί, είναι το σφάλμα γενίκευσης το οποίο είναι πολύ περιορισμένο άρα αποφεύγεται άμεσα η περίπτωση υπερεκπαίδευσης του δικτύου. Τέλος, με την τυχαία επιλογή ταξινόμησης των χαρακτηριστικών δίνει στον αλγόριθμο RF την ιδιότητα αμεροληψίας.

Ένα μειονέκτημα του αλγορίθμου είναι η απαίτηση του σε μεγάλες χρονικές περιόδους για την εκτέλεση του καθώς και computer resources.

Αν και πρόκειται για μια νέα τεχνική έχουν παρατηρηθεί πολλές παραλλαγές της, όπως είναι η Random Jungle (RJ), SNPInterForest κλπ. Επιπλέον, η τεχνική του RF χρησιμοποιήθηκε και σε άλλες εφαρμογές εκτός της βιοπληροφορικής, όπως είναι η αστρονομία.

3.5 Boolean Operation based Screening and Testing (BOOST)

Ο αλγόριθμος BOOST είναι ένας αλγόριθμος ο οποίος είναι μια παραλλαγή του logistic regression. Πρόκειται ωστόσο για μια απλή αλλά ισχυρή μέθοδο που επιτρέπει την εξέταση όλων των κατά ζεύγη αλληλεπιδράσεων. Στην μέθοδο αυτή γίνεται αναπαράσταση των γονοτυπικών δεδομένων σε Boolean μορφή η οποία όχι μόνο προσφέρει χώρο στην μνήμη αλλά και αποτελεσματικότητα στον χρόνο εκτέλεσης του αλγορίθμου, εφόσον χρησιμοποιούνται μόνο Boolean παράμετροι.

Η εξίσωση που στηρίζεται κατά πολύ στον αλγόριθμο Logistic Regression με κύρια χαρακτηριστικά είναι:

$$\log\frac{P(Y=1|Xp=i,Xq=j)}{P(Y=2|Xp=i,Xq=j)} = \beta o + \beta\iota^{Xp} + \beta j^{Xq}$$

Η πλήρης εξίσωση είναι:

$$\log\frac{P(Y=1|Xp=i,Xq=j)}{P(Y=2|Xp=i,Xq=j)} = \beta o + \beta\iota^{Xp} + \beta j^{Xq} + \beta ij^{XpXq}$$

Οι εκθέτες Xp, Xq στα $\beta\iota^{Xp}$ βj^{Xq} αντιπροσωπεύουν δείκτες και όχι εκθέτες. Ο όρος $\beta\iota^{Xp}$ αντιπροσωπεύει την κατηγορία i με συντελεστή Xp. Αυτή η εκπροσώπηση εκτείνεται και στα βj^{Xq} βij^{XpXq}[10].

3.6 Multifactor Dimensionality Reduction (MDR)

Η MDR είναι μια μέθοδος ευρέως διαδεδομένη, η οποία εφαρμόζεται για την αλληλεπίδραση γονιδίου-γονιδίου και γονιδίου-περιβάλλοντος, η οποία δημιουργεί μεγάλες προκλήσεις για πολλές από τις σημερινές στατιστικές προσεγγίσεις. Τα βήματα του αλγορίθμου έχουν ως εξής:

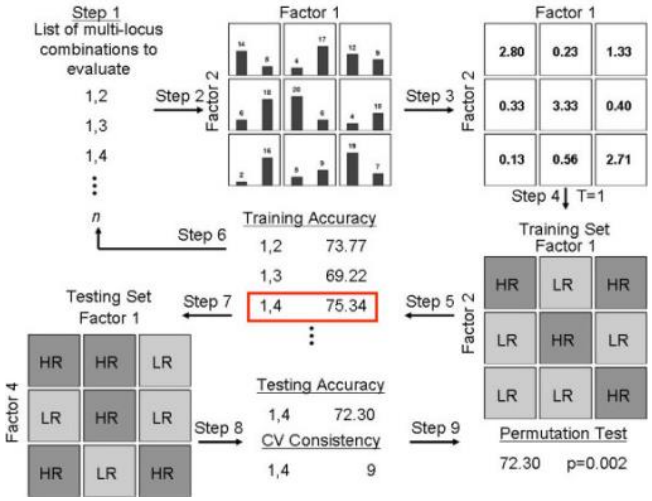
1. Επιλογή των γενετικών/περιβαλλοντικών παραγόντων από τα δεδομένα.
2. Δημιουργία πίνακα n διαστάσεων, ο οποίος περιέχει σε κάθε κελί την αναλογία των cases και controls για κάθε δυνατό συνδυασμό των πιο πάνω παραγόντων.
3. Κάθε κελί χαρακτηρίζεται υψηλού και χαμηλού ρίσκου. Ο χαρακτηρισμός για υψηλό ρίσκο σε ένα συγκεκριμένο κελί γίνεται με την προϋπόθεση ότι στην αναλογία ελέγχου έχει ξεπεραστεί το κατώφλι που έχει ήδη οριστεί.
4. Με έρεισμα το cross validation γίνεται η πρόβλεψη σφάλματος κάθε μοντέλου.

Χρησιμοποιώντας, το 10 fold cross validation επιτυγχάνουμε να διαχωρίζουμε τα δεδομένα μας σε δεδομένα εκπαίδευσης (90%) και δεδομένα επαλήθευσης (10%). Εκπαιδεύοντας το σύστημά αρχικά και μετά επαληθεύοντάς το, υπολογίζουμε την ακρίβεια εκπαίδευσης για κάθε μοντέλο. Με την ψηφοφορία που γίνεται σε σχέση με την πιο πάνω ακρίβεια, επιλέγεται το καλύτερο μοντέλο. Επιπλέον, γίνεται χρήση και του permutation testing στο οποίο έχει γίνει αναφορά στο 3.1.

Εκτός από τα προφανή συμπεράσματα που μπορεί να μας δώσει η MDR, στα θετικά της τεχνικής αυτής εντάσσεται το γεγονός ότι πρόκειται για αλγόριθμο που δεν στηρίζεται σε κάποιο συγκεκριμένο μοντέλο. Αρνητικά της μεθόδου MDR είναι ότι χρειάζεται μεγάλο υπολογιστικό κόστος όσο αφορά χρόνο και υπολογιστικούς πόρους, αν έχει να αξιολογήσει πολλά SNPs μαζί. Αυτό γίνεται λόγω της πολυπλοκότητας στην αναζήτηση των N-SNP ενώσεων. Επιπρόσθετα, υπάρχει η πιθανότητα υπερεκπαίδευσης.

Η MDR τεχνική σε μια συνέλιξη με μια πιο παραδοσιακή τεχνική, pedigree-disequilibrium test (PDT), μπορεί να αξιολογήσει σύνθετα μοντέλα με δεδομένα που προήλθαν από συγγένεια εξ αίματος. Μια εξίσου σημαντική παραλλαγή της MDR είναι η G-MDR

(Generalized MDR). Η G-MDR επιτρέπει τον έλεγχο συσχετίσεων και σε μη κατηγοριοποιημένους φαινοτύπους [8], [11].



Εικόνα 3.1 MDR διαδικασία

(http://openi.nlm.nih.gov/detailedresult.php?img=2837863_1471-2350-11-37-1&req=4)

Κεφάλαιο 4

Μεθοδολογία

4.1 Πλατφόρμα Galaxy

Η πλατφόρμα Galaxy είναι μια ανοιχτή web-based πλατφόρμα για ενοποίηση δεδομένων που χρησιμοποιούνται σε βιοϊατρικές μελέτες. Πρόκειται για μια πλατφόρμα που επιτρέπει στους χρήστες να συλλέγουν και να χειρίζονται τα δεδομένα από τους υπάρχοντες πόρους σε μία ποικιλία τρόπων. Το πιο βασικό χαρακτηριστικό της ανάλυσης στην πλατφόρμα αυτή είναι ότι δεν πρέπει οι χρήστες να είναι προγραμματιστές ούτε χρειάζονται λεπτομέρειες για την εφαρμογή ενός εργαλείου. Ένα επιπλέον βασικό στοιχείο της πλατφόρμας Galaxy είναι ότι αποθηκεύει κάθε ενέργεια του χρήστη στο σύστημα, έτσι οι χρήστες μπορούν να εξάγουν διάφορες πληροφορίες για γονιδιακά δεδομένα και χρησιμοποιώντας την πλατφόρμα μπορούν να συνδυάσουν ή να βελτιώσουν, να εκτελέσουν υπολογισμούς για να εξαχθούν οι αντίστοιχες αλληλουχίες.

Η πλατφόρμα περιέχει τρείς βασικές κατηγορίες: την λειτουργία ερωτημάτων(queries), εργαλεία για ανάλυση αλληλουχίας καθώς και την απεικόνιση της εξόδου. Η πρώτη κατηγορία περιλαμβάνει ένα σύνολο λειτουργιών όπως ένωση, τομή, αφαίρεση. Τα εργαλεία ανάλυσης είναι αυτόνομες ενότητες που έχουν σχεδιαστεί για να εκτελούν υπολογισμούς, όπως την συσχέτιση SNPs και φαινοτύπων που καλούμε να υλοποιήσω. Οι οθόνες επιτρέπουν την ανάκτηση/προβολή των αποτελεσμάτων που παράγονται από το χρήστη σε μια ποικιλία μορφών.

Για ανάλυση από την πλατφόρμα είναι απαραίτητο να υπάρχουν ενσωματωμένα εργαλεία που να μπορεί ο χρήστης να εισάγει τα δεδομένα του και να παίρνει τα αποτελέσματα που επιθυμεί. Ένα εργαλείο μπορεί να είναι ένα οποιοδήποτε κομμάτι λογισμικού (σε οποιανδήποτε γλώσσα), φτάνει να μπορεί να καλεστεί από γραμμή εντολών [14].

Στην υλοποίηση έχουμε προσθέσει συνολικά 5 εργαλεία: Chi-square, Fisher Exact, logistic regression, BOOST, MDR. Ο αλγόριθμος Random Forest αν και έχει μελετηθεί δεν βρίσκεται open source για να τον χρησιμοποιήσουμε.

4.2 Δεδομένα Εισόδου

Τα δεδομένα μας είναι σπασμένα σε δυο αρχεία, στο αρχείο Map και στο αρχείο Ped.

Από το αρχείο Map μας παρέχονται πληροφορίες για το κάθε SNP. Στην πρώτη στήλη του αρχείου μας δίνεται η κωδικοποίηση του χρωμοσώματος, από 1-22, X, Y ή 0 για απροσδιόριστο. Στη δεύτερη στήλη μας παρουσιάζεται το όνομα του SNP, στην τρίτη στήλη το genetic distance καθώς στην τελευταία στήλη μας δίνεται η τοποθεσία του SNP σε αριθμό βάσεων στο strand. Στο πιο κάτω πίνακα δίνουμε ένα παράδειγμα από τα δεδομένα σε map αρχείο.

Chromosome	SNP ID	Genetic distance	Physical Position
21	rs11511647	0	26765
X	rs3883674	0	32380
X	rs12218882	0	48172
9	rs10904045	0	48426
9	rs10751931	0	49949
8	rs11252127	0	52087
10	rs12775203	0	52277
8	rs12255619	0	52481

Πίνακας 3.1 Παράδειγμα από map αρχείο παρθέν από 'GWASPI'

Στο Ped αρχείο υπάρχουν 6 στήλες που μας δίνονται πληροφορίες για τον δότη του δείγματος. Η πρώτη στήλη αντιπροσωπεύει την ταυτότητα οικογένειας, η δεύτερη στήλη είναι η ταυτότητα του δότη, η τρίτη και τέταρτη στήλη περιέχουν την ταυτότητα του πατέρα και της μητέρας αντίστοιχα, η πέμπτη στήλη το φύλο και στην έβδομη ο φαινότυπος. Κάθε άτομο έχει μοναδικό συνδυασμό πληροφοριών. Οι επόμενες στήλες περιέχουν bi-allelic δείκτες που είναι τα SNPs. Οι δείκτες των γονότυπων είναι κωδικοποιημένοι με τα γράμματα "A" για Αδενίνη, "C" για Κυτοσίνη, "T" για Θυμίνη και "G" για Γουανίνη, ένα για κάθε αλληλόμορφο γονίδιο. Για ένα ελλιπή αλληλόμορφο χρησιμοποιείται το 0 ή το -9. Στο πιο κάτω πίνακα δίνουμε ένα παράδειγμα από τα δεδομένα σε ped αρχείο.

Family ID	Sample ID	Paternal ID	Maternal ID	SEX	Affection	Genotype
FAM1	NA06985	0	0	1	1	A T T T G G C C A T T T G G C C
FAM2	NA06991	0	0	1	1	C T T T G G C C C T T T G G C C
0	NA06993	0	0	1	1	C T T T G G C T C T T T G G C T
0	NA06994	0	0	1	1	C T T T G G C C C T T T G G C C
0	NA07000	0	0	2	1	C T T T G G C T C T T T G G C T
0	NA07019	0	0	1	1	C T T T G G C C C T T T G G C C
0	NA07022	0	0	2	1	C T T T G G C T C T T T G G 0 0
0	NA07029	0	0	1	1	C T T T G G C C C T T T G G C C
FAM2	NA07056	0	0	0	2	C T T T A G C T C T T T A G C T
FAM2	NA07345	0	0	1	1	C T T T G G C C C T T T G G C C

Πίνακας 3.2 Παράδειγμα από ped αρχείο παρθέν από ‘GWASPI’

4.3 Προεπεξεργασία

Κάθε αλγόριθμος που θα εισαχθεί στην πλατφόρμα λόγω του ότι υλοποιήθηκε για την εξυπηρέτηση άλλων εργασιών δέχονται διαφορετικό τύπο αναπαράστασης των δεδομένων εισόδου. Για το λόγο αυτό χρειάστηκε να γίνει μια προεπεξεργασία στα αρχικά μας δεδομένα ούτως ώστε να μπορούν οι αλγόριθμοι να αποκριθούν. Συγκεκριμένα, για τον αλγόριθμο BOOST και MDR χρησιμοποιήσαμε 2 ξεχωριστά προγράμματα γραμμένα σε γλώσσα προγραμματισμού C++ που μας δόθηκαν από την διπλωματική του κ. Άρη Αριστοδήμου [6]. Επίσης σε μερικούς αλγορίθμους χρειάστηκε η κωδικοποίηση των SNPs ώστε να ακολουθούν ένα συγκεκριμένο πρότυπο και να διεκπεραιωθεί η λειτουργία του αλγορίθμου.

4.4 Univariate analysis

Στην Univariate analysis αναφερόμαστε στους αλγορίθμους οι οποίοι θα κάνουν έλεγχο 1 προς 1 SNP. Στην κατηγορία αυτή συναντούμε τους αλγορίθμους Chi Square και Fisher Exact Test. Οι αλγόριθμοι αυτοί είναι βασισμένοι στο PLINK toolset [15]. Πρόκειται για μια ανοικτού κώδικα εργαλειοθήκη για ανάλυση συσχέτισης γονιδιώματος, σχεδιασμένη για να εκτελεί μια σειρά αναλύσεων μεγάλης κλίμακας με αποτελεσματικό τρόπο. Το PLINK μπορεί να αναγνωρίζει δεδομένα σε μια ποικιλία μορφών, μπορεί να αναδιατάξει αρχεία και να συγχωνεύσει δύο ή περισσότερα αρχεία. Επιπλέον, μπορεί να συμπιέσει τα δεδομένα από τα αρχεία σε δυαδική μορφή. Το επίκεντρο του PLINK είναι για ανάλυση δεδομένων για γονότυπο και φαινότυπο, έτσι δεν υπάρχει καμία στήριξη για βήματα πριν την χρήση του όπως μελέτη design και planning κλπ. Το PLINK ωστόσο υποστηρίζει μεταγενέστερες απεικονίσεις, σχολιασμό και αποθήκευση αποτελεσμάτων με την ενοποίηση του gPLINK και Haploview. Ωστόσο, θα χρησιμοποιήσουμε μόνο τους αλγορίθμους που έχουν αναφερθεί πιο πάνω.

Στα εργαλεία Chi-square, Fisher Exact ζητάμε από τον χρήστη να εισάγει τα .map και .ped αρχεία. Επιπλέον, εισάγει ένα κατώφλι (threshold), όπου θα εμφανίζονται τα αποτελέσματα του p-value που είναι ίσα και μικρότερα του κατωφλίου. Η τιμή p-value αντιπροσωπεύει την πιθανότητα του αποτελέσματος που βρήκαμε να προκύψει τυχαία η συγκεκριμένη συσχέτιση. Επομένως όσο πιο μικρή τιμή έχουμε τόσο το καλύτερο.

Στην συνέχεια τρέχουμε τον κάθε αλγόριθμο και μπορούμε να δούμε τα αποτελέσματα στον πίνακα εξόδου.

4.5 2-SNP interaction

Στην 2-SNP interaction βασιζόμαστε στον αλγόριθμο logistic regression ο οποίος ελέγχει την αλληλεπίδραση 2 προς 2 SNPs. Ο αλγόριθμος logistic regression θα είναι από το PLINK toolset. Επιπλέον, στην κατά ζεύγη ανάλυση θα ασχοληθούμε με το "Boolean Operation-based Screening and Testing"(BOOST) [10].

Ο αλγόριθμος BOOST δέχεται κωδικοποιημένα δεδομένα εισόδου. Η κωδικοποίηση αυτή χωρίζεται σε 3 κατηγορίες αναπαριστώντας την κάθε κατηγορία με βάση το αλληλόμορφο γονιδίωμα. Συγκεκριμένα, το ΑΑ αναπαριστάτε από τον αριθμό 1, το Αα και αΑ από τον αριθμό 2 και το αα από τον αριθμό 0. Ο BOOST αλγόριθμος δεν κωδικοποιεί missing data. Στο πίνακα αποτελεσμάτων που μας δίνεται από τον εν λόγο αλγόριθμο έχουμε αλλάξει την παράμετρο index και προσθέσαμε το όνομα του SNP που βρίσκουμε από το map αρχείο.

Αλληλόμορφο γονιδίωμα	**Κωδικοποίηση**
ΑΑ	1
Αα	2
Αα	0

Πίνακας 3.3 Κωδικοποίηση αλληλόμορφων

Και σε αυτά τα εργαλεία ζητάμε από τον χρήστη να εισάγει τα .map και .ped αρχεία, καθώς επίσης στο εργαλείο του Regression ζητάμε ένα κατώφλι όπως έχει αναφερθεί στο 4.3.

Για το εργαλείο Logistic Regression υπάρχει η επιλογή για univeriate analysis και 2-SNP interaction. Ορίζουμε από την αρχή ότι η ανάλυση που θα γίνεται είναι για 2 SNPs.

4.6 Multi-SNP interaction

Στην n-SNP interaction η οποία πρόκειται για έλεγχο περισσότερων από δύο SNPs, θα χρησιμοποιήσουμε τον αλγόριθμο MDR [11].

Στο εργαλείο του MDR αλγορίθμου εκτός από τα .map και .ped αρχεία ζητάμε από τον χρήστη να εισάγει 4 επιπλέον παραμέτρους έτσι ώστε να προσδιορίσει α) τον ελάχιστο αριθμό SNPs που θα επιλέγει για έλεγχο (π.χ. αν επιλέξει δυο, θα ελέγχει συσχετίσεις μεταξύ δυο ή περισσότερων SNPs), β) το μέγιστο αριθμό SNPs που θα ελέγχει, γ) τον

αριθμό των folds στο cross validation, δ) τον αριθμό των καλύτερων μοντέλων που θα θέλαμε να εμφανιστούν στην οθόνη.

4.7 Μεταεπεξεργασία

Σε όλα τα εργαλεία χρειάστηκε να επεξεργαστούμε τα δεδομένα εξόδου, ούτως ώστε να παρουσιάζονται στην οθόνη αυτά που χρειάζεται πραγματικά ο χρήστης. Συγκεκριμένα έχουν υλοποιηθεί τρία προγράμματα για τα εργαλεία Chi Square, Fisher's Exact και Regression που παραλείπουν τις στήλες που δεν μας ενδιαφέρουν και αφήνουν μόνο το όνομα του SNP και το p-value. Έχει υλοποιηθεί ακόμη ένα πρόγραμμα για τον αλγόριθμο MDR όπου εμφανίζει στην αρχή τα ορθά SNPs που ανίχνευσε με μεγάλο cross validation και στην συνέχεια τα παρουσιάζονται τα καλύτερα μοντέλα για κάθε επίπεδο.

4.8 Εισαγωγή εργαλείων στην πλατφόρμα Galaxy

Για να προστεθεί ένα καινούργιο εργαλείο στην πλατφόρμα Galaxy, ο προγραμματιστής θα πρέπει να ακολουθήσει τα πιο κάτω βήματα:

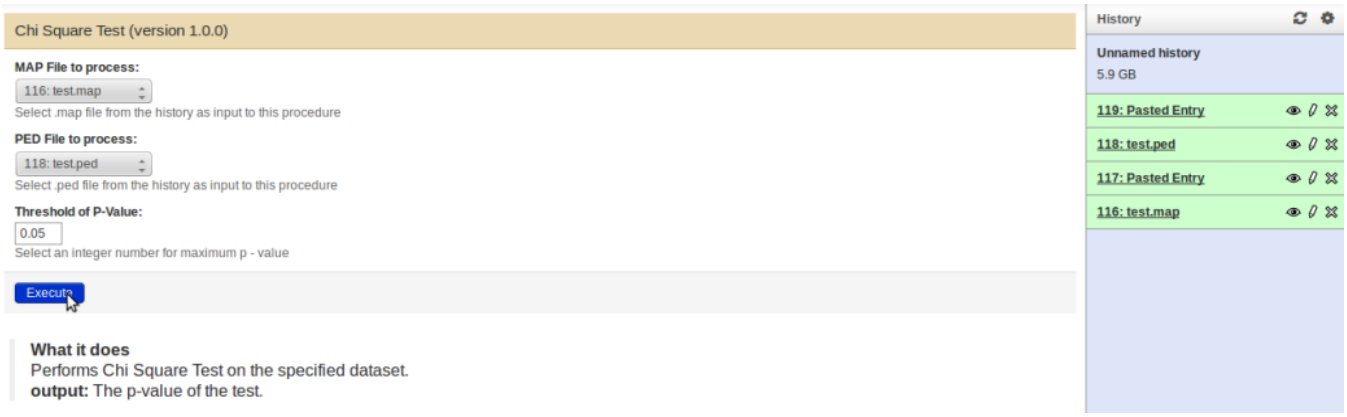
1. Αρχικά δημιούργησε ένα script αρχείο, το οποίο περιέχει τις εντολές που εκτελούν τους αλγορίθμους, και τα βοηθητικά προγράμματα που αναφέρονται για προεπεξεργασία και μεταεπεξεργασία.
2. Δημιουργία καινούργιου φακέλου (π.χ MyTools) κάτω από τον υποκατάλογο tools.
3. Δημιουργία ενός configuration αρχείου (.xml) που περιγραφεί πως εκτελείτε το εργαλείο, συμπεριλαμβανομένων των λεπτομερών περιγραφών των παραμέτρων εισόδων και εξόδου του εργαλείου. Αυτή η περιγραφή γίνεται για την αυτόματη δημιουργία γραφικής διεπαφής του συγκεκριμένου εργαλείου.
4. Εισάγουμε στον φάκελο που δημιουργήσαμε τα script και configuration αρχεία.
5. Επεκτείνουμε τα αρχείο tool_conf.xml που βρίσκεται στο root των αρχείων της πλατφόρμας, ούτως ώστε η πλατφόρμα να γνωρίζει για τα καινούργια tools.

6. Τρέχουμε την Galaxy με την εντολή run.sh, και τα εργαλεία είναι διαθέσιμα για να τα χρησιμοποιήσουμε.

4.8.1 Αλγόριθμος Chi-square

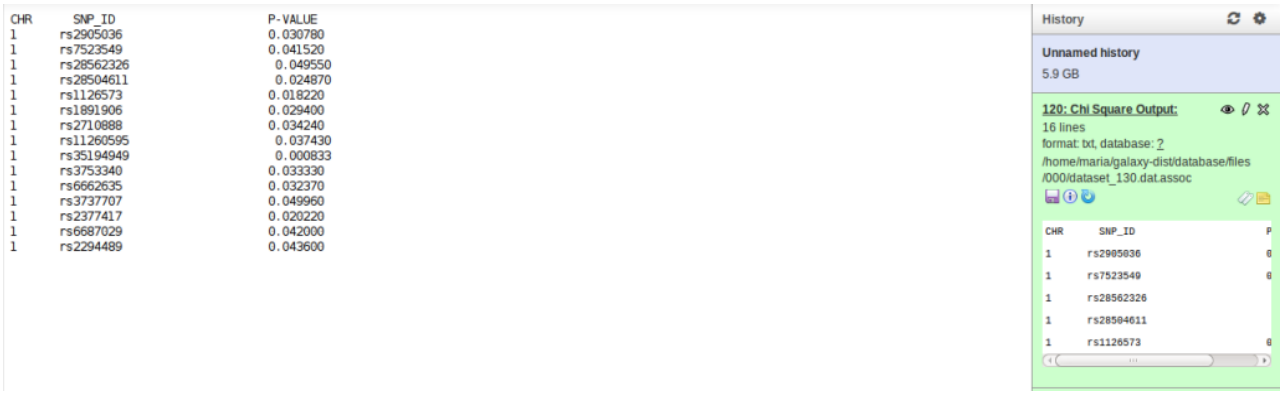
Πιο κάτω περιγράφεται ο τρόπος χρησιμοποίησης του εν λόγο εργαλείου, όπως επίσης επισυνάπτονται τα screenshoots από τα αρχεία .xml (Εικόνα 4.3) και .sh(Εικόνα 4.4) που εκτελούν την γραφική διεπαφή. Στο xml αρχείο μπορούμε να διαχωρίσουμε τις παραμέτρους για είσοδο και παραμέτρους για έξοδο, καθορίζοντας τον τύπο και την μορφή τους στην γραφική διεπαφή. Το script(.sh) αρχείο καλεί αρχικά το plink με τις εξής παραμέτρους: --noweb για την μη χρησιμοποίηση του διαδικτύου, --map $1 όπου καταχωρείτε το όνομα του map αρχείου, --ped $2 όπου καταχωρείτε το όνομα του ped αρχείου, --pfilter $3 όπου καταχωρείτε η τιμή του threshold, --out $4 για το όνομα του αρχείου εξόδου καθώς επίσης το --silent προστίθεται στην γραμμή εντολής του plink για καταστολή του αρχείου εξόδου στην κονσόλα. Τέλος, τρέχουμε το program_chi_square το οποίο είναι υπεύθυνο για τις αλλαγές που θα γίνουν στο αρχείο εξόδου.
Η χρήση του εργαλείου είναι η εξής:

1. Επέλεξε map αρχείο για την ανάλυση. Αυτό γίνεται επιλέγοντας το αρχείο εισόδου από το drop-down μενού κάτω από την ετικέτα Map File to process.
2. Επέλεξε ped αρχείο για την ανάλυση. Αυτό γίνεται επιλέγοντας το αρχείο εισόδου από το drop-down μενού κάτω από την ετικέτα Ped File to process.
3. Όρισε τον αριθμό του κατωφλίου για τον περιορισμό των P-value τιμών.
4. Πάτησε το κουμπί Execute.



Εικόνα 4.8.1.1 Γραφική Διεπαφή του Chi-Square Test

Μόλις εκτελεστεί ο αλγόριθμος στο παράθυρο του History εμφανίζεται η έξοδος του εργαλείου. Πατώντας το κουμπί «μάτι» μπορούμε να δούμε αναλυτικά τα αποτελέσματα στην οθόνη . Η πρώτη στήλη περιγράφει τον αριθμό του χρωμοσώματος, η δεύτερη στήλη είναι το όνομα(id) του SNP και η τρίτη στήλη περιγράφει το P-Value.



Εικόνα 4.8.1.2 Γραφική Διεπαφή των αποτελεσμάτων του Chi-Square Test

```
<!--ChiSquareTest-->
<tool id="chi_square_test" name="Chi Square Test">
        <description></description>
        <command interpreter="bash">
                runChiSquare.sh $testMapFile $testPedFile $threshold $chiSquareOut
        </command>
        <inputs>
                <param
                        name="testMapFile" type="data" format="txt" label=" MAP File to process" >
                        <help>
                        Select .map file from the history as input to this procedure
                        </help>
                </param>
                <param
                        name="testPedFile" type="data" format="txt" label="PED File to process" >
                        <help>
                        Select .ped file from the history as input to this procedure
                        </help>
                </param>
                <param
                        name="threshold" size="4" type="float" value="0.05" label="Threshold of P-Value" >
                        <help>
                        Select an integer number for maximum p - value
                        </help>
                </param>

        </inputs>

        <outputs>
                <data name="chiSquareOut" format="txt" label="Chi Square Output:"/>
        </outputs>


        <help>
        **What it does**

        Performs Chi Square Test on the specified dataset.

        **output:** The p-value of the test.
        </help>

</tool>
```

Εικόνα 4.8.1.3 Αρχείο xml

```
cd $(dirname $0)
plink --noweb --map $1 --ped $2 --pfilter $3 --out $4 --assoc --silent
./program_chi_square $4.assoc $4
rm -f $4.assoc
```

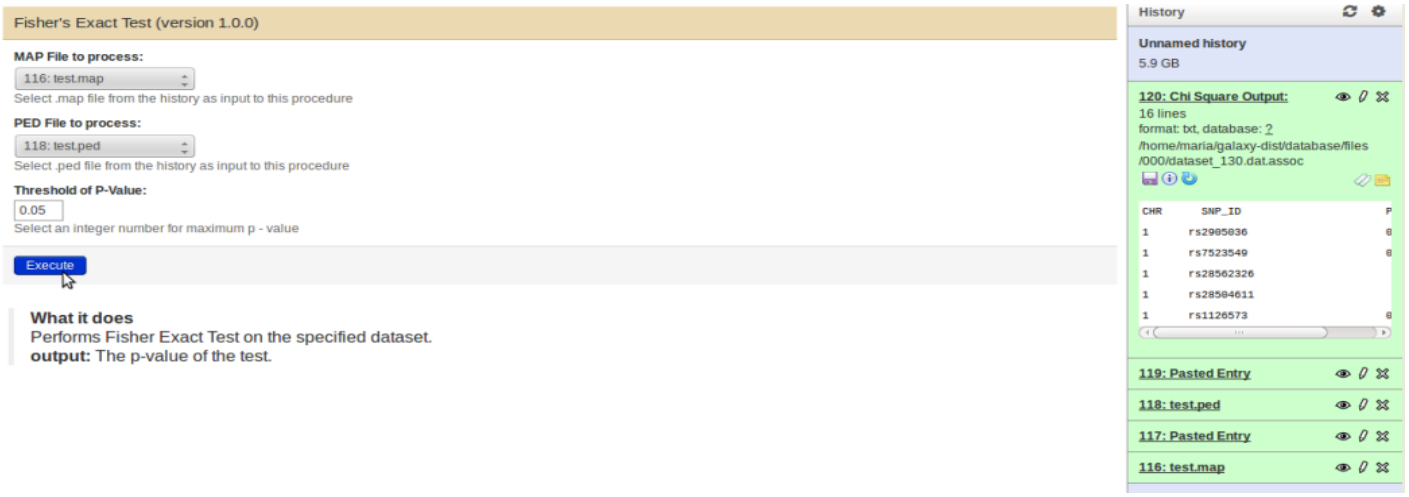
Εικόνα 4.8.1.4 Αρχείο script

4.8.2 Αλγόριθμος Fisher Exact

Το εργαλείο Fisher Exact μπορεί να χρησιμοποιηθεί με τα πιο κάτω βήματα καθώς επισυνάπτονται παράλληλα τα .xml και .sh αρχεία. Οι εντολές στα αρχεία αυτά περιγράφονται στο κεφάλαιο 4.8.1 με διαφορά το γεγονός ότι στο script αρχείο τρέχουμε το plink με –fisher έναντι του --assoc.

Η χρήση του εργαλείου είναι η εξής:

1. Επέλεξε map αρχείο για την ανάλυση. Αυτό γίνεται επιλέγοντας το αρχείο εισόδου από το drop-down μενού κάτω από την ετικέτα Map File to process.
2. Επέλεξε ped αρχείο για την ανάλυση. Αυτό γίνεται επιλέγοντας το αρχείο εισόδου από το drop-down μενού κάτω από την ετικέτα Ped File to process.
3. Όρισε τον αριθμό του κατωφλίου για τον περιορισμό των P-value τιμών.
4. Πάτησε το κουμπί Execute.



4.8.2.1 Γραφική Διεπαφή για Fisher Exact Test

Πατώντας το κουμπί «μάτι» μπορούμε να δούμε τα αποτελέσματα του Fisher Exact Test. Η πρώτη στήλη περιγράφει τον αριθμό του χρωμοσώματος, η δεύτερη στήλη είναι το όνομα(id) του SNP και η τρίτη στήλη περιγράφει το P-Value.

Εικόνα 4.8.2.2 Γραφική Διεπαφή των αποτελεσμάτων του Fisher Exact Test

```
<!--Fisher's Exact Test-->
<tool id="fisher_exact_test" name="Fisher's Exact Test">
        <description></description>
        <command interpreter="bash">
                runFisher.sh $testMapFile $testPedFile $threshold $fisherOut
        </command>

        <inputs>
                <param
                        name="testMapFile" type="data" format="txt" label="MAP File to process" >
                        <help>
                        Select .map file from the history as input to this procedure
                        </help>
                </param>
                <param
                        name="testPedFile" type="data" format="txt" label="PED File to process" >
                        <help>
                        Select .ped file from the history as input to this procedure
                        </help>
                </param>
                <param
                        name="threshold" size="4" type="float" value="0.05" label="Threshold of P-Value" >
                        <help>
                        Select an integer number for maximum p - value
                        </help>
                </param>

        </inputs>

        <outputs>
                <data name="fisherOut" format="txt" label="Fisher Exact Output:"/>
        </outputs>

        <help>
        **What it does**

        Performs Fisher Exact Test on the specified dataset.

        **output:** The p-value of the test.
        </help>

</tool>
```

Εικόνα 4.8.2.3 Αρχείο xml

```
cd $(dirname $0)
plink --noweb --map $1 --ped $2 --pfilter $3 --out $4 --fisher --silent
./program_fisher $4.assoc.fisher $4
rm -f $4.assoc.fisher
```

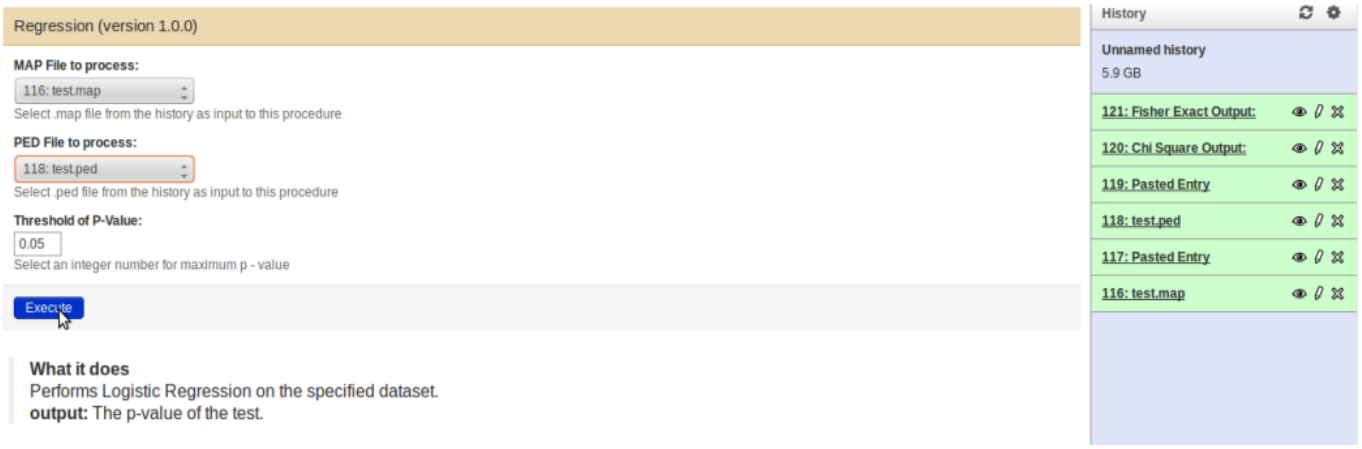
Εικόνα 4.8.2.4 Αρχείο script

4.8.3 Αλγόριθμος Regression

Το εργαλείο Regression μπορεί να χρησιμοποιηθεί με τα πιο κάτω βήματα καθώς επισυνάπτονται παράλληλα τα .xml και .sh αρχεία. Οι εντολές στα αρχεία αυτά περιγράφονται στο κεφάλαιο 4.8.1 με διαφορά το γεγονός ότι στο script αρχείο τρέχουμε το plink με –logistic έναντι του --assoc. Επιπλέον, προσθέτουμε το --epistasis --epi1 0.0001 για regression 2-SNP analysis.

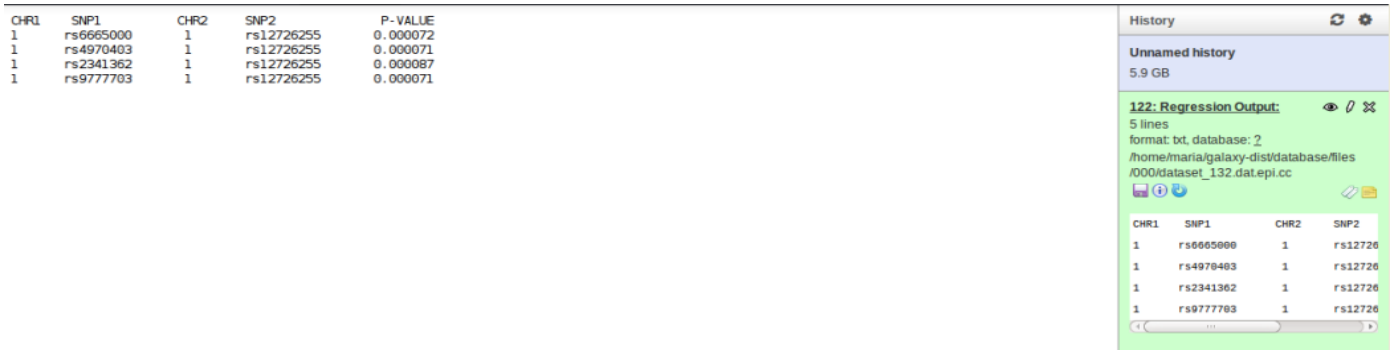
Η χρήση του εργαλείου είναι η εξής:

1. Επέλεξε map αρχείο για την ανάλυση. Αυτό γίνεται επιλέγοντας το αρχείο εισόδου από το drop-down μενού κάτω από την ετικέτα Map File to process.
2. Επέλεξε ped αρχείο για την ανάλυση. Αυτό γίνεται επιλέγοντας το αρχείο εισόδου από το drop-down μενού κάτω από την ετικέτα Ped File to process.
3. Όρισε τον αριθμό του κατωφλίου για τον περιορισμό των P-value τιμών.
4. Πάτησε το κουμπί Execute.



Εικόνα 4.8.3.1 Γραφική Διεπαφή για Regression

Πατώντας το κουμπί «μάτι» μπορούμε να δούμε τα αποτελέσματα του Regression. Η πρώτη και η δεύτερη στήλη αντιπροσωπεύουν το χρωμόσωμα και το όνομα (id) του πρώτου SNP, η τρίτη και τέταρτη στήλη αντιπροσωπεύουν το χρωμόσωμα και το όνομα (id) του δεύτερου SNP και η πέμπτη στήλη είναι το P-value της αλληλεπίδρασης των δυο SNP.



CHR1	SNP1	CHR2	SNP2	P-VALUE
1	rs6665000	1	rs12726255	0.000072
1	rs4970403	1	rs12726255	0.000071
1	rs2341362	1	rs12726255	0.000087
1	rs9777703	1	rs12726255	0.000071

Εικόνα 4.8.3.2 Γραφική Διεπαφή των αποτελεσμάτων του Regression Test

```
<!--Logistic Regression-->
<tool id="regression" name="Regression">
        <description></description>
        <command interpreter="bash">
                runLogistic.sh $testMapFile $testPedFile $threshold $logisticOut
        </command>

        <inputs>
                <param
                        name="testMapFile" type="data" format="txt" label=" MAP File to process" >
                        <help>
                        Select .map file from the history as input to this procedure
                        </help>
                </param>
                <param
                        name="testPedFile" type="data" format="txt" label="PED File to process" >
                        <help>
                        Select .ped file from the history as input to this procedure
                        </help>
                </param>
                <param
                        name="threshold" size="4" type="float" value="0.05" label="Threshold of P-Value" >
                        <help>
                        Select an integer number for maximum p - value
                        </help>
                </param>

        </inputs>

        <outputs>
                <data name="logisticOut" format="txt" label="Regression Output:"/>
        </outputs>


        <help>
        **What it does**

        Performs Logistic Regression on the specified dataset.

        **output:** The p-value of the test.
        </help>

</tool>
```

Εικόνα 4.8.3.3 Αρχείο xml

```
cd $(dirname $0)
plink --noweb --map $1 --ped $2 --pfilter $3 --out $4 --logistic --epistasis --epi1 0.0001 --silent
./program_logistic $4.epi.cc $4
rm -f $4.epi.cc
```

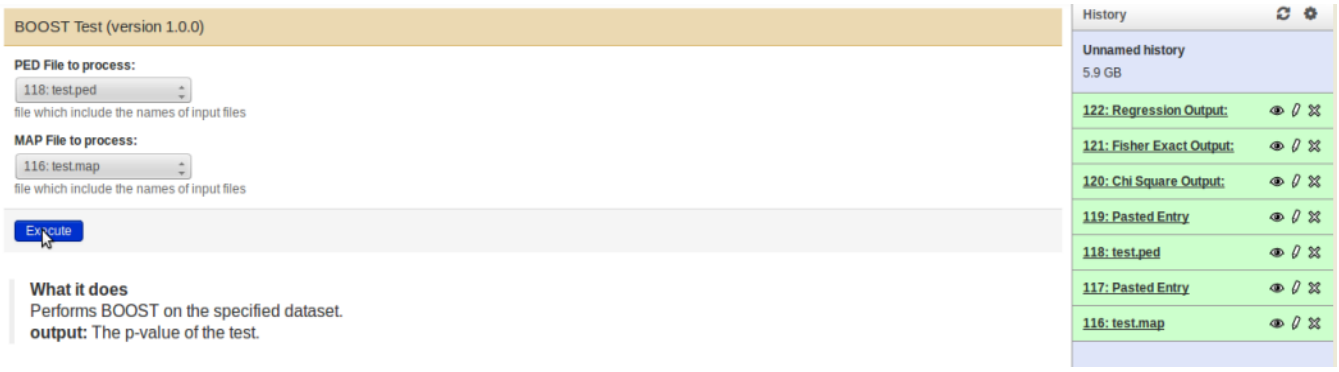
Εικόνα 4.8.3.4 Αρχείο script

4.8.4 Αλγόριθμος BOOST

Το εργαλείο BOOST μπορεί να χρησιμοποιηθεί με τα πιο κάτω βήματα καθώς επισυνάπτονται παράλληλα τα .xml και .sh αρχεία. Στο script αρχείο αρχικά καλούμε το πρόγραμμα που μετατρέπει το ped αρχείο στην μορφή που δέχεται τα δεδομένα ο αλγόριθμος. Στην συνέχεια εισάγουμε τα ονόματα των αρχείων εισαγωγής στο filenamelist, και καλούμε τον αλγόριθμο.

Η χρήση του εργαλείου είναι η εξής:

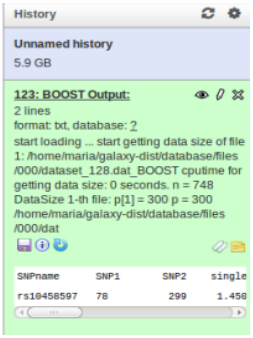
1. Επέλεξε ped αρχείο για την ανάλυση. Αυτό γίνεται επιλέγοντας το αρχείο εισόδου από το drop-down μενού κάτω από την ετικέτα Ped File to process.
2. Επέλεξε map αρχείο για την ανάλυση. Αυτό γίνεται επιλέγοντας το αρχείο εισόδου από το drop-down μενού κάτω από την ετικέτα Map File to process.
3. Πάτησε το κουμπί Execute.



Εικόνα 4.8.4.1 Γραφική διεπαφή για BOOST

Πατώντας το κουμπί «μάτι» μπορούμε να δούμε τα αποτελέσματα του BOOST. Πρώτη στήλη είναι ο αριθμός της πρώτης αλληλεπίδρασης σε αυτή την κατά ζεύγη ανάλυση, η δεύτερη και η τρίτη στήλη περιέχει τα ονόματα των SNPs, η τέταρτη στήλη και πέμπτη στήλη το single locus Association 1 και 2, η έκτη στήλη την αλληλεπίδραση του BOOST και στην έβδομή την αλληλεπίδραση του Plink.

SNPname	SNP1	SNP2	singlelocusAssoc1	singlelocusAssoc2	InteractionBOOST	InteractionPLINK
rs10458597	78	299	1.450511	7.123064	34.380599	3.524782



Εικόνα 4.8.4.2 Γραφική Διεπαφή των αποτελεσμάτων του BOOST

```
<!--BOOST-->
<tool id="boost" name="BOOST Test">
        <description></description>
        <command interpreter="bash">
                runboost.sh $ped $map $boostout
        </command>

        <inputs>
                <param
                        name="ped" type="data" format="txt" label="PED File to process" >
                        <help>
                        file which include the names of input files
                        </help>
                </param>
                <param
                        name="map" type="data" format="txt" label="MAP File to process" >
                        <help>
                        file which include the names of input files
                        </help>
                </param>
        </inputs>

        <outputs>
                <data name="boostout" format="txt" label="BOOST Output:"/>
        </outputs>


        <help>
        **What it does**

        Performs BOOST on the specified dataset.

        **output:** The p-value of the test.
        </help>
</tool>
```

Εικόνα 4.8.4.3 Αρχείο xml

```
#$1-> PED file $2->MAP file $3 output file
cd $(dirname $0)
./ped2BOOST $1 $2 $1_BOOST
echo "$1_BOOST" > $1_filenamelist
./BOOST $1_filenamelist $3 $2
```

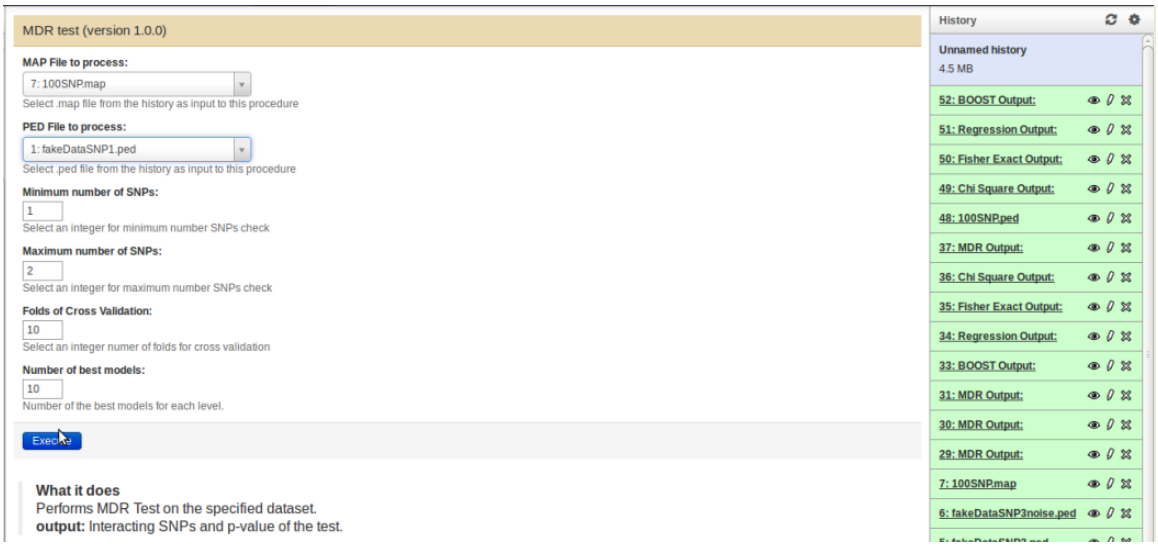
Εικόνα 4.8.4.4 Αρχείο scrip

4.8.5 Αλγόριθμος MDR

Πιο κάτω παρουσιάζονται τα βήματα για τον τρόπο χρησιμοποίησης του εργαλείου MDR. Παράλληλα επισυνάπτονται τα screenshoots από τα αρχεία .xml (Εικόνα 4.8.5.3) και .sh(Εικόνα 4.8.5.4) που εκτελούν την γραφική διεπαφή. Στο xml αρχείο μπορούμε να διαχωρίσουμε τις παραμέτρους για είσοδο και παραμέτρους για έξοδο, καθορίζοντας τον τύπο και την μορφή τους στην γραφική διεπαφή. Αρχικά το script(.sh) αρχείο καλεί το πρόγραμμα που μετατρέπει το ped αρχείο στην μορφή που πρέπει να εισαχθεί στον αλγόριθμο MDR. Ακολούθως, τρέχουμε τον αλγόριθμο με τις εξής παραμέτρους: -min $3 και –max $4, όπου $3 και $4 είναι ο μικρότερος και μεγαλύτερος αριθμός SNP που θα γίνει η multi-univariate ανάλυση, -cv $5 που αντιπροσωπεύει τον αριθμό των folds στο cross validation, -parallel για να τρέχει ο αλγόριθμος παράλληλα, -table_data=true για εμφάνιση των δεδομένων σε μορφή πίνακα, –nolandscape και -top_models_landscape_size=$7 όπου το $7 παίρνει ακέραια τιμή που αναπαριστά τον αριθμό των καλύτερων μοντέλων που θα τυπωθούν στον πίνακα αποτελεσμάτων.

Η χρήση του εργαλείου είναι η εξής:

1. Επέλεξε map αρχείο για την ανάλυση. Αυτό γίνεται επιλέγοντας το αρχείο εισόδου από το drop-down μενού κάτω από την ετικέτα Map File to process.
2. Επέλεξε ped αρχείο για την ανάλυση. Αυτό γίνεται επιλέγοντας το αρχείο εισόδου από το drop-down μενού κάτω από την ετικέτα Ped File to process.
3. Όρισε τον αριθμό του μικρότερου αριθμού των SNP που θα γίνει η ανάλυση.
4. Όρισε τον αριθμό του μεγαλύτερου αριθμού των SNP που θα γίνει η ανάλυση.
5. Όρισε τον αριθμό των folds του cross validation
6. Επέλεξε αριθμό για εμφάνιση καλύτερων μοντέλων για κάθε επίπεδο.
7. Πάτησε το κουμπί Execute.



Εικόνα 4.8.5.1 Γραφική διεπαφή για MDR

Πατώντας το κουμπί «μάτι» μπορούμε να δούμε τα αποτελέσματα του MDR. Η πρώτη στήλη περιγράφει τα ονόματα των SNP, η δεύτερη στήλη περιγράφει το cross validation consistency, η τρίτη στήλη το overall, η τέταρτη το cross validation της εκπαίδευσης, η πέμπτη το cross validation της επαλήθευσης και η έκτη και έβδομη στήλη την εκπαίδευση και επαλήθευση του μοντέλου αντίστοιχα.

Εικόνα 4.8.5.2 Γραφική διεπαφή αποτελεσμάτων MDR

```
<tool id="mdr" name="MDR test">
        <description></description>
        <command interpreter="bash">
                runmdr.sh $testMapFile $testPedFile $min $max $cv $mdrOut $number
        </command>

        <inputs>
                <param
                        name="testMapFile" type="data" format="txt" label="MAP File to process" >
                        <help>
                        Select .map file from the history as input to this procedure
                        </help>
                </param>
                <param
                        name="testPedFile" type="data" format="txt" label="PED File to process" >
                        <help>
                        Select .ped file from the history as input to this procedure
                        </help>
                </param>
                <param
                        name="min" size="4" type="integer" value="1" label="Minimum number of SNPs" >
                        <help>
                        Select an integer for minimum number SNPs check
                        </help>
                </param>
                <param
                        name="max" size="4" type="integer" value="2" label="Maximum number of SNPs" >
                        <help>
                        Select an integer for maximum number SNPs check
                        </help>
                </param>
                <param
                        name="cv" size="4" type="integer" value="10" label="Folds of Cross Validation" >
                        <help>
                        Select an integer numer of folds for cross validation
                        </help>
                </param>
                <param
                        name="number" size="4" type="integer" value="10" label="Number of best models" >
                        <help>
                        Number of the best models for each level.
                        </help>
                </param>
        </inputs>

        <outputs>
                <data name="mdrOut" format="txt" label="MDR Output:"/>
        </outputs>
```

Εικόνα 4.8.5.3 Αρχείο xml

```
1 cd $(dirname $0)
2 ./ped2MDR $2 $1 $6_ped2MDR
3 java -Xmx5g -jar mdr_3.0.2.jar -min=$3 -max=$4 -cv=$5 -parallel -table_data=true -nolandscape -top_models_landscape_size=$7 $6_ped2MDR > $6_output_MDR
4 cat $6_output_MDR
5 ./output_format_mdr $6_output_MDR $6 $6_ped2MDR
6 rm -f $6_ped2MDR $6_output_MDR
```

Εικόνα 4.8.5.4 Αρχείο script

Κεφάλαιο 5

Αποτελέσματα

5.1 Δεδομένα Εισόδου για σύγκριση Αλγορίθμων

Τα δεδομένα που εισήχθησαν για σύγκριση των εργαλείων, δημιουργήθηκαν τυχαία για σκοπούς επαλήθευσης της υλοποίησης. Έχουμε δυο κατηγορίες δεδομένων, αυτά με uniform κατανομή και αυτά χωρίς uniform κατανομή. Τα δεδομένα με uniform κατανομή είχαν την ίδια συχνότητα minor (ελάσσον) αλληλόμορφου γονιδίου με τα major (μείζον) αλληλόμορφου γονιδίου.

Έχουμε στην διάθεση μας 7 .ped αρχεία και ένα .map αρχείο για κάθε κατηγορία. Πιο κάτω παρουσιάζονται οι συναρτήσεις που προσομοιώνει η κάθε κατηγορία – για uniform κατανομή και χωρίς uniform κατανομή, για κάθε υποκατηγορία δεδομένων – για 1 SNP

συσχέτιση, για 2 SNP αλληλεπίδραση, για 3 SNP αλληλεπίδραση. Κάθε συμβολοσειρά που βρίσκεται στην αγκύλη σε κάθε συνάρτηση αντιπροσωπεύει ένα SNP που θα οριστεί σαν interaction SNP σε κάθε σύνολο δεδομένων που προσομοιώνεται η συγκεκριμένη συνάρτηση.

Ακολουθούν οι συναρτήσεις για τα δεδομένα με uniform κατανομή, όπου στα δεδομένα αυτά έχουμε 200 SNPs για 2000 subjects:

Στο πρώτο ped αρχείο, βρίσκουμε δεδομένα με SNP των οποίων το main effect τους συναθροίζεται μεταξύ τους, όπως και στο δεύτερο με διαφορά ότι προστέθηκε θόρυβος (Gaussian noise). Τα αρχεία αυτά προσομοιώνουν την συνάρτηση 5.1.

$$[rs126] + [rs162] - 3 * [rs239] + 2 * \sqrt{[rs289]}$$

5.1. Συνάρτηση για univariate SNP interaction

Στο τρίτο και τέταρτο ped αρχείο, έχουμε δεδομένα με 2 SNP interaction, όπου προστέθηκε θόρυβος στο τέταρτο αρχείο και προσομοιώνουν την συνάρτηση 5.2.

$$2 * [rs126] * [rs162] - 1.5 * [rs239] * [rs289] + [rs322]/[rs360]$$

5.2. Συνάρτηση για 2 SNP interaction

Στο πέμπτο και έκτο ped αρχείο, έχουμε δεδομένα με 3 SNP interaction, όπου προστέθηκε θόρυβος στο έκτο αρχείο και προσομοιώνουν την συνάρτηση 5.3.

$$[rs126] * \frac{[rs162]}{-[rs239]} + [rs289] * [rs322]/[rs360]$$

5.3. Συνάρτηση για 3 SNP interaction

Τέλος, το έβδομο ped αρχείο των αυθεντικών δεδομένων δεν προστέθηκε καμιά αλληλεπίδραση SNP.

Ακολουθούν οι συναρτήσεις για τα δεδομένα χωρίς uniform κατανομή, όπου στα δεδομένα αυτά έχουμε 100 SNPs για 1183 subjects::

Στο πρώτο ped αρχείο, βρίσκουμε δεδομένα με SNP με αθροιστική επίδραση, όπως και στο δεύτερο με διαφορά ότι προστέθηκε θόρυβος (Gaussian noise). Τα αρχεία αυτά προσομοιώνουν την συνάρτηση 5.4. Κάθε συμβολοσειρά που βρίσκεται στην αγκύλη αντιπροσωπεύει ένα SNP που θα οριστεί σαν interaction SNP.

$$[rs28576697] + [rs13303106] - 3 * [rs3128117] + 2 * \sqrt{[rs3737728]}$$

5.4. Συνάρτηση για univariate SNP interaction

Στο τρίτο και τέταρτο ped αρχείο, έχουμε δεδομένα με 2 SNP interaction, όπου προστέθηκε θόρυβος στο τέταρτο αρχείο και προσομοιώνουν την συνάρτηση 5.5.

$$2 * [rs28576697] * [rs13303106] - 1.5 * [rs3128117] * [rs3737728] + \frac{[rs2341354]}{[rs2710888]}$$

5.5. Συνάρτηση για 2 SNP interaction

Στο πέμπτο και έκτο ped αρχείο, έχουμε δεδομένα με 3 SNP interaction, όπου προστέθηκε θόρυβος στο έκτο αρχείο και προσομοιώνουν την συνάρτηση 5.6.

$$-1.5 * [rs28576697] * \frac{[rs13303106]}{[rs3128117]} + [rs3737728] * \frac{[rs2341354]}{[rs2710888]}$$

5.6. Συνάρτηση για 3 SNP interaction

Τέλος, το έβδομο ped αρχείο των αυθεντικών δεδομένων δεν προστέθηκε καμιά αλληλεπίδραση SNP.

Σημειώνεται ότι στα πάρα κάτω αποτελέσματα για τους αλγόριθμους που παρουσιάζουν στα αποτελέσματα τους τιμή p-value όταν τους τρέχουμε βάζουμε ως κατώφλι για τη τιμή του p-value ίσο με 0.05 για να μας εμφανίζονται όλα τα SNPs με τιμή κάτω από αυτό.

5.2 Αποτελέσματα από δεδομένα που είναι ανεξάρτητα από το φαινότυπο

Παρακάτω παρουσιάζεται ο πίνακας αποτελεσμάτων που πήραμε από τα δεδομένα που είναι ανεξάρτητα από το φαινότυπο, δηλ. αυτά που δεν προσομοιώνουν καμία συνάρτηση προσθετικής επίδρασης πάνω σε SNPs. Οι στήλες του πίνακα αναπαριστούν τον κάθε αλγόριθμο ενώ οι υποστήλες χωρίζουν τα αποτελέσματα που πήραμε από uniform (U) και μη uniform κατανομή (NU). Οι πρώτη γραμμή παρουσιάζει τον αριθμό των ορθών SNPs, η δεύτερη τον αριθμό των λανθασμένων SNPs, η τρίτη και τέταρτη γραμμή αναπαριστά το μέσο όρο και την τυπική απόκλιση του p-value για τα ορθά και λανθασμένα SNPs αντίστοιχα. Στους αλγορίθμους BOOST και MDR δεν έχουμε στα αποτελέσματα τιμές p-value, ωστόσο για τον αλγόριθμο BOOST έχουμε την μετρική interaction BOOST και στον MDR την μετρική overall. Η μετρική overall καθορίζεται από την επαλήθευση του μοντέλου καθώς και το cross validation consistency. Στους πίνακες θα εμφανίζεται ο μέσος όρος των μετρικών αυτών για τον κάθε αλγόριθμο.

	Chi-Square		**Fisher-Exact**		**Regression**		**BOOST**		**MDR**	
	U	NU	U	NU	N	NU	U	NU	U	NU
Correctly	0	0	0	0	0	0	0	0	0	0
Incorrectly	4	9	4	9	6	0	3	0	9	10
Correctly AVG (STDEV)	_	_	_	_	_	_	_	_	_	_
Incorrectly AVG (STDEV)	0.02217 (0.0182)	0.02412 (0.0153)	0.02295 (0.0192)	0.02574 (0.0163)	6.33E-05 (2.19E-05)	_	31.468	_	0.5517 (0.020)	0.5371 (0.0045)

Πίνακας 5.1 Συνέπεια Αποτελεσμάτων – Δεδομένα με αθροιστική επίδραση

Παρατηρώντας τον πίνακα 5.1 θα παρατηρήσουμε ότι όλοι οι αλγόριθμοι δεν βρίσκουν κανένα ορθό SNP, για αυτό το λόγο παραλείπονται οι τιμές του μέσου όρου και της τυπικής απόκλισης των μετρικών του κάθε αλγορίθμου.

Θα παρατηρήσουμε ότι οι αλγόριθμοι βρίσκουν ένα μικρό αριθμό από λανθασμένα SNPs με εξαίρεση τους αλγορίθμους Regression και BOOST σε uniform κατανομή που δεν βρίσκουν κανένα SNP. Συνεπώς, καταλαβαίνουμε ότι οι αλγόριθμοι ορθά δεν ανταποκρίνονται στα δεδομένα αυτά αφού δεν υπάρχουν SNP interaction. Συγκρίνοντας τα αποτελέσματα από τις 2 κατανομές θα παρατηρήσουμε ότι οι τιμές των μετρικών που πήραμε από κάθε αλγόριθμο είναι πολύ κοντά.

Θα πρέπει να αναφερθεί ωστόσο ότι στον αλγόριθμο MDR όπως έχουμε ήδη αναφερθεί, τον τρέχουμε με ένα αριθμό που αντιστοιχεί στην εμφάνιση καλύτερων μοντέλων για κάθε επίπεδο. Στα αποτελέσματα που έχουμε στον πίνακα με τα λανθασμένα SNPs συμπεριλαμβάνονται και τα SNPs που δεν έχουν ικανοποιητικό cross validation αλλά εμφανίστηκαν λόγω του ότι θέσαμε τον αριθμό που αντιστοιχεί στην εμφάνιση καλύτερων μοντέλων. Για αυτό το λόγο είναι τόσο χαμηλή η τιμή του μέσου όρου της μετρικής του MDR.

5.3 Αποτελέσματα από δεδομένα σε SNP με αθροιστική επίδραση χωρίς θόρυβο

Παρατηρώντας την συνάρτηση 5.1. μπορούμε να προσδιορίσουμε ότι πρόκειται για μια συνάρτηση προσθετικής επίδρασης πάνω σε ένα SNP με uniform κατανομή. Συγκεκριμένα, διακρίνουμε τέσσερα SNPs : rs126, rs162, rs239, rs289, τα οποία θα θεωρήσουμε ότι θα πρέπει να τα αναγνωρίζει ο κάθε αλγόριθμος.

Εξετάζοντας τώρα την συνάρτηση 5.4. βλέπουμε εξίσου ότι πρόκειται για μια συνάρτηση προσθετικής επίδρασης πάνω σε ένα SNP χωρίς uniform κατανομή. Συγκεκριμένα, διακρίνουμε τα ακόλουθα SNPs: rs28576697, rs13303106, rs3128117, rs3737728 τα οποία θα θεωρήσουμε ότι θα πρέπει να ανιχνεύσει ο κάθε αλγόριθμος.

Πιο κάτω παρουσιάζεται ο πίνακας αποτελεσμάτων που πήραμε από κάθε αλγόριθμο με τα δεδομένα για univariate SNP interaction χωρίς θόρυβο. Οι στήλες του πίνακα αντιστοιχούν σε κάθε αλγόριθμο ξεχωριστά ενώ οι υποστήλες χωρίζουν τα αποτελέσματα που πήραμε από uniform (U) και μη uniform κατανομή (NU). Η πρώτη γραμμή του πίνακα αναπαριστά τον αριθμό τον ορθών SNPs που βρήκε ο αλγόριθμος ενώ η δεύτερη τον αριθμό των λανθασμένων (SNPs που μας παρουσιάζονται χωρίς να ανήκουν σε αυτά που εισάγαμε με την συνάρτηση 5.1). Η τρίτη και τέταρτη γραμμή του πίνακα αναπαριστά το μέσο όρο και την τυπική απόκλιση του p-value για τους αλγορίθμους Chi-Square, Fisher Exact Test και Regression για ορθά και λανθασμένα SNP αντίστοιχα. Για τους αλγορίθμους BOOST και MDR εμφανίζεται ο μέσος όρος των μετρικών τους λόγω του ότι στα αποτελέσματα δεν έχουμε τιμές p-value για τους δυο αυτούς αλγορίθμους.

	Chi-Square		**Fisher-Exact**		**Regression**		**BOOST**		**MDR**	
	U	NU	U	NU	N	NU	U	NU	U	NU
Correctly	4	2	4	2	0	4	0	1	4	1
Incorrectly	14	72	14	71	8	74	0	29	6	19
Correctly AVG (STDEV)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	_	3.57E-06 (1.16E-05)	_	80.111	0.64457 (0.103)	0.901264 (0.0025)
Incorrectly AVG (STDEV)	0.02503 (0.0144)	0.00309 (0.0080)	0.02617 (0.0150)	0.00294 (0.0078)	4.05E-05 (1.38E-05)	1.07E-05 (2.14E-05)	_	55.0848	0.525 (0.0064)	0.720744 (0.0482)

Πίνακας 5.2 Συνέπεια Αποτελεσμάτων - univariate SNP interaction χωρίς θόρυβο

Συγκρίνοντας τα αποτελέσματα των πιο πάνω αλγορίθμων παρατηρούμε ότι ο αλγόριθμος Chi-Square έχει επιτυχία 100% σε uniform κατανομή εφόσον βρίσκει και τα 4 SNPs και 50% σε μη uniform κατανομή εφόσον βρίσκει μόλις 2 ορθά SNPs. Θα παρατηρήσουμε ωστόσο ότι ο μέσος όρος p-value και η τυπική απόκλιση των ορθών SNPs είναι μηδέν, ενώ στα λανθασμένα είναι μεγαλύτερος από μηδέν με καλύτερα αποτελέσματα στα δεδομένα χωρίς uniform κατανομή. Θα πρέπει να αναφέρουμε ότι ίσως με πιο αυστηρό p-value

κατώφλι, θα μπορούσαμε να επιλέξουμε τα ορθά SNPs. Εντούτοις, δεν είναι εύκολο να αποφασιστεί πόσο περιορισμένο θα πρέπει να είναι το κατώφλι, στο οποίο θα αποδεχόμαστε ότι πρόκειται για ορθό SNP.

Ο αλγόριθμος Fisher Exact Test έχει επιτυχία 100% σε uniform κατανομή και 50% σε μη uniform κατανομή σε αυτό το σύνολο δεδομένων. Επιπλέον, όπως και στον Chi-Square ο μέσος όρος και η τυπική απόκλιση των ορθών SNPs είναι μηδέν, ενώ στα λανθασμένα μεγαλύτερή από μηδέν με καλύτερα αποτελέσματα στα δεδομένα χωρίς uniform κατανομή.

Ακολούθως, ο αλγόριθμος Regression έχει επιτυχία 0% σε uniform κατανομή εφόσον δεν βρίσκει κανένα ορθό SNP και 100% σε μη uniform κατανομή εφόσον ανιχνεύει και τα 4 SNPs. Βρίσκει ωστόσο άλλα 8 λανθασμένα SNPs για uniform κατανομή και 74 για μη uniform κατανομή. Ο μέσος όρος και η τυπική απόκλιση των ορθών SNPs σε μη uniform κατανομή αν και δεν είναι μηδέν, έχει πολύ μικρή τιμή και αυτό που παρατηρείται είναι ότι έχει και πάλι μικρότερη τιμή από τον μέσο όρο και την τυπική απόκλιση των λανθασμένων SNPs.

Στην συνέχεια παρατηρούμε ότι ο αλγόριθμος BOOST ανιχνεύει μόλις ένα SNP σε μη uniform κατανομή και κανένα σε uniform, άρα έχει ποσοστό επιτυχίας μόνο 25% στο σύνολο δεδομένων χωρίς uniform κατανομή. Ανιχνεύει ακόμα 29 λανθασμένα SNPs σε μη uniform κατανομή. Θα δούμε επίσης ότι βάσει της μετρικής του αλγορίθμου έχουμε καλύτερα αποτελέσματα στα ορθά SNPs, εφόσον γνωρίζουμε ότι όσο μεγαλύτερη είναι η τιμή τόσο μεγαλύτερη είναι και η συσχέτιση.

Τέλος, ο αλγόριθμος MDR έχει ποσοστό επιτυχίας 100% σε uniform κατανομή αφού βρίσκει και τα 4 SNPs και έχει 25% επιτυχία σε μη uniform κατανομή. Βρίσκει άλλα 6 λανθασμένα SNPs για uniform κατανομή 19 λανθασμένα SNPs και για μη uniform κατανομή. Θα παρατηρήσουμε ότι η μετρική για επαλήθευση του μοντέλου στα ορθά SNPs είναι μεγαλύτερη από την τιμή των λανθασμένων SNPs κάτι που επιθυμούσαμε όπως και η τιμή της τυπικής απόκλισης είναι καλύτερη στα ορθά SNP από ότι στα λανθασμένα.

5.4 Αποτελέσματα από δεδομένα σε SNP με αθροιστική επίδραση με θόρυβο

Τα δεδομένα στηρίζονται και πάλι στις συναρτήσεις 5.1 και 5.4 με διαφορά ότι προστίθεται θόρυβος με την χρήση της Gaussian συνάρτησης. Οι αλγόριθμοι θα πρέπει να ανιχνεύσουν τα τέσσερα αυτά SNPs για uniform κατανομή: rs126, rs162, rs239, rs289 και για μη uniform κατανομή τα SNPS: rs28576697, rs13303106, rs3128117, rs3737728.

Ο πιο κάτω πίνακας αντιστοιχεί στις ίδιες στήλες και γραμμές όπως τον πίνακα 5.2. όπου θα παρουσιαστούν τα αποτελέσματα κάθε αλγορίθμου από τα δεδομένα σε SNP με αθροιστική επίδραση με θόρυβο.

	Chi-Square		**Fisher-Exact**		**Regression**		**BOOST**		**MDR**	
	U	NU	U	NU	N	NU	U	NU	U	NU
Correctly	4	4	4	4	0	2	1	1	4	3
Incorrectly	17	66	16	66	4	69	1	29	6	10
Correctly AVG (STDEV)	0.0 (0.0)	5E-07 (0.0001)	0.0 (0.0)	7.5E-07 (0.0001)	_	1.66E-05 (3.09E-05)	54.34	42.74	0.6408 (0.097)	0.9819 (0.0003
Incorrectly AVG (STDEV)	0.02188 (0.0132)	0.004442 (0.0096)	0.02310 (0.013)	0.004706 (0.0101)	0.00041 (9.9E-06)	1.19E-05 (2.16E-05)	54.34	40.53	0.532 (0.006)	0.7142 (0.0688)

Πίνακας 5.3 Συνέπεια Αποτελεσμάτων - univariate SNP interaction με θόρυβο

Παρατηρώντας τον πίνακα αποτελεσμάτων παρατηρούμε ότι ο αλγόριθμος Chi-Square έχει 100% επιτυχία και σε uniform κατανομή και χωρίς αυτή αφού ανιχνεύει όλα τα ορθά SNPs. Η τιμή του μέσου όρου του p-value και η τυπική απόκλιση των ορθών SNPs είναι 0.0 και 5E-07 για uniform και μη uniform κατανομή αντίστοιχα. Ο αλγόριθμος Chi-Square έχει ανιχνεύσει ακόμα 17 λανθασμένα SNPs για uniform κατανομή και 66 λανθασμένα SNPs για μη uniform κατανομή, με τιμές του μέσου όρου του p-value μεγαλύτερες από αυτές των ορθών SNPs όπως περιμέναμε. Θα παρατηρήσουμε ότι για τα ορθά SNPs

έχουμε καλύτερα αποτελέσματα στην uniform κατανομή ενώ για τα λανθασμένα SNPs έχουμε καλύτερα αποτελέσματα στην μη uniform κατανομή.

Ο αλγόριθμος Fisher Exact Test βρίσκει τα 4 SNPs, άρα έχει 100% επιτυχία και στις 2 κατηγορίες δεδομένων που έχουμε. Ωστόσο ανιχνεύει επιπλέον 16 λανθασμένα SNPs για uniform κατανομή και 66 λανθασμένα SNPs για μη uniform κατανομή. Και πάλι θα παρατηρήσουμε ότι η τιμή του μέσου όρου του p-value και η τυπική απόκλιση είναι μικρότερη στα ορθά SNPs σε σχέση με τα λανθασμένα. Συγκρίνοντας τα δεδομένα με uniform και χωρίς uniform κατανομή θα δούμε ότι παίρνουμε καλύτερα αποτελέσματα ορθών SNPs σε uniform κατανομή ενώ για λανθασμένα SNPs έχουμε καλύτερα αποτελέσματα σε μη uniform κατανομή.

Μελετώντας τα αποτελέσματα του αλγορίθμου Regression βλέπουμε ότι επιτυγχάνει κατά 50% μόνο σε μη uniform κατανομή αφού έχει ανιχνεύσει δυο ορθά SNPs ενώ στην uniform κατανομή έχει 0% επιτυχία εφόσον δεν ανιχνεύει κανένα SNP. Ωστόσο είναι η πρώτη φορά που παρατηρείτε ότι η τιμή του μέσου όρου του p-value και η τυπική απόκλιση για τα ορθά SNPs είναι μεγαλύτερη από αυτήν των λανθασμένων SNPs. Η διαφορά αυτή είναι πολύ μικρή.

Ο αλγόριθμος BOOST έχει βρει ένα SNP ορθό και σε uniform και χωρίς uniform κατανομή καθώς 1 και 29 λανθασμένα αντίστοιχα για κάθε κατηγορία. Θα παρατηρήσουμε ότι η τιμή του μέσου όρου του p-value και η τυπική απόκλιση είναι καλύτερη στα ορθά SNPs σε σχέση με τα λανθασμένα.

Παρατηρώντας τα αποτελέσματα του MDR, παρατηρούμε ότι επιτυγχάνει με 100% επιτυχία για uniform κατανομή και 75% για μη uniform κατανομή. Ωστόσο ανιχνεύει 6 λανθασμένα SNPs για uniform κατανομή και 10 λανθασμένα SNPs χωρίς uniform κατανομή. Θα παρατηρήσουμε και πάλι ότι η μετρική για επαλήθευση του μοντέλου στα ορθά SNPs είναι μεγαλύτερη από την τιμή των λανθασμένων SNPs όπως και η τιμή της τυπικής απόκλισης είναι καλύτερη στα ορθά SNP από ότι στα λανθασμένα.

5.5 Αποτελέσματα από δεδομένα με 2 SNP interaction χωρίς θόρυβο

Μελετώντας την συνάρτηση 5.2. παρατηρούμε ότι πρόκειται για μια συνάρτηση προσθετικής επίδρασης πάνω σε δύο SNP. Συγκεκριμένα, διακρίνουμε 6 SNPs: το rs126 αλληλεπιδρά με το rs162, rs239 αλληλεπιδρά με το rs289 και το rs322 αλληλεπιδρά με το rs360.

Παρατηρώντας τώρα την συνάρτηση 5.5. μπορούμε να προσδιορίσουμε ότι πρόκειται για μια συνάρτηση προσθετικής επίδρασης πάνω σε δύο SNP. Εδώ διακρίνουμε 6 SNPs: το rs28576697 αλληλεπιδρά με το rs13303106, rs3128117 αλληλεπιδρά με το rs3737728 και το rs2341354 αλληλεπιδρά με το rs2710888.

Πάρα κάτω παρουσιάζονται δύο πίνακες αποτελεσμάτων που πήραμε από κάθε αλγόριθμο με τα δεδομένα για 2 SNP interaction χωρίς θόρυβο. Στον πρώτο πίνακα 5.4 αναπαριστούνται τα αποτελέσματα των αλγορίθμων που είναι για univariate SNP ανάλυση, δηλ. Chi Square και Fisher Exact Test, καθώς και τα αποτελέσματα του αλγόριθμου MDR που αναφέρονται για την ανίχνευση ενός SNP. Οι υποστήλες χαρακτηρίζουν τα αποτελέσματα που παίρνουμε από uniform κατανομή (U) και από μη uniform κατανομή (NU). Η πρώτη γραμμή του πίνακα αναπαριστά τον αριθμό τον ορθών SNPs που βρήκε ο αλγόριθμος ενώ η δεύτερη τον αριθμό των λανθασμένων. Η τρίτη και τέταρτη γραμμή του πίνακα αναπαριστά το μέσο όρο και την τυπική απόκλιση του p-value για τους αλγορίθμους Chi Square και Fisher Exact για ορθά και λανθασμένα SNP αντίστοιχα. Για τον αλγόριθμο MDR στην τρίτη και τέταρτη γραμμή αναπαριστάτε ο μέσος όρος την μετρικής overall.

Ο δεύτερος πίνακας 5.5 συμπεριλαμβάνει τα αποτελέσματα των αλγορίθμων Regression, BOOST και τα αποτελέσματα του αλγορίθμου MDR για 2 SNP ανάλυση. Η πρώτη γραμμή του πίνακα αναπαριστά τον αριθμό τον ορθών SNPs που βρήκε ο αλγόριθμος και η δεύτερη γραμμή αναπαριστά τον αριθμό των λανθασμένων. Η τρίτη γραμμή του πίνακα αναπαριστά τον αριθμό των ορθών συνδυασμών, δηλαδή των αλληλεπιδράσεων που γίνονται μεταξύ ορθών SNPs. Η τέταρτη γραμμή παρουσιάζει τον αριθμό των μερικών ορθών αλληλεπιδράσεων, δηλ. όταν ένα ορθό SNP αλληλεπιδρά με ένα λανθασμένο SNP. Στην πέμπτη και έκτη γραμμή παρουσιάζονται οι τιμές του μέσου όρους της μετρικής κάθε

αλγορίθμου για τα ορθά και λανθασμένα SNPs αντίστοιχα. Στην τελευταία γραμμή θα παρουσιάζεται ο συνολικός αριθμός όλων των αλληλεπιδράσεων.

	Chi-Square		**Fisher-Exact**		**MDR**	
	U	NU	U	NU	U	NU
Correctly	6	6	6	6	6	2
Incorrectly	7	83	7	83	4	8
Correctly AVG (STDEV)	0.0007 (0.001818)	3.33E-07 (8.16497E-07)	0.000808 (0.001978)	5E-07 (1.22474E-06)	0.615 (0.0656)	0.7645 (0.00891)
Incorrectly AVG (STDEV)	0.02 (0.0125)	0.001197 (0.005519)	0.0278 (0.01311)	0.001367 (0.006395)	0.528 (0.0017)	0.768163 (0.0163)

Πίνακας 5.4 Συνέπεια Αποτελεσμάτων - 2 SNP interaction χωρίς θόρυβο

	Regression		**BOOST**		**MDR**	
	U	NU	U	NU	U	NU
Correctly	4	6	0	5	4	4
Incorrectly	2	69	2	29	4	10
Correct 2 SNPs	2	5	0	0	6	1
Partly Correct 2 SNPs	0	64	0	5	4	7
Correctly AVG(STDEV)	0.0 (0.0)	2.07E-05 (2.6961E-05)	_	38.7742	0.7462 (0.0234)	0.8658 (0.0068)
Incorrectly AVG(STDEV)	0.000043 (-)	1.49E-05 (2.37425E-05)	36.669541	44.0875	0.684575 (0.000486)	0.8701 (0.009475)
Total 2 SNPs	3	333	1	33	10	10

Πίνακας 5.5 Συνέπεια Αποτελεσμάτων - 2 SNP interaction χωρίς θόρυβο

Παρατηρώντας τα αποτελέσματα από τον πίνακα 5.4 παρατηρούμε ότι ο αλγόριθμος Chi-Square έχει επιτυχία 100% εφόσον βρίσκει και τα 6 SNPs και σε uniform και σε μη uniform κατανομή. Θα παρατηρήσουμε επίσης ότι ο αλγόριθμος ανιχνεύει ακόμη 7 λανθασμένα SNPs για uniform κατανομή και 83 λανθασμένα SNPs για μη uniform κατανομή. Οι τιμές του μέσου όρου του p-value και των τυπικών αποκλίσεων από τα ορθά SNPs είναι καλύτερες από ότι στα λανθασμένα.

Ο αλγόριθμος Fisher Exact Test ανιχνεύει και αυτός 6 ορθά SNPs και σε uniform και σε μη uniform κατανομή, ενώ ανιχνεύει 7 λανθασμένα SNPs για uniform κατανομή και 83 λανθασμένα SNPs για μη uniform κατανομή. Παρατηρούμε ότι οι τιμές στο μέσο όρο του p-value και της τυπικής απόκλισης από τα ορθά SNPs είναι καλύτερες από ότι στα λανθασμένα.

Ακολούθως, παρατηρούμε ότι ο αλγόριθμος MDR ανιχνεύει 6 ορθά SNPs, συνεπώς και 100% επιτυχία στα δεδομένα με uniform κατανομή , και δύο ορθά στα δεδομένα με μη uniform κατανομή, άρα 33,3% επιτυχία. Επιπλέον βρίσκει 4 λανθασμένα για uniform κατανομή και 6 λανθασμένα SNPs για μη uniform κατανομή. Επίσης έχουμε καλύτερα αποτελέσματα στις τιμές του μέσου όρου του p-value και της τυπικής απόκλισης για τα ορθά SNPs.

Μελετώντας τον πίνακα 5.5 μπορούμε να δούμε ότι ο αλγόριθμος Regression έχει επιτυχία 100% σε μη uniform κατανομή, άρα ανιχνεύει και τα 6 SNPs, ενώ σε uniform κατανομή έχει μόνο 66,6% επιτυχία, που σημαίνει ότι ανιχνεύει μόνο 4 SNPs. Επίσης βλέπουμε ότι ανιχνεύει 2 λανθασμένα SNPs για uniform κατανομή και 69 λανθασμένα SNPs για μη uniform κατανομή. Από τα αποτελέσματα ξεχωρίζουμε ότι ο αλγόριθμος κατάφερε να επιτύχει σε 2 ορθούς συνδυασμούς 2 SNPs για uniform κατανομή σε αντίθεση με τους 5 ορθούς συνδυασμούς 2 SNPs που ανιχνεύθηκαν στα αποτελέσματα από τα δεδομένα χωρίς uniform κατανομή. Επίσης, βλέπουμε ότι βρίσκει 64 μερικά ορθές αλληλεπιδράσεις στα δεδομένα χωρίς uniform κατανομή και καμία μερική ορθή αλληλεπίδραση στα δεδομένα με uniform κατανομή.

Παράλληλα, παρατηρούμε ότι ο αλγόριθμος BOOST ανιχνεύει 5 ορθά SNPs για δεδομένα χωρίς uniform κατανομή και κανένα ορθό SNP για uniform κατανομή. Επιπλέον βλέπουμε

ότι βρίσκει 2 και 29 λανθασμένα SNP για uniform και μη uniform κατανομή αντίστοιχα. Δεν βρίσκει κανένα ορθό συνδυασμό 2 SNP σε καμία από τις κατηγορίες αλλά ανιχνεύει 5 μερικά ορθές αλληλεπιδράσεις στη μη uniform κατανομή.

Τέλος, ο αλγόριθμος MDR για 2 SNP ανάλυση ανιχνεύει 4 ορθά SNPs και για τις 2 κατηγορίες. Βλέπουμε ότι βρίσκει 4 λανθασμένα SNPs για uniform κατανομή και 10 λανθασμένα SNPs για μη uniform κατανομή. Θα παρατηρήσουμε ότι βρίσκει μια ορθή συσχέτιση SNPs σε μη uniform κατανομή, σε αντίθεση με την uniform κατανομή που ανιχνεύονται 6 ορθές συσχετίσεις 2 SNP. Στην συνέχεια βλέπουμε ότι ανιχνεύονται 4 μερικά ορθές συσχετίσεις για uniform κατανομή και 7 για δεδομένα χωρίς uniform κατανομή. Ο μέσος όρος της μετρικής του MDR είναι καλύτερος στα ορθά SNP για uniform κατανομή ενώ για μη uniform κατανομή θα παρατηρήσουμε ότι ελάχιστα έχουμε καλύτερα αποτελέσματα στα λάθος SNPs. Ωστόσο αν παρατηρήσουμε την τυπική απόκλιση των τιμών θα δούμε ότι έχουμε καλύτερη τιμή για τα ορθά SNPs.

5.6 Αποτελέσματα από δεδομένα με 2 SNP interaction με θόρυβο

Τα δεδομένα στηρίζονται στην συνάρτηση 5.2 και 5.5 με διαφορά ότι προστίθεται θόρυβος με την χρήση της Gaussian συνάρτησης. Οι αλγόριθμοι θα πρέπει να ανιχνεύσουν τα έξι αυτά SNPs για τα δεδομένα με uniform κατανομή: το rs126 αλληλεπιδρά με το rs162, rs239 αλληλεπιδρά με το rs289 και το rs322 αλληλεπιδρά με το rs360 και τα έξι αυτά SNPs για τα δεδομένα χωρίς uniform κατανομή: το rs28576697 αλληλεπιδρά με το rs13303106, rs3128117 αλληλεπιδρά με το rs3737728 και το rs2341354 αλληλεπιδρά με το rs2710888.

Πιο κάτω θα παρουσιαστούν οι δύο πίνακες με τα αποτελέσματα. Ο πίνακας 5.6. αντιστοιχεί στις ίδιες στήλες και γραμμές όπως τον πίνακα 5.4. όπου θα παρουσιαστούν τα αποτελέσματα των αλγορίθμων Chi-Square, Fisher Exact Test και MDR για ανίχνευση ενός SNP από τα δεδομένα 2 SNP interaction με θόρυβο .

Ακολούθως, ο πίνακας 5.7 αντιστοιχεί στις ίδιες στήλες και γραμμές όπως τον πίνακα 5.5. όπου θα παρουσιαστούν τα αποτελέσματα των αλγορίθμων Regression, BOOST και MDR για 2 SNP ανάλυση.

	Chi-Square		Fisher-Exact		MDR	
	U	NU	U	NU	U	NU
Correctly	6	6	6	6	6	2
Incorrectly	10	83	10	82	4	8
Correctly AVG (STDEV)	0.000687 (0.01683)	0 (0.0)	0.000732 (0.001794)	0 (0.0)	0.6141 (0.0701)	0.76525 (0.006859)
Incorrectly AVG (STDEV)	0.02503 0.014443	0.001086 (0.005985)	0.029768 (0.019118)	0.000507 (0.002582)	0.53 (0.0034)	0.7679 (0.0164)

Πίνακας 5.6 Συνέπεια Αποτελεσμάτων - 2 SNP interaction με θόρυβο

	Regression		BOOST		MDR	
	U	NU	U	NU	U	NU
Correctly	4	6	0	5	6	4
Incorrectly	0	68	0	29	3	10
Correct 2 SNPs	2	6	0	0	7	1
Partly Correct 2 SNPs	0	67	0	5	3	6
Correctly AVG(STDEV)	0.0 (0.0)	1.17E-05 (1.68816E-05)	_	37.69288	0.7436 (0.0426)	0.8652 (0.0064)
Incorrectly AVG(STDEV)	_	1.6E-05 (2.45239E-05)	_	44.04789	0.7036 (0.0023)	0.8645 (0.0111)
Total 2 SNPs	2	331	0	33	10	10

Πίνακας 5.7 Συνέπεια Αποτελεσμάτων - 2 SNP interaction χωρίς θόρυβο

Παρατηρώντας τον πίνακα αποτελεσμάτων παρατηρούμε ότι ο αλγόριθμος Chi-Square και Fisher Exact έχουν 100% επιτυχία αφού ανιχνεύουν όλα τα ορθά SNPs. Οι τιμές του μέσου όρου του p-value και η τυπική απόκλιση των ορθών SNPs είναι καλύτερες από ότι στα

λανθασμένα SNPs και σε uniform και σε μη uniform κατανομή. Ο αλγόριθμος Chi-Square έχει ανιχνεύσει ακόμα 10 λανθασμένα SNPs για uniform κατανομή και 83 λανθασμένα SNPs για μη uniform κατανομή, ενώ ο Fisher Exact Test βρίσκει 10 λανθασμένα SNPs για uniform κατανομή και 82 λανθασμένα για μη uniform κατανομή.

Ακολούθως, παρατηρούμε ότι ο αλγόριθμος MDR ανιχνεύει 6 ορθά SNPs, συνεπώς και 100% επιτυχία στα δεδομένα με uniform κατανομή , και 2 ορθά στα δεδομένα με μη uniform κατανομή, άρα 33,3% επιτυχία. Επιπλέον βρίσκει 4 λανθασμένα για uniform κατανομή και 8 λανθασμένα SNPs για μη uniform κατανομή. Επίσης έχουμε καλύτερα αποτελέσματα στις τιμές του μέσου όρου του p-value και της τυπικής απόκλισης για τα ορθά SNPs στα δεδομένα με uniform κατανομή. Στα δεδομένα χωρίς uniform κατανομή βλέπουμε ότι έχουμε ελάχιστα καλύτερα αποτελέσματα στο μέσο όρο της μετρικής στα λανθασμένα SNPs αλλά βλέπουμε ότι η τυπική απόκλιση είναι καλύτερη για τα αποτελέσματα από τα ορθά SNPs.

Μελετώντας τον πίνακα 5.7 μπορούμε να δούμε ότι ο αλγόριθμος Regression έχει επιτυχία 66,6% σε uniform κατανομή, άρα ανιχνεύει μόνο 4 SNPs, ενώ σε μη uniform κατανομή έχει 100% επιτυχία, που σημαίνει ότι ανιχνεύει και τα 6 SNPs. Επίσης βλέπουμε ότι δεν ανιχνεύει κανένα SNP για uniform κατανομή αλλά βρίσκει 68 λανθασμένα SNPs για μη uniform κατανομή. Από τα αποτελέσματα ξεχωρίζουμε ότι ο αλγόριθμος κατάφερε να επιτύχει σε 2 ορθούς συνδυασμούς 2 SNPs για uniform κατανομή σε αντίθεση με τους 6 ορθούς συνδυασμούς 2 SNPs που ανιχνεύθηκαν στα αποτελέσματα από τα δεδομένα χωρίς uniform κατανομή. Επίσης, βλέπουμε ότι βρίσκει 67 μερικά ορθές αλληλεπιδράσεις στα δεδομένα χωρίς uniform κατανομή και καμία μερική ορθή αλληλεπίδραση στα δεδομένα με uniform κατανομή.

Παρατηρούμε ότι ο αλγόριθμος BOOST ανιχνεύει 5 ορθά SNPs για δεδομένα χωρίς uniform κατανομή και κανένα ορθό SNP για uniform κατανομή. Επιπλέον βλέπουμε ότι βρίσκει 29 λανθασμένα SNP για μη uniform κατανομή και κανένα λανθασμένα SNP για uniform κατανομή. Δεν βρίσκει κανένα ορθό συνδυασμό 2 SNP σε καμία από τις κατηγορίες αλλά ανιχνεύει 5 μερικά ορθές αλληλεπιδράσεις στη μη uniform κατανομή.

Τέλος, ο αλγόριθμος MDR για 2 SNP ανάλυση ανιχνεύει 6 ορθά SNPs για uniform κατανομή και 4 ορθά SNPs για μη uniform κατανομή. Βλέπουμε ότι βρίσκει 3 λανθασμένα SNPs για uniform κατανομή και 10 λανθασμένα SNPs για μη uniform κατανομή. Θα παρατηρήσουμε ότι βρίσκει μια ορθή συσχέτιση SNPs σε μη uniform κατανομή, σε αντίθεση με την uniform κατανομή που ανιχνεύονται 7 ορθές συσχετίσεις 2 SNP. Στην συνέχεια βλέπουμε ότι ανιχνεύονται 3 μερικά ορθές συσχετίσεις για uniform κατανομή και 6 για δεδομένα χωρίς uniform κατανομή. Ο μέσος όρος της μετρικής του MDR είναι καλύτερος στα ορθά SNP.

5.7 Αποτελέσματα από δεδομένα με 3 SNP interaction χωρίς θόρυβο

Μελετώντας την συνάρτηση 5.3. παρατηρούμε ότι πρόκειται για μια συνάρτηση πολλαπλασιαστικής επίδρασης πάνω σε τρία SNP. Συγκεκριμένα, διακρίνουμε 6 SNPs: το rs126 αλληλεπιδρά με το rs162, rs239 αλληλεπιδρά με το rs289 και το rs322 αλληλεπιδρά με το rs360.

Παρατηρώντας τώρα την συνάρτηση 5.6. μπορούμε να προσδιορίσουμε ότι πρόκειται για μια συνάρτηση πολλαπλασιαστικής επίδρασης πάνω σε τρία SNP. Εδώ διακρίνουμε 6 SNPs: το rs28576697 αλληλεπιδρά με το rs13303106, rs3128117 αλληλεπιδρά με το rs3737728 και το rs2341354 αλληλεπιδρά με το rs2710888.

Πάρα κάτω παρουσιάζονται τρείς πίνακες αποτελεσμάτων που πήραμε από κάθε αλγόριθμο με τα δεδομένα για 3 SNP interaction χωρίς θόρυβο. Στον πρώτο πίνακα 5.8 αναπαριστούνται τα αποτελέσματα των αλγορίθμων που είναι για univariate SNP ανάλυση, δηλ. Chi Square και Fisher Exact Test, καθώς και τα αποτελέσματα του αλγόριθμου MDR που αναφέρονται για την ανίχνευση ενός SNP. Ο πίνακας αυτός είναι όμοιος με τον πίνακα 5.6.

Ο δεύτερος πίνακας 5.9 συμπεριλαμβάνει τα αποτελέσματα των αλγορίθμων Regression, BOOST και τα αποτελέσματα του αλγορίθμου MDR για 2 SNP ανάλυση. Ο πίνακας αυτός είναι όμοιος με τον πίνακα 5.7

Ο τρίτος πίνακας 5.10 συμπεριλαμβάνει μόνο τα αποτελέσματα του αλγορίθμου MDR εφόσον είναι και ο μόνος που ανιχνεύει multi SNP interactions. Και ο πίνακας αυτός είναι όμοιος με τον πίνακα 5.7.

	Chi-Square		**Fisher-Exact**		**MDR**	
	U	NU	U	NU	U	NU
Correctly	6	5	6	5	6	4
Incorrectly	10	44	10	42	4	6
Correctly AVG (STDEV)	0.0 (0.0)	0 (0)	0.0 (0.0)	0 (0)	0.6304 (0.0152)	0.671625 (0.028374)
Incorrectly AVG (STDEV)	0.024216 (0.011758)	0.009802 (0.014777)	0.0253 (0.0124)	0.008621 (0.013469)	0.528 (0.0017)	0.631217 (0.021067)

Πίνακας 5.8 Συνέπεια Αποτελεσμάτων - 3 SNP interaction χωρίς θόρυβο

	Regression		**BOOST**		**MDR**	
	U	NU	U	NU	U	NU
Correctly	3	4	0	2	6	2
Incorrectly	9	73	0	40	0	12
Correct 2 SNPs	1	1	0	0	10	1
Partly Correct 2 SNPs	1	128	0	2	0	6
Correctly AVG(STDEV)	0.0 (0.0)	0.0 (0.0)	–	–	0.7077 (0.0117)	0.8534 (-)
Incorrectly AVG(STDEV)	0.000043 (-)	1.27E-05 (2.267E-05)	–	45.5660	–	0.7613 (0.2541)
Total 2 SNPs	12	526	0	132	10	10

Πίνακας 5.9 Συνέπεια Αποτελεσμάτων - 3 SNP interaction χωρίς θόρυβο

	MDR	
	U	NU
Correctly	6	5
Incorrectly	0	5
Correct 2 SNPs	10	3
Partly Correct 2 SNPs	0	7
Correctly AVG(STDEV)	0.7683 (0.0102)	0.879133 (0.006062)
Incorrectly AVG(STDEV)	_	0.871557 (0.002597)
Total 2 SNPs	10	10

Πίνακας 5.10 Συνέπεια Αποτελεσμάτων - 3 SNP interaction χωρίς θόρυβο

Παρατηρώντας τα αποτελέσματα από τον πίνακα 5.8 παρατηρούμε ότι ο αλγόριθμος Chi-Square βρίσκει 6 SNPs με ποσοστό επιτυχίας στα 100% με uniform κατανομή και 5 SNPs με ποσοστό επιτυχίας στα 83.3% στα δεδομένα χωρίς uniform κατανομή. Θα παρατηρήσουμε επίσης ότι ο αλγόριθμος ανιχνεύει ακόμη 10 λανθασμένα SNPs στα δεδομένα με uniform κατανομή και 44 SNPs χωρίς uniform κατανομή. Οι τιμές του μέσου όρου του p-value και της τυπικής απόκλισης είναι καλύτερη στα αποτελέσματα των ορθών SNPs.

Ο αλγόριθμος Fisher Exact Test ανιχνεύει και αυτός 6 ορθά SNPs σε δεδομένα με uniform κατανομή και 5 ορθά SNPs σε δεδομένα χωρίς uniform κατανομή με τιμή μέσου όρου του p-value και τυπική απόκλιση και για τις δυο κατηγορίες ίσο με μηδέν. Ωστόσο, ανιχνεύει 10 uniform κατανομή και 42 χωρίς uniform κατανομή λανθασμένα SNPs με πολύ μεγαλύτερες τιμές στο μέσο όρο του p-value και της τυπικής απόκλισης.

Στην συνέχεια, παρατηρούμε ότι ο αλγόριθμος MDR ανιχνεύει 6 ορθά και 4 λανθασμένα SNPs για την ανίχνευση ενός SNP την φορά για uniform κατανομή καθώς επίσης ανιχνεύει 4 ορθά για uniform κατανομή και 6 λανθασμένα SNPs για μη uniform κατανομή.

Παρατηρούμε ότι η μετρική του MDR είναι μεγαλύτερη στα ορθά SNP παρά στα λανθασμένα, κάτι που αναμέναμε.

Παρατηρώντας τον πίνακα 5.9 μπορούμε να δούμε ότι ο αλγόριθμος Regression ανιχνεύει 3 SNPs, άρα έχει επιτυχία 50% για uniform κατανομή και 4 SNPs, άρα έχει επιτυχία 66,6% για μη uniform κατανομή. Βρίσκει επίσης 73 λανθασμένα SNPs σε μη uniform κατανομή ενώ βρίσκει 9 λανθασμένα σε uniform κατανομή. Επιπλέον παρατηρούμε ότι ο αλγόριθμος κατάφερε να επιτύχει σε μόνο σε 1 ορθό συνδυασμό 2 SNPs και στις δυο κατηγορίες. Επίσης, βλέπουμε ότι βρίσκει 128 μερικά ορθές αλληλεπιδράσεις σε μη uniform κατανομή ενώ βρίσκει 1 σε uniform κατανομή. Ο μέσος όρος της μετρικής του Regression είναι καλύτερος στα ορθά SNPs.

Επιπλέον, βλέπουμε ότι ο αλγόριθμος BOOST ανιχνεύει 2 ορθά SNPs και 40 λανθασμένα σε μη uniform κατανομή και δεν βρίσκει κανένα SNP στην uniform κατανομή. Δεν βρίσκει κανένα ορθό συνδυασμό 2 SNP αλλά ανιχνεύει 2 μερικά ορθές αλληλεπιδράσεις.

Ο αλγόριθμος MDR για 2 SNP ανάλυση ανιχνεύει 6 ορθά SNPs και κανένα λανθασμένο για uniform κατανομή ενώ για μη uniform κατανομή βρίσκει 2 ορθά SNPs και 12 λανθασμένα. Βλέπουμε ότι βρίσκει 10 ορθές συσχετίσεις σε uniform κατανομή και μια ορθή συσχέτιση SNPs σε μη uniform κατανομή. Επιπλέον, ανιχνεύει 6 μερικά ορθές αλληλεπιδράσεις σε μη uniform κατανομή. Ο μέσος όρος της μετρικής του MDR είναι σαφώς καλύτερος στα ορθά SNPs.

Τέλος, για τον αλγόριθμο MDR για 3 SNP interaction παρατηρούμε ότι έχει 100% επιτυχία σε uniform κατανομή αφού ανιχνεύει 6 ορθά SNPs και έχει 83,3% επιτυχία σε μη uniform κατανομή αφού ανιχνεύει 5 ορθά SNPs. Επίσης, βλέπουμε ότι μόνο στην μη uniform κατανομή αναγνωρίζει λανθασμένα SNPs, που είναι συνολικά 5. Όπως παρατηρούμε σε uniform κατανομή βρίσκει 10 ορθούς συνδυασμούς από 3 SNP interaction και 3 για μη uniform κατανομή. Βρίσκει επιπλέον 7 μερικά ορθούς συνδυασμούς, δηλαδή συνδυασμούς που περιέχουν 1 ή 2 ορθά SNPs, σε μη uniform κατανομή και κανένα σε uniform κατανομή εφόσον όλοι οι συνδυασμοί που μας παρουσιάστηκαν ήταν πλήρως ορθοί. Ο μέσος όρος της μετρικής του MDR είναι σαφώς καλύτερος στα ορθά SNPs.

5.8 Αποτελέσματα από δεδομένα με 3 SNP interaction με θόρυβο

Τα δεδομένα στηρίζονται στην συνάρτηση 5.3 και 5.6 με διαφορά ότι προστίθεται θόρυβος με την χρήση της Gaussian συνάρτησης. Οι αλγόριθμοι θα πρέπει να ανιχνεύσουν τα έξι αυτά SNPs για τα δεδομένα με uniform κατανομή: το rs126 αλληλεπιδρά με το rs162, rs239 αλληλεπιδρά με το rs289 και το rs322 αλληλεπιδρά με το rs360 και τα έξι αυτά SNPs για τα δεδομένα χωρίς uniform κατανομή: το rs28576697 αλληλεπιδρά με το rs13303106, rs3128117 αλληλεπιδρά με το rs3737728 και το rs2341354 αλληλεπιδρά με το rs2710888.

Πιο κάτω θα παρουσιαστούν οι τρείς πίνακες με τα αποτελέσματα. Ο πίνακας 5.11. αντιστοιχεί στις ίδιες στήλες και γραμμές όπως τον πίνακα 5.8. όπου θα παρουσιαστούν τα αποτελέσματα των αλγορίθμων Chi-Square, Fisher Exact Test και MDR για ανίχνευση ενός SNP από τα δεδομένα 2 SNP interaction με θόρυβο .

Ακολούθως, ο πίνακας 5.12 αντιστοιχεί στις ίδιες στήλες και γραμμές όπως τον πίνακα 5.9. όπου θα παρουσιαστούν τα αποτελέσματα των αλγορίθμων Regression, BOOST και MDR για 2 SNP ανάλυση.

Τέλος, ο πίνακας 5.13 αντιστοιχεί στις ίδιες στήλες και γραμμές με τον πίνακα 5.10 όπου θα παρουσιαστούν τα αποτελέσματα του αλγορίθμου MDR.

	Chi-Square		**Fisher-Exact**		**MDR**	
	U	NU	U	NU	U	NU
Correctly	6	5	6	5	6	5
Incorrectly	10	54	10	54	4	5
Correctly AVG (STDEV)	0.0 (0.0)	0 (0)	0.0 (0.0)	0 (0)	0.6278 (0.0298)	0.6521 (0.0349)
Incorrectly AVG (STDEV)	0.021827 (0.012143)	0.004852 (0.008602)	0.023061 (0.012976)	0.005417 (0.009601)	0.532 (0.0046)	0.629 (0.314)

Πίνακας 5.11 Συνέπεια Αποτελεσμάτων - 3 SNP interaction με θόρυβο

	Regression		BOOST		MDR	
	U	NU	U	NU	U	NU
Correctly	3	5	0	3	6	3
Incorrectly	7	77	0	40	0	10
Correct 2 SNPs	1	2	0	0	10	2
Partly Correct 2 SNPs	1	118	0	3	0	6
Correctly AVG(STDEV)	0.00019 (-)	0.000002 (2.82843E-06)	–	–	0.7132 (0.0172)	0.7716 (0.0713)
Incorrectly AVG(STDEV)	5.95E-05 (3.52E-05)	1E-05 (2.0081E-05)	–	44.5428	–	0.75 (0.0264)
Total 2 SNPs	5	442	0	124	10	10

Πίνακας 5.12 Συνέπεια Αποτελεσμάτων - 3 SNP interaction με θόρυβο

	MDR	
	U	NU
Correctly	6	5
Incorrectly	0	7
Correct 2 SNPs	10	3
Partly Correct 2 SNPs	0	7
Correctly AVG(STDEV)	0.779 (0.0145)	0.8612 (0.0108)
Incorrectly AVG(STDEV)	–	0.8468 (0.0057)
Total 2 SNPs	10	10

Πίνακας 5.13 Συνέπεια Αποτελεσμάτων - 3 SNP interaction με θόρυβο

Παρατηρώντας τα αποτελέσματα από τον πίνακα 5.11 παρατηρούμε ότι ο αλγόριθμος Chi-Square βρίσκει 6 SNPs με ποσοστό επιτυχίας στα 100% με uniform κατανομή και 5 SNPs με ποσοστό επιτυχίας στα 83.3% στα δεδομένα χωρίς uniform κατανομή. Θα παρατηρήσουμε επίσης ότι ο αλγόριθμος ανιχνεύει ακόμη 10 λανθασμένα SNPs στα δεδομένα με uniform κατανομή και 54 SNPs χωρίς uniform κατανομή. Οι τιμές του μέσου όρου του p-value και της τυπικής απόκλισης είναι καλύτερη στα αποτελέσματα των ορθών SNPs.

Ο αλγόριθμος Fisher Exact Test ανιχνεύει και αυτός 6 ορθά SNPs σε δεδομένα με uniform κατανομή και 5 ορθά SNPs σε δεδομένα χωρίς uniform κατανομή με τιμή μέσου όρου του p-value και τυπική απόκλιση και για τις δυο κατηγορίες ίσο με μηδέν. Ωστόσο, ανιχνεύει 10 uniform κατανομή και 54 χωρίς uniform κατανομή λανθασμένα SNPs με πολύ μεγαλύτερες τιμές στο μέσο όρο του p-value και της τυπικής απόκλισης.

Στην συνέχεια, παρατηρούμε ότι ο αλγόριθμος MDR ανιχνεύει 6 ορθά και 5 λανθασμένα SNPs για την ανίχνευση ενός SNP την φορά για uniform κατανομή καθώς επίσης ανιχνεύει 4 ορθά για uniform κατανομή και 5 λανθασμένα SNPs για μη uniform κατανομή. Παρατηρούμε ότι η μετρική του MDR είναι μεγαλύτερη στα ορθά SNP παρά στα λανθασμένα, κάτι που αναμέναμε.

Παρατηρώντας τον πίνακα 5.12 μπορούμε να δούμε ότι ο αλγόριθμος Regression ανιχνεύει 3 SNPs, άρα έχει επιτυχία 50% για uniform κατανομή και 5 SNPs, άρα έχει επιτυχία 83,3% για μη uniform κατανομή. Βρίσκει επίσης 77 λανθασμένα SNPs σε μη uniform κατανομή ενώ βρίσκει 7 λανθασμένα σε uniform κατανομή. Επιπλέον παρατηρούμε ότι ο αλγόριθμος κατάφερε να επιτύχει σε μόνο σε 1 ορθό συνδυασμό 2 SNPs στην uniform κατανομή και 2 ορθούς συνδυασμούς 2 SNPs χωρίς uniform κατανομή. Επίσης, βλέπουμε ότι βρίσκει 118 μερικά ορθές αλληλεπιδράσεις σε μη uniform κατανομή ενώ βρίσκει 1 σε uniform κατανομή. Ο μέσος όρος της μετρικής του Regression είναι καλύτερος στα ορθά SNPs.

Επιπλέον, βλέπουμε ότι ο αλγόριθμος BOOST ανιχνεύει 3 ορθά SNPs και 40 λανθασμένα σε μη uniform κατανομή και δεν βρίσκει κανένα SNP στην uniform κατανομή. Δεν βρίσκει κανέναν ορθό συνδυασμό 2 SNP αλλά ανιχνεύει 3 μερικά ορθές αλληλεπιδράσεις.

Ο αλγόριθμος MDR για 2 SNP ανάλυση ανιχνεύει 6 ορθά SNPs και κανένα λανθασμένο για uniform κατανομή ενώ για μη uniform κατανομή βρίσκει 3 ορθά SNPs και 10 λανθασμένα. Βλέπουμε ότι βρίσκει 10 ορθές συσχετίσεις σε uniform κατανομή και 2 ορθές συσχετίσεις SNPs σε μη uniform κατανομή. Επιπλέον, ανιχνεύει 6 μερικά ορθές αλληλεπιδράσεις σε μη uniform κατανομή. Ο μέσος όρος της μετρικής του MDR είναι σαφώς καλύτερος στα ορθά SNPs.

Τέλος, για τον αλγόριθμο MDR για 3 SNP interaction παρατηρούμε ότι έχει 100% επιτυχία σε uniform κατανομή αφού ανιχνεύει 6 ορθά SNPs και έχει 83,3% επιτυχία σε μη uniform κατανομή αφού ανιχνεύει 5 ορθά SNPs. Επίσης, βλέπουμε ότι μόνο στην μη uniform κατανομή αναγνωρίζει λανθασμένα SNPs, που είναι συνολικά 7. Όπως παρατηρούμε σε uniform κατανομή βρίσκει 10 ορθούς συνδυασμούς από 3 SNP interaction και 3 για μη uniform κατανομή. Βρίσκει επιπλέον 7 μερικά ορθούς συνδυασμούς, σε μη uniform κατανομή και κανένα σε uniform κατανομή εφόσον όλοι οι συνδυασμοί που μας παρουσιάστηκαν ήταν πλήρως ορθοί. Ο μέσος όρος της μετρικής του MDR είναι καλύτερος στα ορθά SNPs.

Κεφάλαιο 6

Συζήτηση

6.1 Σύγκριση Αποτελεσμάτων από SNP με αθροιστική επίδραση

Συγκρίνοντας, τα αποτελέσματα που πήραμε από τα δεδομένα με αθροιστική επίδραση με θόρυβο και χωρίς, θα παρατηρήσουμε ότι για τους αλγορίθμους Chi-Square, Fisher Exact Test και MDR ο αριθμός των SNPs που βρίσκει στα δεδομένα χωρίς θόρυβο είναι μικρότερος σε σχέση με τον αριθμό των SNPs που ανιχνεύτηκαν στα δεδομένα με θόρυβο. Η προσθήκη θορύβου βοήθησε στην ανίχνευση SNP τα οποία προηγουμένως δεν είχαν ανιχνευθεί το οποίο ίσως να δείχνει ότι ο θόρυβος αύξησε το σήμα των συγκεκριμένων SNP.

Παρόλα αυτά θα παρατηρήσουμε ότι η τιμή του μέσου όρου και της τυπικής απόκλισης είναι πολύ μικρότερη στα δεδομένα χωρίς θόρυβο κάτι που υποδεικνύει ότι τα αποτελέσματα που πήραμε από τα δεδομένα χωρίς θόρυβο είναι καλύτερα.

Ο αλγόριθμος Regression παρατηρούμε ότι ανταποκρίνεται καλύτερα στα δεδομένα χωρίς θόρυβο εφόσον ανιχνεύει τέσσερα SNPs ενώ στα δεδομένα με θόρυβο ανιχνεύει δυο SNPs. Παρατηρούμε επίσης ότι η τιμή του μέσου όρου και της τυπικής απόκλισης είναι πολύ μικρότερη στα δεδομένα χωρίς θόρυβο.

Επιπλέον, ο αλγόριθμος BOOST δεν παρουσιάζει κάποια αλλαγή στα αποτελέσματα μεταξύ των δυο κατηγοριών των αποτελεσμάτων.

6.2 Σύγκριση Αποτελεσμάτων από δεδομένα 2 SNP interaction

Εξετάζοντας τα αποτελέσματα από τα δεδομένα 2 SNP interaction χωρίς θόρυβο και τα δεδομένα 2 SNP interaction με θόρυβο για όλους τους αλγορίθμους παίρνουμε σχεδόν όμοια αποτελέσματα.

Στο αλγόριθμο Chi-Square παρατηρούμε ότι και στις δυο κατηγορίες δεδομένων βρίσκει όλα τα ορθά SNPs ενώ παρατηρείται μια μικρή διαφοροποίηση στις τιμές του μέσου όρου και της τυπικής απόκλισης. Στα δεδομένα χωρίς θόρυβο η τιμές αυτές είναι ελάχιστα μεγαλύτερες του μηδενός ενώ στα δεδομένα με θόρυβο είναι μηδέν.

Επιπλέον, μικρή διαφορά στις τιμές του μέσου όρου και της τυπικής απόκλισης παρατηρείτε και στον αλγόριθμο Fisher Exact όπου και πάλι στα δεδομένα με θόρυβο η τιμές αυτές είναι μηδέν ενώ σε δεδομένα χωρίς θόρυβο είναι ελάχιστα μεγαλύτερες.

6.3 Σύγκριση Αποτελεσμάτων από δεδομένα 3 SNP interaction

Τα αποτελέσματα από τα δεδομένα 2 SNP interaction χωρίς θόρυβο και τα δεδομένα 2 SNP interaction με θόρυβο για όλους τους αλγορίθμους παίρνουμε σχεδόν όμοια αποτελέσματα.

Από τον αλγόριθμο MDR παρατηρούμε ότι στα δεδομένα με θόρυβο έχει καλύτερα αποτελέσματα. Αυτό οφείλεται ότι στο γεγονός ότι ο MDR έχει το μειονέκτημα της

υπερεκπαίδευσης και έτσι με λίγο θόρυβο απαλείφουμε την έννοια αυτή και παίρνουμε καλύτερα αποτελέσματα.

6.4 Σύγκριση univariate αλγορίθμων

Συγκρίνοντας τα αποτελέσματα που πήραμε από τους αλγορίθμους Chi-Square και Fisher Exact Test μπορούμε να παρατηρήσουμε ότι τα αποτελέσματα των δύο αλγορίθμων τις περισσότερες φορές είναι παρόμοια. Βλέπουμε ότι στα δεδομένα univariate SNP interaction χωρίς θόρυβο παίρνουμε ελαχίστως καλύτερες τιμές του μέσου όρου και της τυπικής απόκλισης στον αλγόριθμο Fisher Exact ενώ στα δεδομένα 2 SNP interaction χωρίς θόρυβο η τιμές του μέσου όρου και της τυπικής απόκλισης είναι καλύτερες στον αλγόριθμο Chi-Square. Επιπλέον, μπορούμε να προσθέσουμε το γεγονός ότι πρόκειται για δυο γρήγορους αλγορίθμους και αποδοτικούς στην ανίχνευση univariate SNP interaction.

Συμπερασματικά και οι δυο αλγόριθμοι επιτυγχάνουν χωρίς να διακρίνουμε κάποια ιδιαίτερη διαφορά μεταξύ τους.

6.5 Σύγκριση 2 SNP interaction αλγορίθμων

Για την σύγκριση 2 SNP interaction θα παρατεθούν οι αλγόριθμοι Regression, BOOST και MDR.

Από τα δεδομένα univariate SNP interaction χωρίς θόρυβο καλύτερα ανταποκρίθηκε ο αλγόριθμος Regression. Κατάφερε να βρει τα περισσότερα ορθά SNPs, αλλά θα παρατηρήσουμε ότι στα αποτελέσματα του υπάρχει μεγάλος αριθμός από λανθασμένα SNPs. Οι αλγόριθμοι BOOST και MDR ανιχνεύουν μόλις ένα SNP ωστόσο δεν εμφανίζουν μεγάλο αριθμό λανθασμένων SNPs.

Προσθέτοντας θόρυβο στα δεδομένα θα παρατηρήσουμε ότι ο MDR έχει πολύ καλύτερα αποτελέσματα από τους άλλους δυο αλγορίθμους, ο MDR ανιχνεύει τρία SNPs ενώ οι άλλοι δύο αλγόριθμοι από ένα, συνεπώς καταλαβαίνουμε ότι ο MDR είναι λιγότερο επιρρεπής στον θόρυβο και επίσης φαίνεται ότι ο θόρυβος που προσθέσαμε βοήθησε τον αλγόριθμο να ανιχνεύσει περισσότερα SNPs.

Από τα δεδομένα 2 SNP interaction χωρίς θόρυβο και με θόρυβο διακρίνουμε ότι ο αλγόριθμος Regression είναι αυτός που μπορεί να ανιχνεύσει περισσότερα SNPs, και περισσότερους ορθούς συνδυασμούς. Ωστόσο θα παρατηρήσουμε πάλι ότι ο αλγόριθμος Regression βρίσκει πολλούς συνδυασμούς από λανθασμένα SNPs. Μελετώντας τους BOOST και MDR θα δούμε ότι αν και ο BOOST καταφέρνει να βρει περισσότερα ορθά SNPs από ότι ο MDR δεν καταφέρνει να ανιχνεύσει συνδυασμούς ορθών SNPs. Επιπλέον, ο MDR ανιχνεύει περισσότερες μερικά ορθές αλληλεπιδράσεις από ότι ο BOOST.

Κεφάλαιο 7

Συμπεράσματα

7.1 Γενικά

Μέσα από την διπλωματική αυτή εργασία έχει επιτευχθεί η δημιουργία πλατφόρμας για την σύγκριση αλγόριθμων για την ανάλυση συσχετίσεων μεταξύ σημειακών νουκλεοτιδικών πολυμορφισμών και φαινοτύπων. Αρχικά έγινε μελέτη ξεχωριστά για τον κάθε αλγόριθμο, ούτως ώστε να κατανοήσουμε τον τρόπο λειτουργίας τους καθώς επίσης να παρατηρήσουμε τα πλεονεκτήματα και τα μειονεκτήματα τους. Ακολούθως, έγινε η υλοποίηση των εργαλείων και η εισαγωγή τους στην πλατφόρμα Galaxy. Έτσι τρέξαμε για κάθε σύνολο δεδομένων τους αλγορίθμους για να μπορούμε να τους συγκρίνουμε. Τα δεδομένα μας ήταν πλασματικά και έτσι γνωρίζαμε από πριν ποια SNPs θα έπρεπε να ανιχνευθούν σε κάθε αλγόριθμο για κάθε σύνολο δεδομένων.

Χωρίζοντας τα δεδομένα μας σε 3 κατηγορίες : δεδομένα που είναι ανεξάρτητα από το φαινότυπο, δεδομένα σε SNP με αθροιστική επίδραση με ή χωρίς θόρυβο και δεδομένα με 2 SNP interaction με ή χωρίς θόρυβο πήραμε αποτελέσματα από τον κάθε αλγόριθμο. Στην συνέχεια βάση των αποτελεσμάτων αυτών έγιναν οι συγκρίσεις των αλγορίθμων. Από τα αποτελέσματα που προέκυψαν, ήταν εμφανές ότι οι αλγόριθμοι ανταποκρίνονταν θετικά στα σύνολα δεδομένων που τρέξαμε. Εντούτοις, παρατηρήθηκαν μερικά πλεονεκτήματα και μειονεκτήματα που έχει κάθε αλγόριθμος. Οι αλγόριθμοι Chi-Square και Fisher Exact

αν και ανίχνευαν τις περισσότερες φορές τα ορθά SNPs, έβρισκαν και ένα αριθμό λανθασμένων SNPs. Αυτό οφειλόταν εν μέρει στο γεγονός ότι δεν μπορούμε να προσδιορίσουμε ποιο θα είναι το εύρος της τιμής κατωφλίου του p-value ώστε να περιορίσουμε τον αριθμό των λανθασμένων SNPs. Ο αλγόριθμος Regression αν και αναμέναμε να είχε μεγαλύτερο υπολογιστικό κόστος λόγω του ότι εντάσσετε στους αλγορίθμους 2 SNP interaction ήταν αρκετά γρήγορος κάτι που δεν ίσχυε για τους BOOST και MDR. Ωστόσο παρατηρήθηκε ότι οι αλγόριθμοι BOOST και Regression μερικές φορές κατάφερναν να βρουν περισσότερα ορθά SNPs από ότι ο MDR αλλά στον MDR είχαμε την τιμή του cross validation που μας βοηθούσε να δούμε κατά πόσο σημαντικό ήταν το συγκεκριμένο SNP ή η συγκεκριμένη συσχέτιση.

7.2 Μελλοντική Εργασία

Ένας από τους μελλοντικούς στόχους είναι η ενσωμάτωση περισσότερων αλγορίθμων στην πλατφόρμα Galaxy έτσι ώστε να μπορούμε να έχουμε περισσότερα εργαλεία για την συσχέτιση ενός SNP με την ύπαρξη ή όχι μιας ασθένειας όπως για καρκίνο ύπατος, κατά πλάκα σκλήρυνση κλπ.

Παράλληλα, θα εξεταστούν οι αλγόριθμοι και με άλλα δεδομένα έτσι ώστε να μπορούμε να κατανοήσουμε καλύτερα της αδυναμίες του κάθε αλγορίθμου και να εξάγουμε πιο γενικά συμπεράσματα.

Τέλος, μπορούμε να δημιουργήσουμε ένα καινούργιο αλγόριθμο ο οποίος θα επεξεργάζεται τα αποτελέσματα όλων των αλγορίθμων και να μπορεί να παρουσιάζει αυτόματα τις συγκρίσεις που έκανε δίνοντας τες στον χρήστη ώστε να μπορεί να δει την ανταπόκριση των αλγορίθμων για το συγκεκριμένο σύνολο δεδομένων.

Βιβλιογραφία

[1] H. Schwender and K. Ickstadt, “Identification of SNP interactions using logic regression,” *Biostat*, vol. 9, no. 1, pp. 187–198, Jan. 2008.

[2] “Ηλεκτρονική Πύλη του Ασκληπιακού Πάρκου.” [Online]. Available: http://panacea.med.uoa.gr/topic.aspx?id=626.].

[3] “100 Concepts of Biology.” [Online]. Available: http://userpages.umbc.edu/~farabaug/biologycentury/pages/chromo1.html.

[4] Apostolopoulou M, *Grand Larousse: “Έμβιος Κόσμος”*, Εκδοτικός Οίκος “Ελληνικά Γράμματα”, Αθήνα, 2001

[5] Makris I., Matsopoulos G., *“Στοιχεία Βιοπληροφορικής”*, 2009

[6] Aristodimou A., C.S. Pattichis, “Παραλληλοποίηση και βελτιστοποίηση χαρτών αυτό-οργάνωσης και χρήση τους για κατηγοριοποίηση σημειακών νουκλεοτιδικών πολυμορφισμών”, 2010

[7] Katsanos C., Avouris N., “Στατιστικές Μέθοδοι Ανάλυσης Πειραματικών Δεδομένων Συνεργασίας”

[8] Aristodimou A. C.S. Pattichis, “Machine Learning in Genetic Studies”, 2013

[9] Zaggana E. "*Αναγνώριση και κατάταξη ονομάτων σε ελληνικά κείμενα με την χρήση τυχαίων δασων* ",pp. 45-49, October 2010

[10] Xiang Wan, Can Yang, Qiang Yang, Hong Xue, Xiaodan Fan, Nelson L.S. Tang,

and Weichuan Yu1, "BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies", pp 325-333, 2010

[11] Lance W. Hahn, Marylyn D. Ritchie and Jason H. Moore, "Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions", pp 376-381, 2002

[12] Mueller, J.C., Linkage disequilibrium for different scales and applications. Briefings in Bioinformatics, 2004. 5(4): p. 355-364.

[13] "Fisher's exact test," *Wikipedia, the free encyclopedia*",

[14] "Galaxy wiki" [Online]. Available: https://wiki.galaxyproject.org/Learn

[15] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. De Bakker, M. Daly et al., "PLINK: a tool set for whole-genome association and population-based linkage analyses," The American Journal of Human Genetics, vol. 81, no. 3, pp. 559–575, 2007.