ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΠΛΗΡΟΦΟΡΙΚΗ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

«Κρυπτογράφηση γενετικών δεδομένων για μελέτες γενετικής συσχέτισης»

Ονοματεπώνυμο: Νικόλαος Κυριακάκης

Επιβλέπων καθηγητής: Παντελής Μπάγκος

Συνεπιβλέπων καθηγητής: Γεώργιος Σπαθούλας

Λαμία, 2021

Περίληψη

Στην παρούσα πτυχιακή εργασία αναλύεται ένας τρόπος να επιτευχθεί μεταφορά γενετικών δεδομένων μεταξύ κόμβων που ανήκουν σε ένα δίκτυο με χρήση κρυπτογραφικών πρωτοκόλλων. Βασικές παράμετροι της εργασίας αποτελούν οι μελέτες γενετικής συσχέτισης (GWAS) καθώς και η μέθοδος της μετα-ανάλυσης, τα αποτελέσματα της οποίας χρησιμοποιούνται στο κρυπτογραφικό πρωτόκολλο που παρουσιάζεται. Αναλύεται ο αλγόριθμος που σχεδιάστηκε, παρουσιάζεται η υλοποίηση του σε γλώσσα python και τα αποτελέσματα που παράγει. Στόχος του λογισμικού που αναπτύχθηκε είναι η χρήση του από ερευνητικά κέντρα ή οργανισμούς υγείας που μελετούν το ίδιο ερευνητικό ερώτημα με σκοπό το συνδυασμό αποτελεσμάτων μετα-ανάλυσης για την εξαγωγή ενός ενιαίου συμπεράσματος .

Abstract

The purpose of this thesis is to describe a way to achieve genetic data transfer between interconnected nodes with the usage of cryptographic protocols. The core concepts of the project are genetic association studies (GWAS) and meta-analysis, the results of which are used in the cryptosystem that was developed. To be more specific, the algorithm that was implemented will be analyzed thoroughly along with the results that it produces. The main goal of the software is to be used by research centers and health organizations that study the same scientific question, in order to combine the meta-analysis results and draw on a collective conclusion.

Πίνακας περιεχομένων

Εισαγωγή

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. 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.

Για να επιτευχθεί αυτό γίνεται ενοποίηση και στατιστική ανάλυση δεδομένων προερχόμενων από διαφορετικές έρευνες οι οποίες προκύπτουν από τυχαιοποιημένες κλινικές δοκιμές. Η διαδικασία αυτή ονομάζεται μετα-ανάλυση και 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.

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

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 εξηγείται με το παρακάτω σενάριο. Έστω ότι έχουμε N 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.

Είναι λογικό αυτή η αλληλεπίδραση μεταξύ των οργανισμών να εγείρει προβλήματα καθώς οι συμμετέχοντες πρέπει να εφαρμόσουν μαθηματικές πράξεις και διαδικασίες στα δεδομένα και ταυτόχρονα να μην αποκαλύπτεται η πραγματική τους τιμή σε κανένα εκτός του κατόχου. Για να

λυθεί το πρόβλημα αυτό γίνεται χρήση ιδιοτήτων που παρέχονται από τα ομομορφικά κρυπτοσυστήματα τα οποία επιτρέπουν πράξεις σε κρυπτογραφημένα δεδομένα.

Ένα ομομορφικό κρυπτοσύστημα κάνοντας χρήση ενός αλγορίθμου υπολογίζει ένα κρυπτογραφημένο άθροισμα ή γινόμενο δύο μηνυμάτων τα οποία δίνουν το δημόσιο κλειδί και τα κρυπτογραφημένα μηνύματα αλλά όχι τα ίδια τα μηνύματα. Για να είναι αποτελεσματικό ένα σύστημα κρυπτογράφησης, είναι σημαντικό να διασφαλίσουμε ότι το μέγεθος των κρυπτοκειμένων παραμένει πολυωνυμικά οριοθετημένο στην παράμετρο ασφαλείας κατά τη διάρκεια επαναλαμβανόμενων υπολογισμών.

Ο αλγόριθμος που αναπτύχθηκε ανήκει στο τομέα του ασφαλούς υπολογισμού και κάνει χρήση ομομορφικών πρωτοκόλλων για να πετύχει το διαμοιρασμό δημόσιων παραμέτρων και τιμών χωρίς να αποκαλύπτει τις ίδιες τις τιμές.

Στη δικιά μας περίπτωση κάθε κόμβος έχει στη κατοχή του γενετικά δεδομένα, από τα οποία επιθυμεί να εξάγει συμπεράσματα με τη διαδικασία της μετα-ανάλυσης. Θεωρούμε ότι κάθε κόμβος έχει δύο μυστικές τιμές $w, w * Y$ και η τελική πράξη που επιθυμεί να εκτελέσει είναι $\sum_{i=1}^{n} wi * Yi/wi$ χρησιμοποιώντας τις μυστικές τιμές όλων των συμμετεχόντων.

Αρχικά οι κόμβοι συνδέονται μεταξύ τους και σχηματίζουν ένα δίκτυο, υπολογίζουν την ακρίβεια των δεδομένων τους και πραγματοποιώντας κρυπτογραφημένη ψηφοφορία βρίσκουν συντελεστές a, b που είναι ικανοί να μετατρέψουν τα δεδομένα τους σε θετικούς ακέραιους αριθμούς, μια διαδικασία γνωστή ως upscaling.

Στη συνέχεια υπολογίζουν βάσει αυτών των συντελεστών δύο τιμές που θα χρησιμοποιηθούν ως παράμετροι ασφαλείας για τα δεδομένα τους ως εξής: $a * w + b,\ a * w * Y + b$. Οι τελευταίες πρέπει να λειτουργούν ως άνω φράγματα στις αριθμητικές τιμές των δεδομένων για να λειτουργήσει σωστά το πρωτόκολλο και συνεπώς εκτελούν και πάλι κρυπτογραφημένη ψηφοφορία για να καταλήξουν σε δύο τιμές που καλύπτουν την απαραίτητη συνθήκη για όλους.

Έπειτα προχωρούν σε κρυπτογραφημένο άθροισμα των τιμών αυτών και αποκτώντας τα επαυξημένα αθροίσματα $\sum_{i=1}^{n} a * wi + b$ και $\sum_{i=1}^{n} a * wi * Yi + b$. Σε αυτά εφαρμόζουν τη διαδικασία του downscaling τοπικά και επαναφέρουν τις κανονικές τιμές τους. Τέλος εκτελούν τη διαίρεση των αθροισμάτων όπως στην σχέση που παρατέθηκε νωρίτερα και εξάγουν το τελικό αποτέλεσμα.

Σχετικές Εργασίες

Τίτλος: Achieving GWAS with homomorphic encryption

Μέθοδος ομομορφικής κρυπτογράφησης για την διεξαγωγή GWAS χρησιμοποιώντας τον κρυπτογραφικό αλγόριθμο CKKS. Με τη χρήση του τελευταίου αυξάνεται ο αριθμός των SNPs που μπορούν να ενσωματωθούν μέσα σε ένα κρυπτοκείμενο. Χρησιμοποιεί δύο βιβλιοθήκες ανοικτού κώδικα, HEAAN και SEAL. Η φιλοσοφία της εργασίας είναι η αποστολή κρυπτοκειμένων σε ένα κεντρικό server που γίνονται οι υπολογισμοί και τα αποτελέσματα στέλνονται πίσω. [23]

Τίτλος: Private Genomes and Public SNPs: Homomorphic Encryption of Genotypes and Phenotypes for Shared Quantitative Genetics

Στόχος της εργασίας είναι η ανταλλαγή γονοτύπων και φαινοτύπων με τρόπο που δεν παραβιάζει την ιδιωτικότητα των ατόμων στα οποία ανήκουν αυτά τα δεδομένα. Οι συμμετέχοντες σε αυτό το σύστημα συμφωνούν να μοιραστούν ένα κοινό σετ από γονότυπους πριν τη κρυπτογράφηση. Στη συνέχεια ο κάθε ένας κρυπτογραφεί τα δεδομένα που έχει στη διάθεση του και τα μοιράζεται με τους υπόλοιπους. Η μέθοδος που χρησιμοποιείται ονομάζεται ορθογώνια κρυπτογράφηση.[24]

Τίτλος: Privacy-preserving Genome-wide Association Studies on cloud environment using fully homomorphic encryption

Στη συγκεκριμένη εργασία σκοπός είναι ο ασφαλής διαμοιρασμός γονοτύπων και φαινοτύπων χρησιμοποιώντας ένα πλήρως ομομορφικό σύστημα κρυπτογράφησης και επιτελώντας όλους τους απαραίτητους υπολογισμούς στο cloud. Το πρωτόκολλο που χρησιμοποιείται δέχεται ένα κρυπτογραφημένο πίνακα συχνοτήτων με γενετικά/κλινικά δεδομένα. Προτείνεται μία μέθοδος για αποδοτικό πακετάρισμα των δεδομένων και εκτέλεσης υπολογισμών σε αυτά.[25]

Τίτλος: Optimized homomorphic encryption solution for secure genome-wide association studies

Ομομορφική κρυπτογράφηση η οποία χρησιμοποιεί μια παραλλαγή του αλγορίθμου CKKS που λέγεται RNS (Residue-Number-System) και διατηρεί τα δεδομένα κρυπτογραφημένα καθ' όλη τη διάρκεια της μελέτης. Ένας αριθμός από βελτιστοποιήσεις ως προς τη παράλληλη επεξεργασία των δεδομένων δίνουν χρονικό πλεονέκτημα σε αυτή τη μεθοδολογία σε σύγκριση με άλλες προσεγγίσεις.[26]

Τίτλος: Ultra-Fast Homomorphic Encryption Models enable Secure Outsourcing of Genotype Imputation

Αυτό το εργαλείο χρησιμοποιείται για τη γενετική σύνδεση μονονουκλεοτιδικών πολυμορφισμών χρησιμοποιώντας προηγμένες στατιστικές μεθόδους. Εξαιτίας της ανάγκης για υπολογιστικούς πόρους είναι απαραίτητο να μοιραστούν τα δεδομένα σε διαφορετικούς κόμβους και αυτό εγείρει προβλήματα ιδιωτικότητας. Για να λυθεί το παραπάνω χρησιμοποιούνται αλγόριθμοι ομομορφικής κρυπτογράφησης (BFV, CKKS, TFHE) με τους οποίους επιτυγχάνεται end to end κρυπτογράφηση των δεδομένων, δηλαδή τα δεδομένα είναι κρυπτογραφημένα καθ' όλη τη διάρκεια της μεταφοράς και της ανάλυσης τους. Ο αλγόριθμος πετυχαίνει χαμηλούς χρόνους εκτέλεσης, όμως παρατηρήθηκε μικρή απώλεια στην ακρίβεια των τελικών αποτελεσμάτων.[27]

Τίτλος: iDASH secure genome analysis competition 2018: blockchain genomic data access logging, homomorphic encryption on GWAS, and DNA segment searching

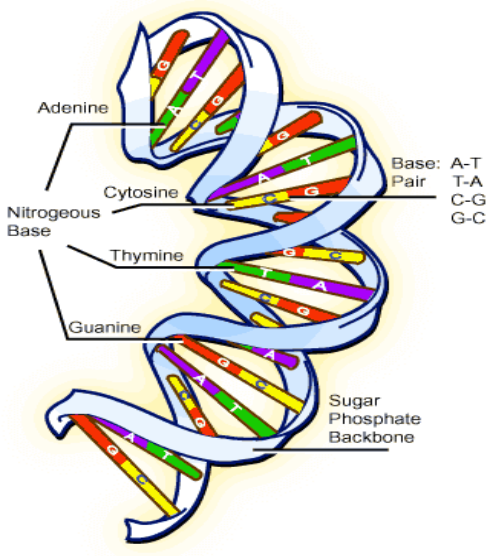
Σε αυτή την εργασία προτείνονται τρόποι καταχώρησης κλινικών/γενετικών δεδομένων χρησιμοποιώντας την τεχνολογία του blockchain καθώς και μέθοδοι ομομορφικής κρυπτογράφησης για την παράλληλη επεξεργασία τέτοιων δεδομένων από πολλούς και διαφορετικούς κόμβους ενός συστήματος. Τέλος γίνεται αναφορά σε τρόπους αφαλούς αναζήτησης τμημάτων DNA από βάσεις δεδομένων.[28]

Γενετικό Υλικό

Το γενετικό υλικό του κυττάρου περιλαμβάνει τα γονίδια που περιέχουν τις πληροφορίες για τη μεταγραφή και τη μετάφραση των γονιδίων σε πρωτεΐνες. Το DNA κωδικοποιεί τις πληροφορίες για την αύξηση και τη διαίρεση των κυττάρων καθώς και τις πληροφορίες για τη διαφοροποίηση των απογόνων ως προς τα εξειδικευμένα κύτταρα τους. Η αποκωδικοποίηση του DNA, η αποσαφήνιση δηλαδή του τρόπου με τον οποίο η δομή του DNA καθορίζει συγκεκριμένες γενετικές επιλογές, επέτρεψε στους επιστήμονες να κατανοήσουν καλύτερα την γενετική της ζωής και την κληρονόμηση ορισμένων χαρακτηριστικών και νόσων. Η ανακάλυψη της δομής του DNA πραγματοποιήθηκε το 1953 από τους Franklin (Φράνκλιν), Τζέιμς Γουότσον (James D. Watson) και Φράνσις Κρικ (Francis Crick). Από πολλούς η ανακάλυψη της διπλής έλικας του DNA θεωρείται ως η μεγαλύτερη βιολογική ανακάλυψη του 20ού αιώνα.

Δομή του DNA

Η διαμόρφωση των μεγάλων μορίων του DNA στο χώρο έχει τη μορφή δύο επιμηκών αλυσίδων, οι οποίες συστρέφονται ελικοειδώς μεταξύ τους. Οι αζωτούχες βάσεις στο DNA είναι τέσσερις: κυτοσίνη C, γουανίνη G, θυμίνη T, αδενίνη A.



Εικόνα 1 : DNA double helix

Οι αζωτούχες βάσεις, ανάλογα με την σειρά αλληλουχίας τους σε τριάδες, κωδικοποιούν το μήνυμα για τη σύνθεση των αμινοξέων του κυττάρου στα ριβοσώματα. Εκεί τα αμινοξέα συνδυάζονται, με τη σειρά κατά την οποία μεταφέρθηκαν στο ριβόσωμα και συντίθενται έτσι οι διαφορετικές πρωτεΐνες [1].

Η Λειτουργία του DNA

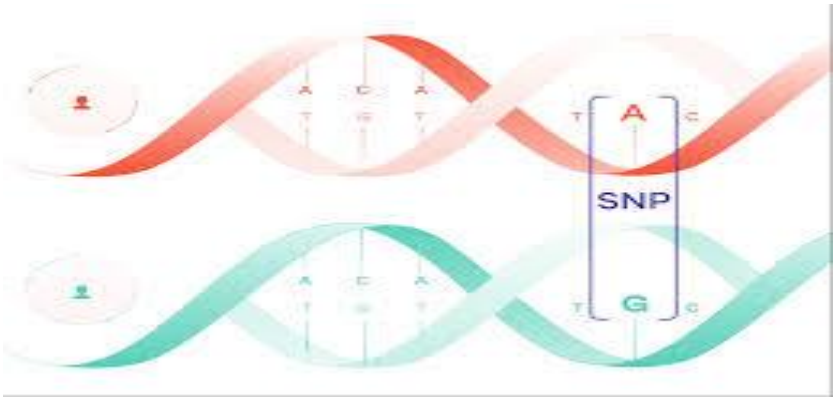
Το DNA βρίσκεται στον πυρήνα των ευκαρυωτικών κυττάρων και άλλων οργανιδίων (π.χ. μιτοχόνδρια) και τους δίνει τη δυνατότητα αυτονομίας στην αναπαραγωγή. Το DNA φέρει τις απαραίτητες οδηγίες προκειμένου ένας οργανισμός να αναπτυχθεί, να επιβιώσει και να αναπαραχθεί. Για να πραγματοποιηθούν οι παραπάνω 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. Μόνο το ένα τοις εκατό (1%) της αλληλουχίας DNA 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 αφορούν τόσο τη μεταβίβαση ιδιοτήτων, αμετάβλητων από γενιά σε γενιά, όσο 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 [2].

Συσχέτιση ολόκληρου του γονιδιώματος

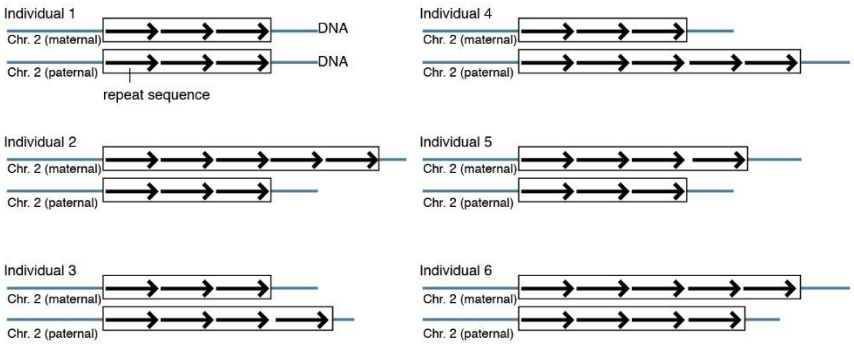
Γενετική ποικιλομορφία

Ο άνθρωπος και η πλειοψηφία των θηλαστικών είναι διπλοειδείς καθώς στο γονιδίωμά τους φέρουν δύο αντίγραφα από κάθε χρωμόσωμα, δηλαδή δύο ομόλογα σημεία του γονιδιώματος υπάρχουν σε κάθε στοιχείο του γονιδιώματος [2][3] και δεν είναι πανομοιότυπα γεγονός στο οποίο οφείλεται η ύπαρξη της γενετικής ποικιλομορφίας. Το ίδιο γονιδιακό χαρακτηριστικό δύο διαφορετικών ατόμων, μπορεί να υπάρχει στο ίδιο γονιδίωμα σε περισσότερες από δύο εκδοχές. Τα γονίδια με περισσότερες από μία μορφές ονομάζονται αλληλόμορφα. 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[4] επαναπροσδιορίζοντας την προσοχή από το επίπεδο των αλληλόμορφων γονιδίων και του γενετικού τόπου. Ο σημειακός νουκλεοτιδικός πολυμορφισμός (SNP) αντιστοιχεί σε μία μοναδική θέση στο απλοειδές γονιδίωμα που διαφέρει μεταξύ των ατόμων του πληθυσμού[5]. 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 [6]. Η πλειονότητα των πολυμορφικών θέσεων έχουν δύο αλληλόμορφα παρότι οι δυνατότητες για ένα SNP είναι μεγαλύτερες από δύο .Εκτός των σημειακών πολυμορφισμών, οι μεγαλύτερες σε μήκος μεταβολές(ποικιλομορφία αριθμού αντιγράφων CNV) μπορούν να οδηγήσουν σε γενετική ποικιλομορφία.



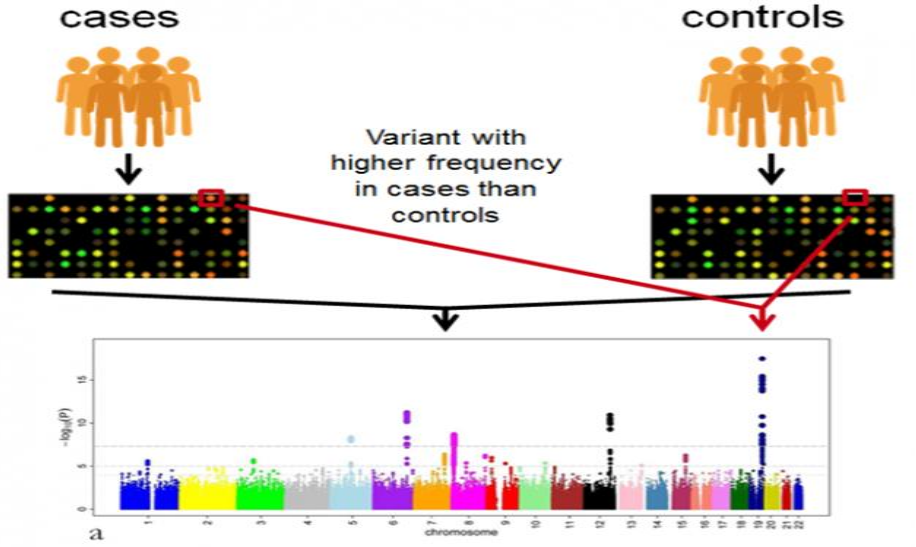
Εικόνα 2 : SNP



Εικόνα 3: CNV

GWAS: Μελέτη συσχετισμού ολόκληρου του γονιδιώματος

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.



Η GWAS ταυτοποιεί SNPs στο DNA που σχετίζονται με μία ασθένεια όμως δεν είναι δυνατόν να συγκεκριμενοποιήσει ποια είναι τα αιτιολογικά γονίδια. 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[7, 8]. Η πρώτη επιτυχής GWAS δημοσιεύτηκε το 2005 και ερευνούσε τους ασθενείς με εκφύλιση της ωχράς κηλίδας σε σχέση με την ηλικία. Κατόπιν σύγκρισης με υγιή δείγματα ελέγχου, βρέθηκαν δύο SNPs με τελείως διαφορετικές συχνότητες αλληλόμορφων[9]. Οι μελέτες GWAS παρουσιάζουν αρκετά προβλήματα και περιορισμούς, παρόλα αυτά μπορούν να αντιμετωπιστούν με τον κατάλληλο έλεγχο ποιότητας και σχεδιασμό μελέτης. Η έλλειψη σαφώς καθορισμένων ομάδων ασθενών και ελέγχου, το ανεπαρκές μέγεθος δείγματος, ο έλεγχος των πολλαπλών δοκιμών και ο έλεγχος της πληθυσμιακής διαστρωμάτωσης είναι κοινά προβλήματα[8].Επιπλέον, η συγκεκριμένη προσέγγιση μπορεί να είναι προβληματική, επειδή ο τεράστιος αριθμός των στατιστικών δοκιμασιών που γίνονται, παρουσιάζουν άνευ προηγουμένου πιθανότητες για ψευδώς θετικά αποτελέσματα[8]. Προσφάτως, η ραγδαία μείωση της τιμής αλληλούχισης ολόκληρου του γονιδιώματος παρέχει επίσης μία ρεαλιστική εναλλακτική των GWAS, που βασίζονται στις γονοτυπικές συστοιχίες.

Η μελέτη συσχέτισης ολόκληρου του γονιδιώματος είναι μία προσέγγιση που περιλαμβάνει σάρωση δεικτών από ολόκληρο το γονιδίωμα σε γονιδιώματα από πολλά άτομα (για παράδειγμα χιλιάδες ασθενείς και χιλιάδες άτομα ελέγχου) για την εύρεση γενετικών παραλλαγών που σχετίζονται με μία ασθένεια [10]. Στις μελέτες συσχέτισης ολόκληρου του γονιδιώματος οι παραλλαγές του γονιδιώματος που βρίσκουμε συχνά σε ένα πληθυσμό μπορεί να είναι υπεύθυνες για τις συχνά παρατηρούμενες ασθένειες.

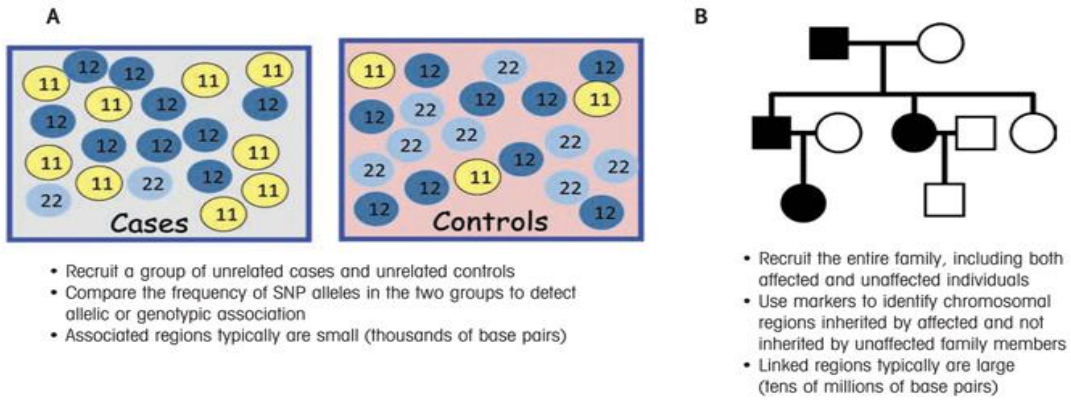
Παράλληλα, αυτές οι μελέτες είναι πλέον δυνατόν να πραγματοποιηθούν εξαιτίας της δομής του γονιδιώματος αλλά και της ιδιότητας ανισορροπίας σύνδεσης που παρατηρείται ανάμεσα στις παραλλαγές του DNA (μη τυχαία συσχέτιση αλληλόμορφων)[11].

Μελέτες συσχέτισης σε οικογένειες

Οι μελέτες συσχέτισης που πραγματοποιούνται σε επίπεδο οικογένειας ερευνούν τη γενετική ποικιλομορφία σε περιορισμένο αριθμό ατόμων. Το πλεονέκτημά τους είναι ότι εκμεταλλεύονται την συνηθισμένη απλοτυπική διάταξη που είναι αναμενόμενο να υφίσταται μεταξύ μελών της ίδιας οικογένειας. Βασικό μειονέκτημά τους είναι ότι ο εντοπισμός συσχετίσεων μεταξύ πολύπλοκων φαινοτύπων και γονοτύπου διενεργείται με δυσκολία.

Μελέτες ασθενών-μαρτύρων (case-control studies)

Οι μελέτες ασθενών μαρτύρων είναι το είδος GWAS που πραγματοποιείται συχνότερα. Βασίζονται στο διαχωρισμό του δείγματος σε δύο κατηγορίες με βάση ένα φαινοτυπικό χαρακτηριστικό δυαδικού τύπου [12]. Πολύ τακτικά ο διττός 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. 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. Στις περιπτώσεις αυτές, διευκολύνεται η ανάλυση των δειγμάτων καθώς χρειάζεται μικρότερης εμβέλειας γονοτύπηση καθώς ο γενετικός τόπος ενδιαφέροντος είναι ήδη γνωστός. 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.



Εικόνα 4: Case-control study – Family study

Μετα–ανάλυση

Η ανάπτυξη της τεχνολογίας έχει ως αποτέλεσμα την αύξηση του όγκου πληροφορίας και έχει οδηγήσει επιστήμονες και ερευνητές στην αναζήτηση μεθόδων που έχουν ως στόχο την καλύτερη διαχείριση 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 ανασκοπήσεις στις οποίες οι συγγραφείς συλλέγουν όλες τις μελέτες με ένα συγκεκριμένο αντικείμενο ενισχύοντας τα πρωταρχικά αποτελέσματα. Οι ανασκοπήσεις χωρίζονται στις περιγραφικές και στις συστηματικές. Οι περιγραφικές ασχολούνται με θεωρητικές απόψεις σχετικά με την νέα και την ήδη υπάρχουσα γνώση ερευνητικών προτάσεων ενώ οι συστηματικές συλλέγουν 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[13]]. 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.

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

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.

Υπολογισμός μεγέθους επίδρασης

Το μέγεθος επίδρασης 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[14]. Η συσχέτιση μίας ασθένειας με ένα παράγοντα γίνεται συνήθως με το odds ratio. 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). Για δύο αλληλόμορφα A, B με A το αλληλόμορφο κινδύνου, οι γονότυποι δομούνται ως εξής για τον επικρατή τύπο κληρονομικότητας AA + AB vs BB, για τον υπολειπόμενο τύπο κληρονομικότητας AA vs AB + BB και για τον συνεπικρατή τύπο A vs B. Οι συγκρίσεις μεταξύ των γονοτύπων συχνά ελαττώνονται σε ένα συγκεκριμένο γενετικό μοντέλο επικρατές και υπολειπόμενο[15].

	Αλληλόμορφο A	Αλληλόμορφο B	N
Ασθενείς	*a*	*b*	*N1*
Υγιείς	*c*	*d*	*N2*

$$OddsRatio = \frac{a * d}{b * c}$$

Οι συνδυασμοί που μπορούν να προκύψουν είναι ο α να είναι ασθενής και να έχει το αλληλόμορφο A, ο b να είναι ασθενής και να έχει το αλληλόμορφο B, ο c να είναι υγιής και να έχει το αλληλόμορφο A και ο συνδυασμός d που είναι υγιής και έχει το αλληλόμορφο B. Η βασική διαδικασία στη μετα-ανάλυση είναι η αντιμετώπιση n μελετών από τις οποίες υπολογίζεται μία κοινή παράμετρος ενδιαφέροντος θi (i = 1, … , 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)[14][16]. Ένα μοντέλο σταθερών επιδράσεων προϋποθέτει ότι όλα τα δείγματα Yi από κάθε μελέτη προέρχονται από έναν ενιαίο πληθυσμό.

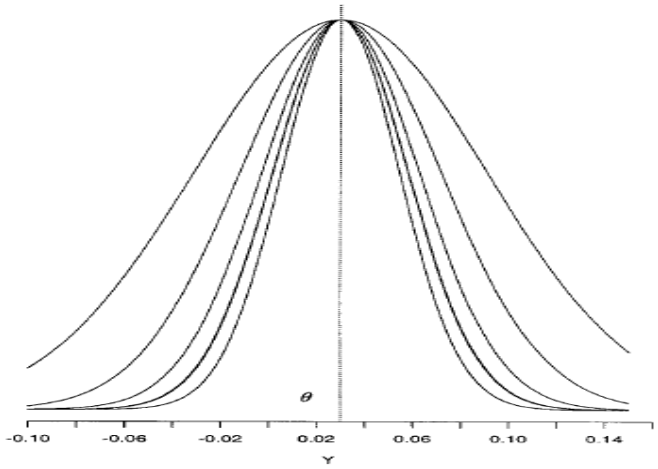
Υποθέτουμε ότι κοινή παράμετρος ενδιαφέροντος είναι η θ, ότι έχουμε 1,2 …n ανεξάρτητες μελέτες και ότι το Yiείναι τέτοιο ώστε $E(Yi)$ = θ και η διακύμανση από κάθε μελέτη si^2 = $var(Yi)$. Για μελέτες μεγάλου μεγέθους, κάθε Yi (δείγματα κάθε μελέτης) πρέπει να ακολουθούν ασυμπωτικά την κανονική κατανομή.

$$\widehat{\boldsymbol{\theta}} = \frac{\sum_{i=0}^{k} \boldsymbol{w_i} * \boldsymbol{Y_i}}{\sum_{i=0}^{k} \boldsymbol{w_i}}, \qquad \boldsymbol{w_i} = \frac{\mathbf{1}}{\boldsymbol{\sigma_i^2}}.$$

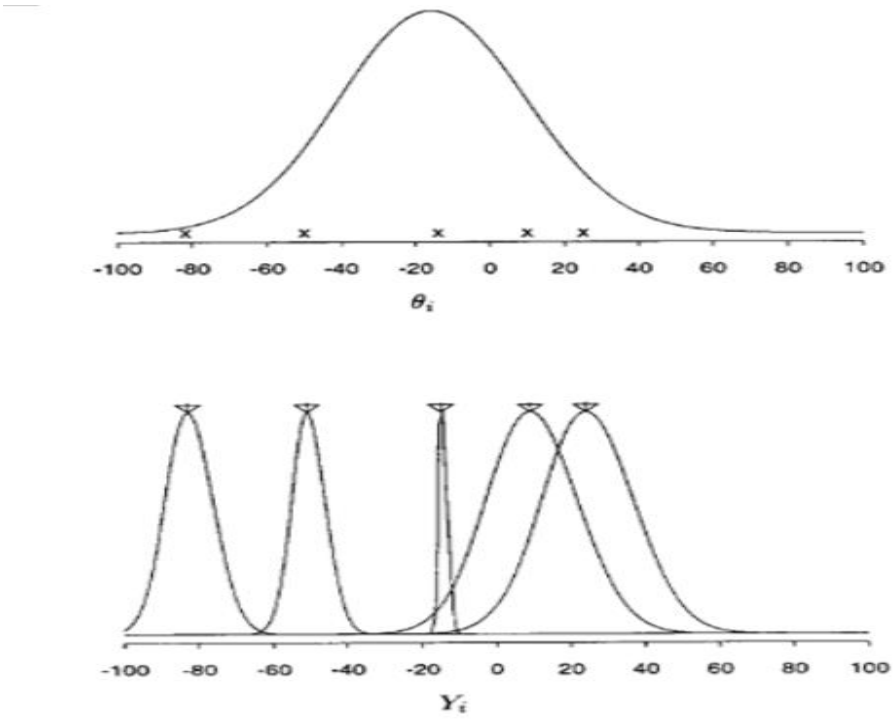
$$\text{όπου } \boldsymbol{Yi = LogOddsRatio}, \qquad \boldsymbol{si} = \frac{\mathbf{1}}{\boldsymbol{A}} + \frac{\mathbf{1}}{\boldsymbol{B}} + \frac{\mathbf{1}}{\boldsymbol{C}} + \frac{\mathbf{1}}{\boldsymbol{D}}$$

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 [13] προέρχονται από μία διανομή πληθυσμού με μέγεθος επίδρασης θi και διακύμανση si^2 . Κάθεθiαπό κάθε μελέτη υποθέτουμε ότι προέρχεται από ανεξάρτητο τυχαίο δείγμα από ένα φυσιολογικό πληθυσμό [16][18] με μέση τιμή θ και τυπική απόκλιση τ^2 με τύπο: $\theta i \sim N(\theta, \tau^2)$ όπου θ και τ^2 αναφέρονται ως υπερπαράμετροι που αντιπροσωπεύουν το κοινό μέγεθος επίδρασης και την διακύμανση αντίστοιχα[17].

$$\widehat{\boldsymbol{\theta(\tau)}} = \frac{\sum_{i=0}^{k} \boldsymbol{w_i}(\boldsymbol{\tau}) * \boldsymbol{Y_i}}{\sum_{i=0}^{k} \boldsymbol{w_i}(\boldsymbol{\tau})}, \qquad \boldsymbol{w_{i(\tau)}} = \frac{\mathbf{1}}{\boldsymbol{\tau^2} + \boldsymbol{\sigma_i^2}}$$



Εικόνα 5: $Y_i \sim N(\theta, \sigma^2)$ για i=1,2,3,…,n. Κατανομή πέντε τυχαίων δειγμάτων για το μοντέλο σταθερών επιδράσεων.



Εικόνα 6: Κατανομή εκτιμητή θi - Κατανομή πέντε τυχαίων δειγμάτων για το μοντέλο τυχαίων επιδράσεων.

Κρυπτογράφηση

Η ανάγκη για κρυπτογράφηση των δεδομένων έχει αυξηθεί ραγδαία τα τελευταία χρόνια, αυτό είναι απόρροια της αύξησης των δεδομένων που παράγει ένας χρήστης και του ηλεκτρονικού αποτυπώματος που δημιουργεί. Η παραπάνω ανάγκη γίνεται επιτακτική όταν μεταφέρονται ιδιωτικά δεδομένα ανάμεσα σε υπολογιστικά συστήματα και οργανισμούς που είναι υπεύθυνοι για τη διαχείριση και τη διασφάλιση τους. Σε συστήματα όπου είναι επιθυμητό να υλοποιηθεί ένας υπολογισμός από περισσότερους από ένα συμμετέχοντες με χρήση των ιδιωτικών δεδομένων του καθενός και ενός κοινού τρόπου υπολογισμού του τελικού αποτελέσματος προκύπτουν προβλήματα εμπιστοσύνης. Αυτό σημαίνει ότι ενδεχομένως ένας συμμετέχοντας σε αυτό τον υπολογισμό επιθυμεί να μάθει τα ιδιωτικά δεδομένα ενός ή περισσοτέρων άλλων συμμετεχόντων. Για να θεωρηθεί ένα σύστημα με τους παραπάνω στόχους ασφαλές, πρέπει κάθε συμμετέχοντας να γνωρίζει μόνο την δική του ιδιωτική τιμή και το κοινό τελικό αποτέλεσμα των πράξεων του υπολογισμού. Κάτι τέτοιο είναι δυνατό να επιτευχθεί με χρήση ομομορφικών συστημάτων κρυπτογράφησης. Αυτό προκύπτει διότι τέτοιου είδους συστήματα επιτρέπουν υπολογισμούς σε κρυπτογραφημένα δεδομένα χωρίς να απαιτείται να αποκρυπτογραφηθούν νωρίτερα.

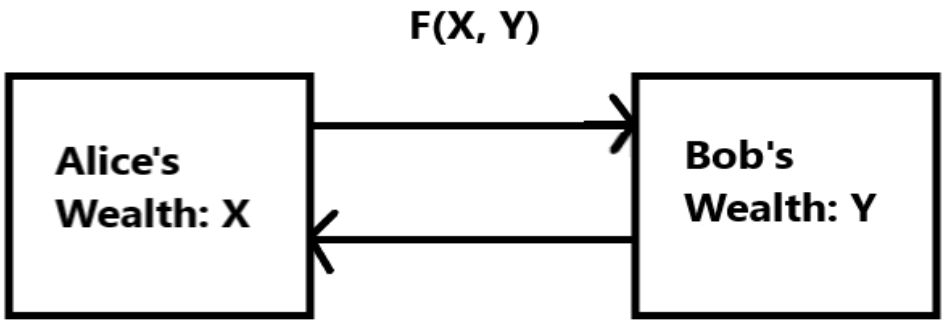
Ασφαλής Υπολογισμός

Ασφαλής Υπολογισμός (Secure Computation) είναι ένας κλάδος της κρυπτογραφίας που ασχολείται με τη δημιουργία μεθόδων για ομάδες που επιθυμούν να υπολογίσουν ένα κοινό αποτέλεσμα χωρίς κανένα μέλος της ομάδας να αποκαλύψει ποτέ τα δεδομένα του στα υπόλοιπα μέλη της ομάδας. [19] Σε τέτοια συστήματα κρυπτογράφησης δεν υπάρχει εγγύηση ότι ένας αντίπαλος δεν βρίσκεται ανάμεσα στα μέλη της ομάδας και συνεπώς σκοπός ενός τέτοιου μοντέλου είναι να διατηρεί την ιδιωτικότητα των δεδομένων κάθε συμμετέχοντα από τους υπόλοιπους. Μία πρώτη προσπάθεια επίτευξης του ασφαλούς υπολογισμού ήταν η ύπαρξη ενός έμπιστου (κεντρικού) διαχειριστή ο οποίος θα διαχειρίζεται τα δεδομένα του εκάστοτε συμμετέχοντα και θα υπολογίζει το αποτέλεσμα των επιθυμητών πράξεων του, χωρίς να αποκαλύπτει τα ιδιωτικά δεδομένα κανενός σε τρίτους. Είναι προφανές πως το προαναφερθέν σενάριο δεν ήταν ρεαλιστικό διότι αν ο κεντρικός διαχειριστής δεν είναι πραγματικά έμπιστος ή υπάρξει κάποια κακόβουλη απόπειρα εναντίον του, τα δεδομένα όλων των συμμετεχόντων θα μπορούσαν να βρεθούν εκτεθειμένα. Η δυσκολία στην υλοποίηση μιας αποτελεσματικής λύσης προκύπτει από το γεγονός ότι οι συμμετέχοντες δεν μπορούν να εμπιστευθούν ο ένας τον άλλον, αλλά ούτε και κάποιο εξωτερικό διαχειριστή. Για παράδειγμα σε μία ηλεκτρονική ψηφοφορία είναι καίριο ζητούμενο να διασφαλιστεί ότι το αποτέλεσμα δε θα τροποποιηθεί και καμία ψήφος δε θα γνωστοποιηθεί σε τρίτους. Ένα άλλο παράδειγμα είναι το πρόβλημα του εκατομμυριούχου. Στο συγκεκριμένο παράδειγμα θεωρούμε ότι έχουμε δύο εκατομμυριούχους, την Alice και τον Bob οι οποίοι επιθυμούν να μάθουν ποιος από τους δύο είναι πιο πλούσιος χωρίς να θέλουν αποκαλύψουν τον προσωπικό πλούτο τους.

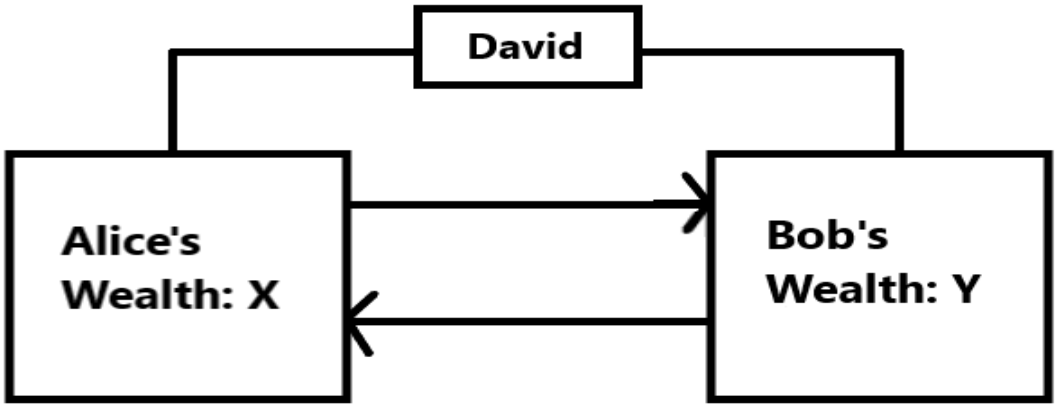
Το τελευταίο παράδειγμα αναλύεται παρακάτω στην απλούστερη εκδοχή του με δύο συμμετέχοντες αλλά μπορεί να γενικευθεί σε Ν συμμετέχοντες.

Ανάλυση προβλήματος του εκατομμυριούχου

Έστω ότι οι Alice και Bob έχουν στη κατοχή τους X και Y τιμές αντίστοιχα και F(X, Y) η συνάρτηση υπολογισμού του μεγαλύτερου εκ των δύο ποσών.

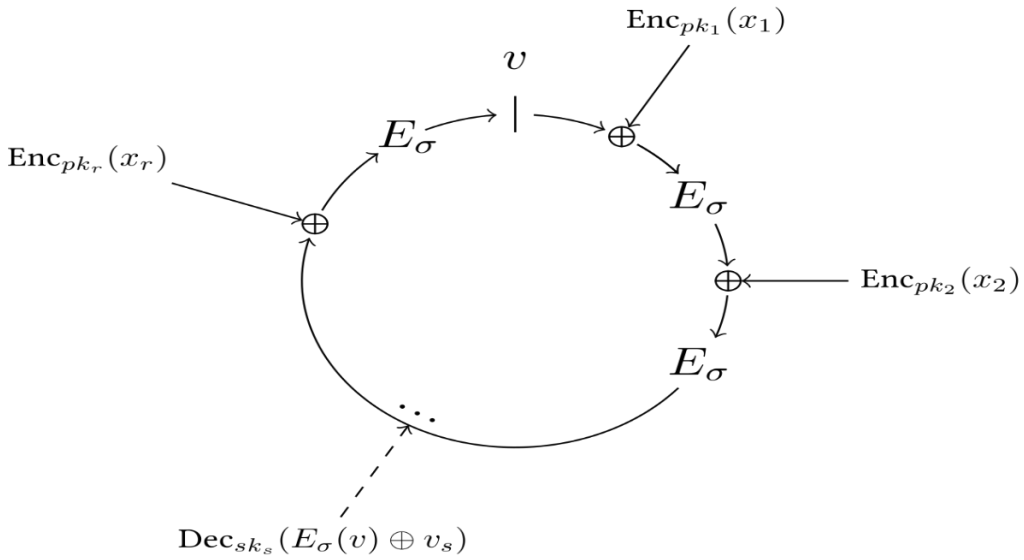


Η δυσκολία στον παραπάνω υπολογισμό έγκειται στο γεγονός ότι ο ένας δεν εμπιστεύεται τον άλλο, οπότε μια απλοϊκή λύση είναι η εισαγωγή ενός τρίτου ατόμου (David) ο οποίος θα εκτελέσει τον υπολογισμό. Συνεπώς η Alice και ο Bob στέλνουν τα X, Y στο David ο οποίος εκτελεί τη πράξη και στέλνει τα αποτελέσματα και στους δύο τους. Ιδανικά ο David δε αποκαλύπτει τις τιμές εισόδου σε κανένα άλλο. Ακόμα και σε αυτό το σενάριο όμως ενδέχεται να υπάρξει διαρροή πληροφορίας. Αυτό είναι δυνατό γιατί για παράδειγμα όταν ο Bob δεχτεί το αποτέλεσμα R = F(X, Y) μπορεί να εκτελέσει R – Y = X.



Κατ' αυτό τον τρόπο γνωρίζει πλέον και τον πλούτο που διαθέτει η Alice και αντίστοιχα η Alice μπορεί να κάνει το ίδιο από μεριά της. Μία πιο αποτελεσματική λύση σε τέτοια περιβάλλοντα είναι η χρήση ομομορφικών κρυπτοσυστημάτων τα οποία αναλύονται αμέσως μετά.

Ομομορφικά Κρυπτοσυστήματα



Εικόνα 7 : Ομομορφικό κρυπτοσύστημα

Ομομορφική κρυπτογράφηση είναι ένα είδος κρυπτογράφησης το οποίο διαθέτει την ικανότητα να επιτελεί υπολογισμούς σε κρυπτογραφημένα δεδομένα χωρίς τη χρήση του μυστικού κλειδιού, με το αποτέλεσμα των υπολογισμών να παραμένει κρυπτογραφημένο.

Όταν τα δεδομένα αποκρυπτογραφηθούν το αποτέλεσμα των πράξεων θα είναι το ίδιο με αυτό που θα είχε παραχθεί αν οι πράξεις είχαν γίνει σε μη κρυπτογραφημένα δεδομένα.

Η γενική ιδέα πίσω από αυτού του είδους τη κρυπτογράφηση είναι η εύρεση μιας συνάρτησης E η οποία μπορεί να υπολογίσει το E(X+Y) και E(X, Y) μόνο με τη χρήση της E(X) και E(Y). Ένα ομομορφικό σύστημα κρυπτογράφησης χρησιμοποιεί έναν αλγόριθμο με σκοπό να υπολογίσει ένα κρυπτογραφημένο άθροισμα ή γινόμενο δύο τιμών. Κατ' αυτό το τρόπο προκύπτει ένα δημόσιο κλειδί και οι κρυπτογραφημένες τιμές που πρόκειται να ανταλλαχθούν μεταξύ των κόμβων του συστήματος. Αυτές οι τιμές πρέπει να παραμείνουν πολυωνυμικά φραγμένες (από υψηλές τιμές ασφαλείας) και μετά από διαδοχικές πράξεις ώστε να διατηρηθεί η ιδιωτικότητα των δεδομένων. Τα ομομορφικά συστήματα βλέπουν κατά κύριο λόγο χρήση σε περιβάλλοντα outsourced storage και cloud computing. Αυτό συμβαίνει διότι τα δεδομένα είναι δυνατόν να διαμοιράζονται σε διαφορετικά μηχανήματα και να επιδέχονται τροποποιήσεις χωρίς να χρειάζεται να αποκρυπτογραφηθούν. [20]

Κύριες κατηγορίες ομομορφικών συστημάτων

Τα ομομορφικά κρυπτοσυστήματα χωρίζονται στις παρακάτω κατηγορίες:

1) Partially homomorphic
2) Somewhat homomorphic
3) Levelled fully homomorphic
4) Fully homomorphic

Από τις παραπάνω κατηγορίες οι πιο διαδεδομένες είναι τα FHE(Fully Homomorphic Encryption) και τα PHE(Partially Homomorphic Encryption).

PHE

Ένα κρυπτοσύστημα το οποίο επιτρέπει μόνο επιλεγμένες μαθηματικές συναρτήσεις να εκτελούνται απεριόριστες φορές σε κρυπτογραφημένες τιμές ονομάζεται PHE. Αυτού του είδους η κρυπτογράφηση είναι εν μέρει το θεμέλιο για την κρυπτογράφηση RSA και έχει εφαρμογή στη δημιουργία ασφαλών συνδέσεων μέσω SSL / TLS.

FHE

Ένα κρυπτοσύστημα που είναι ικανό να εφαρμόζει αυθαίρετους υπολογισμούς σε κρυπτοκείμενα ονομάζεται FHE. Η κατασκευή προγραμμάτων με βάση ένα τέτοιο σχήμα δίνει τη δυνατότητα να εκτελούνται αριθμητικές πράξεις και υπολογισμοί σε κρυπτογραφημένες εισόδους τις οποίες δε χρειάζεται ποτέ να αποκρυπτογραφήσει. Αυτό το χαρακτηριστικό δίνει τη δυνατότητα οι υπολογισμοί να εκτελούνται ακόμα και από μη αξιόπιστους συμμετέχοντες στο σύστημα. Τέτοια είδη συστημάτων απαντώνται πολύ συχνά στο πλαίσιο του cloud computing.

Παραδείγματα και Βιβλιοθήκες ομομορφικών κρυπτοσυστημάτων

PHE[19]

Unpadded RSA: αν το δημόσιο κλειδί RSA έχει $modulus\ n$ και εκθέτη e, η κρυπτογράφηση του m δίνεται από τη σχέση $\varepsilon(m) = m^e \bmod n$
και η ομομορφική ιδιότητα προκύπτει ως εξής:

$$\begin{aligned}\mathcal{E}(m_1) \cdot \mathcal{E}(m_2) &= m_1^e m_2^e \mod n \\ &= (m_1 m_2)^e \mod n \\ &= \mathcal{E}(m_1 \cdot m_2)\end{aligned}$$

ElGamal: σε ένα κυκλικό γκρουπ G με order q και generator g, αν το δημόσιο κλειδί είναι (G, q, g, h) όπου $h = g^x$, x είναι το μυστικό κλειδί. Το μήνυμα m δίνεται από τη σχέση $\varepsilon_{(m)} = (g^r, m \cdot h^r)$ με $r \in \{0, \dots, q-1\}$ και την παρακάτω ιδιότητα

$$\begin{aligned}\mathcal{E}(m_1) \cdot \mathcal{E}(m_2) &= (g^{r_1}, m_1 \cdot h^{r_1})(g^{r_2}, m_2 \cdot h^{r_2}) \\ &= (g^{r_1+r_2}, (m_1 \cdot m_2) h^{r_1+r_2}) \\ &= \mathcal{E}(m_1 \cdot m_2).\end{aligned}$$

- **GoldWasser-Micali**: Η κρυπτογράφηση για ένα bit b προκύπτει
από τη σχέση: $\varepsilon(b) = x^b \cdot r^2 \ mod\ n$, με $r \in \{0, \dots, q-1\}$ και την ιδιότητα

$$\begin{aligned}\mathcal{E}(m_1) \cdot \mathcal{E}(m_2) &= (g^{r_1}, m_1 \cdot h^{r_1})(g^{r_2}, m_2 \cdot h^{r_2}) \\ &= (g^{r_1+r_2}, (m_1 \cdot m_2) h^{r_1+r_2}) \\ &= \mathcal{E}(m_1 \cdot m_2).\end{aligned}$$

FHE[19] (βιβλιοθήκες που υλοποιούν πλήρη ομομορφική κρυπτογράφηση)

- HElib
- Microsoft SEAL
- FHEW

Εφαρμογές ομομορφικών κρυπτοσυστημάτων

Multiparty computations

Διαφορετικοί συμμετέχοντες λαμβάνουν μέρος στον υπολογισμό μίας κοινής συνάρτησης χρησιμοποιώντας μυστικές ιδιωτικές τιμές. Το παραπάνω ανήκει σε τύπους προβλημάτων όπως αυτά που αναφέρθηκαν νωρίτερα π.χ. (πρόβλημα εκατομμυριούχου).

Mobile agent protection

Στη περίπτωση αυτή έχουμε λογισμικό το οποίο σώζει τη τρέχουσα κατάστασή του (process image) και μεταβαίνει από ένα host μηχάνημα σε ένα άλλο συνεχίζοντας την εκτέλεση του από το σημείο που είχε σταματήσει. Κατ' αυτό το τρόπο κάθε μηχάνημα έχει πρόσβαση στα δεδομένα που επεξεργάζεται το λογισμικό. Τα ομομορφικά κρυπτοσυστήματα δίνουν μια πολύ κομψή λύση στο πρόβλημα αυτό διότι κάθε συνδεδεμένο μηχάνημα μπορεί και εκτελεί πράξεις σε κρυπτογραφημένα δεδομένα.

Election protocols

Τέτοιου είδους πρωτόκολλα χρησιμοποιούνται για να υπάρξει συμφωνία μεταξύ πολλών παραγόντων σε ένα σύστημα και να αποφασισθεί μια εκλογική διαδικασία χωρίς να αποκαλυφθεί καμία μεμονωμένη ψήφος.

Watermarking / Fingerprinting

Πραγματεύονται την ενσωμάτωση επιπλέον μοναδικών χαρακτηριστικών σε ψηφιακά δεδομένα, με τέτοιο τρόπο ώστε να μπορεί να μπορεί να ταυτοποιηθεί ο ιδιοκτήτης ή να επιβεβαιωθεί ότι τα δεδομένα αυτά δεν έχουν τροποποιηθεί από κάποιο τρίτο άτομο. Το αποτέλεσμα της λειτουργίας τους είναι όμοιο με μια ψηφιακή υπογραφή που δημιουργεί ένα PGP κλειδί.

Υλοποίηση εφαρμογής

Η ιδέα όπισθεν της εφαρμογής είναι ο υπολογισμός αθροισμάτων και γινομένων χωρίς τη χρήση ασφαλών καναλιών και στηρίζεται στην έρευνα που διεξήγαγε ο Clifton [20] και προτείνει δύο εκδοχές. Η πρώτη εκδοχή περιλαμβάνει ένα κεντρικό κόμβο ο οποίος λειτουργεί ως συλλέκτης (Aggregator) που λαμβάνει αριθμητικές τιμές από τους υπόλοιπους συμμετέχοντες. Οι τελευταίοι δεν έχουν δικαίωμα υπολογισμού και δεν γνωρίζουν τις ιδιωτικές τιμές των άλλων. Η δεύτερη εκδοχή εξαλείφει το κόμβο που λειτουργεί ως συλλέκτης και θεωρεί κάθε συμμετέχοντα ισότιμο ο οποίος έχει το δικαίωμα υπολογισμού του τελικού αποτελέσματος. Παρόλα αυτά είναι απαραίτητο να υπάρχει ένας κόμβος εκκίνησης ο οποίος αρχικοποιεί συγκεκριμένες τιμές τις οποίες διαμοιράζει στους υπόλοιπους και κατά τα άλλα έχει την ίδια λειτουργία με αυτούς.

Περιγραφή αλγορίθμου

Γενικό πρωτόκολλο

Έστω ότι έχουμε N κόμβους οι οποίοι θέλουν να στείλουν κρυπτογραφημένα δεδομένα. Κάθε κόμβος έχει στη κατοχή του δύο μυστικούς αριθμούς w_i και $w_i * Y_i$. Έπειτα πρέπει να βρεθούν δύο αριθμοί a, b τέτοιοι ώστε τα ποσά $(a * w_i + b)$ και $(a * w_i * Y_i + b)$ να είναι θετικοί ακέραιοι αριθμοί.

- Ο αριθμός a πρέπει 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 κάθε συμμετέχοντας να έχει στη διάθεση του ένα ακέραιο αριθμό αφού τον πολλαπλασιάσει με αυτόν.

- Με τη χρήση του πρωτοκόλλου max το οποίο θα αναλυθεί πιο κάτω βρίσκεται το μέγιστο a από όλους τους κόμβους, δηλαδή το ελάχιστο δυνατό πλήθος δεκαδικών ψηφίων που ικανοποιεί τη προηγούμενη συνθήκη για αυτό.

- Στη συνέχεια υπολογίζεται το μέγιστο b πάλι με τη χρήση του max και τελικά κάθε κόμβος έχει στη κατοχή του ένα max_a και ένα max_b με τα οποία εξασφαλίζει ότι κάθε w_i και $w_i * Y_i$ είναι θετικοί ακέραιοι αριθμοί.

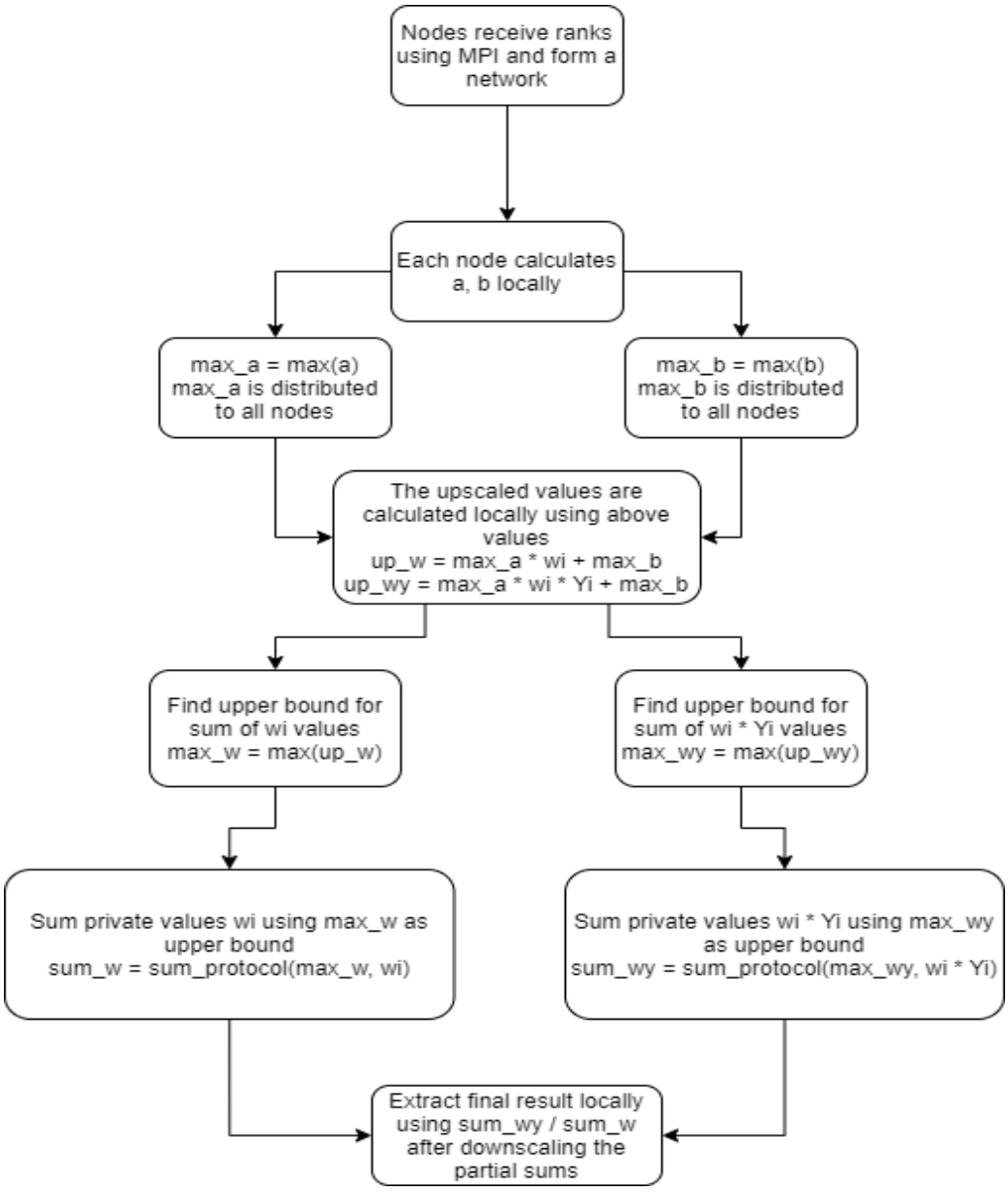
- Έπειτα τρέχει το πρωτόκολλο max άλλες δύο φορές με τιμές εισόδου:

    - $up\_w = max\_a * w_i + b$
    - $up\_wy = max\_a * w_i * Y_i + b$

- Επόμενο βήμα είναι η εκτέλεση του sum πρωτοκόλλου με τιμές αρχικοποίησης up_w, up_wy και τον αριθμό των κόμβων ώστε να βρεθεί ο πρώτος αριθμός p.
- Ακολουθεί η διαδικασία υποβάθμισης του αθροίσματος με τον τύπο: $\frac{(sum - n * \max\_b)}{\max\_a}$ και μετά τον υπολογισμό των επιμέρους αθροισμάτων $\sum_{i=1}^{n} wi$ και $\sum_{i=1}^{n} w_i Y_i$ γίνεται η διαίρεση $\frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} wi}$ που δίνει και το τελικό αποτέλεσμα.

```
start = MPI.Wtime()
max_precision = max(precision(w), precision(wy))

a = pow(10, max_precision)
b = max(-a * w, -a * wy) + 1

max_a = pow(10, len(str(max_protocol(a))) + 1)
max_b = max_protocol(b)

upscale_wy = int(max_a * wy + max_b)
max_wy = max_protocol(upscale_wy)

upscale_w = int(max_a * w + max_b)
max_w = max_protocol(upscale_w)

sum_wy = sum_protocol(max_num=max_wy, data=upscale_wy)
sum_wy = (sum_wy - size * max_b) / max_a

sum_w = sum_protocol(max_num=max_w, data=upscale_w)
sum_w = (sum_w - size * max_b) / max_a
print("Rank: {}, result: {}, execution time: {}".format(rank, sum_wy / sum_w, MPI.Wtime() - start))
```

Εικόνα 1: Στιγμιότυπο του γενικού πρωτοκόλλου



Εικόνα 2: Ροή εφαρμογής

Max πρωτόκολλο

Το πρωτόκολλο max χρησιμοποιείται για να βρεθεί μια τιμή αρκετά μεγάλη ώστε να ικανοποιεί τις ανάγκες κάθε κόμβου. Εσωτερικά χρησιμοποιείται το πρωτόκολλο του γινομένου για να μπορεί κάθε κόμβος να δημιουργεί μια ψήφο η οποία θα σηματοδοτεί εάν η εκάστοτε τιμή που δέχθηκε σαν όρισμα είναι αρκετά υψηλή για τον ίδιο . Για να επιτευχθεί αυτό χρησιμοποιούνται διαδοχικοί γύροι υπολογισμών και επικοινωνίας μεταξύ των κόμβων.

Αρχικά κάθε κόμβος παράγει αυθαίρετα ένα τυχαίο αριθμό δεκαδικών ψηφίων x. Έπειτα ξεκινά μια επαναληπτική διαδικασία και υπολογίζεται ένα max_value = $x \cdot 2^r$, όπου r είναι ο γύρος υπολογισμού. Κάθε κόμβος υπολογίζει την ψήφο του χρησιμοποιώντας το λόγο max_candidate/max_value + 1 και τη διαμοιράζει στους υπόλοιπους. Με τη βοήθεια του πρωτοκόλλου του γινομένου εκτελείται κρυπτογραφημένο γινόμενο των ψήφων. Αν το αποτέλεσμα του γινομένου είναι μεγαλύτερο του 1 ο γύρος υπολογισμού αυξάνεται και συνεχίζεται η διαδικασία ενώ αν το αποτέλεσμα είναι ίσο με μονάδα τότε όλοι οι κόμβοι βρίσκονται σε συμφωνία και η διαδικασία σταματά.

```
cpt_round = 1
while True:
    if rank == 0:
        low, high = prt.product_bounds(size)
        p = prt.prime_p(low, high)
        q = prt.prime_q(p)
        g1 = prt.generator_1(p, q)
        init_data = [p, q, g1]
    else:
        init_data = []
    init_data = comm.bcast(init_data, root=0)
    p, q, g1 = init_data
    r = prt.secret_random(q)
    param_y = pow(g1, r, p)
    prev_param, next_param = exchange_params(param_y)
    R = prt.random_generator(next_param, prev_param, r, p)
    max_value = digits * pow(2, cpt_round)
    vote = int(candidate / max_value) + 1
    if vote > 1:
        vote = 2
    vote = (vote * R) % p
```

```
all_votes = comm.allgather(vote)
final_vote = 1
for v in all_votes:
    final_vote *= v
final_vote %= p
if final_vote == 1:
    break
else:
    cpt_round += 1
```

Εικόνα 3 : Στιγμιότυπο max protocol

Άθροισμα και Γινόμενο δίχως ασφαλές κανάλι

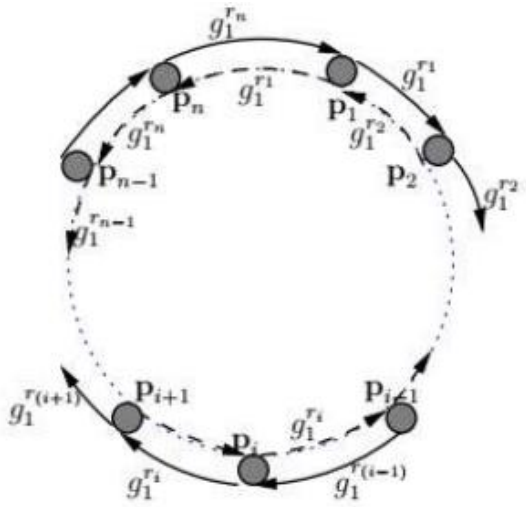
Αρχικά θεωρούμε πως υπάρχουν N-συμμετέχοντες οι οποίοι επιθυμούν να συμμετέχουν σε ένα κοινό υπολογισμό χωρίς να αποκαλύψουν τα ιδιωτικά τους δεδομένα. Στη συνέχεια ορίζονται δύο γκρουπ G1, G2 τα οποία προκύπτουν με τη παρακάτω διαδικασία[22].

1. Υπολογισμός δύο πρώτων αριθμών του ίδιου «μεγέθους» p και q, τέτοιους ώστε ο q να διαιρεί τον $p-1$ ακριβώς.
2. Υπολογισμός ενός τυχαίου αριθμού $h \in [2, p]$
3. Υπολογισμός της τυχαίας γεννήτριας $g_1 = h^{(p-1)/q} \ mod\ p \ \ s.t. \ g_1 \neq 1 \ mod\ p$ η οποία αντιστοιχεί στο γκρουπ G1
4. Υπολογισμός της τυχαίας γεννήτριας $g_2 = g_1^p \ mod\ P^2$

Πρωτόκολλο Γινομένου

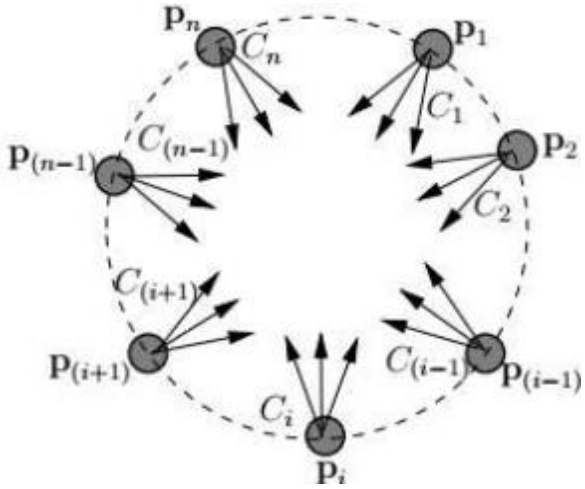
Σε αυτό το πρωτόκολλο οι συμμετέχοντες P_i με $i \epsilon [1, n]$ υπολογίζουν τη συνάρτηση $f(x) = \prod_{i=1}^{n} x_i, x_i \in \mathbb{Z}_P$. Αρχικά πρέπει να υπολογιστούν τυχαίοι ακέραιοι αριθμοί $R_i \epsilon G_1$ τέτοιοι ώστε να ισχύει $\prod_{i=1}^{n} R_i = 1 \, mod \, p$. Η διαδικασία χωρίζεται σε τρία βασικά μέρη και αυτά είναι τα ακόλουθα[22].

1. **Setup** $\rightarrow ri \in \mathbb{Z}q, Ri = \left(\frac{g_1^{r_i+1}}{g_1^{r_{i-1}}}\right)^{r_i} \in G_1$
   - Ας θεωρήσουμε πως οι συμμετέχοντες σχηματίζουν ένα κύκλο και καθένας παράγει ένα τυχαίο αριθμό $r_i \in \mathbb{Z}q$ και υπολογίζει τη παράμετρο $g_1^{r_i}, g_1 \in G_1$
   - Στη συνέχεια κάθε ένας υπολογίζει μια δημόσια παράμετρο $Y_i = g_1^{r_i}, g_1 \in G_1$ και τη μοιράζεται με τους δύο γείτονες του $p_{i-1} \, και \, p_{i+1}$.
   - Εφόσον μετά από ένα γύρο ανταλλαγών κάθε κόμβος έχει στη κατοχή του τις δημόσιες παραμέτρους των γειτόνων του υπολογίζει τον αριθμό $R_i = \left(\frac{Y_{i+1}}{Y_{i-1}}\right)^{r_i} = \left(\frac{g_1^{r_i+1}}{g_1^{r_{i-1}}}\right)^{r_i} \epsilon \, G_1$



Εικόνα 4: Ανταλλαγή δημόσιων παραμέτρων[21]

2. **Encrypt**
   - Σε αυτό το στάδιο κάθε κόμβος δημιουργεί το κρυπτοκείμενο $C_i = x_i R_i = x_i \left(\frac{g_1^{r_i+1}}{g_1^{r_{i-1}}}\right)^{r_i}$ και το στέλνει στους υπόλοιπους.



Εικόνα 5: Διαμοιρασμός κρυπτοκειμένων[21]

3. **Product**
   - Κάθε κόμβος έχει λάβει n – 1 κρυπτοκείμενα από τους υπόλοιπους συμμετέχοντες και υπολογίζει το γινόμενο: $\prod_{i=1}^{n} C_i = \prod_{i=1}^{n} x_i \left(\frac{g_1^{r_i+1}}{g_1^{r_{i-1}}}\right)^{r_i} mod\ p = (\prod_{i=1}^{n} x_i) \prod_{i=1}^{n} \left(\frac{g_1^{r_i+1}}{g_1^{r_{i-1}}}\right)^{r_i} mod\ p = (\prod_{i=1}^{n} x_i)\ g_1^{\sum_{i=1}^{n}(r_{i+1} \cdot r_i - Y_i r_{i-1})} mod\ p = (\prod_{i=1}^{n} x_i)\ mod\ p$

   - Απαραίτητη προϋπόθεση ώστε να διασφαλιστεί σωστό αποτέλεσμα χωρίς τη χρήση υπολοίπου είναι ότι ο αριθμός p που θα επιλεγεί πρέπει να είναι αρκετά μεγάλος, δηλαδή $p \geq M^n$, *όπου Μ είναι το άνω όριο του* x_i

Πρωτόκολλο του Αθροίσματος

Σε αυτή τη ενότητα αναλύεται το πρωτόκολλο του αθροίσματος με συνάρτηση υπολογισμού $f(x) = \sum_{i=1}^{n} x_i$. Ο στόχος του πρωτοκόλλου είναι η μετατροπή του αθροίσματος σε γινόμενο και η βασική του ιδέα βασίζεται στον παρακάτω τύπο[22].

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

Από τη σχέση (1) προκύπτει:

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

Αντίστοιχα με το προηγούμενο πρωτόκολλο και το παρόν αποτελείται από τρία στάδια Setup, Encrypt, Sum.

Setup: $\boldsymbol{r_i} \in \mathbb{Z}q, Ri = \left(g_2^{r_{i+1}} / g_2^{r_{i-1}}\right)^{r_i} \epsilon\, G_2$

Κάθε συμμετέχων επιλέγει τυχαία ένα μυστικό αριθμό $r_i \epsilon\, \mathbb{Z}q$ και υπολογίζει τη δημόσια παράμετρο $Y_i = g_2^{r_i} \in G_2$ και την προωθεί στους δύο γείτονες του p_{i+1} και p_{i-1}. Στη συνέχεια και ύστερα από ένα γύρο ανταλλαγών κάθε συμμετέχοντας υπολογίζει το $Ri = \left(\frac{g_2^{r_i+1}}{g_2^{r_{i-1}}}\right)^{r_i} mod\, p^2$ το οποίο χρησιμοποιεί ως γεννήτρια τυχαιοποίησης.

Encrypt:

Κάθε συμμετέχοντας υπολογίζει τη τιμή του κρυπτοκειμένου $C_i = (1 + x_i * p) * Ri\ mod\, p^2$ και το γνωστοποιεί σε όλους τους υπόλοιπους, οι οποίοι με τη σειρά τους εκτελούν την ακόλουθη πράξη:

$$C = \prod_{L=1}^{n} C_i \bmod p^2$$

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

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

Τέλος υπολογίζει το πηλίκο $(C-1)\ /\ p = \sum_{i=1}^{n} x_i \bmod p$ το οποίο δίνει το τελικό άθροισμα.

```python
def sum_protocol(max_num, data):
    # comm.barrier()
    if rank == 0:
        low, high = prt.sum_bounds(size, max_num)
        p = prt.prime_p(low, high)
        q = prt.prime_q(p)
        g1 = prt.generator_1(p, q)
        g2 = prt.generator_2(g1, p)
        init_data = [p, q, g2]
    else:
        init_data = []
    init_data = comm.bcast(init_data, root=0)
    p, q, g2 = init_data
    r = prt.secret_random(q)
    p_sq = p * p
    param_y = pow(g2, r, p_sq)
    prev_param, next_param = exchange_params(param_y)
    R = prt.random_generator(next_param, prev_param, r, p_sq)
    encrypted_data = ((1 + data * p) * R) % p_sq
    all_data = comm.allgather(encrypted_data)
    final_sum = 1
    for val in all_data:
        final_sum = (final_sum * val) % p_sq
    return (final_sum - 1) / p
```

Εικόνα 6 : Στιγμιότυπο sum protocol

Συγχρονισμός Διεργασιών

Σε ένα σύστημα με πολλούς κόμβους προκύπτει το πρόβλημα του συγχρονισμού και της επικοινωνίας μεταξύ αυτών. Αντίστοιχα και στη παρούσα εργασία λόγω του αυξημένου φόρτου επικοινωνίας που απαιτείται σε πραγματικό χρόνο μεταξύ των συμμετεχόντων προκύπτουν αρκετά προβλήματα που αφορούν στο συγχρονισμό των πράξεων που εκτελεί ταυτόχρονα με τους υπόλοιπους ο κάθε κόμβος. Για παράδειγμα μια διαδικασία που εκτελείται σε έναν από όλους μπορεί να καθυστερήσει να τελειώσει τους υπολογισμούς της ή να υπάρξει κάποιο σφάλμα στο δίκτυο. Η εφαρμογή για να λειτουργεί σωστά πρέπει να διασφαλίζει ότι κάθε κόμβος βρίσκεται σε ένα σημείο εκτέλεσης είτε περιμένοντας τους υπόλοιπους να φτάσουν σε αυτό είτε το ανάποδο.

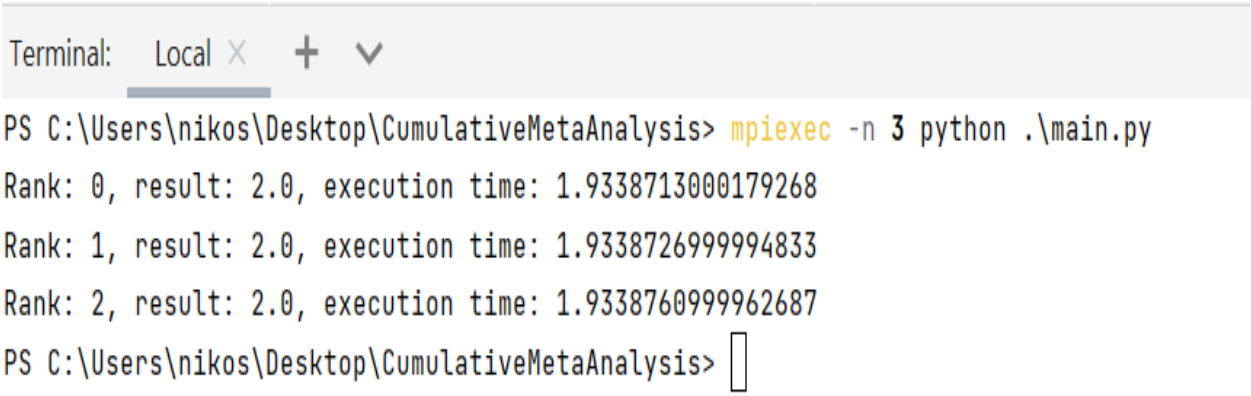
Για να επιτευχθεί αυτό, χρησιμοποιήθηκε η MPI(Message Passing Interface) η οποία απλοποιεί σημαντικά την επικοινωνία μεταξύ των κόμβων και παρέχει μεθόδους συντονισμού μέσω της αποστολής μηνυμάτων μεταξύ των διεργασιών που ανήκουν σε ένα ετερογενές δίκτυο.

Περιγραφή Εκτέλεσης

Αρχικά ο κόμβος εκκίνησης χρίζεται με το βαθμό (rank = 0) από την MPI, δηλαδή μπορεί να λειτουργήσει ως συντονιστής για πολλές διαδικασίες συντονισμού των διεργασιών. Κάθε κόμβος υπολογίζει τοπικά την ακρίβεια σε δεκαδικά ψηφία των δύο μυστικών τιμών που διαθέτει και επιλέγει τη μεγαλύτερη. Η τελευταία χρησιμοποιείται για να υπολογιστούν οι τιμές a, b που εφαρμόζουν τη διαδικασία του upscaling στις μυστικές ιδιωτικές τιμές wi $και$ $wi * Yi$. Στη συνέχεια όλοι οι κόμβοι εκτελούν το πρωτόκολλο max για τις a, b διαδοχικά και εξάγουν τα max_a, max_b τα οποία είναι μέγιστες τιμές για όλους τους.

Οι max_a και max_b χρησιμοποιούνται για τα up_w = max_a * wi + max_b και up_wy = max_a * wi * Yi + max_b. Με τιμές εισόδου στο max πρωτόκολλο τις up_w, up_wy προκύπτουν οι max_w, max_wy. Σε αυτό το σημείο μπορούμε να κάνουμε τα τελικά αθροίσματα χρησιμοποιώντας το πρωτόκολλο του αθροίσματος με άνω φράγμα στα δεδομένα που θα αθροιστούν τις max_w και max_wy και τις ιδιωτικές του τιμές αντίστοιχα και έτσι να υπολογιστούν οι παραστάσεις $\sum_{i=1}^{n} w_i$ $και$ $\sum_{i=1} w_i * Y_i$. Τέλος κάθε κόμβος διαιρεί τις τιμές αυτές και έχει στη κατοχή του το τελικό αποτέλεσμα. Ακολουθεί και το σχετικό διάγραμμα που παρουσιάζει όλα τα παραπάνω.

Εκτέλεση Προγράμματος τοπικά στη γραμμή εντολών και αποτελέσματα



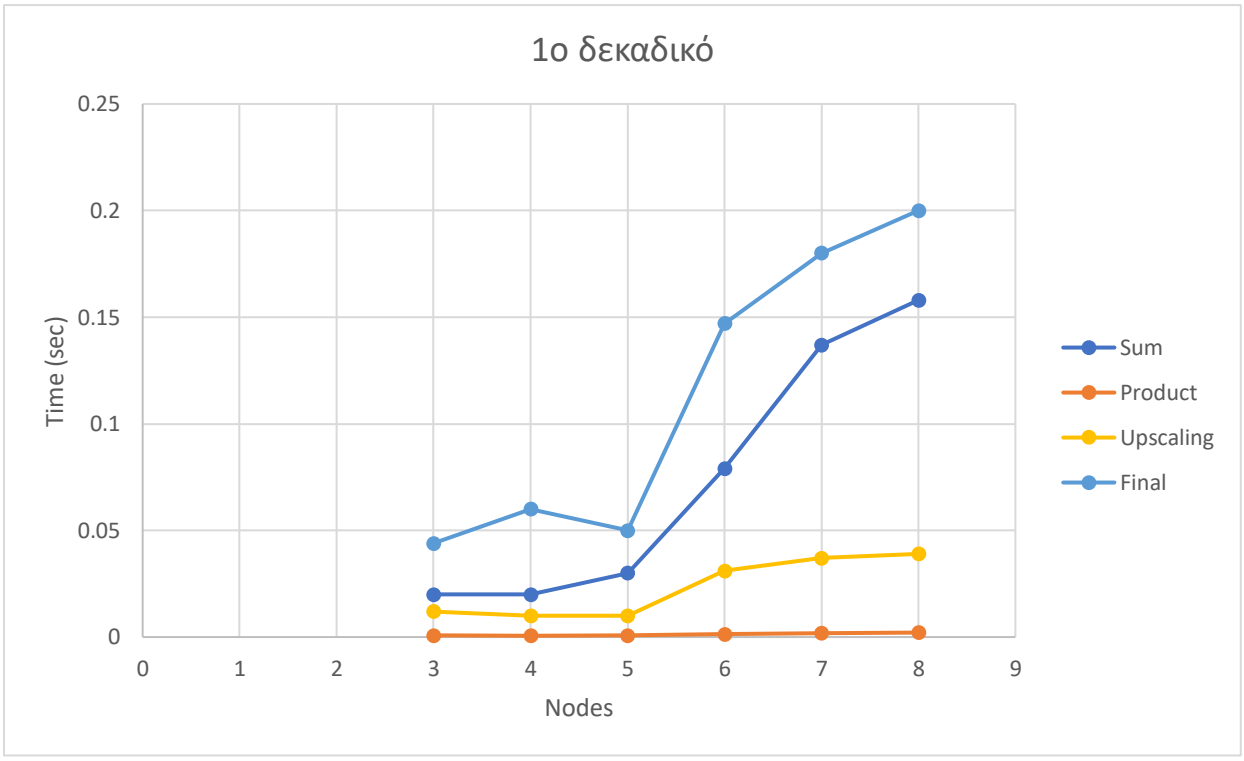
Εικόνα 7: Εκτέλεση προγράμματος για 3 κόμβους

Για να τρέξει το πρόγραμμα σε **διαφορετικούς υπολογιστές** μπορεί να τροποποιηθεί η παραπάνω εντολή ως εξής:

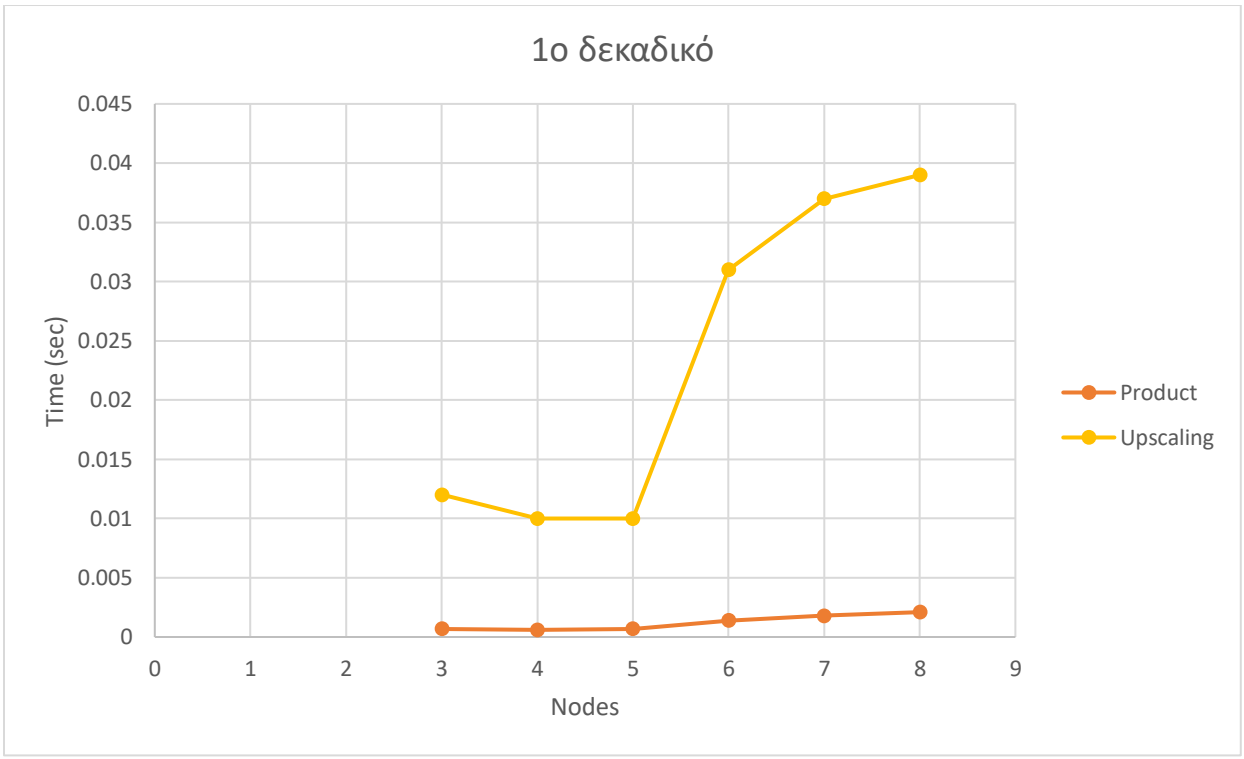
mpiexec -n #total_processes -ppn #processes_per_node -hosts node1, node2, … python ./main.py

Μετρήσεις

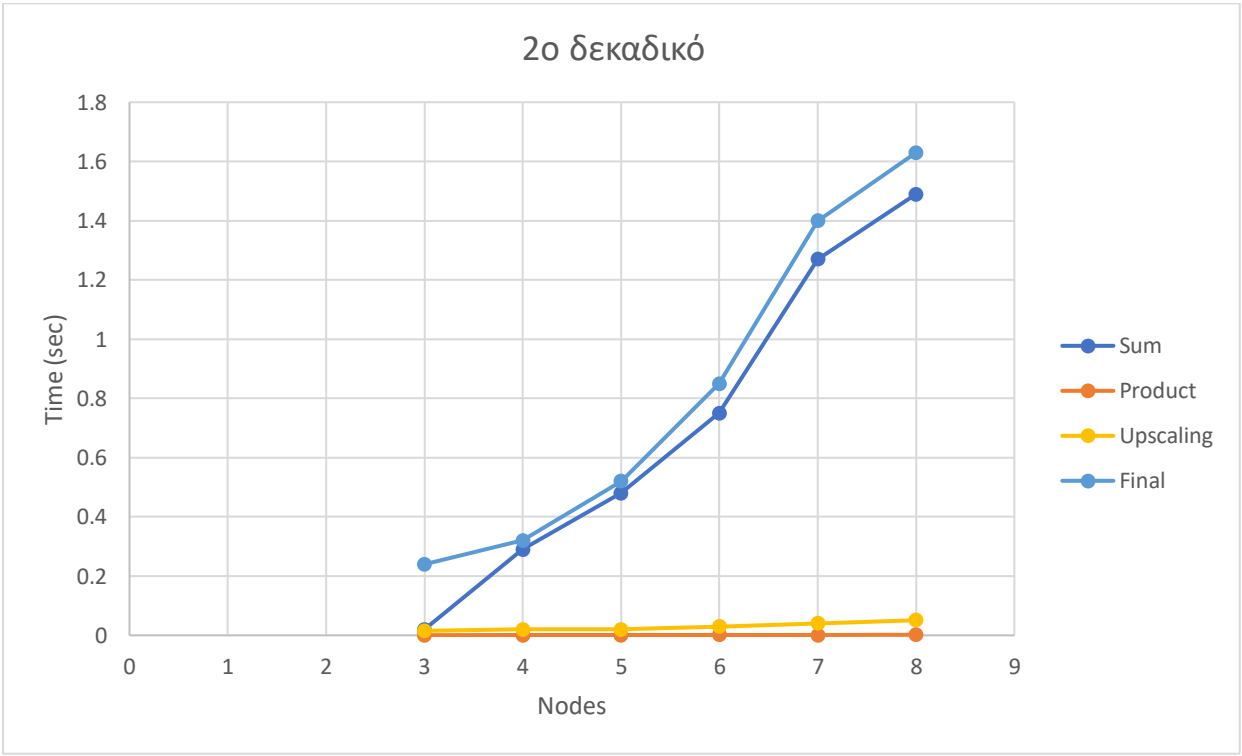
Στη παρούσα ενότητα παρουσιάζονται κάποιες γραφικές παραστάσεις που δημιουργήθηκαν χρησιμοποιώντας τους χρόνους εκτέλεσης του προγράμματος. Στη διαμόρφωση του τελικού χρόνου παίζουν σημαντικό ρόλο δύο παράγοντες. Αυτοί είναι το πλήθος των κόμβων που συμμετέχουν στο σύστημα και το πλήθος των δεκαδικών ψηφίων που έχουν τα δεδομένα. Ο υπολογισμός που πρέπει να επιτευχθεί χωρίζεται σε στάδια. Κάθε στάδιο πρέπει να ολοκληρώνεται από όλους για να προχωρήσουν στο επόμενο. Τα δεκαδικά ψηφία καθώς αυξάνονται αναγκάζουν το πρόγραμμα να εκτελεί χρονοβόρες πράξεις με αρκετά μεγάλους αριθμούς.



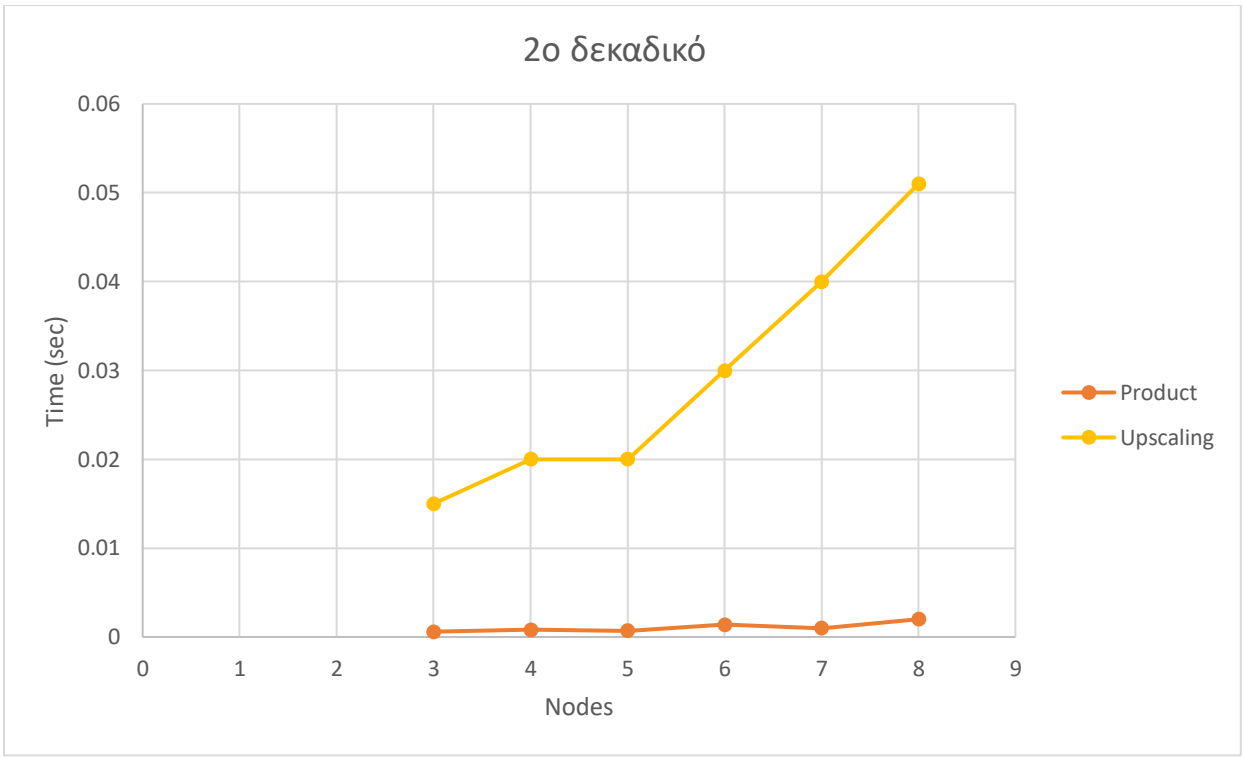
Γράφημα 1



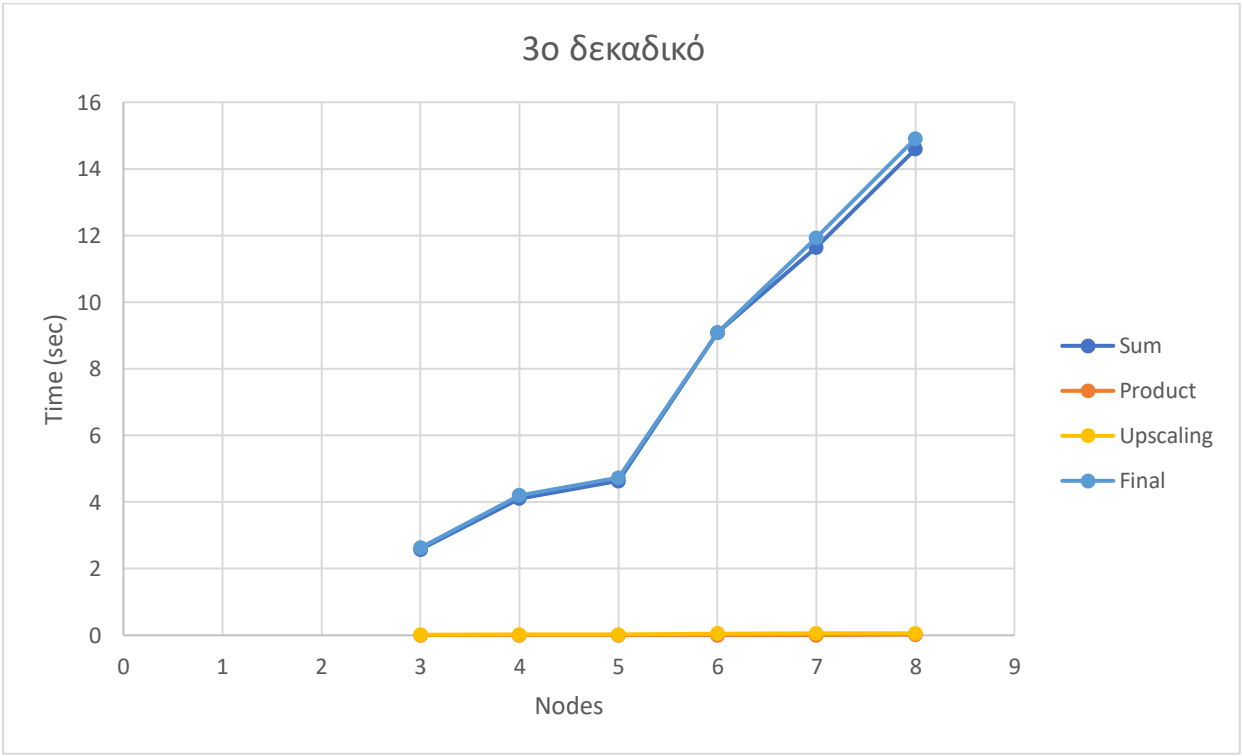
Γράφημα 2



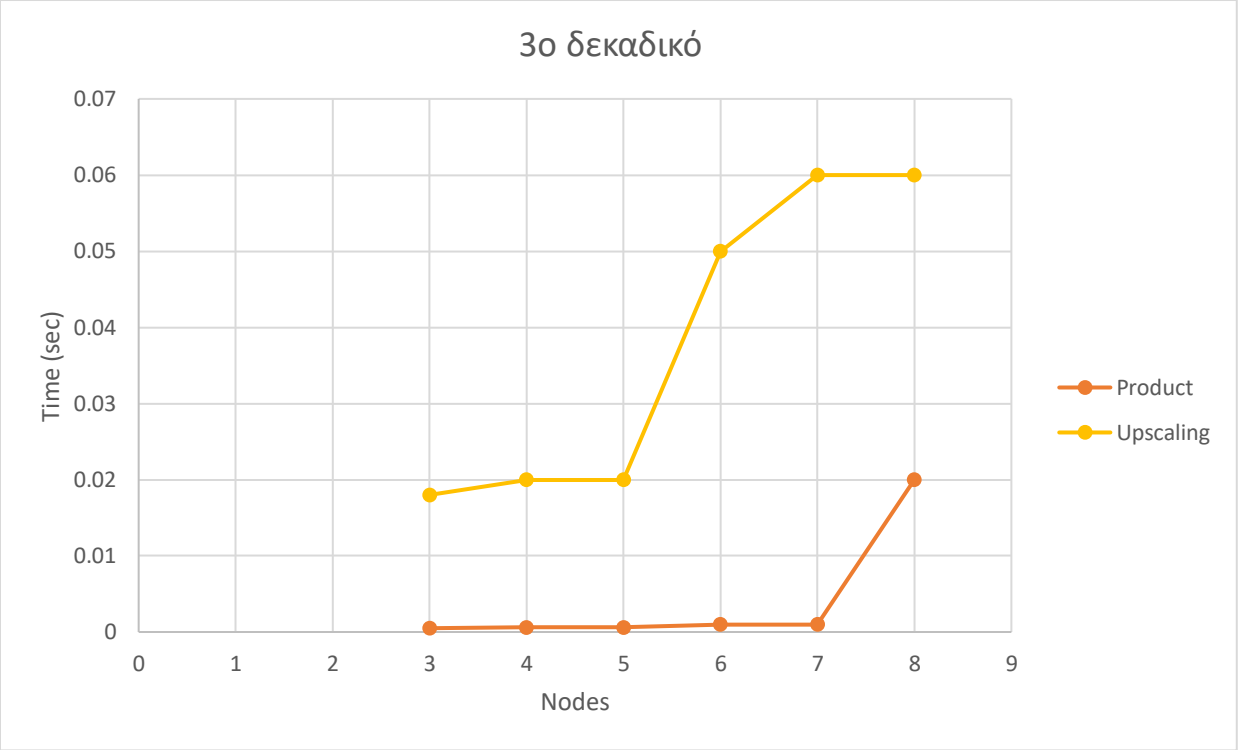
Γράφημα 3



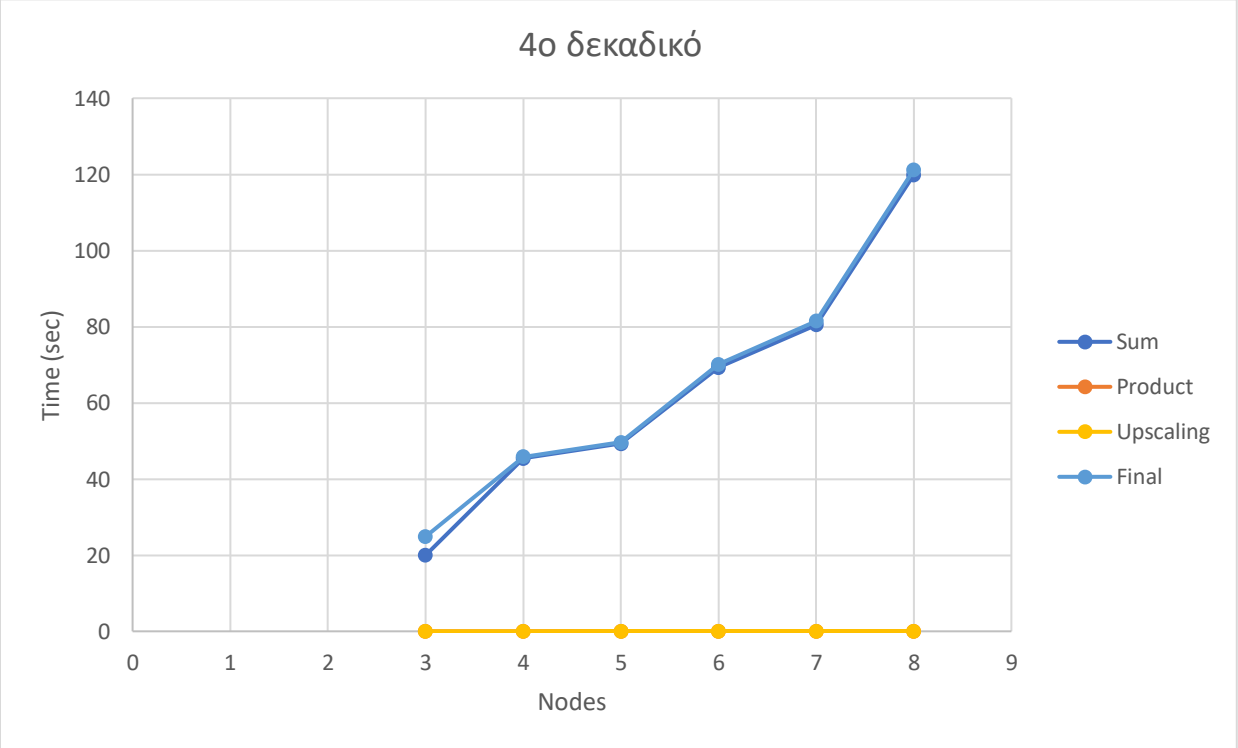
Γράφημα 4



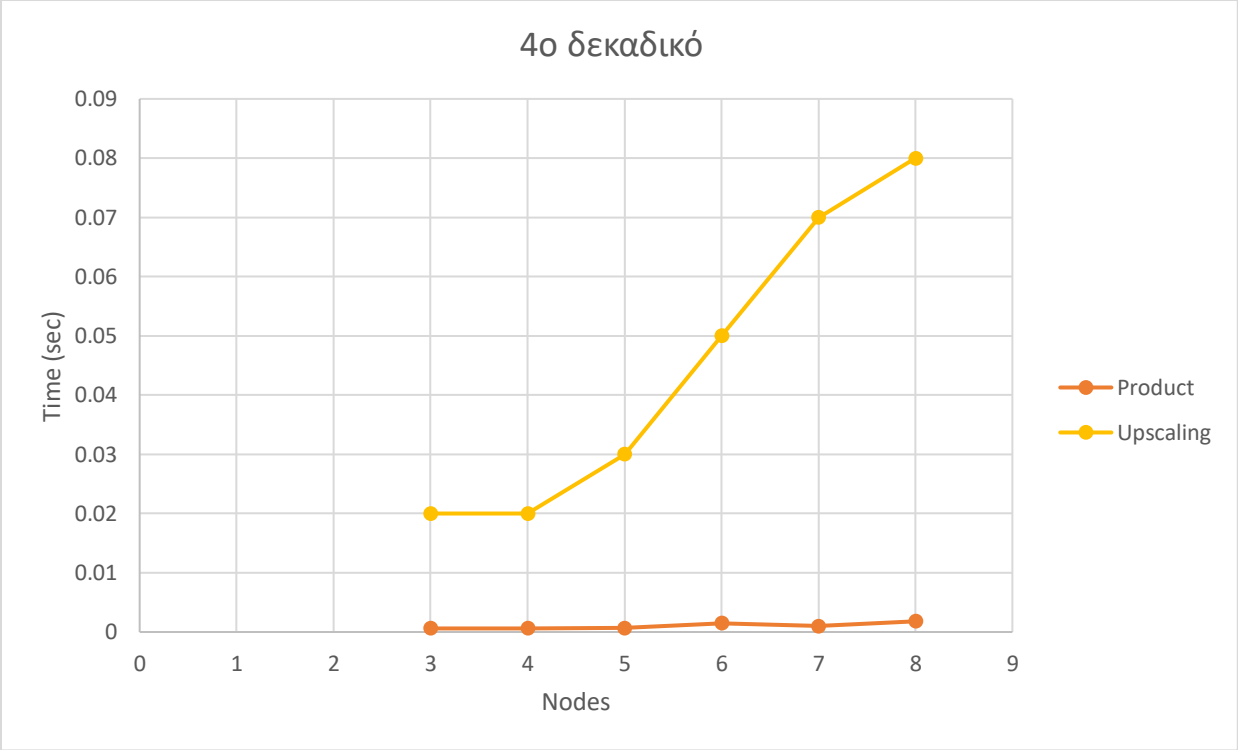
Γράφημα 5



Γράφημα 6



Γράφημα 7



Γράφημα 8

Συμπεράσματα

Τα συμπεράσματα που προέκυψαν από τις μετρήσεις ήταν τα αναμενόμενα. Η απόδοση του αλγορίθμου εξαρτάται από τον αριθμό των κόμβων που θα ανταλλάξουν δεδομένα και από την ακρίβεια των αριθμών που θα ορίσουν οι συμμετέχοντες στο σύστημα. Από τα διαγράμματα που παρουσιάστηκαν στην αντίστοιχη ενότητα προκύπτει ότι το πρωτόκολλο του αθροίσματος είναι ο κύριος παράγοντας που συμβάλλει στην αύξηση του χρόνου εκτέλεσης του προγράμματος.

Επιπροσθέτως, το πρωτόκολλο του γινομένου δεν επηρεάζεται από την αύξηση της ακρίβειας και το ίδιο παρατηρείται και για τη διαδικασία του upscaling. Οι διαδικασίες αυτές επηρεάζονται περισσότερο από τον αριθμό των κόμβων διότι περιλαμβάνουν πολλές ανταλλαγές δεδομένων. Αντίθετα το πρωτόκολλο του αθροίσματος χωλαίνει στο κομμάτι της ακρίβειας επειδή περιλαμβάνει πιο σύνθετες πράξεις όπως ύψωση σε μεγάλες δυνάμεις αριθμών.

Σε περίπτωση που οι κόμβοι θα βρίσκονται σε μεγάλες γεωγραφικά αποστάσεις είναι λογικό να υπάρχουν και μεγαλύτερες καθυστερήσεις. Επίσης κάποιες τιμές ασφαλείας έχουν μειωθεί σχετικά για να γίνουν πιο εύκολα τα διαγράμματα, ικανοποιώντας και πάλι όμως τις απαιτήσεις των πρωτοκόλλων. Η εφαρμογή που αναπτύχθηκε όμως δεν είναι κρίσιμο να τρέχει σε πραγματικό χρόνο, οπότε τυχόν επιβραδύνσεις λόγω τέτοιων παραμέτρων δεν επηρεάζουν σημαντικά το σκοπό μας.

Βιβλιογραφία

1. Βικιπαίδειας, Σ.τ. Νουκλεϊκά οξέα. 2017 [cited 2018 16/08]; Available from: https://el.wikipedia.org/wiki/DNA

2. Αλεπόρου, Β., et al., Βιολογία Θετικής κατεύθυνσης Γ' τάξης Γενικού Λυκείου. 2012, Αθήνα: Οργανισμός Εκδόσεων Διδακτικών Βιβλίων (ΟΕΔΒ).

3. Russell, P.J., iGenetics A Molecular Approach. 2010, San Francisco: Benjamin Cummings.

4. Gunderson, K.L., et al., A genome-wide scalable SNP genotyping assay using microarray technology, in Nature Genetics. 2005. p. 549–554.

5. Gunderson, K.L., et al., A genome-wide scalable SNP genotyping assay using microarray technology. Nature genetics, 2005. 37(5): p. 549.

6. Krawitz, P., et al., Microindel detection in short-read sequence data, in Bioinformatics. 2010. p. 722–729.

7. Manolio, T.A., Genomewide association studies and assessment of the risk of disease, in N Engl J Med. 2010. p. 166-176.

8. Pearson, T.A. and T.A. Manolio, How to Interpret a Genome-wide Association Study, in Jama. 2008. p. 1335-1344.

9. Klein, R.J., et al., Complement factor H polymorphism in age-related macular degeneration, in science. 2005. p. 385-389.

10. Paschou, P., et al., Maritime route of colonization of Europe, in Maritime route of colonization of Europe. 2014. p. 9211-9216. 60

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

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

13. EPEYNA, E.I., Συστηματική ανασκόπηση και μετα-ανάλυση.

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

15. 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.

16. 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.

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

18. Daimonians, R. and N. Laird, Meta-analysis in clinical trials. Controlled clinical trials, 1986. 7(3): p. 177-188. 33. Odlyzko, A.M., Advances in Cryptology-CRYPTO'86: Proceedings. Vol. 263. 2003: Springer.\

19. Wikipedia, S.t Secure Multi-Party Computation, Available from: https://en.wikipedia.org/wiki/Secure_multi-party_computation

20. Wikipedia, S.t Homomorphic encryption, Available from: https://en.wikipedia.org/wiki/Homomorphic_encryption

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

22. Taeho Jung , Xiang-Yang Li, et. al., Collusion-Tolerable Privacy-Preserving Sum and Product Calculation without Secure Channel, From Meng Wan Department of Computer Science, Illinois Institute of Technology, Chicago, IL Department of Computer Science and Technology and TNLIST, Tsinghua University, Beijing Center for Science and Technology Development, Ministry of Education, Beijing

23. Jun Jie Sim, Fook Mun Chan, Shibin Chen, Benjamin Hong Meng Tan, Khin Mi Mi Aung, Achieving GWAS with homomorphic encryption, BMC Medical Genomics, 21 July 2020

24. Private Genomes and Public SNPs: Homomorphic Encryption of Genotypes and Phenotypes for Shared Quantitative Genetics View ORCID Profile Richard Mott, View ORCID Profile Christian Fischer, View ORCID Profile Pjotr Prins and View ORCID Profile Robert William Davies Genetics June 1, 2020 vol. 215 no. 2 359-372; https://doi.org/10.1534/genetics.120.303153

25. Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption Wen-Jie Lu1*, Yoshiji Yamada3, Jun Sakuma 1,2. From 4th iDASH Privacy Workshop San Diego, CA, USA. 16 March 2015

26. Optimized homomorphic encryption solution for secure genome-wide association studies Marcelo Blatt, Alexander Gusev, Yuriy Polyakov , Kurt Rohloff1 and Vinod Vaikuntanathan, From 7th iDASH Privacy and Security Workshop 2018 San Diego, CA, USA. 15 October 2018

27. Ultra-Fast Homomorphic Encryption Models enable Secure Outsourcing of Genotype Imputation, Miran Kim, Arif Harmanci, Jean-Philippe Bossuat, Sergiu Carpov, Jung Hee Cheon, Ilaria Chillotti, Wonhee Cho et. al., From Cell Systems, November 17, 2021

28. iDASH secure genome analysis competition 2018: blockchain genomic data access logging, homomorphic encryption on GWAS, and DNA segment searching, Tsung-Ting Kuo1, Xiaoqian Jiang et.al. From 7th iDASH Privacy and Security Workshop 2018 San Diego, CA, USA. 15 October 2018, BMC Genomics