ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ

ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ

ΒΙΟΪΑΤΡΙΚΗ»

……. Διασφάλιση Ιδιωτικότητας Δεδομένων σε Ευρυγονιδιωματικές Μελέτες ……

Σιούλας Παναγιώτης - Βλάσιος

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Υπεύθυνος

……Παντελής Μπάγκος……..

Λαμία, 18/07 έτος 2018

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

ΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗ ΒΙΟΙΑΤΡΙΚΗ

ΚΑΤΕΥΘΥΝΣΗ

«ΥΠΟΛΟΓΙΣΤΙΚΗ ΙΑΤΡΙΚΗ ΚΑΙ ΒΙΟΛΟΓΙΑ»

«Υπεύθυνη Δήλωση μη λογοκλοπής και ανάληψης προσωπικής ευθύνης»

Με πλήρη επίγνωση των συνεπειών του νόμου περί πνευματικών δικαιωμάτων, και γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα και ενυπογράφως ότι η παρούσα εργασία με τίτλο [«Διασφάλιση Ιδιωτικότητας Δεδομένων σε Ευρυγονιδιωματικές Μελέτες»] αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές από τις οποίες χρησιμοποίησα δεδομένα, ιδέες, φράσεις, προτάσεις ή λέξεις, είτε επακριβώς (όπως υπάρχουν στο πρωτότυπο ή μεταφρασμένες) είτε με παράφραση, έχουν δηλωθεί κατάλληλα και ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή. Αναλαμβάνω πλήρως, ατομικά και προσωπικά, όλες τις νομικές και διοικητικές συνέπειες που δύναται να προκύψουν στην περίπτωση κατά την οποία αποδειχθεί, διαχρονικά, ότι η εργασία αυτή ή τμήμα της δεν μου ανήκει διότι είναι προϊόν λογοκλοπής.

Ο ΔΗΛΩΝ
ΣΙΟΥΛΑΣ ΠΑΝΑΓΙΩΤΗΣ - ΒΛΑΣΙΟΣ
Ημερομηνία
18/07/2018
Υπογραφή

……. Διασφάλιση Ιδιωτικότητας Δεδομένων σε Ευρυγονιδιωματικές Μελέτες ……

Παναγιώτης – Βλάσιος Σιούλας

Τριμελής Επιτροπή:

Ονοματεπώνυμο, Παντελής Μπάγκος…..(επιβλέπων)

Ονοματεπώνυμο, Βασίλειος Δρακόπουλος

Ονοματεπώνυμο, Χαρίλαος Σανδαλίδης

Επιστημονικός Σύμβουλος:

Ονοματεπώνυμο: Γεώργιος Σπαθούλας

Περιεχόμενα

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιστημονικό μου σύμβουλο και καθηγητή μου Σπαθούλα Γεώργιο για την υπομονή που έδειξε και την αμέριστη βοήθεια που μου παρείχε καθ' όλη τη διάρκεια υλοποίησης της διπλωματικής μου εργασίας. Επίσης, θέλω να ευχαριστήσω τη σύζυγο μου, Δέσποινα, η οποία στήριξε τις σπουδές μου ποικιλοτρόπως όπως και τους γονείς μου οι οποίοι συνεχώς με παροτρύνουν να εξελίσσομαι και να γίνομαι καλύτερος.

Εισαγωγή

Στην παρούσα διπλωματική εργασία αναλύεται και υλοποιείται μία μέθοδος κρυπτογραφικής ανταλλαγής δεδομένων μεταξύ οργανισμών που αφορούν δεδομένα ευρυγονιδιωματικών μελετών. Στο πρώτο μέρος της εργασίας περιγράφονται βασικοί όροι της βιολογίας όπως το DNA, το RNA και η αναφορά καταλήγει στην κατάταξη των πρωτεϊνών. Στη συνέχεια γίνεται αναλυτική περιγραφή του Genome Wide Association καθώς αποτελεί βασική παράμετρο της όλης εργασίας. Στη συνέχεια, αναλύεται η μέθοδος της μετα-ανάλυσης και παρουσιάζονται μερικές βασικές μέθοδοι κρυπτογράφησης που εντοπίστηκαν στη βιβλιογραφία. Στο τελευταίο μέρος παρουσιάζεται η ανάλυση, η υλοποίηση και η δοκιμή του αλγορίθμου που δημιουργήθηκε.

ΚΕΦΑΛΑΙΟ 1ο

1. Το γενετικό υλικό

Το σύνολο των οργανισμών αποτελείται από DNA το οποίο έχει όλες τις εντολές για τον τρόπο με τον οποίο δομείται, λειτουργεί, αναπτύσσεται και μορφοποιείται ένας οργανισμός. Είναι ταυτόχρονα μοναδικό για το κάθε άτομο και αποτελεί ουσιαστικά την ταυτότητα του κάθε ανθρώπου[1]. Το 1869 πραγματοποιήθηκε ο εντοπισμός του DNA αλλά παρέμενε άγνωστο έως το 1944 ότι αποτελεί το γενετικό υλικό των οργανισμών. Οι επιστήμονες μέχρι τότε είχαν την άποψη η γενετική πληροφορία μεταφέρεται μέσω των πρωτεϊνών οι οποίες είναι ποικιλόμορφες για το λόγο ότι είναι το αποτέλεσμα είκοσι αμινοξέων [2][1], σε αντίθεση με τα νουκλεϊκά οξέα [3][2] που είναι αποτέλεσμα συνδυασμού τεσσάρων νουκλεοτιδίων.

Ο Griffith το 1928 [4]σε μία από τις μελέτες του κατέληξε στην ταυτοποίηση του DNA ως γενετικό υλικό χρησιμοποιώντας δύο στελέχη του βακτηρίου Diplococcus pneumoniae (πνευμονιόκκος). Το πρώτο, ονομάζεται λείο γιατί έχει μία κάψα που περιβάλλει το κύτταρο και το προστατεύει μέσα στον ξενιστή, ενώ το δεύτερο αδρό γιατί δεν έχει την κάψα και δεν είναι λοιμογόνο. Κατά την διεξαγωγή των πειραμάτων του ο Griffith σκότωσε με θέρμανση βακτήρια που ήταν λεία και με αυτά μόλυνε το ποντίκι το οποίο παρέμεινε ζωντανό. Ωστόσο, όταν το ποντίκι μολύνθηκε με βακτήρια τα οποία ήταν ζωντανά και ευμεγέθη αλλά και νεκρά λεία τότε το ποντίκι πέθανε και στο αίμα του βρέθηκαν πολλά λεία βακτήρια. Από την όλη διαδικασία συμπέρανε ότι αρκετά ευμεγέθη βακτήρια μετασχηματίστηκαν σε λεία λοιμογόνα εφόσον αλληλοεπέδρασαν με βακτήρια που ήταν λεία και νεκρά. Στο επόμενο στάδιο, υπέθεσε ότι ο παράγοντας μετασχηματισμού είναι κάποιο μόριο των νεκρών βακτηρίων το οποίο θεωρείται υπεύθυνο για την αλλαγή του γενετικού υλικού των αδρών βακτηρίων, αλλά δεν κατέστη δυνατόν να εξηγήσει πως από τα νεκρά λεία βακτήρια

[1]Τα αμινοξέα είναι τα δομικά υλικά όλων των πρωτεϊνών. Διέρχονται από το αίμα στον εγκέφαλο και είναι απαραίτητα για την ανάπτυξη και το μεταβολισμό του. Έχουν ιδιαίτερο ρολό ως νευροδιαβιβαστές και τα επίπεδα τους στον εγκέφαλο πρέπει να ρυθμίζονται προσεκτικά.

[2]Τα νουκλεϊκά οξέα ή νουκλεϊνικά οξέα (πυρηνικά οξέα) είναι σύνθετα βιολογικά μακρομόρια, που αποτελούνται από αλυσίδες νουκλεϊδίων που περιέχουν γενετική πληροφορία. Τα πιο κοινά νουκλεϊκά οξέα είναι το Δεοξυριβονουκλεϊκό οξύ (DNA) και το Ριβονουκλεϊκό οξύ (RNA). Τα νουκλεϊκά οξέα υπάρχουν στα κύτταρα όλων των έμβιων οργανισμών.

δημιουργήθηκαν ζωντανά λεία και ποια ήταν η αιτία στην οποία οφειλόταν αυτή η αλλαγή.

Επιβεβαιώθηκε οριστικά το 1952 ότι το DNA αποτελεί το γενετικό υλικό των οργανισμών με τα κλασικά πειράματα των Hershey και Chase οι οποίοι εξέτασαν το πόσο διαρκεί ο κύκλος ζωής ενός βακτηριοφάγου (φάγου) T_2. Οι ερευνητές χρησιμοποίησαν τη σήμανση χημικών μορίων με τη χρήση ραδιενεργού ^{35}S, που ενσωματώνεται μόνο στις πρωτεΐνες αλλά όχι στο DNA, και με ραδιενεργό ^{32}P, που ενσωματώνεται αποκλειστικά στο DNA κι όχι στις πρωτεΐνες προκειμένου να ιχνηθετήσουν τους φάγους. Αμέσως μετά τα βακτήρια μολύνθηκαν με ραδιενεργούς φάγους. Η διαδικασία είχε ως αποτέλεσμα να εντοπιστεί ότι 'οι απαραίτητες εντολές' για τον πολλαπλασιασμό και την παραγωγή των καινούριων φάγων δίνονται εφόσον DNA των φάγων εισέρχεται στα βακτηριακά κύτταρα [5].

1.1.Η Λειτουργία του DNA

Πρόκειται για μία ένωση μεγάλων μορίων που αποτελείται από φωσφορικές ρίζες αζωτούχες - πρωτεϊνικές βάσεις, και την δε(σ)οξυριβόζη που είναι σάκχαρο που αποτελείται από πεντόζη (πέντε άτομα άνθρακα). Βρίσκεται στον πυρήνα των ευκαρυωτικών κυττάρων αλλά και σε κάποια άλλα οργανίδια (π.χ. στα πλαστίδια και τα μιτοχόνδρια) και τους δίνει τη δυνατότητα αυτονομίας στην αναπαραγωγή (ημιαυτόνομα οργανίδια).

Το DNA φέρει τις απαραίτητες οδηγίες προκειμένου ένας οργανισμός να αναπτυχθεί, να επιβιώσει και να αναπαραχθεί. Για να πραγματοποιηθούν οι παραπάνω λειτουργίες, θα πρέπει να χρησιμοποιηθούν οι αλληλουχίες DNA ώστε να γίνουν μηνύματα και να παράγουν πρωτεΐνες, οι οποίες αποτελούν πολύπλοκα μόρια που πραγματοποιούν πολλές λειτουργίες στο σώμα μας..

Κάθε αλληλουχία DNA που φέρει οδηγίες για την σύνθεση μίας πρωτεΐνης ονομάζεται γονίδιο. Το μέγεθός του μπορεί να διαφέρει και να κυμαίνεται από περίπου χίλιες βάσεις έως ένα εκατομμύριο βάσεις. Το ένα τοις εκατό (1%) της αλληλουχίας DNA μόνο αποτελείται από τα γονίδια. Εκτός από αυτό το 1%, οι αλληλουχίες DNA συμμετέχουν στο χρόνο, την ποσότητα και τον τρόπο που μία πρωτεΐνη δημιουργείται.

Όπως αναφέρθηκε παραπάνω, το γενετικό υλικό ενός κυττάρου συνίσταται στο σύνολο των μορίων DNA. Οι γενετικές πληροφορίες του κυττάρου που μεταφέρονται μέσω του DNA δεν αφορούν μόνο τη μεταβίβαση ιδιοτήτων, αμετάβλητων από γενιά σε γενιά, αλλά και τον τρόπο που ρυθμίζεται η μορφή εξειδίκευσης κάθε κυττάρου για

την διενέργεια των συγκεκριμένων λειτουργιών του. Συνεπώς, η δημιουργία γενετικής ποικιλότητας επιτρέπεται εφόσον το DNA υποστεί μεταλλάξεις [5].

1.2. Δομή του DNA

Η διαμόρφωση των μεγάλων μορίων του DNA στο χώρο έχει τη μορφή δύο επιμηκών αλυσίδων, οι οποίες συστρέφονται σε έλικα. Υπάρχουν τέσσερις αζωτούχες βάσεις στη σύνθεση του DNA :

- κυτοσίνη (C)
- θυμίνη (T)
- αδενίνη (A)
- γουανίνη (G)

Ανάλογα με την σειρά εναλλαγής τους σε τριάδες, οι αζωτούχες βάσεις έχουν τον ρόλο της κωδικοποίησης του μηνύματος που αφορά τον τρόπο που συντίθεται τα αμινοξέα των κυττάρων στα ριβοσώματα. Τα αμινοξέα συνδυάζονται στα ριβοσώματα με τη σειρά κατά την οποία μεταφέρθηκαν σε αυτά και σχηματίζονται οι διαφορετικές πρωτεΐνες.

1.3.Η διπλή έλικα του DNA

Το 1953 ένα ‘μοντέλο’ της δομής του DNA, που ονομάστηκε ‘μοντέλο της διπλής έλικας’’[6] παρουσιάστηκε από τους Τζέιμς Γουάτσον (J. Watson), και Φράνσις Κρίκ, (F. Crick)[6], δύο ερευνητές από τη Μ. Βρετανία που εργάζονταν στο Πανεπιστήμιο του Cambridge.

Η διμερής χημική δομή του DNA περιγράφεται με τον όρο ‘διπλή έλικα’. Η εξαιρετική ακρίβεια μεταβίβασης βιολογικών οδηγιών οφείλεται στο σχήμα αυτό της στριμμένης σκάλας. Προκειμένου να γίνει κατανοητή η διπλή έλικα από χημικής άποψης , οι δύο πλευρές της σκάλας απεικονίζονται ως σκέλη εναλλασσόμενης ζάχαρης και φωσφορικών ομάδων που τρέχουν σε αντίθετες κατευθύνσεις. Δύο βάσεις αζώτου, συνδυασμένες με δεσμούς υδρογόνου συνθέτουν κάθε ‘σκαλοπάτι’ της σκάλας. Εξαιτίας της εξαιρετικά ορισμένης φύσης αυτού του τύπου χημικού ζευγαρώματος, η Αδενίνη πάντα ζευγαρώνει με τη Θυμίνη κι αντίστοιχα η Κυτοσίνη με τη Γουανίνη. Συνεπώς, εφόσον γνωρίζουμε την ακολουθία των βάσεων στο ένα σκέλος της διπλής έλικας, είναι εύκολο να βρούμε την ακολουθία των βάσεων στο άλλο.

Κατά τη στιγμή της διαίρεσης του κυττάρου η σύνθεση του DNA επιτρέπει σε αυτό να αντιγραφεί. Κατά τη διάρκεια της κυτταρικής διαίρεσης, η έλικα του DNA σπάει στη μέση και μετατρέπεται σε δύο μονές αλυσίδες. Αυτές με τη σειρά τους χρησιμεύουν ως πρότυπα για την κατασκευή δύο νέων μορίων DNA διπλής έλικας που είναι ένα αντίγραφο του αρχικού μορίου DNA. Κατά τη διάρκεια αυτής της διαδικασίας, μία βάση A προστίθεται σε μία T, μία C σε μία G, μέχρι τη στιγμή που όλες οι βάσεις να έχουν ζευγαρωθεί και πάλι.

Επιπλέον, η διπλή έλικα ξετυλίγεται για να επιτρέψει ένα μοναδικό κλώνο DNA να χρησιμεύσει ως πρότυπο όταν συντίθενται οι πρωτεΐνες. Αυτός ο κλώνος-πρότυπο μεταγράφεται στη συνέχεια σε mRNA. Το mRNA είναι ένα μόριο που μεταφέρει ουσιώδεις οδηγίες για τη δημιουργία των πρωτεϊνών του κυττάρου. Στην ουσία μεταφέρει τη γενετική πληροφορία από το DNA στα ριβοσώματα για τη σύνθεση πρωτεϊνών των κυττάρων. Συνεπώς, το DNA και το RNA αποτελούν το γενετικό υλικό των οργανισμών [5].

1.4. Είδη RNA

Το RNA αποτελείται από πέντε επιμέρους είδη, το πρώτο είδος είναι το Αγγελιοφόρο RNA ή mRNA (Messenger RNA) το οποίο λειτουργεί ως μεταφορέας της γενετικής πληροφορίας από το DNA στα σημεία των ριβοσωμάτων για την πρωτεΐνοσύνθεση των κυττάρων. Στα ευκαρυωτικά κύτταρα το mRNA μεταγράφεται πριν από το DNA και στη συνέχεια γίνεται η επεξεργασία πριν βγει από τον πυρήνα στο κυτταρόπλασμα όπου συνδέεται με τα ριβοσώματα και με τη συνδρομή του tRNA (μεταφορικό RNA) μεταφράζεται στην αντίστοιχη πρωτεΐνη. Το Μεταφορικό RNA ή tRNA (TransferRNA) το οποίο είναι μία μικρή αλυσίδα RNA περίπου 74-95 νουκλεοτιδίων που μεταφέρει ειδικά αμινοξέα στα ριβοσώματα του κυττάρου όπου γίνεται η πρωτεΐνοσύνθεση κατά τη διάρκεια της μετάφρασης του κυττάρου σε μια πολυπεπτιδική αλυσίδα που επεκτείνεται. Ουσιαστικά είναι ένας τύπος μη κωδικοποιούμενου RNA. Το τρίτο είδος ονομάζεται Ριβοσωμικό RNA ή rRNA (Ribosomal RNA) που είναι ένας τύπος RNA των ριβοσωμάτων που καταστρέφει την πρωτεΐνοσύνθεση στο κύτταρο. Η επόμενη κατηγορία ονομάζεται Μη κωδικοποιητικό RNA ή (Non-coding RNA). Είναι γονίδια που κωδικοποιούν RNA που δεν μεταφράζονται σε πρωτεΐνη. Τα πιο χαρακτηριστικά είδη είναι το μεταφορικό RNA (tRNA) και το ριβοσωμικό RNA (rRNA), που διαδραματίζουν ουσιώδη ρόλο στην

μετάφραση. Το τελευταίο είδος ονομάζεται Καταλυτικό RNA ή (Catalytic RNA) το οποίο εμποδίζει μια χημική αντίδραση καθώς ορισμένα αμινοξέα έχουν την δυνατότητα να καταλύουν ειδικές χημικές αντιδράσεις που συμβαίνουν στο κύτταρο.

Οι πρωτεΐνες είναι μακρομόρια διαδεδομένα και πολυδιάστατα και στη λειτουργία και στη μορφή τους. Πληθώρα πρωτεϊνών μπορούν να βρεθούν σε ένα κύτταρο και να παίξουν ένα συγκεκριμένο ρόλο. Συνεπώς, μπορούν να αποτελούν είτε το βασικό συστατικό της δομής του κυττάρου είτε συνδράμουν σε κάποια εξειδικευμένη λειτουργία. Είναι μεγάλα σύνθετα βιομόρια που έχουν μοριακό βάρος από 10.000 έως πάνω από 1 εκατομμύριο αμινοξέα, τα οποία ενώνονται μεταξύ τους με πεπτιδικούς δεσμούς σχηματίζοντας μία γραμμική αλυσίδα, η οποία ονομάζεται αλυσίδα πολυπεπτιδίων. Όλες οι πρωτεΐνες περιέχουν οξυγόνο, άζωτο και άνθρακα κι οι περισσότερες εξ αυτών και θείο. Ένα γονίδιο καθορίζει την ακολουθία των αμινοξέων στις πρωτεΐνες και κωδικοποιείται σύμφωνα με το DNA . Παρότι ο γενετικός κώδικας επιτρέπει την κωδικοποίηση 20 αμινοξέων, που συνθέτουν την πρωτεΐνη, συχνά υφίστανται χημικές αλλαγές κατά τη μεταγραφική κωδικοποίηση: είτε πριν δοθεί στην πρωτεΐνη η δυνατότητα να λειτουργήσει στο κύτταρο είτε ως τμήμα των μηχανισμών ελέγχου. Περισσότερες από μία πρωτεΐνες συχνά συνεργάζονται προκειμένου να επιτύχουν μία συγκεκριμένη λειτουργία, ή μπορεί ακόμα και να συσσωματωθούν για να διαμορφώσουν τα στάδια Σχηματισμός-οργάνωση. Ένα τέτοιο παράδειγμα αποτελεί η αιμοσφαιρίνη.

1.5. Κατάταξη Πρωτεϊνών

Με βάση τη μορφή τους διακρίνουμε τις πρωτεΐνες σε ινώδεις και σε σφαιρικές. Παράλληλα, όσον αφορά τη σύνθεσή τους διακρίνονται σε απλές (όταν αποτελούνται μόνο από αμινοξέα) και σε σύνθετες (όταν στο μόριό τους περιλαμβάνονται και μη πρωτεϊνικά τμήματα όπως μέταλλα, σάκχαρα, λίπη κ.λπ.). Επιπροσθέτως, με κριτήριο τη λειτουργία τους χωρίζονται σε δομικές (όταν αποτελούν τα δομικά υλικά του κυττάρου), και λειτουργικές (όταν συμβάλλουν σε κάποιες λειτουργίες).[7]

Οι διαφορετικές λειτουργίες που έχουν παρατηρηθεί στους οργανισμούς οφείλονται στις πρωτεΐνες. Η τρισδιάστατη δομή τους που είναι συνέπεια της αλληλουχίας των αμινοξέων, καθορίζει το βιολογικό τους ρόλο. Ομοίως με άλλα βιολογικά μακρομόρια (π.χ. οι πολυσακχαρίτες, τα λιπίδια, και τα νουκλεικά οξέα) οι πρωτεΐνες είναι απαραίτητες για όλους τους ζωντανούς οργανισμούς και συμμετέχουν σε κάθε διαδικασία εντός των κυττάρων. Μία μερίδα πρωτεϊνών χρησιμεύουν ως ένζυμα που καταστρέφουν τις βιοχημικές αντιδράσεις και είναι σημαντικές στο μεταβολισμό. Παράλληλα, άλλες πρωτεΐνες έχουν μηχανικές ή δομικές λειτουργίες, όπως για παράδειγμα οι πρωτεΐνες του σκελετού των κυττάρων που βοηθούν στη διατήρηση της μορφής των κυττάρων. Οι πρωτεΐνες είναι εξίσου σημαντικές στον σχηματισμό κυτταρικών ιστών, την επικοινωνία μεταξύ των κυττάρων, τη δράση του ανοσοποιητικού συστήματος και στον κυτταρικό κύκλο. Αποτελούν απαραίτητα συστατικά στη διατροφή μας, δεδομένου ότι τα ζώα δεν μπορούν να συνθέσουν όλα τα αμινοξέα, αλλά πρέπει να τα λάβουν από τα τρόφιμα. Μέσω της διαδικασίας της πέψης, τα ζώα αποικοδομούν την πρωτεΐνη στα ελεύθερα αμινοξέα που μπορούν να χρησιμοποιηθούν για την πρωτεϊνική σύνθεση.

Συμβολοσειρά στην επιστήμη της πληροφορικής καλείται μία σειρά διαδοχικών συμβόλων τα οποία αποτελούν τα στοιχεία ενός πεπερασμένου συνόλου. Κάθε συμβολοσειρά διαφέρει ως προς το περιεχόμενο της σε αναλογία με τα στοιχεία που διαθέτει το αλφάβητο[8]. Όταν το αλφάβητο περιορίζεται σε αριθμούς και γράμματα, η συμβολοσειρά καλείται και αλφαριθμητικό. Στον προγραμματισμό στην πλειονότητα των περιπτώσεων τον ορισμό συμβολοσειρά τον χρησιμοποιούμε όταν επιθυμούμε να αναφερθούμε σε έναν τύπο δεδομένων με τον οποίο ορίζονται ακολουθίες χαρακτήρων. Οι χαρακτήρες αναπαρίστανται με μία συγκεκριμένη κωδικοποίηση χαρακτήρων.

2. Genome Wide AssociationΣυσχέτιση ολόκληρου του γονιδιώματος

2.1 Γενετική ποικιλομορφία

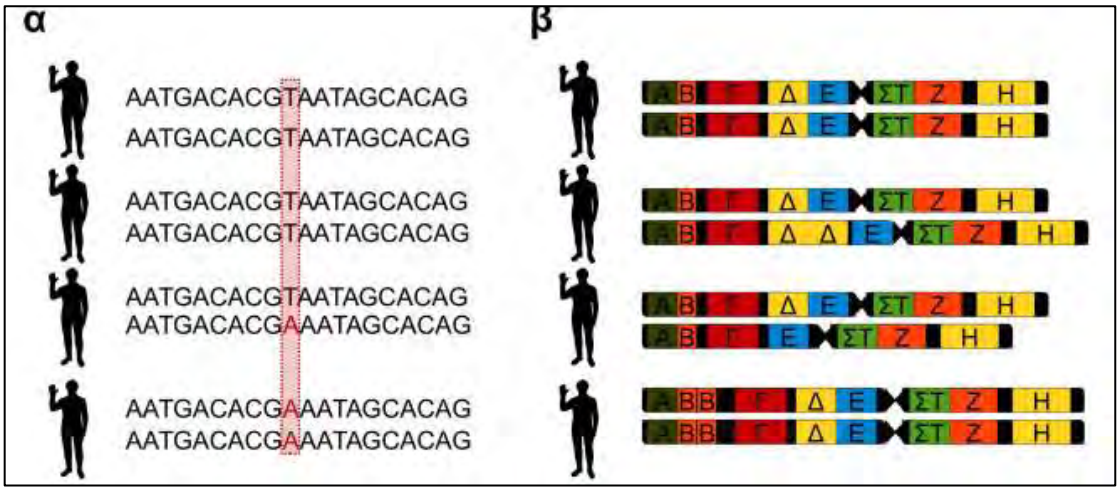
Οργανισμοί όπως ο άνθρωπος και η πλειοψηφία των θηλαστικών είναι διπλοειδείς καθώς στο γονιδίωμά τους φέρουν δύο αντίγραφα από κάθε χρωμόσωμα. Συνεπώς, δύο ομόλογα σημεία του γονιδιώματος ενυπάρχουν σε κάθε γονιδιωματικό στοιχείο [5, 9]. Η γενετική ποικιλομορφία οφείλεται ακριβώς στο γεγονός ότι δεν είναι πανομοιότυπα τα δύο ομόλογα στοιχεία. Ωστόσο, αν κανείς εξετάσει δύο διαφορετικά άτομα, το ίδιο γονιδιακό χαρακτηριστικό μπορεί να υπάρχει στο ίδιο γονιδίωμα σε περισσότερες από δύο εκδοχές. Σε γονιδιακό επίπεδο, αλληλόμορφα (alleles) καλούνται τα γονίδια με περισσότερες από μία μορφές. Όταν τα δύο αντίγραφα του ίδιου γενετικού τόπου είναι διαφορετικά μεταξύ των χρωμοσωμάτων του ίδιου οργανισμού, λέμε ότι ο οργανισμός είναι ‘ετερόζυγος’ για το συγκεκριμένο γενετικό τόπο, ενώ στην περίπτωση που είναι όμοια ο οργανισμός είναι ‘ομόζυγος’.

Η δυνατότητα μελέτης της ποικιλομορφίας με ευρύτερη ικανότητα διάκρισης μέσω των προηγμένων μεθόδων γονιδιωματικής ανάλυσης μας οδηγούν να στην επέκταση των όρων ομοζυγωτίας και ετεροζυγωτίας σε επίπεδο μοναδικών νουκλεοτιδίων [10] επαναπροσδιορίζοντας την προσοχή από το επίπεδο των αλληλόμορφων γονιδίων και του γενετικού τόπου. Ο σημειακός νουκλεοτιδικός πολυμορφισμός (single nucleotide polymorphisms SNP) αντιστοιχεί σε μία μοναδική θέση στο απλοειδές γονιδίωμα που διαφέρει μεταξύ των ατόμων του πληθυσμού[11]. Σε αυτή την περίπτωση, η θέση αυτή καλείται πολυμορφική και κατά συνέπεια ένα άτομο μπορεί να είναι είτε ετεροζυγώτης είτε ομοζυγώτης σε σχέση με τη συγκεκριμένη θέση. Με τη χρήση του όρου SNP αναφερόμαστε τόσο στις σημειακές νουκλεοτιδικές αντικαταστάσεις και στις μονονουκλεοτιδικές ενθέσεις και απαλοιφές [12].

Σε αυτές τις περιπτώσεις αναφερόμαστε με τον όρο ‘δομική ποικιλομορφία’ (structural variation) [13] με την ποικιλομορφία αριθμού αντιγράφων (copy number variation ή CNV) να αποτελεί χαρακτηριστικό παράδειγμα. Η πλειονότητα των πολυμορφικών θέσεων έχουν δύο αλληλόμορφα παρότι οι δυνατότητες για ένα SNP είναι μεγαλύτερες από δύο. Στην Εικόνα 1 φαίνονται τα δύο βασικά είδη γενετικής

ποικιλομορφίας στους πληθυσμούς. Εκτός των σημειακών πολυμορφισμών, οι μεγαλύτερες σε μήκος μεταβολές μπορούν να οδηγήσουν σε γενετική ποικιλομορφία.

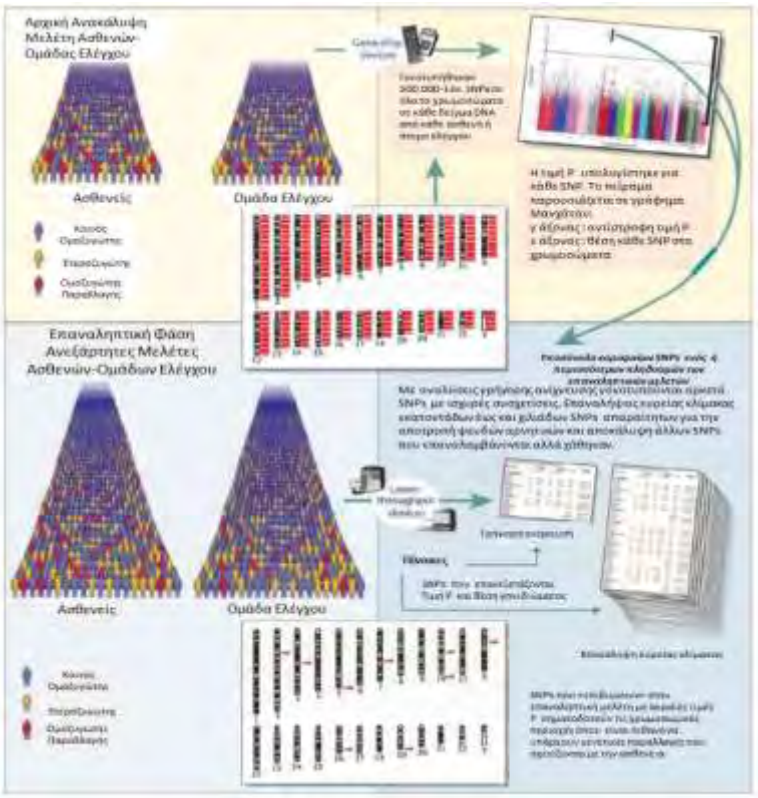
a) Σημειακοί πολυμορφισμοί (Single nucleotide polymorphisms, SNP). Στο κόκκινο πλαίσιο η θέση διαφέρει μεταξύ των ατόμων του πληθυσμό, η πλειοψηφία των οποίων είναι ομόζυγοι (T:T) ενώ κάποια φέρουν έναν πολυμορφισμό (A) είτε σε ετεροζυγωτία (T:A) ή σε ομοζυγωτία (A:A).
b) Ποικιλομορφία αριθμού αντιγράφων (Copy number variation, CNV). Μεγαλύτερες περιοχές του γονιδιώματος (από μερικές βάσεις έως μερικές χιλιάδες βάσεις) μπορούν να διπλασιαστούν ή να απαλείφουν σε ομοζυγωτία ή σε ετεροζυγωτία. [14]



Εικόνα 1: Τα δύο βασικά είδη γενετικής ποικιλομορφίας στους πληθυσμούς

2.2 Μελέτη συσχετισμού ολόκληρου του γονιδιώματος (GWAS)

Η έρευνα προκειμένου να γίνει συσχετισμός του γονιδιώματος (GWAS) αναλύει και υπολογίζει τις εναλλακτικές στην αλληλουχία του DNA στον άνθρωπο για να αναγνωριστούν οι γενετικοί παράγοντες που μπορούν να οδηγήσουν σε ασθένειες συνηθισμένες για τον άνθρωπο.



Εικόνα 2:Μελέτη συσχετισμού ολόκληρου του γονιδιώματος (GWAS)

Η ύπαρξη της GWAS σχετίζεται με τον στόχο να χρησιμοποιηθούν οι γενετικοί παράγοντες κινδύνου εμφάνισης μίας ασθένειας για να προβλέψουμε την επικινδυνότητα και πιθανότητα από άνθρωπο σε άνθρωπο και να συντελεστεί προσπάθεια προσδιορισμού των θεμελίων της ευαισθησίας στη νόσο προκειμένου να γίνει προσπάθεια για ανάπτυξη στρατηγικών πρόληψης και θεραπείας [15, 16]. Η GWAS είναι μία μη κατευθυνόμενη από υποψήφια γονίδια (non-candidate - driven) μελέτη, αυτό σημαίνει πρακτικά ότι ακόμη κι αν εντοπίσει SNPs στο DNA που έχουν σχέση με μία ασθένεια, δεν είναι εφικτό να εντοπίσει συγκεκριμένα ποια είναι τα γονίδια που ευθύνονται ως αιτίες για την ασθένεια.

Η υλοποίηση της ιδέα της GWAS πραγματοποιήθηκε μέσω της λειτουργίας των βιοτραπεζών του προγράμματος HapMap. Από το 2003 αυτό το πρόγραμμα συνέβαλε στην ταυτοποίηση του μεγαλύτερου μέρους των συνηθέστερων SNPs που πραγματεύεται μία GWAS και τέλος η ανάπτυξη γονοτυπικών μεθόδων αυτών των SNPs με συστοιχίες συνέδραμε στην κατεύθυνση αυτή [17] (π.χ. Πλατφόρμα DMET).

Το 2005 δημοσιεύτηκε η πρώτη επιτυχημένη GWAS που αφορούσε έρευνα στους ασθενείς με εκφύλιση της ώχρας κηλίδας συσχετιζόμενη με την ηλικία. Εφόσον πραγματοποιήθηκε σύγκριση με δείγματα ελέγχου που ήταν υγιή, εντοπίστηκαν δύο SNPs που έφεραν ολοκληρωτικά διαφορετικές συχνότητες αλληλόμορφων [18]. Τα

τελευταία 7 χρόνια, ξεκινώντας από το 2011 διενεργήθηκαν εξετάσεις σε εκατοντάδες ή ακόμα και χιλιάδες άτομα και πάνω από 1200 GWAS εξέτασαν περισσότερες από 200 ασθένειες και χαρακτηριστικά. Αξίζει να σημειωθεί ότι βρέθηκαν περίπου 4000 συσχετίσεις SNP[19].

2.3 Περιορισμοί και προβλήματα GWAS

Αρκετά προβλήματα αλλά και περιορισμοί φαίνεται να παρουσιάζονται στις μελέτες GWAS τα οποία μπορούν να αποφευχθούν ή να αντιμετωπιστούν επαρκώς με σχεδιασμό μελέτης και έλεγχο ποιότητας. Κάποια από τα προβλήματα που χρήζουν αντιμετώπισης είναι η έλλειψη ομάδων ασθενών αυστηρά καθορισμένων και ελέγχου αυτών, ο έλεγχος της πληθυσμιακής διαστρωμάτωσης, το ανεπαρκές μέγεθος δείγματος όπως και ο έλεγχος των πολλαπλών δοκιμών [16]. Παράλληλα, μπορεί να δημιουργηθούν προβλήματα εξαιτίας της ίδιας της φύσης της προσέγγισης καθώς ο αριθμός των στατιστικών δοκιμασιών είναι μεγάλος και για το λόγο αυτό παρουσιάζουν εξαιρετικά μεγάλες πιθανότητες για θετικά αποτελέσματα που τελικά αποδεικνύονται ψευδή [16]. Παρόλα αυτά, τα παραπάνω προβλήματα φαίνεται να είναι σχετικά εύκολα στην επίλυσή τους. Υπάρχουν , ωστόσο, και κάποια άλλα πιο λεπτά και εξίσου σημαντικά θέματα είναι δυνατό να προκύψουν. Εξαιτίας, συνεπώς των προαναφερθέντων περιορισμών και δυσκολιών υπάρχει έντονη η πεποίθηση ότι ίσως δεν αξίζουν τα έξοδα προκειμένου να υλοποιηθούν οι μελέτες αυτές [20] και για το λόγο αυτό προτείνονται εναλλακτικές στρατηγικές. Για παράδειγμα η κλιμακούμενη μείωση της τιμής αλληλούχισης ολόκληρου του γονιδιώματος (Εικόνα 2) παρέχει μία εναλλακτική των GWAS καθώς βασίζονται στις γονοτυπικές συστοιχίες που ανταποκρίνεται στην πραγματικότητα.

Year	1990	2002	2006	2008	2012	2014
Cost	$3 billion	$300 million	$20 million	$2 million	$5,000	$1,000

Εικόνα 3:Η ραγδαία μείωση τιμής αλληλουχίας ολόκληρου του γονιδιώματος στη διάρκεια των χρόνων

2.4 Μελέτες Συσχέτισης Ολόκληρου του Γονιδιώματος

Μία μελέτη συσχέτισης ολόκληρου του γονιδιώματος (Genome-Wide Association Study - GWAS) είναι μία συγκριτική προσέγγιση που περιλαμβάνει τη σύγκριση μέσω της σάρωσης δεικτών του DNA (για παράδειγμα ≈0,5 εκατομμύριο ή ένα εκατομμύριο) σε σχέση με γονιδιώματα από πολλά άτομα (π.χ. χιλιάδες ασθενείς και χιλιάδες άτομα

ελέγχου) σε μία προσπάθεια εύρεσης παραλλαγών που σχετίζονται με την γενετική και αφορούν μία ασθένεια (genetic variations) [21].

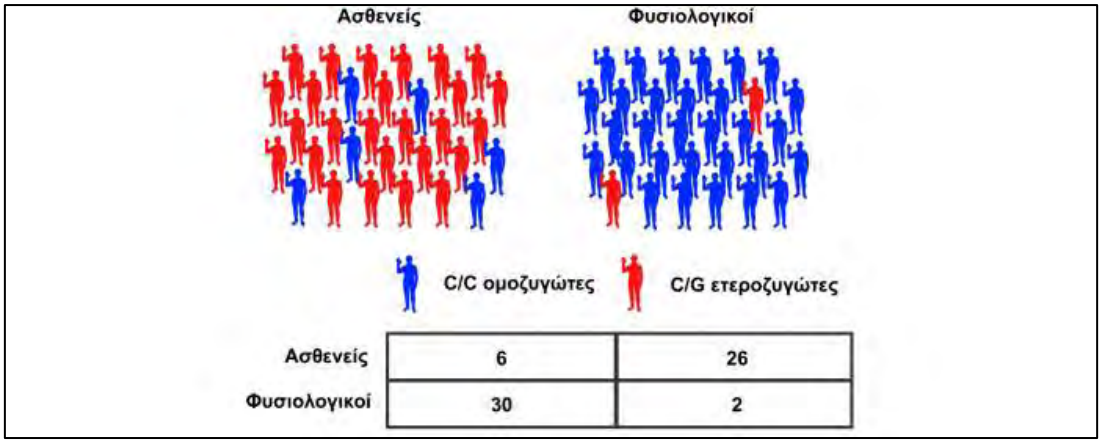
Η βάση των μελετών συσχέτισης ολόκληρου του γονιδιώματος (Genome-Wide Association Studies - GWAS) είναι η «κοινή παραλλαγή-κοινή ασθένεια» (common disease-common variant). Αυτό πρακτικά σημαίνει ότι συγκεκριμένες παραλλαγές του DNA είναι συχνά υπεύθυνες για τις κοινές ασθένειες που παρατηρούνται σε ένα πληθυσμό. Πρέπει στο σημείο αυτό να τονιστεί ότι οι μελέτες αυτές καθίστανται δυνατό να διενεργηθούν λόγω της γονιδιωματικής δομής αλλά και της ιδιότητας ανισορροπίας σύνδεσης των παραλλαγών του DNA (μη τυχαία συσχέτιση αλληλόμορφων)[22].

2.4.1. Μελέτες συσχέτισης σε οικογενειακό επίπεδο

Οι μελέτες συσχέτισης που πραγματοποιούνται σε επίπεδο οικογένειας ερευνούν την ποικιλομορφία του γονιδιώματος σε συγκεκριμένο αριθμό ατόμων. Το μεγαλύτερο όφελος είναι ότι χρησιμοποιούν την συνηθισμένη απλοτυπική διάταξη που είναι αναμενόμενο να υφίσταται μεταξύ αυτών που αποτελούν μέλη της ίδιας οικογένειας. Βασικό μειονέκτημά τους είναι ότι ο εντοπισμός συσχετίσεων μεταξύ πολύπλοκων φαινοτύπων και γονοτύπου διενεργείται με δυσκολία. Στον αντίποδα, γονοτυπικές αλλαγές που έχουν πολύ μεγάλη επίδραση στο φαινότυπο μελετώνται κατά προτίμηση μεταξύ των μελών της ίδιας οικογένειας [23] και κατά συνέπεια αναμένεται να είναι σπανιότερες.

2.4.2. Μελέτες ασθενών-μαρτύρων (case-control studies)

Οι μελέτες ασθενών μαρτύρων βασίζονται στο διαχωρισμό του δείγματος σε δύο κατηγορίες με βάση ένα φαινοτυπικό χαρακτηριστικό δυαδικού τύπου (π.χ. ασθενείς-υγιείς)[24] και είναι το είδος GWAS που πραγματοποιείται συχνότερα. Συχνά, ο εναλλακτικός χαρακτήρας του φαινοτύπου δεν προσδιορίζεται παρά μόνο κατά προσέγγιση. Για παράδειγμα, ένας ασθενής διαφέρει σε σχέση με έναν υγιή οργανισμό σε σχέση με τα επίπεδα έκφρασης μίας πρωτεΐνης κι όχι στην έκφραση ή μη-έκφρασή της. Παρόλα αυτά, όταν είμαστε γνώστες του προς μελέτη συστήματος, συμβάλλει στον σημαντικό περιορισμό τους είδους της μελέτης. Σε περιπτώσεις κατά τις οποίες οι ασθένειες είναι πολύπλοκες ή αποτελούν συνθήκες που αναμένεται να εξαρτώνται από περισσότερα από ένα χαρακτηριστικά, η δυαδική προσέγγιση είναι αυτή που επιλέγεται αναγκαστικά.



Εικόνα 4:Γραφική αναπαράσταση μίας γενετικής ανάλυσης ασθενών-μαρτύρων (case-controlstudy).

Σε περιπτώσεις κατά τις οποίες ο γενετικός χαρακτήρας που επηρεάζει το φαινότυπο είναι γνωστός προσπαθούμε να προσδιορίσουμε την ποσοτική τους σχέση και κατά συνέπεια ο σχεδιασμός της μελέτης αλλάζει. Σε αυτή την περίπτωση, έχουμε ποσοτική μελέτη (quantitative study design). Στις περιπτώσεις αυτές, διευκολύνεται η ανάλυση των δειγμάτων καθώς χρειάζεται μικρότερης εμβέλειας γονοτύπηση καθώς ο γενετικός τόπος ενδιαφέροντος είναι ήδη γνωστός. Από την άλλη πλευρά, μεγαλύτερο εύρος δειγμάτων απαιτείται για την εξαγωγή ποσοτικών συσχετίσεων έτσι ώστε να φτάσουμε σε ικανοποιητική στατιστική ισχύ [25].

Το σύνολο των μελετών ασθενών-μαρτύρων έχει ως κοινό χαρακτηριστικό την ανάγκη να τυποποιηθούν σε ικανοποιητικό βαθμό τα φαινοτυπικά κριτήρια. Η λανθασμένη απόδοση ατόμων μεταξύ των κατηγοριών είναι ένα από το κυριότερα εμπόδια που ανακύπτουν κατά τη διάρκεια της διαδικασίας ανάλυσης των δεδομένων καθώς συχνά ο χαρακτηρισμός ενός ασθενούς γίνεται με μη ικανοποιητικά κριτήρια. Εξειδικευμένο ιατρικό προσωπικό είναι επιφορτισμένο να σταθμίσει όλα τα απαραίτητα κριτήρια προκειμένου να υπάρξει αυστηρός έλεγχος στον καθορισμό του δείγματος στην περίπτωση που οι ασθενείς κι οι μάρτυρες έχουν διαφορετικά κέντρα αναφοράς ή προέρχονται από διαφορετικές χώρες.

2.4.3. Στοιχεία γενετικής ανάλυσης

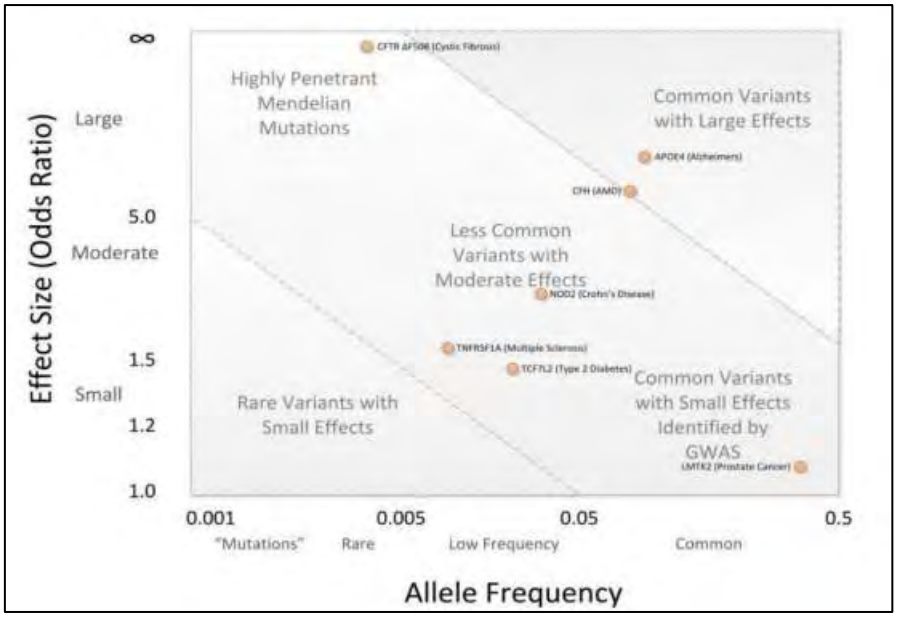
Τα δεδομένα σχετικών συχνοτήτων αλληλόμορφων των διαφορετικών κατηγοριών του δείγματος εξετάζονται σε μία γενετική ανάλυση μεγάλης κλίμακας. Άρα, σε μία μελέτη GWAS ασθενών-μαρτύρων με την οποία διερευνάται ένα εκατομμύριο SNP, τα τελικά αποτελέσματα αναπαρίστανται με τη μορφή πίνακα 2X106 τιμών συχνοτήτων για κάθε πολυμορφική, διαλληλική θέση. Υπάρχουν δύο επίπεδα ανάλυσης :

1. Στην εκτίμηση της στατιστικής σημασίας των παρατηρούμενων διαφορών μεταξύ των δύο κατηγοριών του δείγματος.
2. Στην εκτίμηση της διεισδυτικότητας (penetrance) κάθε πολυμορφισμού, του βαθμού δηλαδή στον οποίο ο πολυμορφισμός είναι επιδραστικός για τον παρατηρούμενο φαινότυπο.

2.4.4. Συσχέτιση σπανιότητας αλληλόμορφων και ασθενειών

Η υπόθεση της συσχέτισης σπανιότητας αλληλόμορφων και ασθενειών (rare disease -rare variant hypothesis) περιγράφεται στην Εικόνα 5 [26]. Σύμφωνα με τη θεωρία της σπανιότητας αλληλόμορφων , υπάρχει συνολικά μία αντίστροφη σχέση μεταξύ της συχνότητας ενός αλληλόμορφου και της πληθυσμιακής του διεισδυτικότητας [22]. Αυτό το αποτέλεσμα είναι αναμενόμενο καθώς ένα αλληλόμορφο που έχει πολύ μεγάλη επιρροή είναι απαραίτητο εξ ορισμού να είναι σπάνιο εξαιτίας του μειονεκτήματος που φέρει σε εξελεγκτικό επίπεδο ως προς την μετάδοσή του από τους φορείς του. Τα πολύ σπάνια διεισδυτικά αλληλόμορφα θεωρούνται χαρακτηριστικά μεντελιανών μονογονιδιακών παθήσεων όπως η κυστική ίνωση.

Αντίστοιχα, στο μέσο της Εικόνας 5 βρίσκονται αλληλόμορφα με σπανιότερη εμφάνιση και κατά συνέπεια με ενδιάμεση επιρροή στο φαινότυπο. Στο επίπεδο αυτό μέσω GWAS με ευκολία μπορούν να διενεργηθούν συσχετίσεις. Παρόλα αυτά οι περιπτώσεις που φαίνονται στο δεξί μέρος του διαγράμματος χρήζουν μεγαλύτερης προσοχής. Αυτές οι περιπτώσεις αντιστοιχούν σε αλληλόμορφα με μεγάλη συχνότητα που όμως έχουν μικρή διεισδυτικότητα και είναι αποτελούν ιδιότητες πολυπαραγοντικών νόσων στις οποίες δεν μπορούμε να αποδώσουμε γενετική προδιάθεση σε ένα μοναδικό γενετικό τόπο, αλλά αποδίδεται σε μεγαλύτερο αριθμό γενετικών ιδιοτήτων [27].



Εικόνα 5: Διάγραμμα της σχέσης μεταξύ διεισδυτικότητας (ως odds-ratio) και της συχνότητας αλληλόμορφων στον πληθυσμό

Ακόμα και με πολύ μικρό βαθμό διεισδυτικότητας προσεγγίσεις GWAS με μεγάλο μέγεθος δείγματος μπορούν να προσδιορίσουν αυτές τις περιπτώσεις. Μέσω GWAS προσεγγίσεων γενικά μπορούμε σήμερα να εντοπίσουμε γονοτυπικές φαινοτυπικές συσχετίσεις για συχνότητες αλληλόμορφων που είναι μεγαλύτερες από 0.05 ακόμα κι αν ο βαθμός διεισδυτικότητας είναι ελάχιστα >1. Αντίστροφα, εφόσον η διεισδυτικότητά τους είναι μεγάλη, μπορούν να εντοπιστούν ακόμα και πολύ σπάνια αλληλόμορφα. Τα επίπεδα αποτελεσματικότητας των GWAS ορίζονται από τις διακεκομμένες διαγώνιες γραμμές που φαίνονται στην Εικόνα 5. Η προσπάθεια που καταβάλλεται είναι να διευρυνθεί όσο το δυνατόν περισσότερο η περιοχή μεταξύ των δύο αυτών γραμμών με την ολοένα και μεγαλύτερη αύξηση της διακριτικής ικανότητας και (κυρίως) του μεγέθους των δειγμάτων.

3. Μετα – ανάλυση

Η εξέλιξη της επιστήμης σε συνδυασμό με την ανάπτυξη της τεχνολογίας έχει ως αποτέλεσμα την αύξηση και συνάμα την πρόσβαση σε τεράστιο όγκο δεδομένων. Η διαχείριση ενός τέτοιου όγκου πληροφορίας είναι αρκετά δύσκολη και έχει οδηγήσει επιστήμονες και ερευνητές στην αναζήτηση μεθόδων που έχουν ως στόχο την καλύτερη διαχείριση αλλά συνάμα και το καλύτερο φιλτράρισμα των δεδομένων. Για λόγους αξιοπιστίας δημιουργήθηκαν κανόνες συγγραφής ερευνητικών εργασιών και γενικότερης επιστημονικής αρθρογραφίας. Την αξιολόγηση των δημοσιεύσεων έχουν βοηθήσει οι ανασκοπήσειςστις οποίες οι συγγραφείς συλλέγουν όλες τις μελέτες με ένα συγκεκριμένο αντικείμενο ενισχύοντας τα πρωταρχικά αποτελέσματα.

Οι ανασκοπήσεις χωρίζονται στις περιγραφικές και στις συστηματικές. Η πρώτη ασχολείται με θεωρητικές απόψεις σχετικά με την νέα και την ήδη υπάρχουσα γνώση ερευνητικών προτάσεων ενώ η δεύτερη συλλέγει κυρίως ποσοτικά και διακριτά χαρακτηριστικά μελετών και με υπόβαθρο την στατιστική θεωρία προσπαθεί να βγάλει στατιστικά σημαντικά αποτελέσματα με σαφή τεκμηρίωση. Τα βήματα της επιστημονικής μεθοδολογίας, τα οποία χρησιμοποιούν μαθηματική απόδειξη για τα αποτελέσματά τους στη συστηματική ανασκόπηση ονομάζονται μετα-ανάλυση[28].

Με την μετα-ανάλυση γίνεται η ενοποίηση και η στατιστική ανάλυση δεδομένων προερχόμενων από διαφορετικές έρευνες οι οποίες προκύπτουν από τυχαιοποιημένες κλινικές δοκιμές. Η μετα-ανάλυση αποτελεί ένα χρήσιμο εργαλείο για την διεξαγωγή μελετών σε διάφορους επιστημονικούς κλάδους για την διατύπωση και διασταύρωση ενός συνολικού συμπεράσματος ανάμεσα από πληθώρα αντιφατικών μελετών και μη.

3.1. Βασικές αρχές συστηματικής ανασκόπησης

Η επιστημονική κοινότητα έχει καταλήξει στα παρακάτω επτά βήματα που αποτελούν τη συστηματική ανασκόπηση.

1. Διατύπωση επιστημονικής υπόθεσης.
2. Αναζήτηση βιβλιογραφίας.
3. Καθορισμός κριτηρίων επιλογής και απόρριψης μελετών
4. Αξιολόγηση και καθορισμός των μελετών που εμπίπτουν στα προηγούμενα βήματα.
5. Καταγραφή και σύνθεση όλων των δεδομένων
6. Στατιστική ανάλυση

7. Παρουσίαση και ερμηνεία των αποτελεσμάτων

Από το 5ο βήμα ακολουθεί η μετα-ανάλυση στη συστηματική ανασκόπησηγια την οποία διευκρινίζεται ότι δεν πρέπει να συγχέεται με την ανασκόπηση της βιβλιογραφίας. Πριν την μετα-ανάλυση θα πρέπει να γίνει έλεγχος του συστηματικού σφάλματος δημοσίευσης που μειώνει τη εγκυρότητά της και επί τούτου θα είναι καλό να είναι ενήμερος ο μελετητής. Μία άλλη παράμετρος, η οποία λαμβάνεται υπόψη, είναι αν ο πληθυσμός όλων των μελετών είναι ομοιογενής ή ετερογενής, επειδή αυτό θα συντελέσει στην επιλογή του καταλληλότερου μοντέλου διεξαγωγής της μετα-ανάλυσης.

3.2. Μέγεθος επίδρασης (size effect)

Το μέγεθος επίδρασης (size effect) είναι ένα μέγεθος που προσδιορίζει την ένταση της σχέσης μεταξύ δυο μεταβλητών ή διαφορετικά μία τυποποιημένη εκτίμηση του μεγέθους επίδρασης της έκθεσης και του αποτελέσματος[29]. Η συσχέτιση μίας ασθένειας με ένα παράγοντα γίνεται συνήθως με το odd, το οποίο ορίζεται ως η πιθανότητα να συμβεί ένα γεγονός προς την πιθανότητα να μην συμβεί δηλαδή:

$$\frac{p}{1-p}, \text{όπου } p \text{ η πιθανότητα επιτυχίας}$$

Για την αξιολόγηση της συσχέτισης γονιδίων-ασθενειών, οι επιστήμονες συλλέγουν πληροφορίες σχετικά με τον κίνδυνο της νόσου σε συνδυασμό διαφορετικών γονοτύπων. Υπάρχουν τουλάχιστον τρεις πιθανοί γονότυποι (δύο ομόζυγοι και ένας ετερόζυγος). Οι συγκρίσεις μεταξύ των γονοτύπων συχνά ελαττώνονται σε ένα συγκεκριμένο γενετικό μοντέλο επικρατές και υπολειπόμενο[30].

Στην παρούσα εργασία θεωρούμε ότι έχουμε δείγματα δύο ομάδων ανθρώπων (ασθενείς-υγιείς) με δύο πιθανές καταστάσεις αυτής της μη μετάλλαξης και της μετάλλαξης όπως φαίνεται στον παρακάτω πίνακα.

	Ασθενείς	Υγιείς
Μετάλλαξη	α	β
Όχι μετάλλαξη	γ	δ

Πίνακας 1: Πιθανές καταστάσεις ασθενών

Οι συνδυασμοί που μπορούν να προκύψουν είναι ο α να είναι ασθενής και να έχει μετάλλαξη, ο β να είναι υγιής και να έχει μετάλλαξη, ο γ να είναι ασθενής και να μην έχει μετάλλαξη κι ο συνδυασμός δ που είναι υγιής και δεν έχει μετάλλαξη. Έχοντας ως οδηγό τα παραπάνω, ο στόχος της παρούσας εργασίας είναι ο υπολογισμός δεδομένων

τα οποία προκύπτουν μέσω τύπων που θα δούμε στη συνέχεια και η αξιοποίησή τους χωρίς, όμως, αυτά τα δεδομένα να αποκαλύπτονται σε τρίτους. Πιο αναλυτικά, έστω ότι έχουμε Ν οργανισμούς οι οποίοι τοπικά υπολογίζουν κάποια δεδομένα, όταν αυτά τα δεδομένα θα πρέπει να χρησιμοποιηθούν «δημόσια» και να αλληλοεπιδράσουν με δεδομένα άλλων οργανισμών πρέπει να υπάρχει η διασφάλιση ότι δεν θα γίνουν γνωστά σε τρίτους. Αυτός είναι ο δεύτερος στόχος της διπλωματικής.

Η βασική διαδικασία στη μετα-ανάλυση είναι η αντιμετώπιση n μελετών από τις οποίες υπολογίζεται μία κοινή παράμετρος ενδιαφέροντος θ_i $(i = 1, \dots, n)$. Στην περίπτωση της ομοιογένειας, το στατιστικό μοντέλο που πρέπει να χρησιμοποιηθεί για τον συνδυασμό των μελετών και την εξαγωγή του αποτελέσματος είναι το μοντέλο σταθερών επιδράσεων(fixed-effect model)[29, 31]. Ένα μοντέλο σταθερών επιδράσεων προϋποθέτει ότι όλα τα δείγματα Y_i από κάθεμελέτη προέρχονται από έναν ενιαίο πληθυσμό. Υποθέτουμε ότι κοινή παράμετρος ενδιαφέροντος είναι η θ, ότι έχουμε 1,2 …n ανεξάρτητες μελέτες και ότι το Y_iείναι τέτοιο ώστε $E(Y_i) = \theta$ και η διακύμανση από κάθε μελέτη${s_i}^2 = var\ Y_i$. Για μελέτες μεγάλου μεγέθους, κάθε Y_i (δείγματα κάθε μελέτης) πρέπει να ακολουθούν ασυμπτωτικά τηνκανονική κατανομή.Στη συνέχεια αποτυπώνονται όλοι οι τύποι που χρησιμοποιούνται για τον υπολογισμό των απαραίτητων μεγεθών που χρησιμοποιούνται[32]

$$Y_i = \log OR_i = \log\left(\frac{\alpha\delta}{\beta\gamma}\right), s_i = \sqrt{\frac{1}{\alpha}+\frac{1}{\beta}+\frac{1}{\gamma}+\frac{1}{\delta}}$$

$$\widehat{\theta}_{MLE} = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i} \text{με } W_i = \frac{1}{s_i^2}$$

Ένα δεύτερο στατιστικό μοντέλο που χρησιμοποιείται για τον υπολογισμό δεδομένων είναι το μοντέλο τυχαίων επιδράσεων (Random Effects Model), στο οποίο η μεταβλητότητα του αποτελέσματος οφείλεται τόσο στη μεταβλητότητα που παρουσιάζει η κάθε μελέτη εξαιτίας της χρήσης διαφορετικών «δειγμάτων» πληθυσμού όσο και στη μεταβλητότητα μεταξύ των διαφόρων μελετών. Στο μοντέλο αυτό είναι δυνατή η γενίκευση των αποτελεσμάτων. Το μοντέλο τυχαίων επιδράσεων προϋποθέτει ότι τα δείγματα που συμπεριλαμβάνονται στην μετα-ανάλυση προέρχονται από μία διανομή πληθυσμού με μέγεθος επίδρασηςθ_i και διακύμανσηs_i^2. Κάθεθ_iαπό κάθε μελέτη υποθέτουμε ότι προέρχεται από ανεξάρτητο τυχαίο δείγμα από

ένα φυσιολογικό πληθυσμό [31, 33] με μέση τιμή θ και τυπική απόκλιση τ2 με τύπο:$\theta_i \sim N(\theta, \tau^2)$ όπου θ και τ2 αναφέρονται ως υπερπαράμετροι που αντιπροσωπεύουν το κοινό μέγεθος επίδρασης και την διακύμανση αντίστοιχα. Οι τύποι που χρησιμοποιούνται είναι οι παρακάτω:[32]

$Y_i \mid \theta_i, s_i^2$

$\hat{\theta}_{MLE} = \frac{\Sigma_i W_i(\tau) Y_i}{\Sigma_k W_i(\tau)}$ με $W_i(\tau) = \frac{1}{s_i^2 + \tau^2}$

η παράμετρος T^2είναι η διακύμανση μεταξύ των μελετών.

Όπου $T^2 = \frac{Q - df}{C}$

Όπου Q = $\sum_{i=1}^{k} W_i Y_i - \frac{\left(\sum_{i=1}^{k} W_i Y_i\right)^2}{\sum_{i=1}^{k} W_i}$,

$df = k - 1$όπου κείναι ο αριθμός των μελετών

$$C = \sum W_i - \frac{\sum w_i^2}{\sum W_i}$$

Στην εργασία εφαρμόζουμε την πρώτη μέθοδο, ωστόσο βασιζόμενοι στους παραπάνω τύπους με πολύ μικρές αλλαγές στην κωδικοποίηση μπορούμε να εφαρμόσουμε και τη δεύτερη μέθοδο.

4. Μέθοδοι κρυπτογράφησης

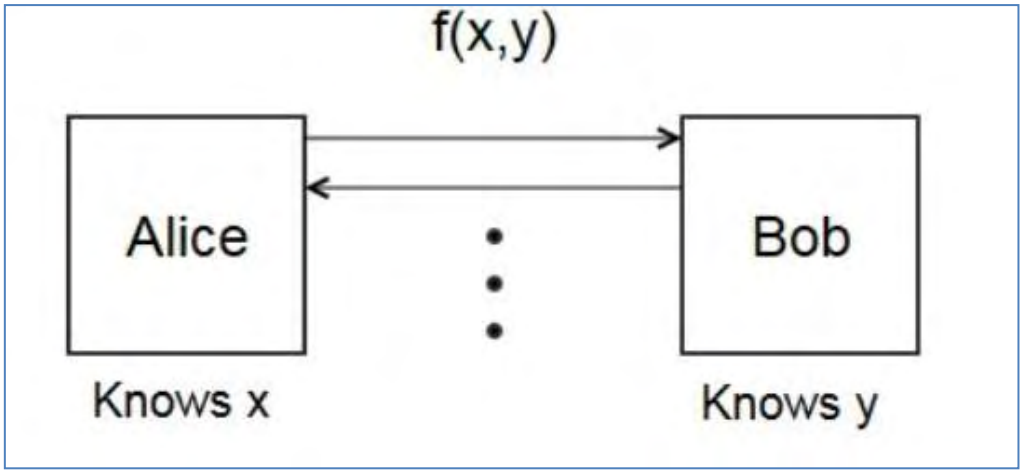
Η ανάγκη για ασφαλέστερη αποθήκευση και μετάδοση της πληροφορίας μέσω των ηλεκτρονικών δικτύων έχει γίνει επιτακτική την τελευταία δεκαετία. Για το λόγο αυτό η κρυπτογράφηση και η χρησιμοποίηση ανθεκτικών υλικών στα φυσικά μέσα αποθήκευσης έχει καταστεί αναγκαία προκειμένου να πραγματοποιείται ασφαλής πρόσβαση και αποθήκευση στα δεδομένα. Παρόλα αυτά, το βασικό πρόβλημα στην αυξημένη ζήτηση υπολογισμού και μεταφοράς ιδιωτικών δεδομένων ή στην περίπτωση που είναι αναγκαία η αλλαγή των αλγορίθμων που ήδη υπάρχουν σε βαθμό τέτοιο που δεν θα επηρεαστεί η λειτουργία τους αλλά την ίδια στιγμή θα εξασφαλίζεται η ιδιωτικότητα.

4.1. Ασφαλής υπολογισμός (Secure Computation)

Η μέθοδος του Ασφαλούς Υπολογισμού (Secure Computation) αφορά στην κατάληξη σε ένα αποτέλεσμα σε ένα ερώτημα με την προϋπόθεση ότι αυτό θα επιτευχθεί με το μεγαλύτερο δυνατό βαθμό ιδιωτικότητας. Για παράδειγμα, όταν διενεργείται μία ηλεκτρονική ψηφοφορία τα βασικά ζητούμενα είναι το αποτέλεσμά της να είναι διασφαλισμένο και η επιλογή του κάθε ψηφοφόρου να μην μπορεί να αποκαλυφθεί σε τρίτους. Ένα άλλο παράδειγμα αποτελεί μία ηλεκτρονική δημοπρασία κατά την οποία ένας πλειοδότης δεν θα πρέπει να γνωρίζει την προσφορά κάποιου άλλου.

4.1.1.Περιγραφή Προβλήματος

Ας υποθέσουμε ότι η ο Bob και η Alice είναι δύο συμμετέχοντες που έχουν στην κατοχή τους ιδιωτικές πληροφορίες. Έστω ότι η Alice είναι X και ο Bob είναι Y. Επιθυμούν μέσω της χρήσης της συνάρτησης $f(X,Y)$ να ανταλλάξουν πληροφορίες δίχως τα δεδομένα που ο καθένας κατέχει να αποκαλυφθούν.

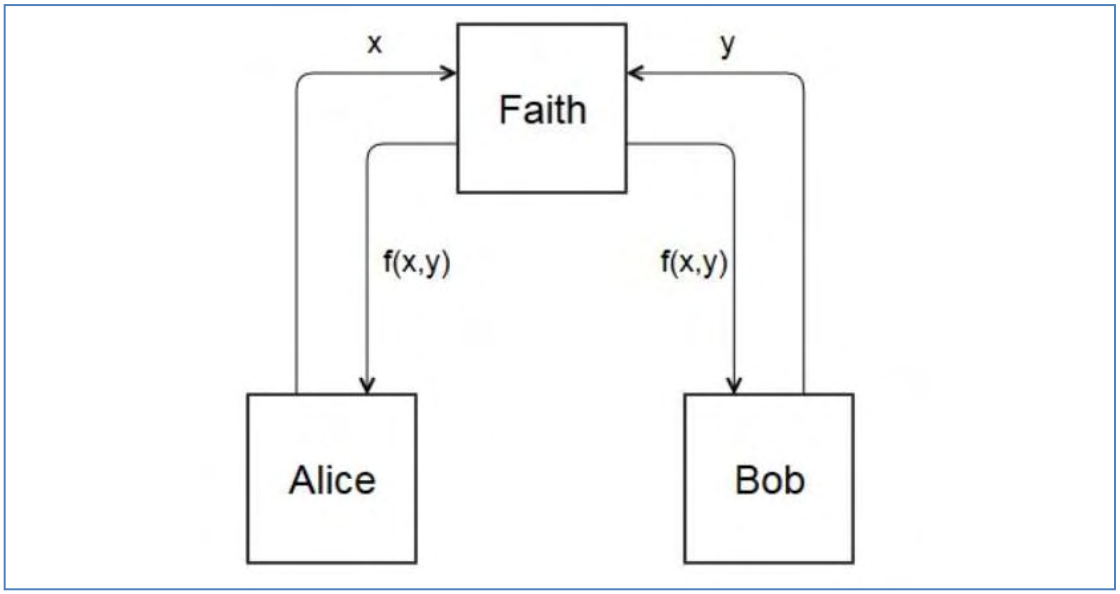


Εικόνα 6:Σχηματική αναπαράσταση της εξίσωσης

Στη συνέχεια θα αναλυθεί το πρόβλημα του εκατομμυριούχου. Η Alice και ο Bob επιθυμούν χωρίς να αποκαλύψουν το ποσό που κατέχει ο καθένας να μάθουν ποιος έχει περισσότερα χρήματα. Συνεπώς, ορίζουμε ως X τον πλούτο της Alice και Y τον πλούτο του Bob. Οι διαφορετικές περιπτώσεις που μπορούν να προκύψουν φαίνονται παρακάτω και εξαιτίας αυτών μπορούμε να μάθουμε ποιος είναι ο πιο πλούσιος

$$f(X,Y) = \begin{cases} Alice\ \epsilon\acute{\alpha}\nu\ X > Y \\ Bob\ \epsilon\acute{\alpha}\nu\ X < Y \\ \text{Ίδιος πλούτος } X = Y \end{cases}$$

Το απαιτητικό στοιχείο στην περίπτωση αυτή είναι ότι δεν μπορεί να υπάρξει αμοιβαία εμπιστοσύνη άρα και να δημιουργηθεί ένα ασφαλές και παράλληλα αξιόπιστο πρωτόκολλο. Κατά συνέπεια, οφείλουμε να βρούμε έναν εναλλακτικό τρόπο κατά τον οποίο τα δύο μέλη δεν θα χρειάζεται να επικοινωνήσουν προκειμένου να αποστείλουν τις πληροφορίες. Διευρύνοντας το αρχικό μοντέλο μπορούμε να υποθέσουμε ότι υπάρχει ένα τρίτο μέρος που φέρει το όνομα Faith το οποίο υπολογίζει το αποτέλεσμα της εξίσωσης $f(X,Y)$. Στην περίπτωση αυτή, η Alice στέλνει το X στη Faith και ο Bob στέλνει το Y. Ο ρόλος της Faith είναι να υπολογίζει το αποτέλεσμα $f(X,Y)$ και να το στέλνει στα δύο μέρη. Η διαδικασία αυτή είναι ιδανική καθώς περιλαμβάνεται στην επικοινωνία ένα τρίτο μέλος το οποίο διενεργεί υπολογισμούς χωρίς, ωστόσο, να αποκαλύψει τα δεδομένα που εισάγονται.



Εικόνα 7:Σχηματική αναπαράσταση ιδανικού σεναρίου

Παρόλα αυτά, ακόμη και στο παραπάνω ιδανικό σενάριο ανταλλαγής πληροφορίας μπορεί να υπάρξει διαρροή και ένα ή και τα δύο μέλη να μπορέσουν τελικά να έχουν πρόσβαση σε πληροφορίες που αφορούν τον άλλο.

Πιο συγκεκριμένα, το αποτέλεσμα της εξίσωσης $f(X,Y)$ έχει υπολογιστεί από τη Faith. Κατά τη λήξη της επικοινωνίας, η Alice γνωρίζει $το\ X, και\ το\ X+Y$ ενώ ο Bob γνωρίζει $το\ Y, και\ το\ X+Y$. Εάν η Alice κάνει την πράξη $(X+Y)-X$ θα έχει τη δυνατότητα να μάθει την τιμή του Y. Αντίστοιχα, αν ο Bob κάνει την πράξη $(X+Y)-Y$ μπορεί να έχει πρόσβαση στην τιμή του X. Άρα, ακόμα και στο ιδανικό σενάριο υπάρχει διαρροή πληροφορίας και στην περίπτωση του πρωτοκόλλου επικοινωνίας μεταξύ Alice και Bob θεωρείται ασφαλές μόνο όταν η Alice γνωρίζει μόνο το αποτέλεσμα των$X\ και\ f(X,Y)$ και ο Bob από την πλευρά του γνωρίζει το αποτέλεσμα μόνο $Y\ και\ f(X,Y)$.[34]

4.2. Homomorphic encryption schemes (Ομομορφικά συστήματα κρυπτογράφησης)

Προκειμένου να επιλυθεί το παραπάνω πρόβλημα ασφάλειας απαιτείται η χρήση ομομορφικών συστημάτων (homomorphic systems). Βασικό τους πλεονέκτημα είναι ότι επιτρέπουν τους υπολογισμούς και την επικοινωνία μέσω κρυπτογραφημένων δεδομένων. Από τον Rivest [35] παρουσιάζεται για πρώτη φορά ένα σχήμα κρυπτογράφησης το 1978. Παρόλα αυτά, το 1980 οι Brickell και Yaccobi [36] καταφέρνουν να το ‘σπάσουν’ ενώ οι Feigenbaum και Merritt το 1991 μέσω της ερώτησης [37] για το κατά πόσο μπορεί να υπάρξει μία συνάρτηση κρυπτογράφησης

(E) ώστε τόσο το $E(X+Y)$ και το $E(X,Y)$να μπορούν να υπολογιστούν μέσω των $E(X)$και $E(Y)$.

Ολοένα και περισσότερα ομομορφικά συστήματα (homomorphic systems) έχουν μελετηθεί τα τελευταία χρόνια εξαιτίας του γεγονότος ότι παίζουν σημαντικό ρόλο στην κρυπτογραφία. Ένα ομομορφικό κρυπτοσύστημα χρησιμοποιεί έναν αλγόριθμο για να υπολογίσει το γινόμενο δύο μηνυμάτων που δίνουν το δημόσιο κλειδί και τα κρυπτογραφημένα μηνύματα αλλά σε καμία περίπτωση τα ίδια μηνύματα ή το κρυπτογραφημένο τους άθροισμα. Προκειμένου ένα σύστημα κρυπτογράφησης να είναι αποτελεσματικό οφείλουμε να διασφαλίσουμε ότι το μέγεθος των κρυπτοκειμένων παραμένει πολυωνυμικό στα όρια της παραμέτρου ασφαλείας κατά τη διάρκεια των επαναλαμβανόμενων υπολογισμών.

4.2.1.Παραδείγματα ομομορφικών (homomorphic) συστημάτων κρυπτογράφησης

Σε αυτή την ενότητα γίνεται μία σύντομη αναφορά σε ομομορφικά συστήματα που απαντώνται στη βιβλιογραφία.

Goldwasser-Micali: Είναι εξαιρετικά σημαντικό καθώς το RSA και άλλα συστήματα έχουν βασιστεί σε αυτό. Το RSA χρησιμοποιεί υπολογισμούς υπολοίπου (modulo) n = p*q καθώς κι ένα γινόμενο δύο μεγάλων πρώτων αριθμών. Η χρησιμοποίηση ενός γινομένου και ενός τετραγώνου καθιστούν την διαδικασία κρυπτογράφηση απλή. Αντίθετα, η αποκρυπτογράφηση είναι μία διαδικασία περισσότερο πολύπλοκη.[38]

Benaloh: Αποτελεί ένα γενικευμένο μοντέλο του GMscheme και επιτρέπει την διαχείριση των εισόδων bits (k). Ο k πρέπει να είναι ένας πρώτος αριθμός που φέρει κάποιους περιορισμούς.[39]

Naccache-Stern: Είναι βελτιωμένη εκδοχή του σχεδίου του Benaloh. Επιτυγχάνεται μικρότερη επέκταση και κατ’ επέκταση ανώτερη απόδοση με τη χρήση μίας τιμής της παραμέτρου k μεγαλύτερης από αυτή που χρησιμοποιείται στο σχέδιο Benaloh. Το βήμα κρυπτογράφησης είναι ακριβώς το ίδιο με το σχέδιο του Benaloh [40].

Paillier - scheme: Πρόκειται για εκ νέου βελτιωμένη έκδοση συγκριτικά με τα προηγούμενα συστήματα υπό το πρίσμα ότι είναι ικανό να μειώσει την αξία της επέκτασης από 3 σε 2 [41] και οφείλεται στον Paillier.

Galbraith - scheme: Με βάση τις ελλειπτικές καμπύλες προσαρμόστηκαν τα υπάρχοντα ομομορφικά σχήματα κρυπτογραφίας. Το ουσιαστικότερο όφελος αυτού

του νέου σχήματος είναι η μείωση των πόρων αποκρυπτογράφησης και κρυπτογράφησης χρησιμοποιώντας μεγαλύτερες τιμές s[42].

Castagnos - scheme: Η χρήση των τετραγωνικών τιμών συνέβαλε στην βελτίωση των επιδόσεων των ομοιομορφικών σχημάτων κρυπτογράφησης. Αυτό το σχήμα επιτυγχάνει μία τιμή επέκτασης 3 και το κλάσμα του κόστους κρυπτογράφησης / αποκρυπτογράφησης με s = 1 πάνω από το σχέδιο Paillier και μπορεί να υπολογιστεί περίπου 2 [43].

4.2.2.Εφαρμογές και ιδιότητες των ομομορφικών συστημάτων (homomorphic encryption schemes)

Προστασία κινητών πρακτόρων (Protection of mobile agents**)**: η χρήση της ομομορφικής κρυπτογράφησης στην προστασία των κινητών πρακτόρων αποτελεί μία από τις πιο ενδιαφέρουσες εφαρμογές της. Τα συστήματα αυτά προσφέρουν τη δυνατότητα κρυπτογράφησης αλλά και εκτέλεσης ολόκληρου του προγράμματος έτσι ώστε να είναι ακόμα εκτελέσιμο. Η προστασία των κινητών παραγόντων με ομοιομορφική κρυπτογράφηση μπορεί να χρησιμοποιηθεί είτε εφαρμόζοντας κρυπτογραφημένες συναρτήσεις είτε κρυπτογραφημένα δεδομένα.

Multi-party computation: μία κοινή συνάρτηση που χρησιμοποιείται από τα μέρη που συμμετέχουν και υπολογίζει την τιμή των στοιχείων που δίνονται κρατώντας, ωστόσο, μυστικά τα επιμέρους στοιχεία. Το συγκεκριμένο πρόβλημα ανήκει στον κλάδο που ασχολείται με τον υπολογισμό μίας τιμής με τη χρήση κρυπτογραφημένων δεδομένων.

Zero-knowledge proofs: Αποτελεί ένα βασικό πρωτόκολλο και παράδειγμα θεωρητικής εφαρμογής των ομομορφικών κρυπτοσυστημάτων (homomorphic encryption schemes). Χρησιμοποιούνται έτσι ώστε να αποδείξουν τη γνώση ορισμένων ιδιωτικών πληροφοριών. Οι Cramer και Damgard [44] αναφέρουν τέτοια παραδείγματα στο άρθρο τους.

Election schemes: Δίνεται η δυνατότητα της αντιστοιχίας με τη χρήση εργαλείου των κρυπτογραφημένων ψήφων χωρίς, ωστόσο, να γίνεται αποκρυπτογράφηση των μεμονωμένων ψήφων.

Water marking and finger printing schemes: Ενσωματώνουν πρόσθετες πληροφορίες σε ψηφιακά δεδομένα. Στην περίπτωση αυτή η ομομορφική ιδιότητα χρησιμοποιείται για να προσθέσει ένα σημάδι (mark) σε δεδομένα τα οποία έχουν προηγουμένως κρυπτογραφηθεί. Σε γενικές γραμμές, τα υδατογραφήματα χρησιμοποιούνται προκειμένου να εντοπιστεί ο ιδιοκτήτης / πωλητής ψηφιακών

αγαθών με σκοπό να εξασφαλιστούν τα δικαιώματα πνευματικής ιδιοκτησίας. Σε αυτά τα συστήματα, ο αγοραστής των δεδομένων είναι απαραίτητο να γνωρίζει ότι τα δεδομένα του δεν έχουν διανεμηθεί παράνομα. Στο άρθρο των Pfitzmann και Waidner [45] μπορούν να εντοπιστούν επιπλέον ιδιότητες αυτών των σχημάτων.

Lottery protocols: Στην περίπτωση μίας κρυπτογραφικής λαχειοφόρου αγοράς, ο τυχερός λαχνός επιλέγεται τυχαία από ένα άτομο. Αν χρησιμοποιηθεί ένα ομομορφικό σχήμα κρυπτογράφησης, αυτό γίνεται ως εξής: κάθε παίκτης επιλέγει έναν τυχαίο αριθμό τον οποίο κρυπτογραφεί. Στη συνέχεια, χρησιμοποιώντας την ιδιότητα του υπολογισμού του κρυπτογραφημένου αθροίσματος σε συνδυασμό με τη χρησιμοποίηση ενός συστήματος αποκρυπτογράφησης threshold οδηγεί στην επιθυμητή λειτουργικότητα.

5. Σχεδιασμός και υλοποίηση της εφαρμογής

Στο συγκεκριμένο κεφάλαιο παρουσιάζεται ο σχεδιασμός και η ανάλυση της εφαρμογής που δημιουργήθηκε. Το πρόβλημα της ιδιωτικότητας στη μετάδοση δεδομένων αποτελεί ένα από τα πιο σπουδαία θέματα έρευνας σε παγκόσμιο επίπεδο καθώς όλο και περισσότερες εφαρμογές ζητούν προσωπικά μας στοιχεία με «αντάλλαγμα» την προσφορά των υπηρεσιών τους.

Για τη διεξαγωγή μελετών γενετικής συσχέτισης σε πολλαπλά επίπεδα μία μέθοδος που χρησιμοποιείται συχνά είναι η μετα-ανάλυση, η οποία όπως αναφέρθηκε είναι μία τεχνική στατιστικής ανάλυσης κι ενοποίησης πολλών μελετών με στόχο στη δημιουργία ενός κοινού συμπεράσματος [46].

Έστω ότι έχουμε οργανισμούς υγείας που υπολογίζουν τοπικά την επίδραση μιας μετάλλαξης στην εμφάνιση μιας ασθένειας. Ο υπολογισμός των δεδομένων βασίζεται σε δείγματα τα οποία έχουν συλλέξει. Στόχος είναι η βελτίωση του υπολογισμού συνδυάζοντας τα δεδομένα όλων των υπολογισμών, ωστόσο σε αυτή την περίπτωση υπάρχει ζήτημα παραβίασης της ιδιωτικότητας. Στην παρούσα εργασία προτείνεται μία μεθοδολογία για τον παραπάνω υπολογισμό, ενώ προστατεύεται η ιδιωτικότητα των δεδομένων.

Η βασική ιδέα της υλοποίησης βασίζεται στον υπολογισμό του αθροίσματος και του γινομένου πολλών αριθμών χωρίς τη χρήση κρυπτογραφημένου καναλιού. Η βασική ιδέα προέρχεται από τον Clifton *et al.*[47] και περιλαμβάνει δύο εκδοχές. Στην πρώτη εκδοχή υπάρχει ένας συλλέκτης (Aggregator) έστω A που λαμβάνει αριθμούς από διαφορετικούς συμμετέχοντες έστωp_i, οι οποίοι στέλνουν δεδομένα χωρίς να έχουν δικαίωμα υπολογισμού και χωρίς να γνωρίζουν τις τιμές των δεδομένων που έχουν στείλει οι υπόλοιποι συμμετέχοντες υπολογίζοντας την τιμή της συνάρτησης $f(\mathbf{x})$.

Στη δεύτερη εκδοχή, η οποία είναι παρόμοια με την πρώτη, υπάρχουν μόνο οι συμμετέχοντεςp_iχωρίς την ύπαρξη συλλέκτη κατά την οποία ο κάθε συμμετέχων είναι ισότιμος και μπορεί να υπολογίσει την τελική τιμή της συνάρτησης $f(x)$.

5.1 Υπολογισμός Αθροίσματος και Γινομένου με απουσία Ασφαλούς Καναλιού (Sum and product calculation without secure channel)

Στη συνέχεια παρουσιάζονται τα δύο πρωτόκολλα υπολογισμού αθροίσματος και γινομένου. Η βασική μεθοδολογία υλοποίησης έγκειται στο ότι υπάρχουν N συμμετέχοντες οι οποίοι δίνουν δεδομένα χωρίς όμως να γίνονται γνωστά.

Τα γκρουπ G1, G2προκύπτουν με την παρακάτω διαδικασία.

- Επιλέγονται δύο πρώτοι αριθμοί ίδιου μεγέθους (p, q)έτσι ώστε ο αριθμός qνα διαιρεί ακριβώς τον αριθμό p-1.
- Ο αριθμός heίναι ένας τυχαίος πρώτος αριθμός στο διάστημα [2,p] $h \in \mathbb{Z}p$.
- Ο αριθμός g1 που αντιστοιχεί στο G1 προκύπτει από τον τύπο

$$g_1 = h^{(p-1)/q} \, mod \, p \text{s.t.} g_1 \neq 1 \, mod \, p$$

- αριθμός g2 που αντιστοιχεί στο G2 προκύπτει από τον τύπο

$$g_2 = g_1^p \, mod \, p^2$$

5.1.1. Product protocol – Participants only model

Στη συγκεκριμένη μεθοδολογία οι συμμετέχοντες $(p_i \ldots p_n)$ υπολογίζουν την τιμή της παράστασης$f(\mathbf{x}) = \Pi_i x_i$.

- Η τιμή $x_i \in \mathbb{Z}$p είναι μυστική για κάθε συμμετέχοντα
- Ο αριθμόςpείναι ένας μεγάλος πρώτος αριθμός.

Η βασική ιδέα του πρωτόκολλου είναι **η εύρεση τυχαίων ακέραιων αριθμών** $R_i \in G_1$όπου θα ισχύει$\Pi_i R_i = 1 \, mod \, p$και ο κάθε χρήστης p_i με τη σειρά του μπορεί να υπολογίζει την τιμή του R_i.

Το πρωτόκολλο αποτελείται από τρεις επιμέρους διαδικασίες, τη δημιουργία (Setup), την κρυπτογράφηση (encrypt) και το γινόμενο (product)

1. **Setup →$r_i \in \mathbb{Z}_q, R_i = (g_1{}^{r_{i+1}} / g_1{}^{r_{i-1}})$ $r_i \in G_1$**
   - Για την κατανόηση της μεθόδου φανταζόμαστε ότι οι συμμετέχοντες σχηματίζουν έναν κύκλο.
   - Ο κάθε συμμετέχων$\boldsymbol{p}_i$(i ∈ {1, … , n}) δημιουργεί έναν τυχαίο ακέραιο αριθμό $r_i \in \mathbb{Z}_q$ και υπολογίζει την παράμετρο$g_1{}^{r_i}$, $g_1 \in G_1$
   - Στη συνέχεια ο κάθε συμμετέχων (p_i) μοιράζεται την τιμή του Y_i η οποία προκύπτει από τον τύπο

$$Y_i = g_1{}^{r_i}, g_1 \in G1$$

   με τους δύο γείτονες του p_{i-1} και p_{i+1} (όπου $p_{n+1} = p_1$ και $p_0 = p_n$).

- Ύστερα από ένα γύρο ανταλλαγών ο κάθε κόμβος (p_i) υπολογίζει και κρατά τον αριθμόR_iαπό τον τύπο,

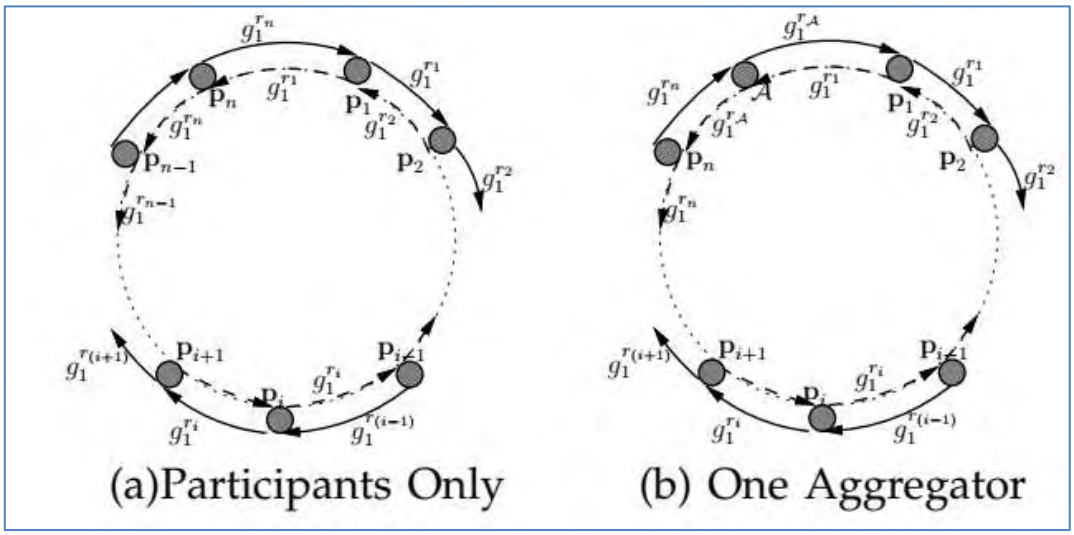
$$R_i = (Y_{i+1}/\ Y_{i-1})^{r_i} = (g_1{}^{r_{i+1}}/\ g_1{}^{r_{i-1}})^{r_i} \in G_1,$$

- Ο αριθμός p_1προκύπτει από την παράσταση

$$(g_1{}^{r_2}/g_1{}^{r_n})^{r_1}$$

- και ο p_nπροκύπτει από την τιμή της παράστασης

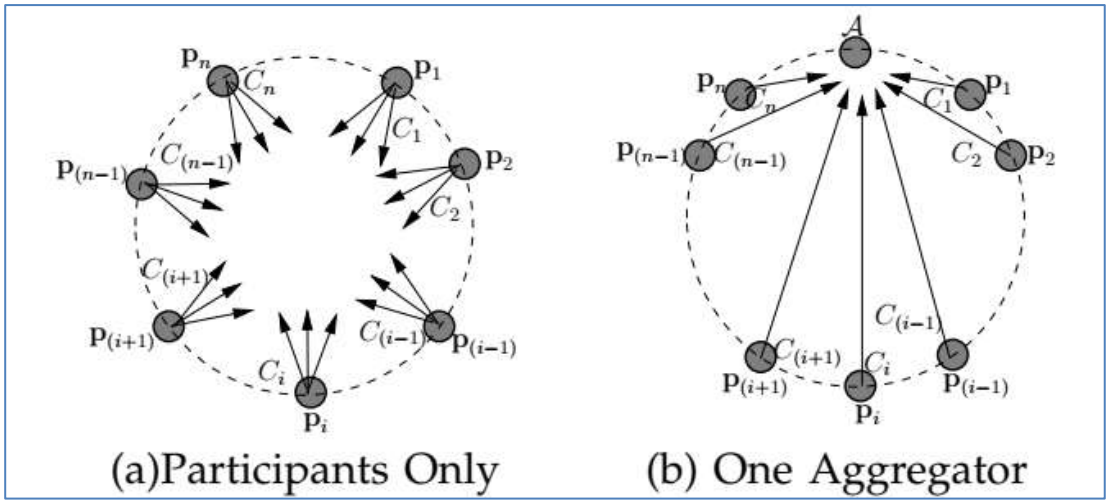
$$(g_1)^{r_2}/\ (g_1{}^{r_{(n-1)}})^{r_n}$$



Εικόνα 8: Τα δύο μοντέλα

2. Encrypt (x_i) →$C_i \in \mathbb{Z}_p$

- Στο στάδιο της κρυπτογράφησης ο κάθε κόμβος p_i δημιουργεί το κρυπτοκείμενο χρησιμοποιώντας τον τύπο:
  $C_i = x_i . R_i = x_i . (g_1{}^{r_{i+1}}/\ g_1{}^{r_{i-1}})^{r_i} \bmod p$, *όπου* $x_i \in \mathbb{Z}_p$ (χρησιμοποιώντας τα δικά του δεδομένα) το οποίο στη συνέχεια τη γνωστοποιεί στους υπόλοιπους κόμβους.



Εικόνα 9: Τα δύο μοντέλα

3. **Product ({C_1, C_2, …, $Cn2$}) →$\prod_{i=1}^{n} xi \in \mathbb{Z}p$**

- Στο στάδιο υπολογισμού του γινομένου κάθε κόμβος p_iαφού έχει λάβει n-1 κρυπτογραφημένες τιμές από τους υπόλοιπους κόμβους υπολογίζει το παρακάτω γινόμενο

$$\prod_{i=1}^{n} C_i = \prod_{i=1}^{n} (x_i (g_1{}^{r_{i+1}} / \ g_1{}^{r_{i-1}})^{r_i}) \ mod \ p$$

$$= (\prod_{i=1}^{n} x_i) \prod_{i=1}^{n} ((g_1{}^{r_{i+1}} / g_1{}^{r_{i-1}})^{r_i}) \ mod \ p$$

$$= (\prod_{i=1}^{n} x_i) g_1^{\sum_{i=1}^{n} (r_{i+1} r_i - r_i r_{i-1})} \ mod \ p$$

$$= \prod_{i=1}^{n} x_i \ mod \ p$$

όπου $r_{n+1} = r_1$, $r_0 = r_n$.

- Για τη διασφάλιση σωστού αποτελέσματος του $\prod_{i=1}^{n} x_i$ χωρίς τη χρήση υπολοίπου ο αριθμός p που επιλέγεται πρέπει να είναι αρκετά μεγάλος, έστω $p \geq M^n$, όπου ο Mείναι το άνω όριο του x_i

5.1.2 Πρωτόκολλο Γινόμενου – Μοντέλο Ενός Συλλέκτη (Product Protocol-One Aggregator Model)

Ο υπολογισμός του γινομένου στη μέθοδο του ενός συλλέκτη (aggregator) είναι παρόμοιος με αυτό που περιγράφηκε στην προηγούμενη ενότητα με την πρώτη διαφορά να έγκειται στο ότι ο συλλέκτης (aggregator) A λειτουργεί όπως ο (n+ 1) συμμετέχοντας (p_{n+1}) και τη δεύτερη στο ότι ο κάθε συμμετέχοντας p_iστέλνει το κρυπτογραφημένο αποτέλεσμα C_i στον συλλέκτη (aggregator), ενώ στην προηγούμενη υλοποίηση το γνωστοποιούσε σε όλους τους συμμετέχοντες.

Ο κάθε συμμετέχων $\mathbf{p}_{i\in[1,n]}$ στέλνει το κρυπτογραφημένο αποτέλεσμα $C_i = R_i x_i$στον συλλέκτη ο οποίος υπολογίζει

$$R_{n+1}\prod_{i=1}^{n} C_i = \prod_{i=1}^{n} x_i \,\mathrm{modp}$$

Για να φτάσει στο τελικό αποτέλεσμα- γινόμενο όπου το R_{n+1}κρατείταιμυστικό σε κάθε συμμετέχοντα p_i

5.1.3 Sum Protocol-Participants Only Model

Σε αυτή την ενότητα αναλύεται το πρωτόκολλο υπολογισμού του αθροίσματος

$$f(x) = \sum_{i=1}^{n} x_i$$

για τις ιδιωτικές τιμές $x_i \epsilon$ G_1 που υπολογίζονται σε κάθε συμμετέχοντα.

Η βασική ιδέα της μεθοδολογίας είναι η μετατροπή του αθροίσματος των επιμέρους αριθμών σε γινόμενο [37]. Όπως θα δούμε στη συνέχεια η συγκεκριμένη μέθοδος βασίζεται στην ιδιότητα του υπόλοιπου (mod) με στόχο την ιδιωτικότητα του πρωτοκόλλου του υπολογισμού του αθροίσματος.

$$(1+p)^m = \sum_{i=0}^{m} \binom{m}{i} p^i = 1 + mp \; mod \; p^2 \; (2)$$

Από την εξίσωση 2 προκύπτει

$$\prod_{i=1}^{n}(1+p)^{x_i} = \prod_{i=1}^{n}(1+p\,.\,x_i) = \left(1 + p\sum_{i} x_i\right) mod\; p^2$$

Το πρωτόκολλο αποτελείται από τρία επιμέρους στάδια, Setup, Encrypt, Sum.

Setup→$r_i \epsilon \mathbb{Z}_q$, $R_i = (g_2{}^{r_{i+1}} / g_2{}^{r_{i-1}})^{r_i} \epsilon$G2

- Ο κάθε συμμετέχων p_i επιλέγει τυχαία έναν μυστικό αριθμό $r_i \epsilon \mathbb{Z}_q$
- υπολογίζει μία δημόσια παράμετρο $g_2{}^{r_i} \bmod p^2$.
- Στη συνέχεια μοιράζει την τιμή Y= $g_2{}^{r_i} \epsilon$G2 την οποία γνωστοποιεί στους δύο γείτονες του p_{i+1} και p_{i-1}.
- Ύστερα από ένα γύρο ανταλλαγών τιμών κάθε συμμετέχοντας p_i υπολογίζει το $R_i = (\frac{g_2{}^{r_{i+1}}}{g_2{}^{r_{i-1}}})^{r_i} mod p^2$το οποίο το κρατάει ως φυσική τιμή τυχαιοποίησης.

Encrypt (x_i,R_i) →Ci $\epsilon \mathbb{Z}_p{}^2$

- Ο κάθε συμμετέχων p_iυπολογίζει την τιμή της παράστασης $(1+x_i \cdot p)$
- στη συνέχεια πολλαπλασιάζει την μυστική παράμετρο $R_i = (g_2{}^{r_{i+1}} / g_2{}^{r_{i-1}})^{r_i}$.για να προκύψει το κρυπτοκείμενο. $C_i = (1 + x_i \cdot p) \cdot R_i \bmod p^2$.
- στο επόμενο βήμαοκάθε p_i κάνει γνωστό τον κρυπτοκείμενο Ciσε κάθε p_i.

  Sum ($\{C_1, C_2, \cdots, C_n\}$) →$\Sigma_{k=1}^{n} x_i \epsilon \mathbb{Z}p$
- στη συνέχεια ο κάθε p_i έχοντας λάβει τα Ciόλων των p_iυπολογίζει την τελική τιμή του C$\epsilon \mathbb{Z}_p{}^2$

$$C = \prod_{i=1}^{n} C_i \ mod\ p^2$$

$$= \prod_{i=1}^{n} (1 + x_i p)(g_2{}^{r_{i+1}} / g_2{}^{r_{i-1}})^{r_i} \ mod\ p^2$$

$$= \left(1 + p \sum_{i=1}^{n} x_i\right) g_2{}^{\sum_{i=1}^{n} r_{i+1} r_i - r_i r_{i-1}} \ mod\ p^2$$

$$= \left(1 + p \sum_{i=1}^{n} x_i\right) mod\ p^2$$

- στο τελικό βήμα υπολογίζει

  $(C\text{-}1) / p = \sum_{i=1}^{n} x_i \bmod p$, έτσι ώστε να προκύψει το τελικό άθροισμα, όπου η διαίρεση δεν είναι $(C\text{ -}1)$ φορές το $p^{-1} \bmod p$ αλλά το πηλίκο $(C - 1) / p$.

5.1.4. Sum Protocol–One Aggregator Model

Η συγκεκριμένη μεθοδολογία είναι παρόμοια με την προηγούμενη με τη διαφορά ότι ο συλλέκτης (aggregator) Α λειτουργεί όπως ο (n+1) –th συμμετέχων (p_i).

- Οι συμμετέχοντες στέλνουν το κρυπτοκείμενο στον συλλέκτη Ακαι αυτός με τη σειρά του υπολογίζει το

$$C = R_{n+1} \prod_{i=1}^{n} C_i = \left(1 + p \sum_{i=1}^{n} x_i\right) mod\ p^2$$

- υπολογίζει το τελικό άθροισμα $\Sigma_{i=1}^{n} x_i$

5.2. Περιγραφή Αλγορίθμου

Η εφαρμογή μας έχει την εξής λογική, έστω ότι έχουμε 10 κόμβους (νοσοκομεία) τα οποία θέλουν να στείλουν κρυπτογραφημένα δεδομένα.

- Ο κάθε κόμβος έχει στην κατοχή του δύο μυστικούς αριθμούς τους οποίους τους ονομάζουμεw_i $και$ $w_i * Y_i$.
- Στη συνέχεια πρέπει να βρεθούν δύο ακέραιοι αριθμοί a, b τέτοιοι ώστε τα γινόμενα $a * w_i + b$ $και$ $a * w_i * Y_i + b$ του κάθε κόμβου να είναι θετικοί και ακέραιοι αριθμοί.

- Για να βρεθεί ο κατάλληλος αριθμός a ουσιαστικά χρειαζόμαστε ο αριθμός a να ικανοποιεί την παράσταση $a \geq 10^x$ όπου x ο μέγιστος αριθμός δεκαδικών ψηφίων των w_i και $w_i * Y_i$ έτσι ώστε ο κάθε κόμβος να μετατρέπεται σε ακέραιο θετικό αριθμό.
- Για την εύρεση του μεγίστου των δεκαδικών ψηφίων καλείται το πρωτόκολλο (συνάρτηση) find_maxκαι στην μεταβλητή a αποθηκεύεται η μέγιστη τιμή των δεκαδικών ψηφίων.
- Στο επόμενο βήμα ο κάθε κόμβος εκτελεί τη συνάρτηση εκ νέου και η συνάρτηση με τη σειρά της υπολογίζει και επιστρέφει την τιμή της μεταβλητής b.
- Η τιμή του κάθε κόμβου $a * w_i + b$ και $a * w_i * Y_i + b$ αναβαθμίζεται χρησιμοποιώντας τις νέες τιμές και η τιμή του κάθε κόμβου πλέον είναι θετικός ακέραιος αριθμός.
- Έπειτα εκτελείται η συνάρτηση find_max δύο φορές,
    - την πρώτη για την παράσταση $a * w_i + b$
    - τη δεύτερη για την $a * w_i * Y_i + b$.
- Στη συνέχεια με τη βοήθεια της συνάρτησης find_sumυπολογίζονται οι τιμέςροι οποίες προκύπτουν
    - από τον μέγιστο αριθμό κόμβων και
    - τις μεγαλύτερες τιμές που έχουν υπολογισθεί στο προηγούμενο βήμα των αθροισμάτων $a * w_i + b$ και $a * w_i * Y_i + b$.
- Στα δύο τελευταία βήματα το άθροισμα των αριθμών υποβαθμίζεται χρησιμοποιώντας τον τύπο $(sum - n * b)/a$και τέλος
- αφού έχουν υπολογιστεί τα επιμέρους αθροίσματα $\sum w_i * Y_i$ και $\sum w_i$ υπολογίζεται το πηλίκο $\sum w_i * Y_i / \sum w_i$

Το πρωτόκολλο (συνάρτηση) find_max ακολουθεί την παρακάτω λειτουργία:

- Στην αρχή ο κάθε κόμβος έχει ένα αριθμό x_i.
- Στη συνέχεια αρχικοποιείται μία μεταβλητή rκαι ξεκινά μία επαναληπτική διαδικασία κατά την οποία
- ο κάθε κόμβος υπολογίζει το πηλίκο της ακεραίας διαίρεσης $c_i = x_i / x^r$ όπου
    - αν το αποτέλεσμα είναι μεγαλύτερο από το 0 τότε επιστρέφεται η τιμή1
    - διαφορετικά επιστρέφεται η τιμή 0.

- Στη συνέχεια, υπολογίζεται το γινόμενο όλων των κόμβων Πc_iμε τη χρήση της συνάρτησης find_product.
    - Αν το αποτέλεσμα είναι 0 τότε ο αριθμός ραυξάνεται κατά ένα και επαναλαμβάνεται η παραπάνω διαδικασία,
    - αντίθετα αν η τιμή είναι 1 τότε η διαδικασία σταματά καθώς βρέθηκε η επιθυμητή τιμήx^r

Τα δύο παραπάνω βήματα χρήζουν ιδιαίτερης προσοχής. Γενικά ισχύουν τα εξής:

$c_i = \frac{x_i}{y*2^r} + 1$ (ακέραια διαίρεση)

$\begin{cases} \alpha\nu\ x_i < y * 2^r \rightarrow c_i = 1 \\ \alpha\nu\ x_i > y * 2^r\ \rightarrow\ c_i > 1 \end{cases}$

Συνεπώς αν όλοι οι κόμβοι ικανοποιηθούν (δηλαδή $x_i < y * 2^n\ \forall\ i$)και βρούν ένα επιθυμητό αριθμό τότε $το\ \Pi c_i = 1\ αλλιώς\ το\ \Pi c_i > 1$

Καθώς το product protocol για να λειτουργήσει σωστά απαιτεί τη μέγιστη πιθανή τιμή των όρων του γινομένου. Για το λόγο αυτό, θέτουμε μία σχετικά μεγάλη τιμή $maxc$. Το πρωτόκολλο κινδυνεύει να αποτύχει στην περίπτωση που$\Pi c_i = maxc + 1$οπότε θα υπολογιστεί ως αποτέλεσμα εσφαλμένα το 1 με αποτέλεσμα να σταματήσει εσφαλμένα η όλη διαδικασία.

$\Pi c_i = \kappa * maxc + 1$

$\Pi c_i{}' = \lambda * maxc + 1$

$\Pi\left(\frac{x_i}{y * 2^r} + 1\right)$

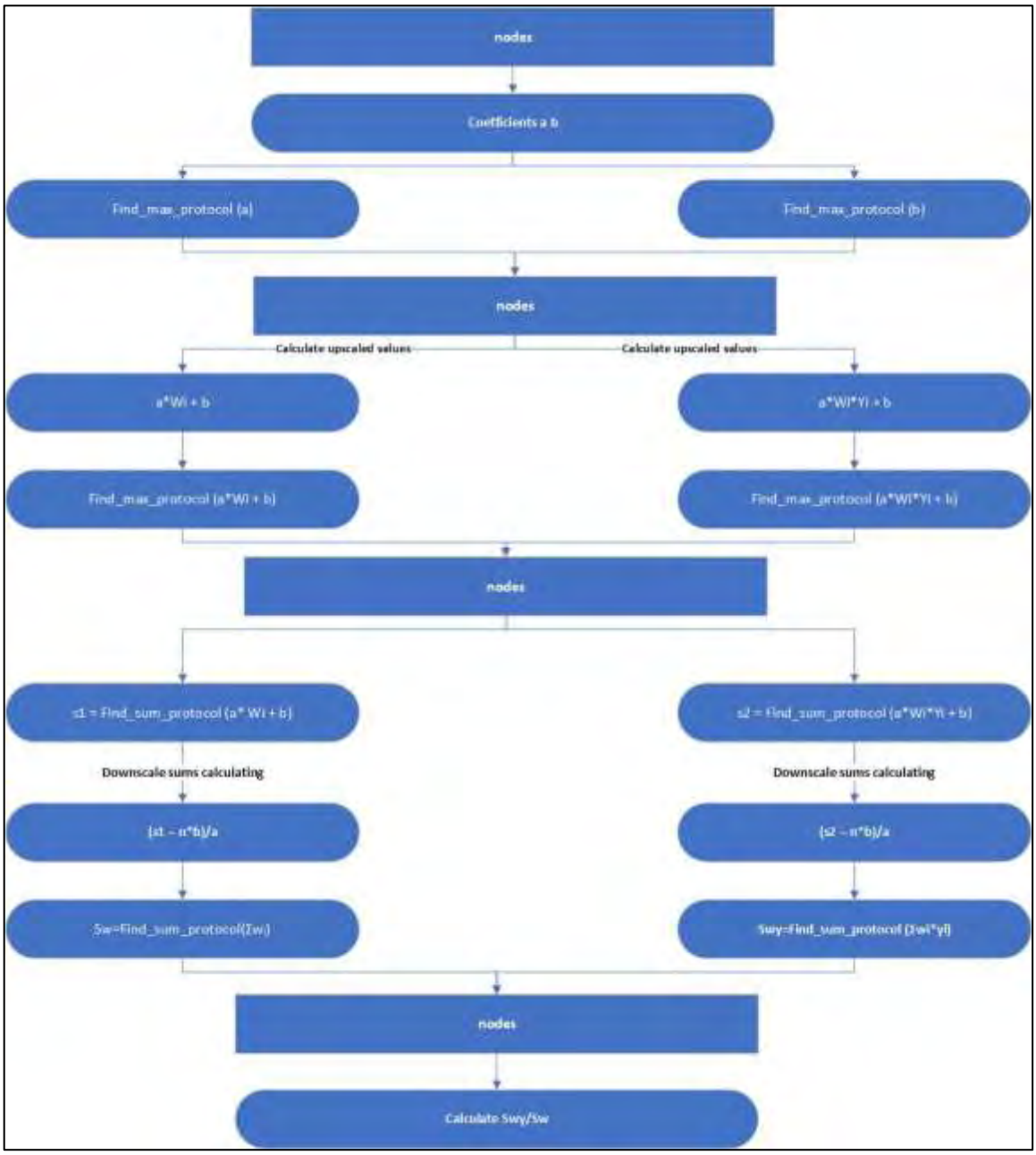
$\Pi\left(\frac{x_i}{y * 2^{(r+1)}} + 1\right)$

$\Pi\ \frac{x_i + y * 2^{r+1}}{y * 2^{r+1}} -\ \Pi\ \frac{x_i + y * 2^r}{y * 2^r} = \kappa * maxc$

$\frac{\Pi(x_i + y * 2^{r+1})}{\Pi y * 2^{r+1}} -\ \frac{\Pi(x_i + y * 2^r)}{y * 2^r} = \kappa * maxc$

$\frac{\Pi(x_i + y * 2^{r+1})}{y^n * 2^{(r+1)*n}} -\ \frac{\Pi(x_i + y * 2^r)}{y^n * 2^{r*n}} = \kappa * maxc$

$\frac{\Pi(x_i + y * 2^{r+1}) - 2^n \Pi(x_i + y *\ 2^n)}{y^n * 2^{r*n} * 2^n} = \kappa * maxc$



Σχήμα1: Σχηματική αναπαράσταση εφαρμογής

5.3 Αρχιτεκτονική Εφαρμογής

Η υλοποίηση βασίστηκε σε τέσσερα πρωτόκολλα με τις εξής ονομασίες, (general protocol, find_max_protocol, find_sum_protocol και find_product_protocol) τα οποία αναλύονται στη συνέχεια.

General Protocol

- Ο κάθε κόμβος (i) κατέχει δύο μυστικές τιμές w_iκαι w_i*Y_i
- Ο κάθε κόμβος είναι απαραίτητο να έχει δύο συντελεστές aκαι bγια τους οποίους υπολογίζονται οι τιμές a*w_i+bκαι a*w_i*Y_i + bοι οποίοι πρέπει να είναι θετικοί ακέραιοι αριθμοί
- Για τον συντελεστή aπρέπει να ισχύει $\alpha \geq 10^x$όπου το xείναι ο μέγιστος αριθμός δεκαδικών ψηφίων των w_i, $w_i * Y_i$ και b χρειάζεται ώστε να ισχύει $b \geq \max(0, -a * w_i, -a * w_i * Y_i)$ για κάθε κόμβο.
- Ο κάθε κόμβος «τρέχει» το πρωτόκολλο find_maxμε τιμές εισόδου $\max(1, -a * w_i, -a * w_i * Y_i)$έτσι ώστε να υπολογίσει την τιμή του συντελεστή b.
- Ο κάθε κόμβος υπολογίζει τις αναβαθμισμένες τιμές $a * w_i + b$ και $a * w_i * Y_i + b$οι οποίες πρέπει να είναι θετικοί ακέραιοι.
- Στη συνέχεια ο κάθε κόμβος «τρέχει» το πρωτόκολλο find_maxδύο φορές με τιμές εισόδου $a * w_i + b$ και $a * w_i * Y_i + b$
- Στο επόμενο βήμα ο κάθε κόμβος «τρέχει» το πρωτόκολλο find_sumδύο φορές με εισόδους $a * w_i + b$ και $a * w_i * Y_i + b$ υπολογίζοντας τις τιμές pκαι τις μέγιστες τιμές που υπολογίστηκαν στο προηγούμενο βήμα.
- Στη συνέχεια μειώνουν τα αθροίσματα υπολογίζοντας την τιμή
$$(sum - n * b)/a$$
- Τέλος με τη χρήση των τιμών του αθροίσματος υπολογίζεται η τιμή της παράστασης $\sum w_i * Y_i / \sum w_i$

Find_Max Protocol

- Ένας κόμβος εκκινητής δημιουργεί έναν τυχαίο αριθμό xτον οποίο διαμοιράζει στους υπόλοιπους κόμβους.
- Στη συνέχεια η παραπάνω διαδικασία επαναλαμβάνεται για συνεχόμενους γύρους με αρχική τιμή r =1.
- Κάθε κόμβος υπολογίζει το αποτέλεσμα της ακέραιας διαίρεσης $c_i = x * r/x_i$. Αν το αποτέλεσμα είναι μεγαλύτερο του μηδέν επιστρέφεται η τιμή 1 διαφορετικά επιστρέφεται η τιμή 0.
- Στη συνέχεια ο κάθε κόμβος «τρέχει» το πρωτόκολλο find_productγια να υπολογίσει την τιμή $\prod c_i$
- Αν η παραπάνω τιμή είναι 0 τότε η τιμή του ραυξάνεται κατά 1 (r = r+1) και επαναλαμβάνεται η διαδικασία.
- Εάν η τιμή είναι 1 η διαδικασία σταματά και η επιθυμητή τιμή καθορίζεται x^r

5.4. Υλοποίηση

Στην παρούσα ενότητα θα αναλυθούν τα βασικά κομμάτια του κώδικα της εφαρμογής. Το συγκεκριμένο project έχει γραφτεί σε γλώσσα python χρησιμοποιώντας το εργαλείο pycharm.

Η συνάρτηση prot_setup χρησιμοποιείται για την αρχικοποίηση των μεταβλητών p, q,h, g1, g2 οι οποίες με τη σειρά τους χρησιμοποιούνται για την εύρεση του αθροίσματος.

```
# SYNARTHSH GIA TO prot_setup
def prot_setup(maxn, nodes):
    # setup p
    p = random.randint(pow(maxn, nodes), 2 * (pow(maxn, nodes)))
    while pyprimes.isprime(p):
        p = random.randint(pow(maxn, nodes), 2 * pow(maxn, nodes))
    print("o ari8mos p einai: ", p)
    # setup q
    q = (random.choice(primefac.factorint(p - 1).keys()))
    while q == 2: #o ari8mos mas dhmiourgei problima ston ypologismo
        q = (random.choice(primefac.factorint(p - 1).keys()))
    print("o ari8mos q einai: ", q)
    # setup h
    h = (random.randint(2, p))
    while not coprime(h, p):
        h = (random.randint(2, p))
    print("o ari8mos h einai: ", h)
    # setup g1
    g1 = (pow(h, int((p - 1) / (q)), p))
    print("o ari8mos g1 einai: ", g1)
    # setup g2
    g2 = (pow(g1, p, p ** 2))
    print("o ari8mos g2 einai: ", g2)
    return p, q, h, g1, g2
```

Η συνάρτηση sum_setup χρησιμοποιείται για την αρχικοποίηση των μεταβλητών p, q, h, g1, g2

```
# SYNARTHSH GIA TO sum_setup
def sum_setup(maxn, nodes):
    # setup p
    p = random.randint(maxn * nodes, 2 * (maxn * nodes))
    while pyprimes.isprime(p):
        p = random.randint(maxn * nodes, 2 * (maxn * nodes))
    print("o ari8mos p gia to sum einai: ", p)
    # setup q
    q = (random.choice(primefac.factorint(p - 1).keys()))
    while q == 2:
        q = (random.choice(primefac.factorint(p - 1).keys()))
    print("o ari8mos q gia to sum einai: ", q)
    # setup h
```

```
    h = (random.randint(2, p))
    while not coprime(h, p):
        h = (random.randint(2, p))
    print("o ari8mos h gia to sum einai: ", h)
    # setup g1
    g1 = (pow(h, int((p - 1) / (q)), p))
    print("o ari8mos g1 gia to sum einai: ", g1)
    # setup g2
    g2 = (pow(g1, p, p ** 2))
    print("o ari8mos g2 gia to sum einai: ", g2)
    return p, q, h, g1, g2
```

Η συνάρτηση product χρησιμοποιείται για τον υπολογισμό του γινομένου το οποίο επιστρέφεται μέσω της μεταβλητής στο κυρίως πρόγραμμα.

```
# H SYNARTHSH GIA TO product
def product(p, q, h, g1, g2, x):
    # lista r
    r = []
    for i in range(0, nodes):
        t = random.randint(2, q)
        while not coprime(t, q):
            t = random.randint(2, q)
        r.append(t)
    # lista y
    y = []
    for i in range(0, nodes):
        y.append(pow(g1, r[i], p))
    # lista R
    R = []
    for i in range(0, nodes):
        if i == 0:
            a = y[nodes - 1]
            zp = xgcd(a, p)[1]
            zp = zp % p
            zn = (y[i + 1]) % p
        elif i == nodes - 1:
            a = y[i - 1]
            zp = xgcd(a, p)[1]
            zp = zp % p
            zn = (y[0]) % p
```

```
            else:
                a = y[i - 1]
                zp = xgcd(a, p)[1]
                zp = zp % p
                zn = (y[i + 1]) % p
            z = pow(zn * zp, r[i], p)
            R.append(z)
        # apotelesma c
        c = 1
        for i in range(0, nodes):
            c = (c * ((x[i] * R[i]) % p))
        c = c % p
        return c


    # SYNARTHSH GIA TO SUM_FUN
    def sum_fun(p, q, h, g1, g2, x):
        # lista r
        r = []
        for i in range(0, nodes):
            t = random.randint(2, q)
            while not coprime(t, q):
                t = random.randint(2, q)
            r.append(t)
        print("H LISTA TWN r tou sum")
        print(r)
        # lista y
        y = []
        for i in range(0, nodes):
            y.append(pow(g2, r[i]) % (p ** 2))
        print("H LISTA TWN y tou sum")
        print(y)
        # lista R
        R = []
        for i in range(0, nodes):
            if i == 0:
                a = y[nodes - 1]
                zp = xgcd(a, p ** 2)[1]
                zp = zp % p ** 2
                zn = (y[i + 1]) % p ** 2
            elif i == nodes - 1:
```

```
            a = y[i - 1]
            zp = xgcd(a, p ** 2)[1]
            zp = zp % p ** 2
            zn = (y[0]) % p ** 2
        else:
            a = y[i - 1]
            zp = xgcd(a, p ** 2)[1]
            zp = zp % p ** 2
            zn = (y[i + 1]) % p ** 2
        z = (((zn * zp) ** (r[i])) % (p ** 2))
        R.append(int(z))
    print("H LISTA TWN R toy sum")
    print(R)

    # APOTELESMA sum
    c = 1
    for i in range(0, nodes):
        ci = ((1 + x[i] * p) * R[i]) % (p ** 2)
        c = c * ci % (p ** 2)
    c = c % (p ** 2)
    sum_up = (c - 1) / p
    return sum_up
```

Η συνάρτηση find_sum χρησιμοποιείται για τον υπολογισμό και την επιστροφή της μέγιστης τιμής

```
def find_max(maxn, nodes, x):
    maxc = 1000
    p, q, h, g1, g2 = prot_setup(maxc, nodes)
    y = 5  # random.randint(10, 20)
    print("y", y)
    r = 1
    pl_syn_1 = 0  # PLH8OS SYNEXOMENWN 1
    while pl_syn_1 != 2:
        c = []
        for i in range(0, nodes):
            c.append((x[i] // (y * (2 ** r))) + 1)
            # c.append((x[i] // (y * r)) + 1)
        res = product(p, q, h, g1, g2, c)
        if res == 1:
            pl_syn_1 = pl_syn_1 + 1
            if pl_syn_1 == 2:
```

```
                r = r - 1
            else:
                r = r + 1
        else:
            pl_syn_1 = 0
            r = r + 1
    print("O ARI8MOS R EINAI: ", r)
    return y * (2 ** r)
```

Η συνάρτηση general_protocol αποτελεί τη βασική συνάρτηση μέσω της οποίας υπολογίζονται όλες οι απαραίτητες τιμές για την εκτέλεση της εφαρμογής.

```
def general_protocol(nodes, maxn, minim, maxim, decimal_numbers):
    ## 1. Arxikopoihsh
    ola_dyo = False
    while ola_dyo == False:
        ola_dyo = True
        w = []
        for i in range(0, nodes):
            w.append(round(random.uniform(minim, maxim),
decimal_numbers))
        copy_w = w  # ANTIGRAFW TH LISTA
        d_w = []  # DHMIOYRGW TH LISTA ME TO PLH8OS TWN DEKADIKWN
PSHFIWN
        for i in range(0, nodes):
            pl_dek = 0  # METRHTHS DEKADIKWN PSHFIWN
            a = copy_w[i]
            b = int(a)
            while (a - b) != 0:
                pl_dek = pl_dek + 1
                a = a * 10
                b = int(a)
            d_w.append(pl_dek)
            if d_w[i] > decimal_numbers: ola_dyo = False
    print("LISTA W", w, "A8ROISMA STOIXEIWN ", N(sum(w), 3))
    print("LISTA D_W", d_w)

    ## 2. Find max decimnal digits
    a = find_max(maxn, nodes, d_w)
    print("TO APOTELESMA TOY FIND MAX GIA A EINAI: ", a)
```

```
## 3. Find max number to add, in order to make them all positive
# DHMIOYRGW TH NEA LISTA w
for i in range(0, nodes):
    w[i] = int(w[i] * pow(10, a))
print ("NEA LISTA w[i]", w)
# DHMIOYRGW TH LISTA a_w
a_w = []
for i in range(0, nodes):
    if w[i] < 0:
        a_w.append(w[i] * (-1))
    else:
        a_w.append(0)
print ("LISTA a_w[i]", a_w)
b = find_max(maxn, nodes, a_w)
print("TO APOTELESMA TOY FIND MAX GIA B EINAI: ", b)
## 4. Find max of resulting numbers
# KANW UPDATE TH LISTA PROS8ETONTAS TO B
for i in range(0, nodes):
    w[i] = w[i] + b
print("H NEA LISTA W EINAI: ", w)
# BRISKW
t = find_max(maxn, nodes, w)
print("TO APOTELESMA TOY FIND MAX GIA TO t EINAI: ", t)
## 5. Sum protocol
p, q, h, g1, g2 = sum_setup(t, nodes)
print("p, q, h, g1, g2", p, q, h, g1, g2)
# KLHSH THS sum
s,gin= sum_fun(p, q, h, g1, g2, w)
print("TO APOTELESMA TOY A8ROISMATOS EINAI: ", s)
print("TO GINOMENO EINAI: ",gin)
return gin
# TELIKO APOTELESMA
a = float(a)
apot = (s - (nodes * b)) / a
print("TO TELIKO APOTELESMA EINAI: ", apot / (pow(10, a - 1)))
```

6. Μετρήσεις

Σε αυτή την ενότητα παρουσιάζονται οι γραφικές παραστάσεις από τις μετρήσεις που πραγματοποιήθηκαν στον υλοποιημένο αλγόριθμο. Η απόδοση του αλγορίθμου εξαρτάται από δύο βασικούς παράγοντες. Ο πρώτος είναι ο αριθμός των κόμβων που θα ανταλλάξουν δεδομένα κι ο δεύτερος είναι η ακρίβεια των αριθμών που θα θελήσουν να ορίσουν οι κόμβοι και καθορίζεται από τον αριθμό των δεκαδικών ψηφίων.

Για τον υπολογισμό των μετρήσεων «τρέξαμε» τον κώδικα με παραμέτρους τον αριθμό των κόμβων και τον αριθμό των δεκαδικών ψηφίων που θα είχαν οι αριθμοί. Από τις μετρήσεις που προέκυψαν με τη χρήση τεταρτημόριων απορρίψαμε τις τιμές στο πρώτο και στο τελευταίο τεταρτημόριο. Όπως είναι γνωστό από τη στατιστική σε σύνολο διατεταγμένων παρατηρήσεων, ορίζουμε το πρώτο τεταρτημόριο Q1 ως την τιμή, αριστερά της οποίας βρίσκεται το πολύ το 25% του συνολικού αριθμού των δεδομένων. Το δεύτερο τεταρτημόριο Q2 ορίζεται αναλόγως και συμπίπτει με τη διάμεσο δύο τιμών δεδομένων. Τέλος, το τρίτο τεταρτημόριο Q3 ορίζεται ως η τιμή, αριστερά της οποίας βρίσκεται το πολύ το 75% του συνολικού αριθμού των δεδομένων[48]. Τα αποτελέσματα που προέκυψαν φαίνονται στα παρακάτω γραφήματα.

Γράφημα 1:Μέσος χρόνος - 1 δεκαδικό

Γράφημα 2: Μέσος χρόνος - 2 δεκαδικά

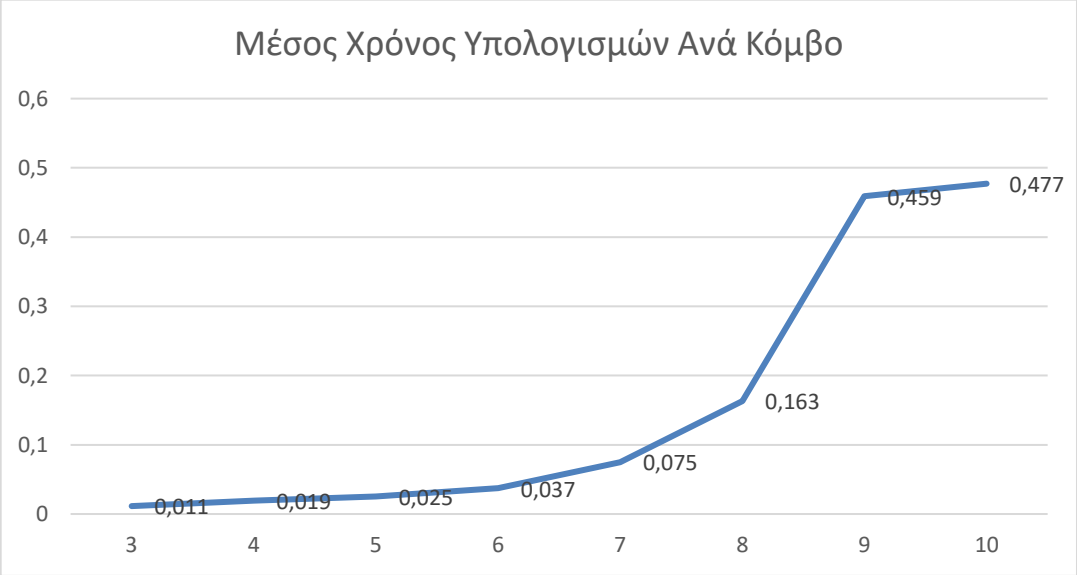
Γράφημα 3: Μέσος χρόνος - 3 δεκαδικά

Γράφημα 4: Μέσος χρόνος - 4 δεκαδικά

Γράφημα 5: Μέσος χρόνος - 5 δεκαδικά

Παρατηρούμε ότι ο μέσος χρόνος παρουσιάζει συνεχώς αυξητική τάση όσο αυξάνουν τα δεκαδικά ψηφία παράλληλα με το όσο αυξάνουν οι αριθμοί των κόμβων για τους οποίους υπολογίζουμε τα δεδομένα. Κάτι τέτοιο είναι απολύτως φυσιολογικό καθώς όσο αυξάνουν είτε τα δεκαδικά ψηφία είτε ο αριθμός των κόμβων, αυξάνουν οι απαιτήσεις για υπολογιστική ισχύ.

Στο παρακάτω διάγραμμα απεικονίζεται ο μέσος χρόνος του υπολογισμού των μετρήσεων ανά κόμβο. Όπως είναι φανερό όσο ανεβαίνει ο αριθμός των κόμβων τόσο αυξάνεται ο χρόνος υπολογισμού των αποτελεσμάτων.



Γράφημα 6:Μέσος χρόνος ανά κόμβο

7. Ανακεφαλαίωση - Συμπεράσματα

Στην παρούσα διπλωματική εργασία αναλύεται και υλοποιείται μία μέθοδος κρυπτογραφικής ανταλλαγής δεδομένων μεταξύ οργανισμών που αφορούν δεδομένα ευρυγονιδιωματικών μελετών. Στο πρώτο μέρος παρουσιάστηκαν μερικοί βασικοί όροι της επιστήμης της βιολογίας όπως το DNA το οποίο πρόκειται για μία μεγαλομοριακή ένωση που συγκροτείται από αζωτούχες - πρωτεϊνικές βάσεις, φωσφορικές ρίζες και ένα σάκχαρο με πέντε άτομα άνθρακα (πεντόζη), την δε(σ)οξυριβόζη και οι αζωτούχες βάσεις του είναι η κυτοσίνη C, η γουανίνη G, η θυμίνη T και η αδενίνη A.

Στη συνέχεια έγινε εκτενής περιγραφή του Genome Wide Associationκαθώς αποτελεί βασική παράμετρο της εργασίας. Η μελέτη συσχετισμού ολόκληρου του γονιδιώματος (GWAS) υπολογίζει και αναλύει τις παραλλαγές στην αλληλουχία του DNA σε ολόκληρο το ανθρώπινο γονιδίωμα σε μία προσπάθεια να ταυτοποιηθούν οι γενετικοί παράγοντες κινδύνου για ασθένειες κοινές στον πληθυσμό. Ο απώτερος στόχος της GWAS είναι η χρήση αυτών των γενετικών παραγόντων κινδύνου για να προβλεφθεί ποιος βρίσκεται σε κίνδυνο και να προσδιοριστούν τα βιολογικά «θεμέλια» της ευαισθησίας της νόσου για την ανάπτυξη νέων στρατηγικών πρόληψης και θεραπείας.

Μία ακόμη μεθοδολογία που ήταν απαραίτητη για την πραγματοποίηση της εργασίας ήταν αυτή της μετα-ανάλυσης. Με την μετα-ανάλυση γίνεται η ενοποίηση και η στατιστική ανάλυση δεδομένων προερχόμενων από διαφορετικές έρευνες οι οποίες προκύπτουν από τυχαιοποιημένες κλινικές δοκιμές. Το πρόβλημα της ιδιωτικότητας στη μετάδοση δεδομένων αποτελεί ένα από τα πιο σπουδαία θέματα έρευνας σε παγκόσμιο επίπεδο. Για τη διεξαγωγή μελετών γενετικής συσχέτισης σε πολλαπλά επίπεδα μία μέθοδος που χρησιμοποιείται συχνά είναι η μετα-ανάλυση, η οποία όπως αναφέρθηκε είναι μία τεχνική στατιστικής ανάλυσης κι ενοποίησης πολλών μελετών με στόχο στη δημιουργία ενός κοινού συμπεράσματος.

Σε ότι αφορά στην εφαρμογή, όπως επισημάνθηκε υπάρχουν δύο μέθοδοι που βασίζονται στον υπολογισμό του αθροίσματος και του γινομένου πολλών αριθμών χωρίς τη χρήση κρυπτογραφημένου καναλιού. Στην πρώτη εκδοχή υπάρχει ένας συλλέκτης (Aggregator) που λαμβάνει αριθμούς από διαφορετικούς συμμετέχοντες οι οποίοι στέλνουν δεδομένα χωρίς να έχουν δικαίωμα υπολογισμού και χωρίς να γνωρίζουν τις τιμές των δεδομένων που έχουν στείλει οι υπόλοιποι συμμετέχοντες υπολογίζοντας την τιμή της συνάρτησης $f(\mathbf{x})$. Στη δεύτερη εκδοχή, υπάρχουν μόνο οι

συμμετέχοντες p_iχωρίς την ύπαρξη συλλέκτη κατά την οποία ο κάθε συμμετέχων είναι ισότιμος και μπορεί να υπολογίσει την τελική τιμή της συνάρτησης *f(x)*.Η υλοποίηση βασίστηκε σε τέσσερα πρωτόκολλα και έχει γραφτεί εξολοκλήρου σε γλώσσα python χρησιμοποιώντας το εργαλείο pycharm.

Τα συμπεράσματα που προέκυψαν από τις μετρήσεις μπορούμε να επισημάνουμε ότι ήταν τα αναμενόμενα. Η απόδοση του αλγορίθμου εξαρτάται από δύο βασικούς παράγοντες. Ο πρώτος είναι ο αριθμός των κόμβων που θα ανταλλάξουν δεδομένα κι ο δεύτερος είναι η ακρίβεια των αριθμών που θα θελήσουν να ορίσουν οι κόμβοι και καθορίζεται από τον αριθμό των δεκαδικών ψηφίων. Από τα διαγράμματα που παρουσιάστηκαν στην αντίστοιχη ενότητα βγαίνει το συμπέρασμα ότι οι διάφορες στους χρόνους εκτέλεσης του αλγορίθμου έγκειται στο γεγονός ότι οι δοκιμές έγιναν σε ένα συμβατικό υπολογιστή που προορίζεται για προσωπική χρήση. Όπως γίνεται αντιληπτό σε πραγματικές συνθήκες ένας οργανισμός που επιθυμεί να ανταλλάξει δεδομένα και να προβεί σε αυτούς τους υπολογισμούς είναι προφανές ότι θα χρησιμοποιήσει υπολογιστές με μεγαλύτερη υπολογιστική ισχύ κάνοντας με αυτό τον τρόπο πιο γρήγορη την εκτέλεση και τη παραγωγή αποτελεσμάτων από τον συγκεκριμένο αλγόριθμο.

8. Ευρετήριο Όρων

7. Βιβλιογραφία

1. Identification, E.C.f.G.d. *Τι είναι ένα γενετικό αποτύπωμα*. 2012 [cited 2018 16/08]; Available from: https://www.dnalogy.eu/content/%CF%84%CE%B9-%CE%B5%CE%B9%CE%BD%CE%B1%CE%B9-%CF%84%CE%BF-dna.
2. Σούλπη, ε., *συμπεριφορά των ανομινοξέων κατά την είσοδο τους στον εγκέφαλο μέσω του αιματοεγκεφαλικού φραγμού*. Ιατρική, 1992. **62**: p. 081-83.
3. Βικιπαίδειας, Σ.τ. *Νουκλεϊκά οξέα*. 2017 [cited 2018 16/08]; Available from: el.wikipedia.org/w/index.php?title=%CE%9D%CE%BF%CF%85%CE%BA%CE%BB%CE%B5%CF%8A%CE%BA%CE%AC_%CE%BF%CE%BE%CE%AD%CE%B1&oldid=6384174.
4. Griffith, F., *The significance of pneumococcal types.* Epidemiology & Infection, 1928. **27**(2): p. 113-159.
5. Αλεπόρου, Β., et al., *Βιολογία Θετικής κατεύθυνσης Γ' τάξης Γενικού Λυκείου*. 2012, Αθήνα: Οργανισμός Εκδόσεων Διδακτικών Βιβλίων (ΟΕΔΒ).
6. Watson, J.D. and F.H. Crick, *Molecular structure of nucleic acids.* Nature, 1953. **171**(4356): p. 737-738.
7. Γιαλούρης, Π., Κ. Μποσινιάκου, and Δ. Σιδέρης, *ΒΙΟΧΗΜΕΙΑ Τεχνολογικής κατεύθυνσης Γ' ΤΑΞΗΣ ΓΕΝΙΚΟΥ ΛΥΚΕΙΟΥ*. 2008, Αθήνα: Οργανισμός Εκδόσεων Διδακτικών Βιβλίων.
8. Magoutis, K., et al., *Συμβολοσειρές, λίστες, πλειάδες, λεξικά.* 2015.
9. Russell, P.J., *iGenetics A Molecular Approach*. 2010, San Francisco: Benjamin Cummings.
10. Gunderson, K.L., et al., *A genome-wide scalable SNP genotyping assay using microarray technology*, in *Nature Genetics*. 2005. p. 549–554.
11. Gunderson, K.L., et al., *A genome-wide scalable SNP genotyping assay using microarray technology.* Nature genetics, 2005. **37**(5): p. 549.
12. Krawitz, P., et al., *Microindel detection in short-read sequence data*, in *Bioinformatics*. 2010. p. 722–729.
13. Flicek, P., et al., *Ensembl 2011*. Nucleic acids research, 2010. **39**(suppl_1): p. D800-D806.
14. Nikolaou, C., et al., *Ανάλυση της Γενετικής Ποικιλομορφίας.* 2015.
15. Manolio, T.A., *Genomewide association studies and assessment of the risk of disease*, in *N Engl J Med*. 2010. p. 166-176.
16. Pearson, T.A. and T.A. Manolio, *How to Interpret a Genome-wide Association Study*, in *Jama*. 2008. p. 1335-1344.
17. Greely, H.T., *The Uneasy Ethical and Legal Underpinnings of Large-Scale Genomic Biobanks*, in *Annu Rev Genomics Hum Genet*. 2007. p. 343-364.
18. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration*, in *Science*. 2005. p. 385-389.
19. Johnson, A. and C. O' Donnell, *An open access database of genome-wide association results*, in *BMC Med Genet*. 2009. p. 6.
20. Paynter, N.P., et al., *Association between a literature-based genetic risk score and cardiovascular events in women*, in *JAMA*. 2010. p. 631-637.

21. Paschou, P., et al., *Maritime route of colonization of Europe*, in *Maritime route of colonization of Europe*. 2014. p. 9211-9216.

22. Bush, W. and J. Moore, *Genome-Wide Association Studies*, in *PLOS Computational*. 2012. p. 7-24.

23. Marenberg, M.E., et al., *Genetic susceptibility to death from coronary heart disease in a study of twins,* in *The New England Journal of Medicine*. 1994. p. 1041-1046.

24. Sawcer, S., et al., *Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis,* in *Nature*. 2011. p. 214-219.

25. Stranger, B.E., et al., *Relative impact of nucleotide and copy number variation on gene expression phenotypes*, in *Science (New York, N.Y.)*. 2007. p. 848-853.

26. Schork, N.J., et al., *Common vs. rare allele hypotheses for complex diseases.* , in *Current Opinion in Genetics & Development*. 2009. p. 212-219.

27. Stranger, B.E., E. Stahl , and T. Raj *Progress and promise of genome-wide association studies for human complex trait genetics.*, in *Genetics*. 2011. p. 367-383.

28. EPEYNA, E.I., *Συστηματική ανασκόπηση και μετα-ανάλυση*.

29. Delgado-Rodríguez, M., *Glossary on meta-analysis.* Journal of Epidemiology & Community Health, 2001. **55**(8): p. 534-536.

30. Minelli, C., et al., *The choice of a genetic model in the meta-analysis of molecular association studies.* International journal of epidemiology, 2005. **34**(6): p. 1319-1328.

31. Van Houwelingen, H.C., L.R. Arends, and T. Stijnen, *Advanced methods in meta-analysis: multivariate approach and meta-regression.* Statistics in medicine, 2002. **21**(4): p. 589-624.

32. Borenstein, M., et al., *Introduction to meta-analysis*. 2011: John Wiley & Sons.

33. DerSimonian, R. and N. Laird, *Meta-analysis in clinical trials.* Controlled clinical trials, 1986. **7**(3): p. 177-188.

34. Odlyzko, A.M., *Advances in Cryptology-CRYPTO'86: Proceedings*. Vol. 263. 2003: Springer.

35. Rivest, R.L., L. Adleman, and M.L. Dertouzos, *On data banks and privacy homomorphisms.* Foundations of secure computation, 1978. **4**(11): p. 169-180.

36. Brickell, E.F. and Y. Yacobi. *On privacy homomorphisms*. in *Workshop on the Theory and Application of of Cryptographic Techniques*. 1987. Springer.

37. Feigenbaum, J. and M. Merritt, *Open questions, talk abstracts, and summary of discussions.* 1991.

38. Shafi, G. and S. Micali, *Probabilistic encryption.* Journal of computer and system sciences, 1984. **28**(2): p. 270-299.

39. Benaloh, J. *Dense probabilistic encryption*. in *Proceedings of the workshop on selected areas of cryptography*. 1994.

40. Naccache, D. and J. Stern. *A new public-key cryptosystem*. in *International Conference on the Theory and Applications of Cryptographic Techniques*. 1997. Springer.

41. Paillier, P. *Public-key cryptosystems based on composite degree residuosity classes*. in *International Conference on the Theory and Applications of Cryptographic Techniques*. 1999. Springer.

42. Galbraith, S.D., *Elliptic curve Paillier schemes.* Journal of Cryptology, 2002. **15**(2): p. 129-138.

43. Castagnos, G., *An efficient probabilistic public-key cryptosystem over quadratic fields quotients.* Finite Fields and Their Applications, 2007. **13**(3): p. 563-576.

44. Cramer, R. and I. Damgård. *Zero-knowledge proofs for finite field arithmetic, or: Can zero-knowledge be for free?* in *Annual International Cryptology Conference*. 1998. Springer.

45. Pfitzmann, B. and M. Waidner. *Anonymous fingerprinting*. in *International Conference on the Theory and Applications of Cryptographic Techniques*. 1997. Springer.

46. Ingram, O., *Statistical methods for meta-analysis*. 1985.

47. Clifton, C., et al., *Tools for privacy preserving distributed data mining*, in *SIGKDD Explorations Newsletter*. 2002. p. 28–34.

48. Αλεξανδρής, N., et al., *Παρατηρήσεις-προτάσεις πάνω στα βιβλία μαθηματικών της Γ΄ Λυκείου, Στατιστική-Συνδυαστική-Πιθανότητες.* Ευκλείδης Γ, 1985(6): p. 7-13.