Ατομική Διπλωματική Εργασία

ΕΡΕΥΝΑ ΤΗΣ ΧΡΗΣΗΣ ΤΩΝ ΝΕΥΡΩΝΙΚΩΝ ΔΙΚΤΥΩΝ ΣΕ ΜΗ – ΕΠΙΒΛΕΠΟΜΕΝΟΥΣ ΤΥΠΟΥΣ ΓΕΝΕΤΙΚΩΝ ΣΥΣΧΕΤΙΣΕΩΝ

Λοΐζος Λοΐζου

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μάρτιος 2010

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΥΠΡΟΥ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Έρευνα της χρήσης των Νευρωνικών Δικτύων σε μη – επιβλεπόμενους τύπους γενετικών συσχετίσεων

Λοΐζος Λοΐζου

Επιβλέπων Καθηγητής
Κωνσταντίνος Παττίχης

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

Μάρτιος 2010

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κύριο Άθω Αντωνιάδη για την πολύτιμη υποστήριξη και καθοδήγηση που προσέφερε κατά την διάρκεια εκπόνησης της συγκεκριμένης διπλωματικής εργασίας. Ήταν ο άνθρωπος που με βοήθησε στην κατανόηση των διάφορων βιολογικών εννοιών που σχετίζονται με το θέμα και όχι μόνο, μου προσέφερε πάρα πολλές γνώσεις τόσο στην βιολογία αλλά κυρίως στον κλάδο της πληροφορικής μέσω της ιδιότητας του ως επιστήμονας πληροφορικής, με έφερε σε επικοινωνία με την Glaxo Smith Kline (GSK), και γενικότερα την άψογη συνεργασία που είχαμε τα τελευταία 3 χρόνια ως μέλη του ερευνητικού εργαστηρίου Medical Informatics Research Team στο Τμήμα Πληροφορικής του Πανεπιστημίου Κύπρου.

Τις ευχαριστίες μου, θα ήθελα να εκφράσω και στον συμφοιτητή και φίλο Άριστο Αριστοδήμου για την άψογη συνεργασία που είχαμε κατά την διάρκεια της διπλωματικής αυτής. Τον ευχαριστώ για την πολύτιμη συνεισφορά του, καθώς δεν ήταν δυνατή η ολοκλήρωση της μελέτης αυτής χωρίς την συμμετοχή του.

Επίσης θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κύριο Κωνσταντίνο Παττίχη που μου έδωσε την ευκαιρία να γίνω μέλος στο προαναφερθέν ερευνητικό εργαστήριο στο οποίο διευθύνει, καθώς και για τη συνεχή επίβλεψη και υποστήριξη του. Υπήρξε βασικός αρωγός στην προσπάθεια ολοκλήρωσης της μελέτης αυτής, συνεπώς η συνεισφορά του ήταν πολύ σημαντική.

Επιπλέον θα ήθελα να δώσω τις θερμές μου ευχαριστίες στους επιστήμονες από την GSK για την παροχή των δεδομένων που χρειάστηκαν για την εκπόνηση της συγκεκριμένης μελέτης, μέσω του κυρίου Άθου Αντωνιάδη.

Περίληψη

Στόχος της μελέτης αυτής ήταν να γίνει προσπάθεια κατηγοριοποίησης γενετικών δεδομένων στην περιοχή HLA, με ασθενείς που παρουσιάζουν στον φαινότυπο τους την ασθένειας της κατά πλάκα σκλήρυνσης. Αποφασίστηκε η χρήση των χαρτών αυτό-οργάνωσης με αλγόριθμο μάθησης όπως προτάθηκε από τον T. Kohonen. [15].

Αρχικά, παρουσιάζεται ένας αλγόριθμος συμπίεσης γενετικών δεδομένων. Αλγόριθμος, ο οποίος σε αντίθεση με άλλους παρόμοιους αλγόριθμους, διατηρεί όλες τις πληροφορίες που είναι σημαντικές σε όλες τις σχεδόν τις αναλύσεις. Είναι αρκετά ευέλικτος αλγόριθμος, είναι εφαρμογή ανοικτού κώδικα και έχει ήδη δημοσιευτεί. [17]

Επιπρόσθετα, προτείνεται στα πλαίσια της μελέτης αυτής, τρόπος προσαρμογής categorical δεδομένων σε αλγόριθμους που μπορούν να επεξεργαστούν μόνο αριθμητικά δεδομένα. Αναπαράσταση των categorical δεδομένων σε δυαδική μορφή ειδικά για τα γενετικά δεδομένα τα οποία επεξεργάζεται η μελέτη αυτή. Προσέγγιση, η οποία αποδεικνύει ότι δεν υπάρχει καθόλου απώλεια πληροφοριών. Προτείνεται επιπλέον διαχείριση των missing δεδομένων κατά την μετατροπή αυτή, καθώς σε πάρα πολλές εφαρμογές τα αποτελέσματα που παράγονται, επηρεάζονται σε μεγάλο βαθμό από τα δεδομένα αυτά.

Επιπλέον, προτείνεται και ένας παράλληλος αλγόριθμος μάθησης χαρτών αυτό-οργάνωσης σε αρχιτεκτονικές κοινής μνήμης, ο οποίος μπορεί εύκολα να μετατραπεί και σε αρχιτεκτονικές ανταλλαγής μηνυμάτων, με πολλές δυνατότητες αναβαθμίσεων. Μπορεί εύκολα να χρησιμοποιηθεί ως βάση για περαιτέρω μελέτη τρόπων παραλληλισμού των χαρτών αυτών. Έχουν γίνει και αρκετές βελτιστοποιήσεις, οι οποίες επίσης προτείνονται στα πλαίσια της μελέτης αυτής

Σαν τελευταίο στάδιο, εκτός από την ανάπτυξη διαφόρων μικρών αλγορίθμων για την αναπαράσταση δεδομένων σε δισδιάστατο χώρο και περαιτέρω επεξεργασία των αποτελεσμάτων, παρουσιάζεται ένας αλγόριθμος ο οποίος αποδεικνύει την συνέπεια του παράλληλου αλγορίθμου, μέσω των διάφορων εκτελέσεων που έγιναν.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγή

Τα τελευταία χρόνια, η ανάπτυξη της βιολογίας και πιο συγκεκριμένα του κλάδου της γενετικής ήταν και είναι τεράστια ειδικά μετά την ανακάλυψη του δισόξυριβοζονουκλεϊκού οξέος το 1953. Το έναυσμα που οδήγησε στην πρόσφατη πρόοδο που παρατηρήθηκε στις βιολογικές επιστήμες είναι η ολοκλήρωση της αλληλουχίας του ανθρώπινου γονιδιώματος που οδήγησε στην αναγνώριση περίπου 35 χιλιάδων γονιδίων. [1] Σχεδόν καθημερινά, νέες πληροφορίες συλλέγονται καταλήγοντας στην δημιουργία ενός τεράστιου όγκου που χρήζουν επεξεργασίας. Η συλλογή στηρίζεται σε τεχνολογίες πληροφορικής και ηλεκτρολογίας. Όσο αφορά την επεξεργασία των πληροφοριών αυτών, γίνεται δυνατή μόνο μέσω τεχνολογιών πληροφορικής. Εδώ είναι που καλούνται οι επιστήμονες της πληροφορικής να συνεισφέρουν με τις γνώσεις τους σε συνδυασμό με τις τεράστιες υπολογιστικές δυνατότητες που προσφέρουν οι σύγχρονοι υπολογιστές. [3]

Ήδη έχουν αναπτυχθεί πάρα πολλές εφαρμογές που προσφέρουν διαφόρων ειδών αναλύσεων σε γενετικά δεδομένα όπως για παράδειγμα η εφαρμογή PLINK. [8] Όμως συνήθως, οι εφαρμογές που υπάρχουν στηρίζονται για τις αναλύσεις τους σε καθαρά στατιστικά μοντέλα. Μέσω όμως της μελέτης αυτής, έγινε προσπάθεια χρησιμοποίησης

μεθόδων υπολογιστικής νοημοσύνης και πιο συγκεκριμένα των τεχνητών νευρωνικών δικτύων που δουλεύουν σε ένα διαφορετικό επίπεδο από την στατιστική, για την ανάλυση γενετικών δεδομένων με στόχο να εξακριβωθεί κατά πόσο ευρετικές μέθοδοι όπως τα νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν για παρόμοιους σκοπούς, κάτι που αναμένεται να μειώσει σε μεγάλο βαθμό τις ανάγκες επεξεργασίας μέσω πολύπλοκων στατιστικών μοντέλων αλλά και να δώσει την δυνατότητα στους επιστήμονες να γνωρίζουν εκ των προτέρων πληροφορίες σχετικά με τα δεδομένα που συλλέγουν, χωρίς καν να απαιτείται η επεξεργασία αυτών.

Ήδη έχει γίνει αρκετή έρευνα σε συναφή θέματα για την επίτευξη του πιο πάνω στόχου. Νευρωνικά δίκτυα χρησιμοποιούνται, για παράδειγμα, για την αναγνώριση σημειακών νουκλεοτιδικών πολυμορφισμών (SNP). [2] Η προσέγγιση που έχει ακολουθηθεί στην μελέτη αυτή όμως, δεν έχει ξαναπροταθεί αφού στοχεύει στην κατηγοριοποίηση των γενετικών σημείων με βάση κοινά πρότυπα (patterns) που υπάρχουν στα διάφορα άτομα. Με την κατηγοριοποίηση που θα γίνει και πιο συγκεκριμένα τα αποτελέσματα της κατηγοριοποίησης, θα δοθούν στους επιστήμονες με απώτερο σκοπό με την συνεισφορά μελλοντικής έρευνας, να μπορούν να αναγνωρίζουν πιθανά προβλήματα – ασθένειες με βάση χαρακτηριστικά πάνω στο γονιδίωμα οποιοδήποτε ατόμου.

Η βάση δεδομένων που χρησιμοποιήθηκε στα πλαίσια της μελέτης αυτής, πρόηλθε από προηγούμενη έρευνα τα αποτελέσματα της οποίας δόθηκαν και αφορούν γενετικά σημεία στην περιοχή HLA. Το πιο σημαντικό γεγονός στο οποίο οφείλεται και η επιλογή της συγκεκριμένης βάσης δεδομένων ως δεδομένα εισόδου, είναι ότι, σύμφωνα με τα αποτελέσματα της έρευνας, προκύπτουν πολλαπλές αλληλεπιδράσεις συσχετιζόμενες με την κατά πλάκα σκλήρυνση, μεταξύ των γενετικών σημείων στην περιοχή αυτή του ανθρώπινου γονιδιώματος.

Με βάση τα πιο πάνω, θα γίνει προσπάθεια κατηγοριοποίησης του DNA συγκεκριμένων ατόμων (subjects), έτσι ώστε να ανακαλυφθεί αν άτομα των οποίων η συσχέτιση συγκεκριμένων γονιδίων μας δίνουν σημαντική στατιστική πληροφορία ότι μπορεί να ευθύνονται για την εμφάνιση ασθένειας στο γονότυπο τους με την χρήση ενός στατιστικού measure (επίσταση), κατηγοριοποιούνται μαζί έχοντας κοινά χαρακτηριστικά – γενετικά σημεία. Με αυτό τον τρόπο θα μπορέσουν να εξαχθούν

συμπεράσματα για τα κοινά γενετικά σημεία που παρατηρούνται με πιθανή συσχέτιση τους με διάφορες ασθένειες που συσχετίζονται με την συγκεκριμένη γονιδιακή περιοχή.

Όπως γίνεται εύκολα αντιληπτό, η προσέγγιση μπορούν να εφαρμοστεί γενικότερα σε οποιαδήποτε ασθένεια με πολύ καλά αποτελέσματα, αν και σίγουρα απαιτείται μεγάλη έρευνα. Συνεπώς, η μελέτη αυτή μπορεί να χρησιμοποιηθεί ως βάση για μελλοντική έρευνα, καθώς τα αποτελέσματα είναι ιδιαίτερα ενθαρρυντικά.

1.2 Παρουσίαση Προβλήματος

Όπως έχει ήδη προαναφερθεί, στόχος της μελέτης αυτής ήταν, με την χρήση μεθόδων υπολογιστικής νοημοσύνης και πιο συγκεκριμένα την χρήση νευρωνικών δικτύων, να γίνει προσπάθεια κατηγοριοποίησης διαφόρων ατόμων (subjects) με βάση την γενετική τους πληροφορία – γονότυπο χωρίς να δίνεται οποιαδήποτε άλλη πληροφορία στο νευρωνικό δίκτυο όσο αφορά την κατάσταση του κάθε ατόμου.

Η λογική πίσω από την ιδέα αυτή στηρίζεται στην υπόθεση ότι λόγω της ύπαρξης γενετικών συσχετίσεων, ένας άνθρωπος για να είναι ασθενής πρέπει να υπάρχουν κάποιες συγκεκριμένες συσχετίσεις μεταξύ των γενετικών του σημείων και πιο συγκεκριμένα σε σημειακούς γενετικούς πολυμορφισμούς. Η ανάλυση και ανακάλυψη αυτών των συσχετίσεων λόγω του μεγάλου αριθμού γενετικών σημείων μετατρέπεται σε ένα πολύ δύσκολο πρόβλημα που απαιτεί τεράστια επεξεργασία.

Για την περιοχή, από την οποία προέρχονται τα δεδομένα είναι ήδη γνωστό μέσω στατιστικής ανάλυσης ότι ευθύνεται για την παρουσία πολλών ασθενειών. Όμως η περιοχή είναι ιδιαίτερα περίπλοκη. Παρατηρούνται πολλές και συνάμα πολύπλοκες συσχετίσεις μέσα στα γενετικά σημεία – δεδομένα. Έτσι με την κατηγοριοποίηση με βάση τα κοινά χαρακτηριστικά, στόχος θα είναι να διαχωριστούν διάφορα άτομα ανάλογα με την ομοιότητα τους, να βρεθούν τα όμοια αυτά γενετικά στοιχεία και να μελετηθούν μόνο αυτά για τις συσχετίσεις, με στόχο την ανακάλυψη των συσχετίσεων που οφείλονται για διάφορες ασθένειες. Όπως έχει προαναφερθεί, αρχικά θα δοκιμαστεί η ιδέα αυτή με δεδομένα ατόμων που πάσχουν από την κατά πλάκα σκλήρυνση.

Για την επιλογή του νευρωνικού δικτύου, μετά από βιβλιογραφική επισκόπηση και εκτεταμένη έρευνα καταλήξαμε στην επιλογή ενός πλέγματος με εκμάθηση με βάση τον αλγόριθμο του Kohonen, δηλαδή σε χάρτες αυτό-οργάνωσης όπως τους ονόμασε ο T. Kohonen, ο οποίος τους έχει προτείνει. [14] Χρησιμοποιήθηκε μη – επιβλεπομένη μάθηση καθώς στόχος ήταν η ανακάλυψη κοινών προτύπων πάνω στον γονότυπο διαφόρων ατόμων, ανεξάρτητα αν φέρουν την ασθένεια της κατά πλάκα σκλήρυνσης, και ανάλογα με τις ομάδες (clusters) που προκύπτουν από τα αποτελέσματα, να εξαχθούν διάφορα συμπεράσματα όσο αφορά τις συσχετίσεις μεταξύ γενετικών σημείων και να διαπιστωθεί κατά πόσο παρόμοιες κατηγοριοποιήσεις μπορούν να χρησιμοποιηθούν για ανακάλυψη συσχετίσεων που οφείλονται για την παρουσία διαφόρων ασθενειών.

Τα αποτελέσματα είναι ιδιαίτερα θετικά, αν και απαιτείται σίγουρα αρκετή μελλοντική έρευνα για καλύτερη αξιοποίηση της ιδέας αυτής.

Όσο αφορά την δομή που θα ακολουθηθεί, στο κεφάλαιο 2 γίνεται μια σύντομη αναφορά στο θεωρητικό υπόβαθρο, εξηγώντας τις θεωρίες που χρησιμοποιούνται στα πλαίσια της διπλωματικής αυτής. Στο κεφάλαιο 3, αναλύεται η μεθοδολογία που ακολουθήθηκε όπως η περιγραφή των αλγορίθμων και των μεθόδων προεπεξεργασίας και μεταεπεξεργασίας. Στο κεφάλαιο 4 παρουσιάζονται όλα τα αποτελέσματα της μελέτης αυτής μέσω γραφικών παραστάσεων και άλλων μεθόδων. Στο κεφάλαιο 5, γίνεται μια εκτενής συζήτηση γύρω από τα αποτελέσματα ώστε να επεξηγηθούν και να αναλυθούν. Τέλος, στο κεφάλαιο 6 παρουσιάζονται τα γενικά συμπεράσματα και οι μελλοντικοί στόχοι.

1.3 Ανάλυση Τρόπου Εργασίας

Για την ολοκλήρωση της μελέτης αυτής, λόγω του μεγάλου όγκου εργασιών που έπρεπε να γίνουν, αποφασίστηκε ο διαμοιρασμός των εργασιών. Συγκεκριμένα, όταν αποφασίστηκε ο στόχος και το τι θα αναπτυσσόταν σε ολόκληρο το έργο (project), λόγω μεγάλης πολυπλοκότητας και πολλών θεμάτων που έπρεπε να γίνουν, δημιουργήθηκε ένα διάγραμμα με τις υπο-εργασίες (tasks) που περιλαμβάνει το έργο (project). Με αυτό τον τρόπο, διασπάστηκε σε πολλά υπό-μέρη. Στην κάθε υπό-εργασία

ορίστηκε ένας επικεφαλής υπό-εργασίας (task leader), ο οποίος θα ήταν υπεύθυνος για την ολοκλήρωση της συγκεκριμένης υπό-εργασίας, δηλαδή θα εκτελούσε χρέη διαχειριστή υπό-εργασίας (task manager). Ουσιαστικά, θα ήταν υπεύθυνος για το πώς θα γίνει η χρονοδρομολόγηση της υπό-εργασίας, για τυχόν προβλήματα που θα παρουσιαστούν κ.τ.λ. Ο διαχωρισμός που έγινε παρουσιάζεται στον πίνακα 1.1. Στις εργασίες που παρουσιάζονται περισσότερο από ένα άτομα υπεύθυνα αφορά εργασίες οι οποίες έπρεπε να αναπτυχθούν ξεχωριστά ή περισσότερες από μια φορές είτε λόγω διαφορετικών δεδομένων εισόδου είτε επειδή μπορούσαν να ξαναμοιραστούν σε μικρότερες υπό-εργασίες που δεν αναφέρονται στον πίνακα.

Τα άτομα που συνεισέφεραν στην μελέτη αυτή, ήταν τρία και λειτουργούσαν σαν ομάδα αναλαμβάνοντας ο καθένας σημαντικά τμήματα ολόκληρου του έργου. Σαν χρέη συμβούλου (consultant) , γενικού επόπτη (supervisor) και καθοδηγητή εκτελούσε ο κύριος Άθως Αντωνιάδης. Ήταν ο γενικός διαχειριστής ολόκληρου του έργου (project leader) και η συνεισφορά του και η εμπειρία του ήταν καθοριστικοί παράγοντες για την ολοκλήρωση ολόκληρου του έργου. Στον πίνακα 1.1, ο κύριος Άθως Αντωνιάδης παρουσιάζεται σε όλες τις υπό-εργασίες λόγω του ρόλου του. Το δεύτερο άτομο ήταν ο κύριος Άριστος Αριστόδημου, που βοήθησε σημαντικά σαν διαχειριστής πολλών υπό-εργασιών, στην ανάπτυξη λογισμικών, στην ανασκόπηση και μελέτης βιβλιογραφίας κτλ. Όποιες υπό – εργασίες ανέλαβε, τις ολοκλήρωσε με μεγάλη επιτυχία χωρίς ιδιαίτερα προβλήματα, προσφέροντας σημαντική βοήθεια. Το τρίτο μέλος ήταν ο κύριος Λοΐζος Λοΐζου και συγγραφέας της διπλωματικής αυτής. Προσωπικά, οι ευθύνες που ανέλαβα ήταν ίδιες με τον φίλο και συνεργάτη Άριστο Αριστοδήμου. Πιστεύω ότι χωρίς την συνεργασία μας, δεν ήταν δυνατή η ολοκλήρωση όλου του έργου.

Επιπλέον, πρέπει να σημειωθεί είναι ότι η κάθε υπό-εργασία (sub task) αναπτύχθηκε μέσω συνεργασίας. Δηλαδή δεν είχε σημασία το ποιος ήταν ο διαχειριστής του έργου. Για την κάθε εργασία, τα άτομα που συνέβαλαν στην ανάπτυξη όλου του έργου συνεργάζονταν τόσο στην ανάπτυξη κώδικα, όσο και στην ανασκόπηση της βιβλιογραφίας (literature review). Δεν ήταν δυνατό να αναπτυχθεί η κάθε υπό-εργασία από ένα άτομο καθώς σε κάθε εργασία έπρεπε να γίνουν πολλά πράγματα είτε μελέτη προηγούμενης εργασίας στο θέμα, είτε αναγνώριση και ανακάλυψη τρόπων

ολοκλήρωσης της κάθε υπό-εργασίας είτε εκμάθηση λογισμικών που ήταν απαραίτητα για την ολοκλήρωση της.

	Άθως Αντωνιάδης	Άριστος Αριστοδήμου	Λοΐζος Λοΐζου
Προεπεξεργασία – Αλγόριθμος Συμπίεσης – Μετατροπή .ped αρχείων σε δυαδική αποθήκευση .bin	X		X
Προεπεξεργασία – Αλγόριθμος Συμπίεσης – Μετατροπή .bin αρχείων σε .ped αρχείων	X	X	
Προεπεξεργασία – Φιλτράρισμα Δεδομένων	X	X	X
Χάρτες Αυτό-οργάνωσης – Σειριακός Αλγόριθμος με Categorical Data	X		X
Χάρτες Αυτό-οργάνωσης – Σειριακός Αλγόριθμος με Linear Data	X	X	
Χάρτες Αυτό-οργάνωσης – Παραλληλοποίηση και Βελτιστοποιήσεις	X	X	
Χάρτες Αυτό-οργάνωσης – Προσαρμογή σε γενετικά δεδομένα – Missing Data	X		X
Σφάλμα Κβαντισμού (Quantization Error)	X	X	
Συνέπεια Αλγορίθμου (Consistency Checking)	X	X	X
Γραφικές Παραστάσεις Δισδιάστατου Χώρου	X	X	X
Γραφικές Παραστάσεις Βαρών	X		X

Πίνακας 1.1 Διάσπαση Έργου σε μικρότερες υπό – εργασίες και αναθέσεις τους

Κεφάλαιο 2

Θεωρητικό Υπόβαθρο

2.1 Βασικές Έννοιες Μοριακής Βιολογίας

2.1.1 Δισόξυριβοζονουκλεϊκό οξύ

Το δεσόξυριβοζονουκλεϊκό οξύ ή DNA, αποτελεί μια από τις σημαντικότερες ανακαλύψεις της ανθρωπότητας (20ος αιώνας), καθώς έδωσε την δυνατότητα στον

άνθρωπο να «ξεκλειδώσει» την ιστορία, την κληρονομικότητα και γενικότερα το μυστήριο το οποίο ονομάζεται ζωή. Είναι ένα νουκλεϊνικό οξύ που περιέχει τις γενετικές πληροφορίες όλων των κυτταρικών οργανισμών και καθορίζει την βιολογική και εξελικτική ανάπτυξη τους και βρίσκεται στο πυρήνα των κυττάρων.

Όπως είναι ευρέως γνωστό, το ανθρώπινο σώμα και γενικότερα σχεδόν όλων των οργανισμών, αποτελείται από τρισεκατομμύρια μικροσκοπικές μονάδες που είναι ενωμένες μεταξύ τους, τα κύτταρα. Πολλά κύτταρα μαζί αποτελούν ένα όργανο. Υπάρχουν πολλά είδη κυττάρων τα οποία διαχωρίζονται κυρίως ως προς την λειτουργία την οποία εκτελούν σε κάθε οργανισμό. Εντούτοις παρά τον διαχωρισμό τους, όλα τα κύτταρα παρουσιάζουν κάποια κοινά χαρακτηριστικά κυρίως ως προς την δομή τους. Υπάρχει ένα κεντρικό μέρος που ονομάζεται πυρήνας. Στο πυρήνα περιέχεται όλος ο γενετικός κώδικας δηλαδή το DNA. Το κύτταρο πραγματοποιεί τις λειτουργίες του με βάση τον γενετικό κώδικα (σειρά πληροφοριών) που έχει κληρονομήσει από τους προγόνους του.

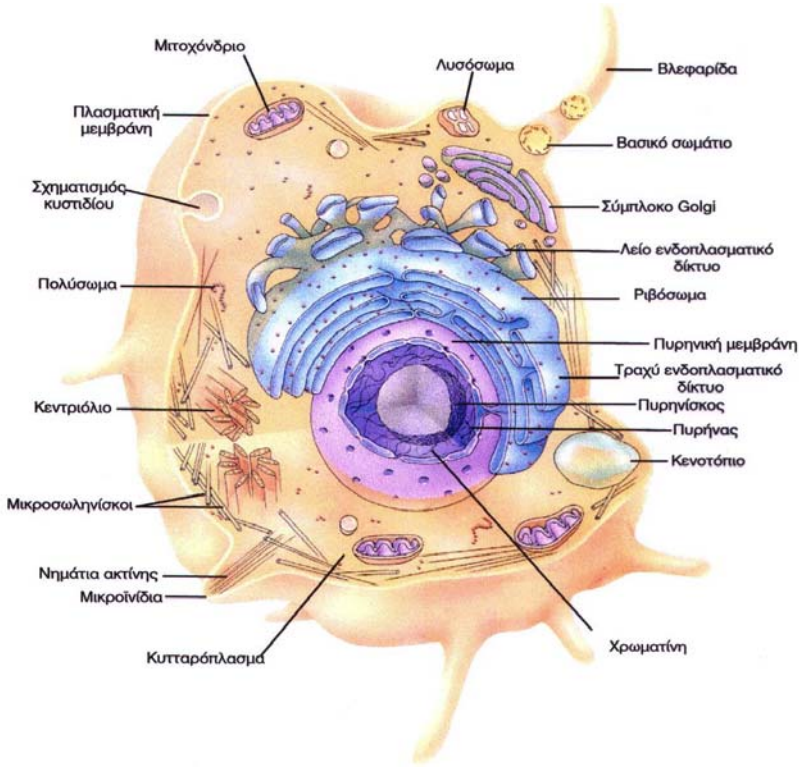
Η πρώτη ανακάλυψη για την ύπαρξη του DNA, έγινε για πρώτη φορά το 1869 από τον Ελβετό ιατρό Friedrich Miescher ο οποίος το ονόμασε νουκλεΐνη. Εντούτοις χρειάστηκαν σχεδόν 100 χρόνια για να ανακαλυφθεί ο ρόλος του ως γενετικός κώδικας. Το 1953 οι James Watson και Francis Crick πρότειναν το μοντέλο της διπλής έλικας για την δομή του DNA. Σύμφωνα με τους δύο επιστήμονες, το DNA αποτελείται από δυο πολυνουκλεϊνικές αλυσίδες που σχηματίζουν στο χώρο μια δεξιόστροφη διπλή έλικα και είναι συνδεδεμένες μεταξύ τους. Η κάθε αλυσίδα αποτελείται από νουκλεοτίδια τα οποία δομούνται από μια πεντόζη, τη δεσοξυριβόζη η οποία είναι ενωμένη με μια φωσφορική ομάδα και μια αζωτούχο βάση. Η μόνη διαφορά που ξεχωρίζει το ένα νουκλεοτίδιο από το ένα άλλο είναι η αζωτούχα βάση με την οποία είναι συνδεδεμένο. Στα νουκλεοτίδια του DNA η αζωτούχος βάση μπορεί να είναι: Α-Αδενίνη(Α), Γουανίνη(G), Θυμίνη(Τ), Κυτοσίνη(C). Συνεπώς μπορούμε να αναπαραστήσουμε τις αλυσίδες του DNA ως μια ακολουθία συμβόλων που αντιστοιχούν στις αζωτούχες βάσεις A-T-C-G. [3,4]

Κάθε αλυσίδα ή κλώνος έχει αρχή και τέλος, με την αρχή να συμβολίζεται ως το 5΄ άκρο και το τέλος ως 3΄ άκρο. Μάλιστα οι δύο αλυσίδες έχουν αντίθετη φορά, καθώς η

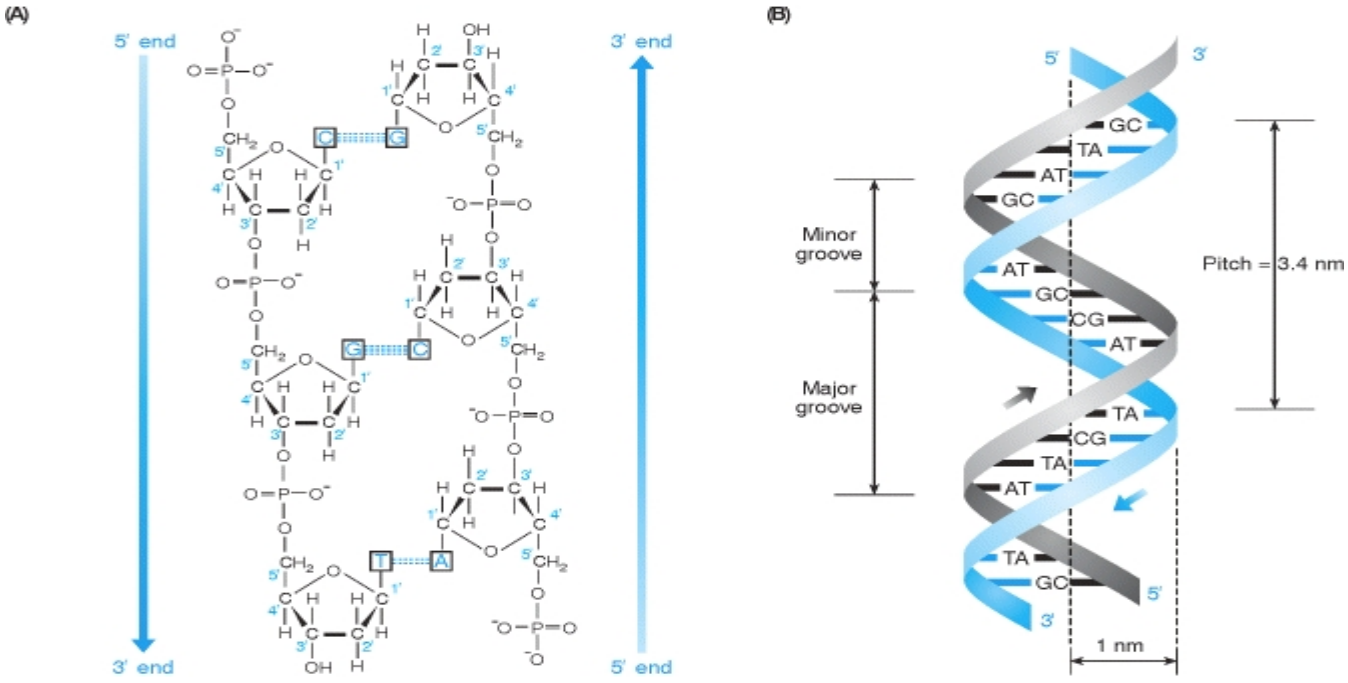
αρχή της μίας είναι απέναντι από το τέλος της άλλης και αποκαλούνται αντιπαράλληλες (και οι δύο έχουν κατεύθυνση από το 5΄ στο 3΄ άκρο με το 5΄ άκρο της μίας αλυσίδας να βρίσκεται απέναντι από το 3΄ άκρο της άλλης και αντίστροφα) .

Όσο αφορά την σύνδεση των δύο αλυσίδων μεταξύ τους, οι αλυσίδες συγκρατούνται μεταξύ τους με δεσμούς υδρογόνου που σχηματίζονται μεταξύ των αζωτούχων βάσεων τους. Οι αζωτούχες βάσεις, ανάμεσα στις οποίες μπορούν να σχηματιστούν δεσμοί υδρογόνου είναι καθορισμένες. Ο κανόνας με τον οποίο συνδέονται ονομάστηκε από τους James Watson και Francis Crick ως ο κανόνας της συμπληρωματικότητας, με την αδενίνη (A) να ενώνεται πάντοτε με τη θυμίνη (T) με δύο δεσμούς υδρογόνου και την κυτοσίνη (C) να ενώνεται πάντοτε με τη γουανίνη (G) με τρείς δεσμούς υδρογόνου (συνεπώς οι αλυσίδες ονομάζονται συμπληρωματικές).

Οι σημαντικότερες διαδικασίες που παρατηρούνται στο DNA είναι η αντιγραφή, η μεταγραφή και η μετάλλαξη. [3,4]



Εικόνα 2.1 Το κύτταρο (http://users.kor.sch.gr/dgspanos/kyttaro/zk.jpg)



Εικόνα 2.2 Η δομή του DNA (A) και η ιδιότητα της συμπληρωματικότητας (B) παρθέν από 'NCBI Human Molecular Genetics 2' (http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=hmg&part=A29&rendertype=figure&id=A59)

2.1.2 Χρωμοσώματα

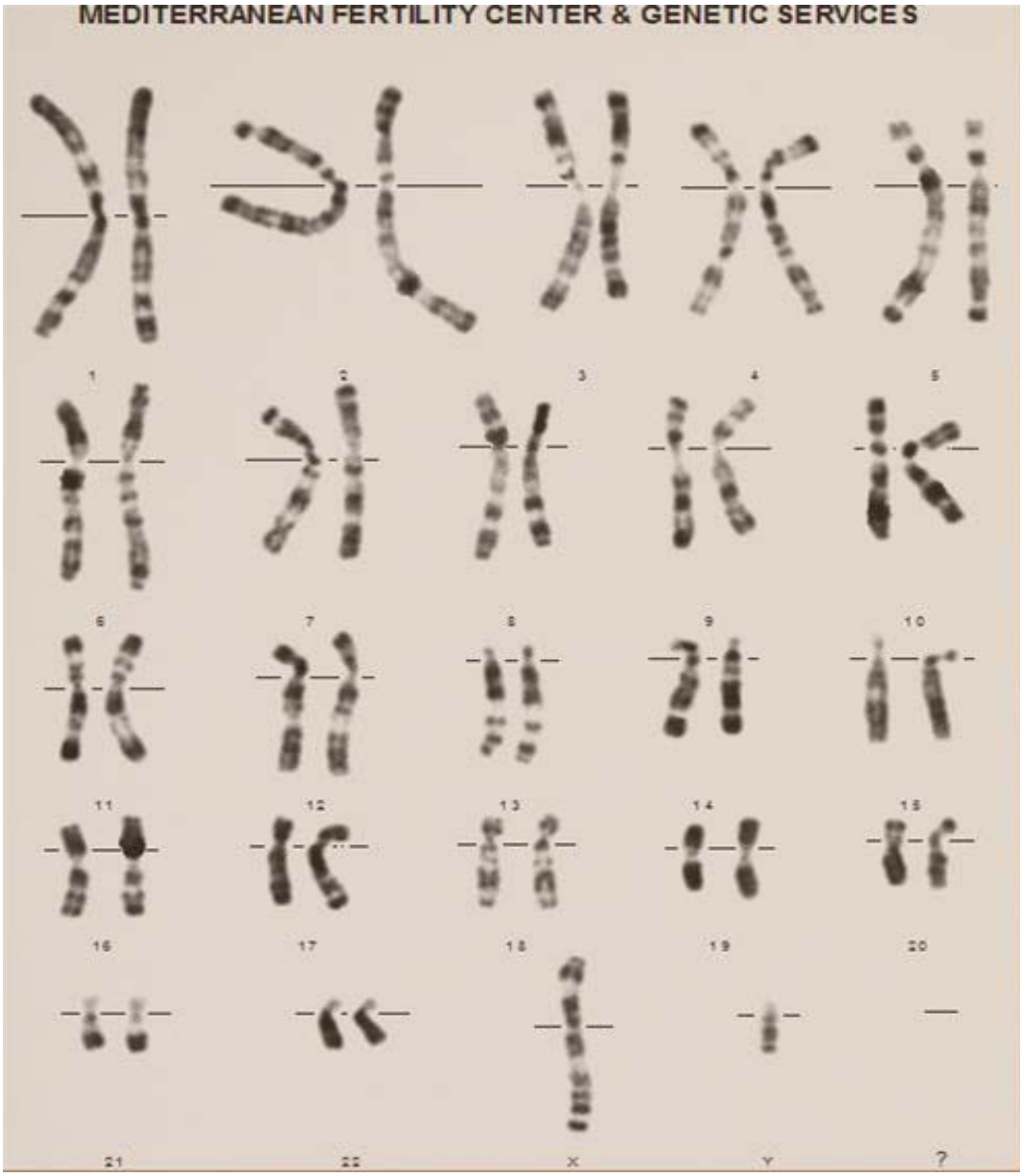
Μέσα στο πυρήνα του κυττάρου, το DNA είναι οργανωμένο σε διάφορες δομές που ονομάζονται χρωμοσώματα. Είναι ένα μοναδικό κομμάτι περιελιγμένου DNA που περιλαμβάνει πολλά γονίδια και άλλες ακολουθίες νουκλεοτιδίων. Είναι, ουσιαστικά, σωματίδια που βρίσκονται στο πυρήνα του ατόμου και μεταφέρουν τις γενετικές πληροφορίες. Κάθε χρωμόσωμα έχει μερικές χιλιάδες γονίδια, το καθένα από τα οποία είναι υπεύθυνο για κάποια λειτουργία. Τα χαρακτηριστικά που κληρονομούνται στους διάφορους οργανισμούς μεταφέρονται από γενιά σε γενιά με τα χρωμοσώματα ή αλλιώς χρωματοσώματα. Γενικότερα ο χρωματοσωματικός χάρτης κάθε οργανισμού λέγεται καρυότυπος.

Για την συλλογή των χρωμοσωμάτων στο εργαστήριο, διασπώνται σε μεγάλα κομμάτια και κλωνοποιούνται τα υπομέρη τους. Στην συνέχεια δημιουργούνται ένας «χάρτης» των κλώνων που δημιουργούνται, και καταγράφονται οι κλώνοι που υπερκαλύπτονται. Η μέθοδος αυτή όμως απαιτεί πολύ χρόνο. Γι' αυτό το λόγο, ακολουθείται μια

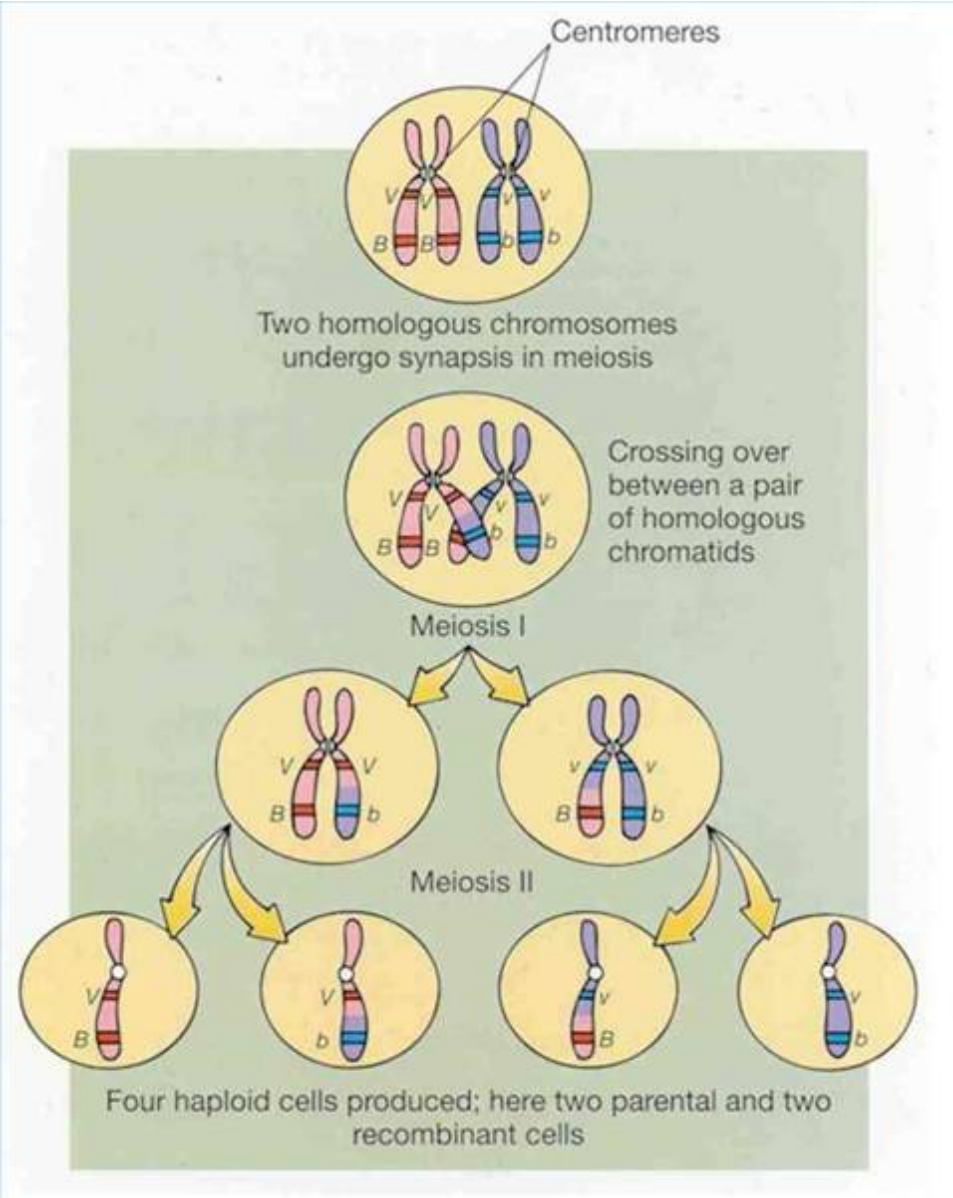
διαφορετική διαδικασία. Κόβονται τα χρωμοσώματα σε κομμάτια των 100000 βάσεων, τα οποία εισάγονται σε τεχνικά βακτηριακά χρωμοσώματα, τα λεγόμενα BACs. Αποτελούν ουσιαστικά κλώνους των κομματιών που δημιουργήθηκαν από τα χρωμοσώματα, έχοντας την ικανότητα να αυτοδιπλασιάζονται ανεξάρτητα. Στη συνέχεια, τα BACs διασπώνται ακόμα περισσότερο με την βοήθεια ενζύμων και οι επικαλυπτόμενες ακολουθίες επιλέγονται και συναρμολογούνται έτσι ώστε να φτιαχτεί η αρχική DNA ακολουθία. Τα αποτελέσματα αποθηκεύονται σε διάφορα αρχεία κειμένου σύμφωνα με διάφορα formats/ πρότυπα τα οποία υπάρχουν.

Ο άνθρωπος έχει 46 χρωμοσώματα σε κάθε σωματικό του κύτταρο. Μισά από αυτά τα έχει κληρονομήσει από τη μητέρα του και τα άλλα μισά από τον πατέρα του. Κατά τη δημιουργία των σπερματοζωαρίων και των ωαρίων ακολουθείται μια διαδικασία όπου το κάθε κύτταρο διασπάται και δημιουργεί δύο κύτταρα των 23 χρωμοσωμάτων, ούτως ώστε όταν ενωθεί το ωάριο (θηλυκός γαμέτης) με το σπερματοζωάριο (αρσενικός γαμέτης) να δημιουργείται και πάλι ένα κύτταρο (ζυγωτό) με 46 χρωμοσώματα. Η διαδικασία αυτή ονομάζεται μείωση (meiosis).

Τα 46 χρωμοσώματα διακρίνονται σε 22 ζεύγη ομόλογων χρωμοσωμάτων (ή αυτοσωμικών) και σε 1 ζεύγος φυλετικών χρωμοσωμάτων. Με τον όρο ομόλογα χρωματοσώματα, ορίζουμε ένα ζευγάρι χρωμοσωμάτων που έχουν το ίδιο σχήμα και μέγεθος, περιέχουν την ίδια σειρά γονιδίων, ελέγχουν την ίδια ιδιότητα με διαφορετικό ενδεχομένως τρόπο και φέρουν και αντίστοιχες γονιδιακές θέσεις. Το ένα είναι πατρικής και το άλλο μητρικής προέλευσης. Με τον όρο φυλετικά χρωματοσώματα, ορίζουμε το ζευγάρι χρωμοσωμάτων που ορίζουν το φύλο σε ένα οργανισμό. Στον άνθρωπο, για παράδειγμα το 23ο χρωμόσωμα καθορίζει το φύλο. Τα φυλετικά χρωμοσώματα στον άνθρωπό είναι δύο τύπων Χ και Υ. Ο άνδρας έχει ένα Χ και ένα Υ χρωμόσωμα ενώ η γυναίκα έχει δύο Χ χρωμοσώματα. Με βάση τα προαναφερθέντα, είναι ξεκάθαρο ότι ο άνδρας καθορίζει το φύλο, ανάλογα με το ποιο σπερματοζωάριο θα γονιμοποιήσει κάποιο ωάριο. Αν το σπερματοζωάριο περιέχει το Υ, τότε το ζυγωτό θα είναι αρσενικού γένους αλλιώς αν περιέχει το Χ θα είναι γυναικείου γένους. [3,4]



Εικόνα 2.3 Ανθρώπινα Χρωμοσώματα παρθέν από ‘Mediterranean Fertility Center & Genetic Services, Chania’ (http://www.fertilitycenter-crete.gr/page.aspx?id=97&lang=el)



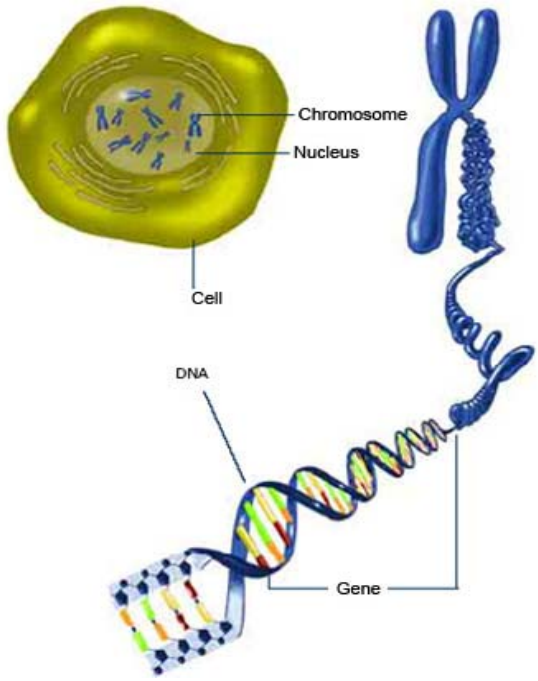
Εικόνα 2.4 Ανθρώπινα Χρωμοσώματα παρθέν από 'Janice Creneti biology tutorials' (http://www.ehow.com/video_4984760_how-many-chromosomes-does-each.html)

2.1.3 Γονίδια

Τα γονίδια, όπως έχει προαναφερθεί, αποτελούν πολλές βάσεις DNA. Αν σκεφτούμε δομικά, τότε έχουμε τα χρωμοσώματα τα οποία αποτελούνται από διάφορα γονίδια, τα οποία αποτελούνται από διάφορες ακολουθίες βάσεων DNA. Θα μπορούσαμε να ορίσουμε ένα γονίδιο ως ένα τμήμα του μορίου του DNA που μπορεί να μεταγραφεί στη μορφή ενός μορίου RNA. Με βάση τον ορισμό, τα γονίδια ελέγχουν την δομή όλων

των πρωτεϊνών ενός οργανισμού.

Τα γονίδια αποτελούν ουσιαστικά το γονιδίωμα των οργανισμών. Είναι η βασική φυσική μονάδα κληρονομικότητας (μεταφέρονται με τα χρωμοσώματα τα οποία αποτελούν μέρος). Όλα τα χαρακτηριστικά μεταφέρονται με τα γονίδια και συνεισφέρουν στο καθορισμό ενός γενετικού χαρακτηρισμού. Συνήθως η πλειοψηφία των γονιδίων κωδικοποιούν πρωτεΐνες, οι οποίες είναι υπεύθυνες για όλες τις λειτουργίες που εκτελεί ένα συγκεκριμένο κύτταρο, καθορίζοντας έτσι το ρόλο του κάθε κυττάρου σε ένα οργανισμό. Υπάρχουν και γονίδια τα οποία δεν κωδικοποιούν πρωτεΐνες όμως διαδραματίζουν βασικούς ρόλους στο τρόπο σύνθεσης των πρωτεϊνών και στον έλεγχο της γονιδιακής έκφρασης. Τα πιο σημαντικά παραδείγματα τμημάτων DNA που δεν μεταφράζονται σε πρωτεΐνες αποτελούν τα ιντρόνια ή εσώνια (εξώνια οι περιοχές που μεταφράζονται). Οι περιοχές αυτές διαγράφονται κατά την διαδικασία της μεταγραφής και χρησιμεύουν για να γνωρίζει ο οργανισμός μέχρι ποιο σημείο θα σταματήσει η κωδικοποίηση ενός εξωνίου. [3,4]



Εικόνα 2.5 Από το DNA στα γονίδια, στα χρωμοσώματα και στο κύτταρο παρθέν από 'National Institute of General Medical Sciences' (http://publications.nigms.nih.gov/thenewgenetics/chapter1.html)

2.1.4 Βασικές Διεργασίες που Παρατηρούνται στο DNA

Οι βασικές διαδικασίες που παρατηρούνται στο DNA είναι η αντιγραφή, η μεταγραφή, και η μετάφραση. Με βάση τις διαδικασίες αυτές, ο F. Crick διατύπωσε το κεντρικό δόγμα της μοριακής βιολογίας όσο αφορά την ροή της γενετικής πληροφορίας. Ουσιαστικά περιγράφει την διαδικασία με την οποία πρωτεΐνες παράγονται από το DNA του πυρήνα. Το DNA, με την διαδικασία της μεταγραφής, μεταγράφει την γενετική πληροφορία του στο RNA,. Το RNA μεταφέρει την πληροφορία στον τόπο παραγωγής των πρωτεϊνών καθορίζοντας τι είδους πρωτεΐνη θα παραχθεί., με την διαδικασία της μετάφρασης. Με αυτό τον τρόπο, αφού οι πρωτεΐνες καθορίζουν την δομή και λειτουργία των κυττάρων και συνεπώς των οργανισμών, το DNA καθορίζει την λειτουργία όλων των ειδών. [3,4]

2.1.4.1 Αντιγραφή

Κατά την διαδικασία αυτή, οι δύο αλυσίδες του DNA ανοίγουν/ ξετυλίγονται, με αποτέλεσμα να μείνουν ελεύθερες οι βάσεις των νουκλεοτιδίων και με την βοήθεια νέων συμπληρωματικών αλυσίδων δημιουργούνται εντελώς όμοια καινούργια μόρια.

Όπως έχει ήδη προαναφερθεί, το αρχικό μόριο αποτελείται από δύο συμπληρωματικές αλυσίδες DNA συνδεδεμένες με δεσμούς υδρογόνου. Αρχικά χωρίζουν οι δύο αλυσίδες και ακολούθως με την βοήθεια κατάλληλων ενζύμων φέρνουν απέναντι από κάθε βάση την συμπληρωματική της, συνθέτοντας μια νέα αλυσίδα. Με αυτό τον τρόπο, κάθε μητρική αλυσίδα λειτουργεί ως καλούπι για την δημιουργία νέας συμπληρωματικής αλυσίδας. Τέλος, τα νουκλεοτίδια συνδέονται μεταξύ τους για να δημιουργηθούν δύο νέες αλυσίδες. Κάθε μία από τις θυγατρικές αλυσίδες αποτελείται από μια μητρική και μία νέα αλυσίδα. Κάθε νέο μόριο, συνεπώς αποτελείται από μία αλυσίδα του προηγούμενου μορίου και μία νέα αλυσίδα. Η διαδικασία λοιπόν της αντιγραφής, έχει ως αποτέλεσμα την δημιουργία δύο νέων μορίων, απολύτως όμοιων και μεταξύ τους και με το παλιό μόριο. [3,4]

2.1.4.2 Μεταγραφή

Η διαδικασία αυτή είναι υπεύθυνη για την σύνθεση του RNA από το DNA. Σ' αυτή την φάση, παράγονται οι τρεις διαφορετικοί τύπου του RNA, το mRNA, το tRNA, το rRNA και το snRNA. Ο πιο σημαντικός τύπος RNA είναι σίγουρα το mRNA το οποίο θα περιέχει την μεταγραφόμενη γενετική πληροφορία του DNA. Κάθε mRNA είναι ένα αντίγραφο ενός γονιδίου. Προκύπτει με διαδικασία πολύ παρόμοιας με την αντιγραφή. Συγκεκριμένα, αρχικά ξετυλίγεται η διπλή έλικα του DNA στην περιοχή του γονιδίου, που πρόκειται να γίνει η μεταγραφή, και ανοίγει με σπάσιμο των δεσμών υδρογόνου που συγκρατούν τις δύο αλυσίδες. Ένα ειδικό ένζυμο (RNA πολυμερές) ακολούθως συνδέεται με τον υποκινητή, μια μικρή ακολουθία νουκλεοτιδίων στο DNA που δρα σαν σημείο εκκίνησης. Ακολούθως, η έναρξη της δημιουργίας της αλυσίδας του RNA σηματοδοτείται από την συνένωση των πρώτων δύο ριβονουκλεοτιδίων. Έπειτα, παρατηρείται η επιμήκυνση της αλυσίδας με την RNA πολυμερές, να τοποθετεί διαδοχικά τα συμπληρωματικά ριβονουκλεοτίδια απέναντι από κάθε νουκλεοτίδιο του DNA, ενώνοντας τα μεταξύ τους με ένα φωσφορικό δεσμό. Καθώς το mRNA επιμηκύνεται, αποσυνδέεται από την αλυσίδα του DNA και το μόριο του DNA ξανατυλίγεται. Τέλος έχουμε την φάση της λήξης όπου η RNA πολυμεράση αντιλαμβάνεται ένα μήνυμα τερματισμού από μια ειδική αλληλουχία βάσεων και έτσι ένα ολοκληρωμένο μόριο mRNA ελευθερώνεται πλήρων από το καλούπι του DNA.

Συνεπώς η αλυσίδα mRNA που δημιουργείται, έχει ως βάση το DNA, χρησιμοποιώντας ως βάση την μία εκ των δύο αλυσίδων προσκολλώντας τις συμπληρωματικές βάσεις του RNA (A → U, C → G, T → A).

Με παρόμοια διαδικασία παράγονται και οι άλλοι τρεις τύποι RNA. [3,4]

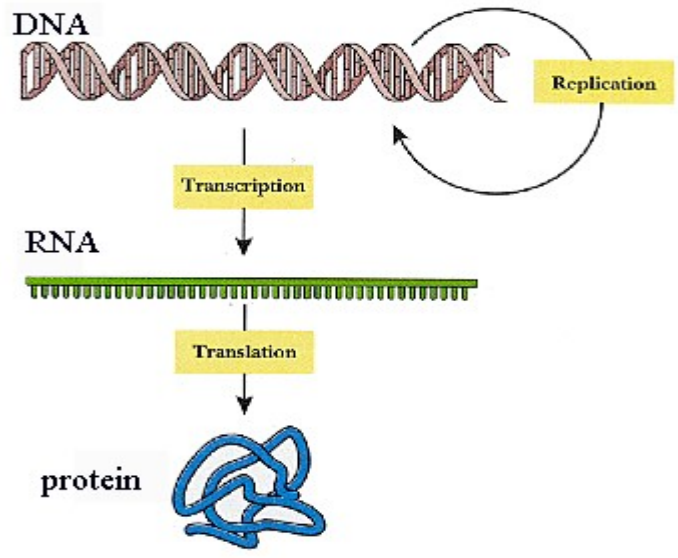
2.1.4.3 Μετάφραση

Στο τελευταίο στάδιο έχουμε ουσιαστικά την παραγωγή των πρωτεϊνών με την βοήθεια των τεσσάρων τύπων του RNA. Η συμμετοχή του DNA έγκειται στην σύνθεση του mRNA το οποίο και κωδικοποιεί την δημιουργία των αμινοξέων και κατά συνέπεια των πρωτεϊνών. Η βασική ιδέα είναι ότι κάθε τρεις βάσεις του mRNA (κωδικόνια),

κωδικοποιούν ένα συγκεκριμένο αμινοξύ. Με αυτό τον τρόπο έχοντας 3 βάσεις και αφού κάθε βάση μπορεί να είναι A,T,C,G τότε μπορούμε να έχουμε μέχρι 43 δυνατά αμινοξέα. Τα 20 επομένως αμινοξέα που υπάρχουν για την κατασκευή των πρωτεϊνών, μπορούν εύκολα να κωδικοποιηθούν με την χρήση ενός κώδικα βασισμένου σε τριπλέτες βάσεων.

Αρχικά το mRNA, αμέσως μετά την σύνθεση του, μεταφέρεται στα ριβοσώματα του κυτταροπλάσματος. Στο ριβόσωμα έρχεται και το tRNA με το αντίστοιχο αμινοξύ, αναγνωρίζοντας ταυτοχρόνως την τριάδα νουκλεοτιδίων του mRNA που κωδικοποιεί το αμινοξύ αυτό. Το ριβόσωμα αλλάζει συνεχώς θέσεις στο mRNA και επομένως και τριπλέτες που αναγνωρίζονται από τα αντίστοιχα tRNA. Έτσι μπαίνουν σε σειρά τα αμινοξέα, σύμφωνα με τις οδηγίες που μεταφέρει το mRNA από το πυρήνα, ενώνονται μεταξύ τους και σχηματίζουν τις πρωτεΐνες.

Είναι σημαντικό να παρατηρηθεί για να κωδικοποιείται μια πρωτεΐνη, πρέπει να υπάρχει ένα σημείο έναρξης και ένα σημείο λήξης στο γενετικό κώδικα. Το σημείο έναρξης αναφέρεται στο κωδικόνιο έναρξης και το σημείο λήξης ως το κωδικόνιο λήξης. Μόλις βρεθεί το κωδικόνιο έναρξης ξεκινά η παραγωγή μια νέας πρωτεΐνης ενώ μόλις βρεθεί το κωδικόνιο λήξης, τα αμινοξέα που έχουν παραχθεί ως εκείνο το σημείο ενώνονται σχηματίζοντας την συγκεκριμένη πρωτεΐνη. [3,4]



Εικόνα 2.6 Το κεντρικό δόγμα της Βιολογίας παρθέν από ‘CytoChemistry’ (http://www.cytochemistry.net/cell-biology/ribosome.htm)

1st position (5' end)	2nd position U	2nd position C	2nd position A	2nd position G	3rd position (3' end)
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

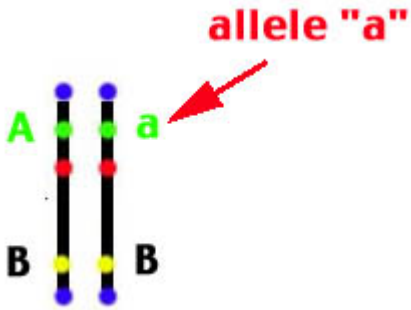
Εικόνα 2.7 Τα αμινοξέα που παράγονται από τα mRNA κωδικόνια παρθέν από 'Biolibogy' (http://biolibogy.com/chemistryoflife.html)

2.1.5 Αλληλόμορφο (Allele)

Ένα αλληλομορφο (allele) είναι μια εναλλακτική μορφή ενός γονιδίου ή ακολουθίας DNA που βρίσκονται σε μια συγκεκριμένη θέση ή ακόμα και ολόκληρου χρωμοσώματος. Συνήθως alleles είναι ακολουθίες κώδικα για ένα γονίδιο, αλλά μερικές φορές ο όρος χρησιμοποιείται για να αναφερθεί σε μια μη-γονιδιακή αλληλουχία.

Οργανισμοί, με δύο σετ χρωμοσωμάτων (διπλοειδή), όπως τα ζώα και τα φυτά, έχουν χρωμοσώματα που βρίσκονται ως ζεύγη στον πυρήνα του κάθε κυττάρου. Αυτό σημαίνει ότι θα υπάρχουν πάντα δύο γονίδια ή ακολουθία DNA για ένα χαρακτηριστικό σε ένα πυρήνα. Αν το ίδιο allele εμφανίζεται δύο φορές, ο οργανισμός θεωρείται ότι είναι ομόζυγος για αυτό το χαρακτηριστικό. Εάν, ωστόσο, ένα χρωμόσωμα περιέχει ένα allele και το άλλο χρωμόσωμα παρουσιάζει ένα διαφορετικό allele, ο οργανισμός ορίζεται ως ετερόζυγος.

Σε ετερόζυγους οργανισμούς, ο φαινότυπος του μπορεί να καθορίζεται από ένα allele και όχι από το άλλο. Το allele που καθορίζει το φαινότυπο λέγεται ότι υπερισχύει στην έκφραση (dominant). Δείχνει ότι υπερτερεί σε σχέση με τα άλλα alleles. Η έκφραση του άλλου allele περιγράφεται ως υποτελή. [4]



Εικόνα 2.8 Alleles

παρθέν από 'Catholic University of louvain la neuve – School of Medicine' (http://www.icampus.ucl.ac.be/courses/SBIM2520/document/genemol/electrophorese/chromo2a.jpg)

2.1.6 Σημειακοί Νουκλεοτιδικοί Πολυμορφισμοί (SNP)

Οι σημειακοί νουκλεοτιδικοί πολυμορφισμοί (single nucleotide polymorphism SNP) είναι μικρές αλλαγές στην αλληλουχία των βάσεων του DNA. Είναι διάσπαρτοι σε όλο το γονιδίωμα των οργανισμών. Βρίσκονται και κοντά και μέσα στα γονίδια. Συνήθως αφορά αλλαγή, έλλειψη ή εισδοχή μιας βάσης πάνω στην γενετική ακολουθία. Πιο συγκεκριμένα, είναι μια παραλλαγή μιας ακολουθίας DNA που συμβαίνει όταν ένα μόνο νουκλεοτίδιο (A,T,C,G) στο γονιδίωμα, διαφέρει σε άτομα ενός συγκεκριμένου είδους, π.χ. δύο τμήματα DNA από δύο διαφορετικά άτομα AGCGATC σε AGCAATC, διαφέρουν σε ένα μόνο νουκλεοτίδιο. Σ' αυτή την περίπτωση έχουμε δύο alleles το G και το A.

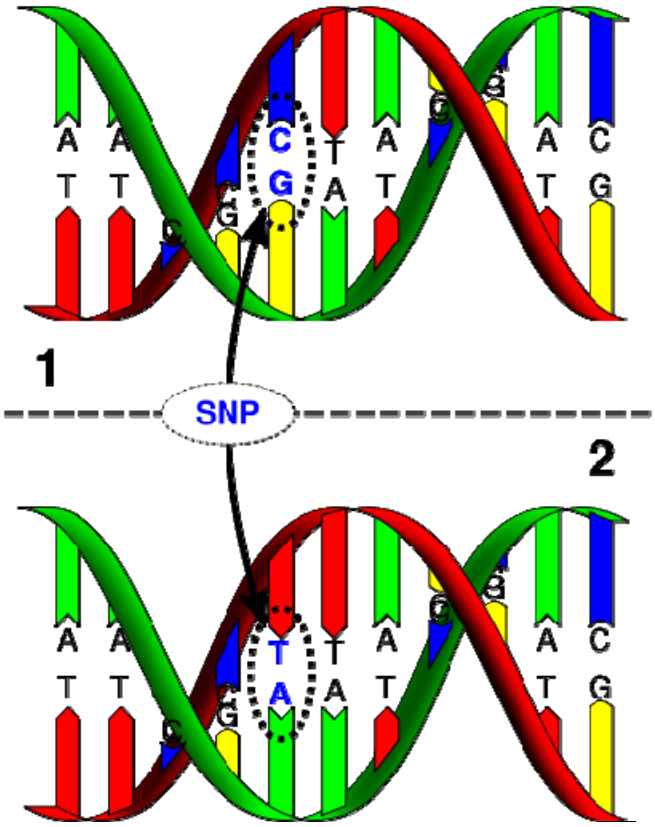
Παρατηρείται πολύ συχνή δημιουργία SNPs, αποτελώντας έτσι τους πιο κοινούς πολυμορφισμούς του DNA (Υπάρχουν και άλλου είδους πολυμορφισμοί). Το σημαντικό σημείο που πρέπει να αναφερθεί, είναι το γεγονός ότι τα SNPs τα οποία παρατηρούνται σε ένα πληθυσμό, έχουν δημιουργηθεί κατά την εξέλιξη ενός συγκεκριμένου οργανισμού. Συνήθως ένα SNP παρουσιάζεται σε ένα σημαντικό

ποσοστό (τουλάχιστον 1%) ενός πληθυσμού. Ειδικά στον άνθρωπο, τα SNPs καλύπτουν 90% των γενετικών παραλλαγών που παρατηρούνται στον άνθρωπο. Ένας τέτοιος πολυμορφισμός παρατηρείται κάθε 100 μέχρι και 300 βάσεις κατά μήκος των 3 δισεκατομμυρίων βάσεων του ανθρώπινου γονιδιώματος.

Ένα SNPs μπορεί να βρίσκεται σε ακολουθίες των γονιδίων που κωδικοποιούν πρωτεΐνες, σε ακολουθίες των γονιδίων που δεν κωδικοποιούν πρωτεΐνες, ή σε ενδογενείς περιοχές (intergenic regions) μεταξύ των γονιδίων. Η παρουσία ενός SNP σε μια περιοχή κωδικοποίησης δεν σημαίνει απαραίτητα την αλλαγή του αμινοξέως που αποτελεί μέρος μια πρωτεΐνης η οποία παράγεται. Ένα SNP που παράγει την ίδια πολυπεπτιδική αλυσίδα με την αρχική ακολουθία ονομάζεται συνώνυμο (ή σιωπηλή μετάλλαξη). Αν παράγεται διαφορετική πολυπεπτιδική αλυσίδα τότε το SNP ονομάζεται μη συνώνυμο. SNPs που δεν βρίσκονται σε περιοχές που κωδικοποιούνται, πάλι μπορεί να προκαλούν διάφορα συνέπειες όπως για παράδειγμα στην ακολουθία του μη – κωδικοποιημένου RNA.

Όπως έχει προαναφερθεί, τα SNPs μπορούν να έχουν πολλές συνέπειες τόσο στο γονότυπο όσο και στο φαινότυπο ενός οργανισμού. Ειδικότερα στο άνθρωπο, στα SNPs μπορεί να οφείλονται πολλές διαφορές που παρουσιάζουν οι άνθρωποι στα χαρακτηριστικά τους, όπως για παράδειγμα το χρώμα των ματιών. Επιπλέον, μπορεί να είναι βοηθήσουν στο καθορισμό της πιθανότητας ανάπτυξης διάφορων ασθενειών καθορίζοντας την ευαισθησία κάποιου ατόμου σε μια ασθένεια, ή την πιθανότητα αντίδρασης του οργανισμού του σε περίπτωση χορήγησης κάποιου συγκεκριμένου φαρμάκου.

Τα περισσότερα κοινά SNPs αποτελούνται από 2 μόνο alleles. Για την συγκεκριμένη μελέτη, τα δεδομένα αφορούσαν δι – αλληλόμορφα (biallelic)SNPs. Συνεπώς υπάρχουν τέσσερις δυνατές περιπτώσεις για ένα SNP, οι οποίες είναι ΑΑ, Αα, αα και η τέταρτη περίπτωση αυτή στην οποία παρατηρούνται missing data δηλαδή έλλειψη δεδομένων, διαφορετικά έλλειψη νουκλεοτιδίων σε κάποιες θέσεις των ακολουθιών DNA. [3,4,5]



Εικόνα 2.9 Ένας σημειακός νουκλεοτιδικός πολυμορφισμός
παρθέν από 'SNiP Screen' (http://www.snipscreen.com/what_we_do.php)

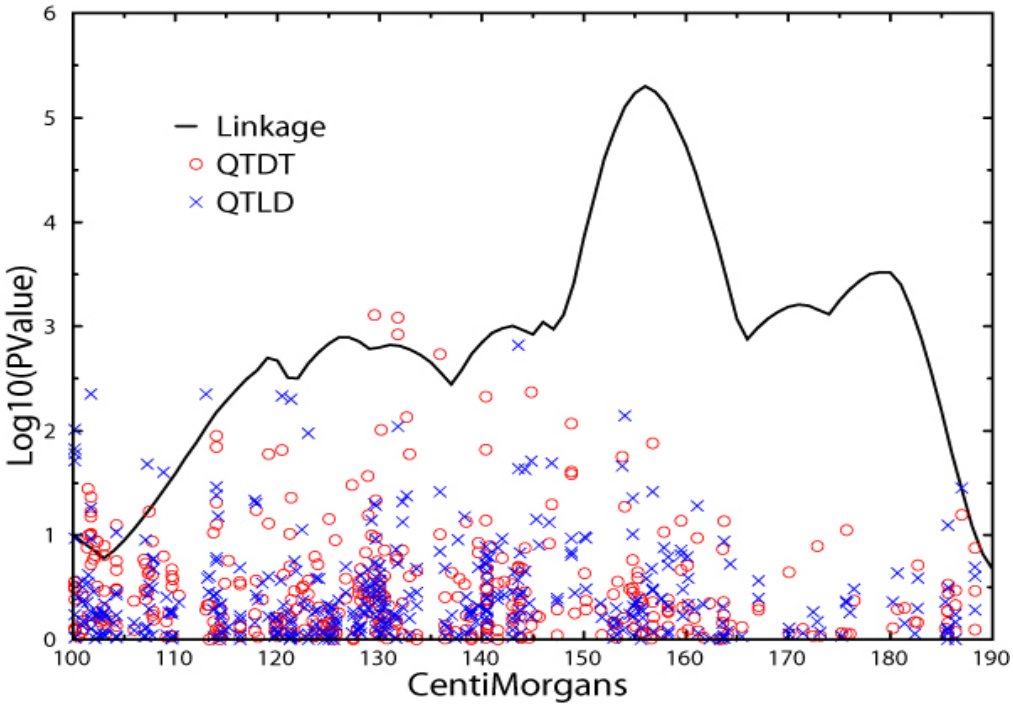
2.1.7 Ανισορροπία Σύνδεσης (Linkage Disequilibrium - LD)

Σαν Ανισορροπία Σύνδεσης, ορίζονται οι μη τυχαίες συσχετίσεις μεταξύ alleles σε διαφορετικά loci (locus ορίζεται η θέση ενός γονίδιου ή μιας άλλης DNA ακολουθίας σε ένα χρωμόσωμα). Παρατηρείται όταν οι φαινότυποι σε δύο διαφορετικά loci, δεν είναι ανεξάρτητα μεταξύ τους. Ουσιαστικά περιγράφει μια κατάσταση κατά την οποία συνδυασμός allele (ειδικότερα για σημειακούς νουκλεοτιδικούς πολυμορφισμούς) συμβαίνουν λιγότερο ή περισσότερο συχνά σε ένα πληθυσμό, από ότι κάποιος θα ανέμενε από ένα τυχαίο σχηματικό allele βασισμένων στις δικές του συχνότητες.

Είναι σημαντικό να αναφερθεί ότι οι συσχετίσεις αυτές μπορούν να παρατηρηθούν ακόμα και αν τα δύο alleles σε δύο διαφορετικά loci δεν είναι συνδεδεμένα. Επιπρόσθετα, δεν σημαίνει ότι επειδή δύο loci είναι συνδεδεμένα, συνεπάγεται απαραίτητα ότι βρίσκονται σε Ανισορροπία Σύνδεσης (LD). Συνεπώς, μπορεί να

χρησιμοποιηθεί και ως μέτρο σύγκρισης για τους σημειακούς νουκλεοτιδικούς πολυμορφισμούς.

Η μέτρηση της Ανισορροπίας Σύνδεσης είναι πολύ σημαντική καθώς αν όλοι οι πολυμορφισμοί ήταν ανεξάρτητοι σε ένα πληθυσμό, μελέτες συσχετίσεις (association studies) θα έπρεπε να εξετάσουν όλους του πολυμορφισμούς ένα προς ένα για εξαγωγή συμπερασμάτων. Η Ανισορροπία Σύνδεσης μειώνει τα κόστη των μελετών συσχετίσεων. Επιπρόσθετα μπορεί να δείξει ότι η παρουσία δύο σημειακών νουκλεοτιδικών πολυμορφισμών μπορεί να ευθύνεται για την εμφάνιση μίας ασθένειας, ενώ η παρουσία ενός από των δύο πολυμορφισμών να μην συνεπάγεται την εμφάνιση της. [6]



Εικόνα 2.10 Linkage Disequilibrium παρθέν από 'BioMedCentral France' (http://biomedcentral.inist.fr/images/1471-2156-6-S1-S91-1.jpg)

2.1.8 Επίσταση (Epistasis)

Ως επίσταση χρησιμοποιείται ως όρος για να περιγράψει ένα φαινόμενο απόκρυψης όπου μια παραλλαγή ενός allele σε ένα συγκεκριμένη θέση (locus) εμποδίζει μια άλλη παραλλαγή σε μια άλλη θέση να εκδηλώσει τη δράση του. Ήρθε για να ενισχύσει την έννοια του επικρατούμενου allele. Έστω ότι έχουμε δυο loci, B και G, που επηρεάζουν σε μια φυλή ένα συγκεκριμένο χαρακτηριστικό. Το locus B μπορεί να έχει δύο πιθανά alleles B ή b. Αντίστοιχα, ισχύει και για το G. Όπως φαίνεται και στον πίνακα 2.1, άσχετα με τον γονότυπο του b, άτομα με οποιοδήποτε αντίγραφο από το allele G, έχουν γκρίζα μαλλιά, δηλαδή το G επικρατεί του allele g. «αποκρύπτοντας» την δράση του g. Τότε, το G επικρατεί (dominant) του g. Αν όμως, ο γονότυπος στο locus G δεν είναι g/g, τότε η δράση του locus B δεν εμφανίζεται, καθώς άτομα με οποιοδήποτε αντίγραφο του allele G έχει γκρίζα μαλλιά ανεξάρτητα του γονότυπου στο locus B. Συνεπώς η δράση του locus B «καλύπτεται» από το locus G και το G ονομάζεται επιστατικό ως προς το B. [7]

Όταν έχουμε επιστατική επίδραση η ύπαρξη των allele που είναι υπεύθυνα για ένα χαρακτηριστικό είναι απαραίτητη για την εμφάνιση του. Αν δεν υπάρχει έστω ένα από τα συγκεκριμένα γονίδια τότε το συγκεκριμένο χαρακτηριστικό δεν εμφανίζεται στο φαινότυπο. Συνήθως, η επίσταση ορίζει τις στατικές ιδιότητες του φαινομένου. Όπως γίνεται εύκολα αντιληπτό υπάρχει μεγάλη συσχέτιση με την Ανισορροπία Σύνδεσης (συγκεκριμένα υπάρχει η καταλληλότητα επίστασης (fitness epistasis) που αποτελεί συσχέτιση μεταξύ δύο ή περισσοτέρων loci που παρουσιάζει μη – γραμμικά αποτελέσματα στην καταλληλότητα και είναι η αιτία της Ανισορροπία Σύνδεσης). Αν η Ανισορροπία Σύνδεσης είναι ελάχιστη ή μη παρούσα, τότε μπορούμε να συμπεράνουμε ότι η επίσταση δεν είναι σημαντική σε κάποιο πληθυσμό. Αντίθετα αν η Ανισορροπία Σύνδεσης είναι μεγάλη, τότε η επίσταση μπορεί να είναι κοινή [8]

	Genotype at locus G		
Genotype at locus B	*g/g*	*g/G*	*G/G*
b/b	White	Grey	Grey
b/B	Black	Grey	Grey
B/B	Black	Grey	Grey

Πίνακας 2.1 Παράδειγμα φαινοτύπων (χρώμα μαλλιών) από διαφορετικούς γονότυπους παρθέν από 'Bateson' ([7])

2.2 Χάρτες αυτό – οργάνωσης – Θεωρία

2.2.1 Εισαγωγή

Τα δίκτυα Kohonen μας προσφέρουν ένα τρόπο αναπαράστασης πολυδιάστατων δεδομένων, δηλαδή δεδομένων που αναπαριστώνται σε περισσότερες από μία διαστάσεις σε πολύ μικρότερες διαστάσεις. Ουσιαστικά αποτελεί ένα τρόπο συμπίεσης δεδομένων. [9] Τα δίκτυα SOM όπως ονομάζονται διαφορετικά τα δίκτυα Kohonen, είναι εμπνευσμένα από γνωστές τοπολογικές ιδιότητες του εγκέφαλου. [12]

Επιπλέον μας δίνουν την δυνατότητα ομαδοποίησης (clustering) δεδομένων (έρχεται με την ιδιότητα της μετατροπής της αναπαράστασης σε χαμηλότερες διαστάσεις) [9]. Τα σχέδια (patterns) δηλαδή τα δεδομένα εισόδου, που είναι κοντά το ένα στο άλλο που συνδέονται δηλαδή με κάποια συγκεκριμένη σχέση πρέπει να είναι στενά συνδεδεμένα το ένα άλλο στο άλλο στο χάρτη: πρέπει να διαταχτούν τοπολογικά. Αυτό μας προσφέρεται από ένα SOM δίκτυο (Self Organizing Map ή αλλιώς Kohonen Network). [10] Ένα SOM δίκτυο αποθηκεύει πληροφορίες με ένα τρόπο έτσι ώστε να διατηρούνται οι τοπολογικές σχέσεις στο σύνολο των διανυσμάτων εισόδου με την χρήση. Συνεπώς και η δυνατότητα ομαδοποίησης.

Τα Kohonen Networks ανήκουν στην κατηγορία των νευρωνικών δικτύων και πιο συγκεκριμένα στην κατηγορία των competitive networks without supervision. Δηλαδή στην κατηγορία των νευρωνικών δικτύων που η κάθε μονάδα συναγωνίζεται την άλλη και έχει την ιδιότητα να μην χρειάζεται «δάσκαλο» για να διδάξει το δίκτυο. Δεν

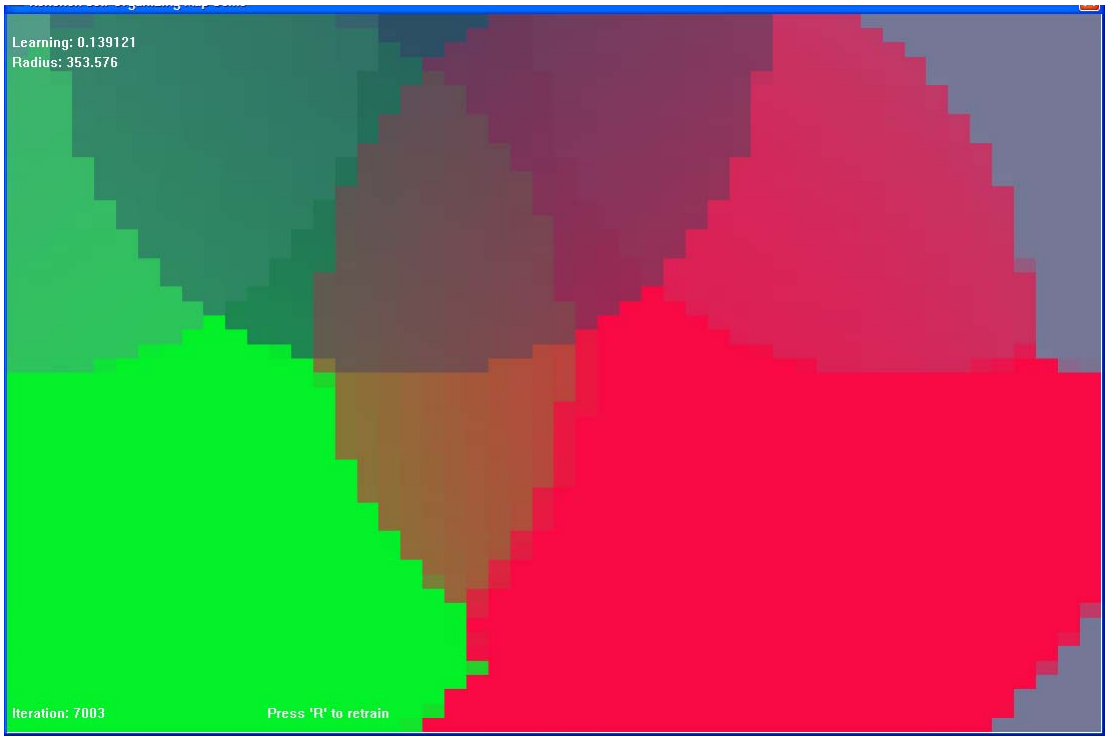
υπάρχει η ανάγκη ύπαρξης δεδομένων για training set δηλαδή των δεδομένων που θέλουμε να παράγεται από το δίκτυο. Στην κατηγορία των δικτύων με supervision υπάρχει πάντα ένα διάνυσμα εισόδου (input vector) και ένα διάνυσμα στόχου (target vector). Με την χρήση του αλγόριθμου back – propagation διορθώνεται σε κάθε επανάληψη τα βάρη των μονάδων του δικτύου έτσι ώστε το αποτέλεσμα που παράγεται να φτάσει να μοιάσει στο διάνυσμα στόχου. Στα Kohonen δίκτυα δεν υπάρχει η ανάγκη χρήσης του διανύσματος στόχου (Η διαδικασία μάθησης σε ένα Kohonen δίκτυο θα εξηγηθεί ακολούθως). [11]

Ένα Kohonen Network αποτελείται από ένα πλέγμα (grid) από μονάδες εξόδου και από μονάδες εισόδου. Η κάθε τιμή του διανύσματος εισόδου εισάγεται σε μια μονάδα εισόδου και ακολούθως σε κάθε μονάδα εξόδου. Σε κάθε μονάδα εξόδου αντιστοιχεί ένα διάνυσμα ίσου μεγέθους με το διάνυσμα εισόδου που αναπαριστούν τα βάρη κάθε μονάδας εξόδου (αρχικοποιούνται τυχαία σε μικρούς αριθμούς). [10]

2.2.2 Ταξινόμηση Δεδομένων

Όπως έχει ήδη αναφερθεί, η ομαδοποίηση/ ταξινόμηση δεδομένων (data classification) αποτελεί την κύρια αιτία χρήσης των SOM δικτύων (σε συνδυασμό με την χαρτογράφηση(mapping) από *n*- διαστάσεις σε μικρότερες καταστάσεις) . Έστω ότι υπάρχει ένα σύνολο από *n*- διαστάσεων διανυσμάτων που περιγράφουν αντικείμενα. Κάθε στοιχείο του κάθε διανύσματος αναπαριστά μία ιδιότητα για το συγκεκριμένο αντικείμενο. Το κάθε διάνυσμα είναι μοναδικό όπως και τα αντικείμενα που αναπαριστούν. Η ομαδοποίηση των αντικειμένων αυτών γίνεται πολύ εύκολα με την χρήση SOM. Τα αντικείμενα αυτά θα ομαδοποιηθούν σε διακριτές περιοχές με τέτοιο τρόπο, έτσι ώστε στην κάθε περιοχή θα βρεθούν αντικείμενα με όμοιες ιδιότητες. [11]

Ένα βασικό παράδειγμα χρήσης είναι η χαρτογράφηση (mapping) των χρωμάτων (που αναπαριστώνται με βάση τα τρία χρώματα που αποτελούν το κάθε χρώμα <κόκκινο, πράσινο, μπλε>) σε <μήκος, πλάτος> δηλαδή από 3-D σε 2-D. Με την χρήση των SOM, τα χρώματα θα ομαδοποιηθούν σε διάφορες περιοχές ανάλογα με τις τιμές που δίνονται στις παραμέτρους (RGB Coloring) όπως φαίνεται και στην πιο κάτω εικόνα. [9]



Εικόνα 2.11 Ένα τυχαίο παράδειγμα με ένα τυχαίο σύνολο στοιχείων σε ένα δίκτυο Kohonen παρθέν από 'AI Junkie' ([9])

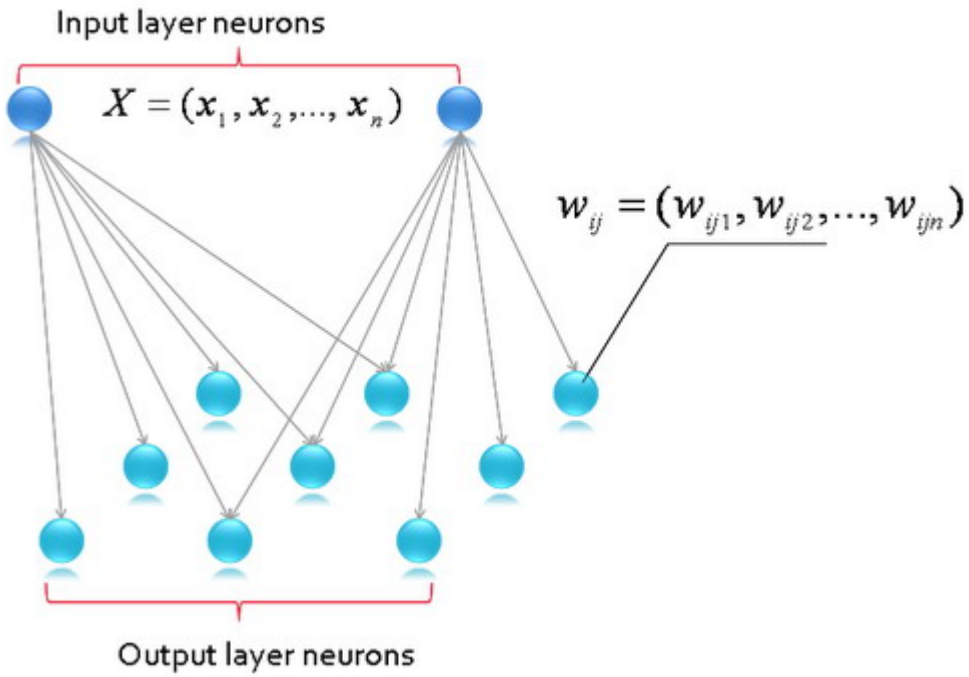
2.2.3 Αρχιτεκτονική Δικτύων Kohonen

Έχουμε ήδη αναφέρει ότι ένα Kohonen δίκτυο αποτελείται από ένα πλέγμα με μονάδες εισόδου και εξόδου.

Συγκεκριμένα υπάρχουν δύο επίπεδα νευρώνων. Το πρώτο επίπεδο αποτελεί το επίπεδο εισόδου και ουσιαστικά αποτελείται από *n*- νευρώνες – μονάδες εισόδου. Ουσιαστικά ο αριθμός των νευρώνων αυτών καθορίζει το μέγεθος της διάστασης των διανυσμάτων εισόδου (input vectors). (Στο παράδειγμα με τα χρώματα θα έχουμε τρεις νευρώνες εισόδου – ο πρώτος για το κόκκινο, ο δεύτερος για το πράσινο και ο τρίτος για το μπλε). Κάθε νευρώνας – μονάδα εισόδου συνδέεται με όλους τις μονάδες εξόδου στο δεύτερο επίπεδο.[10,11]

Το δεύτερο επίπεδο είναι συνήθως συνδεδεμένο σαν δισδιάστατο πλέγμα. Υπάρχουν και πιο δύσκολες εφαρμογές που απαιτούν διαφορετικό σχεδιασμό του δευτέρου επιπέδου όμως τις πλείστες εφαρμογές λόγω απλότητας και χρησιμότητας προτιμείται

αυτός ο σχεδιασμός. Κάθε νευρώνας εξόδου (ή νευρώνας δευτέρου επιπέδου) , όπως έχουμε ήδη πει είναι συνδεδεμένος με όλους τους νευρώνες εισόδου. Επιπλέον σε κάθε νευρώνα εξόδου υπάρχει ένα διάνυσμα *n*- μεγέθους με αριθμούς που αποτελούν τα βάρη για το συγκεκριμένο νευρώνα και έχει το ίδιο μέγεθος με τον αριθμό των νευρώνων εισόδου. Επιπρόσθετα ο κάθε νευρώνας εξόδου συσχετίζεται με άλλους παρακείμενους νευρώνες εξόδου καθορίζοντας την τοπολογία/δομή του δικτύου (2-D, Hex κ.τ.λ.). Η συσχέτιση αυτή καθορίζει γειτνίαση των νευρώνων αυτών και καθορίζεται με βάση μια συνάρτηση που ονομάζεται topological neighborhood (Βλέπε Εικόνα 3.14).



Εικόνα 2.12 Δομή ενός Kohonen Δικτύου παρθέν από 'The Code Project' ([11])

Η τοπολογική ιδιότητα – δηλαδή η διατήρηση των σχέσεων μεταξύ των δεδομένων διανυσμάτων εισόδου διατηρείται με την έννοια της «γειτονιάς» (neighbourhood) κατά την διαδικασία εκμάθησης. [11] Ουσιαστικά η «γειτονιά» είναι ένα σύνολο νευρώνων εξόδου που συσχετίζονται με βάση μια συνάρτηση (topological neighborhood). Συνήθως η συνάρτηση αυτή καθορίζεται με βάση την απόσταση στο τοπολογικό χάρτη δηλαδή στην τοπολογία του δικτύου μας (γραμμική απόσταση). Όπως θα εξηγηθεί πιο

κάτω θα γίνεται χρήση της γραμμικής απόστασης. [12] Το μέγεθος της γειτονιάς μειώνεται ανάλογα με τον αριθμό των επαναλήψεων της διαδικασίας εκμάθησης.

2.2.4 Εκμάθηση στα δίκτυα Kohonen (Learning σε Kohonen Networks)

Τα δίκτυα Kohonen ανήκουν στην κατηγορία των νευρωνικών δικτύων και πιο συγκεκριμένα στην κατηγορία των competitive networks without supervision. Δηλαδή δεν υπάρχει διάνυσμα στόχου. Αντίθετα η εκμάθηση του δικτύου γίνεται με διαφορετικό τρόπο. Αρχικά έχουμε μια τυχαία κατανομή των βαρών κάθε κόμβου εξόδου. Ακολούθως η μονάδα εξόδου, της οποίας τα βάρη ταιριάζουν με το διάνυσμα εισόδου, και η περιοχή του πλέγματος, δηλαδή οι γειτονικοί κόμβοι εξόδου που υπολογίζονται με βάση την συνάρτηση topological neighborhood και αποτελούν τις γειτνιάσεις της μονάδας αυτής, επιλέγονται έτσι ώστε να μοιάσουν με τα στοιχεία της κατηγορίας στην οποία ανήκει το διάνυσμα εισόδου.

Πιο συγκεκριμένα η διαδικασία εκμάθησης που ακολουθείται είναι η εξής:

1. Τα βάρη κάθε νευρώνα εξόδου αρχικοποιούνται τυχαία.
2. Επιλέγεται τυχαία ένα διάνυσμα εισόδου από το σύνολο διανυσμάτων που θεωρούνται τα δεδομένα που θέλουμε να κατηγοριοποιηθούν
3. Συγκρίνονται τα βάρη κάθε κόμβου με τα δεδομένο διάνυσμα εισόδου. Ο κόμβος εξόδου του οποίου τα βάρη «μοιάζουν» περισσότερο από κάθε άλλου κόμβου με τα δεδομένο εισόδου, επιλέγεται (best matching unit BMU) (ο τρόπος επιλογής γίνεται με βάση μια συνάρτηση συνήθως γίνεται χρήση της Ευκλείδειος απόστασης)
4. Ακολούθως υπολογίζονται οι γειτονικοί κόμβοι ανάλογα με το μέγεθος της ακτίνας. Υπολογίζονται με βάση μια τιμή που ονομάζεται ακτίνα της γειτονιάς του επιλεγόμενου κόμβου. Η τιμή αυτή μειώνεται σε κάθε επανάληψη και όσοι κόμβοι βρίσκονται μέσα στην τιμή αυτή ανήκουν στην περιοχή αυτή
5. Τα βάρη των κόμβων στην περιοχή αυτή αλλάζουν έτσι ώστε να μοιάζουν στο διάνυσμα εισόδου. Όσο πιο κοντά είναι κάποιος κόμβος στο κόμβο BMU, τόσο μεγαλύτερη αλλαγή υφίστανται τα βάρη της. [9]

Η διαδικασία επαναλαμβάνεται (από το b) για ένα αριθμό επαναλήψεων που λέγονται εποχές (epochs) που καθορίζονται από τον χρήστη του δικτύου. Όπως γίνεται αντιληπτό, λόγω της συχνής χρήσης τυχαιοποίησης (randomization), όσο μεγαλύτερος είναι ο αριθμός των εποχών τόσο καλύτερη κατηγοριοποίηση (classification) θα γίνει από το δίκτυο. Όμως είναι σημαντικό να παρατηρηθεί ότι αν ξεπεραστεί ένα αρκετά μεγάλο αριθμό εποχών τότε υπάρχει περίπτωση να πλησιάζει το δίκτυο στο σημείο του over learning δηλαδή στο σημείο όπου το δίκτυο πλέον θα μπορεί να κατηγοριοποιεί μόνο πολύ συγκεκριμένα δεδομένα εισόδου. Σκοπός των δικτύων Kohonen είναι η γενική κατηγοριοποίηση δεδομένων όχι συγκεκριμένων δεδομένων.

Για να γίνει καλύτερα κατανοητός ο αλγόριθμος θα δοθεί βαρύτητα στο τρόπο επιλογής του BMU , το τρόπο που υπολογίζεται η ακτίνα του BMU με την βοήθεια της συνάρτησης topological neighborhood, όπως επίσης και το πώς αλλάζουν τα βάρη στους κόμβους που βρίσκονται μέσα στην ακτίνα του BMU σε κάθε επανάληψη.

Η επιλογή του BMU μπορεί να γίνει με την βοήθεια πολλών συναρτήσεων. Συνήθως όμως γίνεται η χρήση της Ευκλείδειας απόστασης. Δηλαδή για κάθε μονάδα – νευρώνα εισόδου υπολογίζεται η Ευκλείδεια απόσταση του διανύσματος εισόδου με το διάνυσμα των βαρών του. Ο κόμβος με την μικρότερη απόσταση θεωρείται ο BMU. Η Ευκλείδεια απόσταση δίνεται από το μαθηματικό τύπο:

$$\sqrt{\sum_{j=0}^{j=n} (V_j - W_j)^2}$$

Εξίσωση 1

όπου

n μέγεθος του διανύσματος εισόδου και βαρών του κόμβου εξόδου που ελέγχεται,

Vj το j-οστό στοιχείο του διανύσματος εισόδου,

Wj το j-οστό στοιχείο του διανύσματος των βαρών του κόμβου εξόδου που ελέγχεται

Για τον υπολογισμό της ακτίνας του BMU και πιο συγκεκριμένα των κόμβων – νευρώνων εξόδου που ανήκουν στην «γειτονιά» (neighbourhood) του BMU χρησιμοποιείται συνήθως η ακόλουθη συνάρτηση (μπορεί ο χρήστης του δικτύου να επιλέξει άλλη συνάρτηση):

$$r(\iota) = r_0 \times \exp\left(-\left(\frac{\iota}{\lambda}\right)\right)$$

Εξίσωση 2

ι = 0, 1 , 2 ,…

όπου

ι ο αριθμός της εποχής που βρίσκεται το δίκτυο μας,

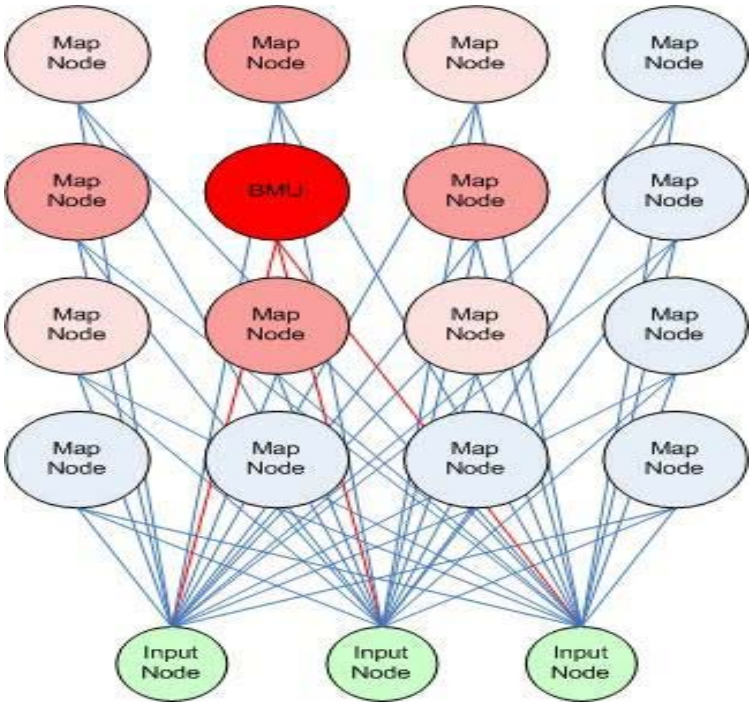
r0 είναι η αρχική τιμή της ακτίνας (συνήθως έχει το μέγεθος του πλέγματος),

r(i) είναι η τιμή της ακτίνας κατά την εποχή i,

λ μια σταθερά που εξαρτάται από την ακτίνα (αριθμός εποχών που επιλέχθηκε/ log (r0)) [9]

Με βάση την τιμή της συνάρτησης αυτής, αυτό που μένει είναι ο έλεγχος για το ποιοι νευρώνες εξόδου ανήκουν σ' αυτή την ακτίνα. Είναι σημαντικό να παρατηρήσουμε ότι και εδώ μετριέται η ευκλείδεια απόσταση στη τοπολογία μας (π.χ. σε ένα δισδιάστατο πλέγμα η απόσταση ενός κόμβου από τον άλλο μπορεί να υπολογιστεί και με βάση το Πυθαγόρειο Θεώρημα). Έτσι έχοντας αυτή την τιμή από τον BMU και την τιμή της ακτίνας μπορούμε εύκολα να προσδιορίσουμε το ποιοι κόμβοι ανήκουν στην «γειτονιά» του BMU. Επιπλέον, ένα μοναδικό χαρακτηριστικό των δικτύων Kohonen είναι ότι η ακτίνα σε κάθε εποχή (επανάληψη) μειώνεται και αυτό επιτυγχάνεται προσθέτοντας στον υπολογισμό της ακτίνας το μέρος exp(- i) . Αν ο χρήστης επέλεξε αρκετό αριθμό εποχών τότε η ακτίνα θα καταλήξει με μέγεθος 1 που θα είναι μόνο ο νευρώνας BMU κάτι που θα σημαίνει ότι έχουμε πετύχει πολύ καλή κατηγοριοποίηση (Βλέπε Εικόνα 3.14). [9,10,14] Επίσης πρέπει να γίνει ιδιαίτερη αναφορά στην χρήση της σταθεράς λ. Η σταθερά λ δίνει την δυνατότητα στο χρήστη να ελέγξει το ρυθμό της πτώσης της ακτίνας. Με τον προσδιορισμό της σταθεράς αυτής (μπορεί να οριστεί απευθείας και όχι με βάση το τύπο) ο χρήστης του δικτύου μπορεί να ελέγξει την πτώση της ακτίνας έτσι ώστε να γίνει με πιο ομαλό τρόπο. Με πολύ μεγάλη τιμή στο λ η ακτίνα μειώνεται πιο αργά, το δίκτυο αργεί να συγκλίνει, χρειάζεται μεγαλύτερος αριθμός εποχών αλλά

υπάρχει μεγαλύτερη πιθανότητα εύρεσης της βέλτιστης λύσης δηλαδή της κατηγοριοποίησης. Μικρότερη τιμή σημαίνει πιο απότομη πτώση στην ακτίνα, λιγότερος υπολογισμός, γρηγορότερη σύγκλιση αλλά πιθανώς να μην βρεθεί η βέλτιστη λύση.



Εικόνα 2.13 Επιλογή BMU και ακτίνας ενός Kohonen Δικτύου παρθέν από 'Generation5' ([14])

Έχοντας υπολογίσει το σύνολο των κόμβων που βρίσκονται εντός της ακτίνας του BMU (συμπεριλαμβανομένου και του BMU), θα γίνουν οι αλλαγές στα βάρη με βάση την πιο κάτω συνάρτηση:

$$\overrightarrow{W_{i+1}} = \overrightarrow{Wi} + \Theta(i)L(i)\left(\overrightarrow{V_i} - \overrightarrow{W_i}\right)$$

Εξίσωση 3

ι = 0, 1 , 2 ,…

όπου i ο αριθμός των εποχών που βρισκόμαστε

W_i το παλιό διάνυσμα βαρών

W_{i+1} το νέο διάνυσμα βαρών

Vi το δεδομένο διάνυσμα εισόδου

$L(i)$ που αποτελεί το learning rate

$\Theta(i)$ που αποτελεί το πόσο η απόσταση από το BMU επηρεάζει την μάθηση του [9]

Το learning rate $L(i)$ είναι συνήθως μια μικρή τιμή συνήθως μεταξύ 0.1 – 0.9 που μειώνεται όσο αυξάνεται ο αριθμός των επαναλήψεων. Ουσιαστικά αποτελεί το μέγεθος τροποποίησης των βαρών των διανυσμάτων «νικητών». Με μικρή τιμή στο learning rate το δίκτυο αργεί να συγκλίνει και συνεπώς απαιτείται μεγαλύτερος αριθμός εποχών που κοστίζει αρκετά. Συνήθως αυτή την διαδικασία την επιθυμούμε όταν το δίκτυο κάνει την «γενική» κατηγοριοποίηση του και επιθυμούμε το πλήρη διαχωρισμό των κατηγοριών και εμφάνιση των κοινών τους χαρακτηριστικών. Γενικά θέτοντας μεγάλη τιμή στο learning rate επιτρέπουμε στο δίκτυο να προχωρήσει/ συγκλίνει πιο γρήγορα, όμως οι πιθανότητες για να μην βρεθεί η βέλτιστη λύση αυξάνονται δηλαδή το clustering που θα γίνει δεν θα είναι το ιδανικό αφού θα βρει μεν τις κατηγορίες αλλά δεν θα μπορεί να τις ξεκαθαρίσει. Επιπλέον απαιτείται σε κάθε εποχή η μείωση του learning rate έτσι ώστε οι αλλαγές να γίνονται λεπτότερες, οι διακρίσεις μέσα στις περιοχές του χάρτη να γίνουν πιο ξεκάθαρες οδηγώντας σε ένα καλύτερο συντονισμό των νευρώνων και τελικά σε καλύτερες λύσεις λαμβάνοντας έτσι τα πλεονεκτήματα τόσο μιας μεγάλης τιμής στο learning rate(clustering) τόσο και μιας μικρής τιμής(fine tuning). [13]

$$L(i) = L_0 \times \exp\left(-\frac{i}{\lambda}\right)$$

ι = 0, 1 , 2 ,…

Εξίσωση 4

όπου i ο αριθμός των εποχών που βρισκόμαστε

λ μια σταθερά που εξαρτάται από την ακτίνα (αριθμός εποχών που πέρασαν/αριθμός εποχών

L0 η αρχική learning rate

Η τιμή $\Theta(i)$ εξαρτάται από την γραμμική απόσταση (απόσταση στην τοπολογία) μεταξύ του BMU και του κάθε νευρώνα. Ουσιαστικά καθορίζει το πόσο επηρεάζει η γραμμική απόσταση που έχει ένας κόμβος που ανήκει στην ακτίνα από τον BMU, δηλαδή πόσο πρέπει να επηρεαστεί η εκμάθηση και συνεπώς η αλλαγή των βαρών του συγκεκριμένου κόμβου.

$$\Theta(i) = \exp\left(-\frac{d^2}{2 \times r^2(i)}\right)$$

Εξίσωση 5

ι = 0, 1 , 2 ,…

όπου i ο αριθμός των εποχών που βρισκόμαστε

r(i) η τιμή που υπολογίζεται στην συνάρτηση 2

d η απόσταση του συγκεκριμένου κόμβου από τον BMU που υπολογίζεται στο βήμα d του αλγορίθμου

[9,11,15]

Κεφάλαιο 3

Μεθοδολογία

3.1 Βάσεις Δεδομένων

3.1.1 Βάση Δεδομένων κατά πλάκα σκλήρυνσης

Τα δεδομένα που χρησιμοποιήθηκαν στα πλαίσια της μελέτης αυτής, αποτελούντο όπως έχει προαναφερθεί, από διάφορα γενετικά σημεία στην περιοχή HLA από την οποία υπάρχει γνώση από προηγούμενη έρευνα, ότι προκύπτουν πολλαπλές αλληλεπιδράσεις συσχετιζόμενες με την κατά πλάκα σκλήρυνση. Τα δεδομένα ανήκουν στην μορφή του Linkage Pedigree Format ή Merlin Format και αποτελούνται από δύο διαφορετικά αρχεία τα .ped και .map αρχεία. [22]

3.1.1.1 Περιγραφή Βάσης Δεδομένων

Η βάση η οποία χρησιμοποιήθηκε στα πλαίσια της μελέτης αυτής αφορούσε την κατά πλάκα σκλήρυνση (multiple sclerosis). Είναι χρόνια ασθένεια που συχνά απενεργοποιεί νευρώνες και άξονες του εγκεφάλου προκαλώντας πολλά και σοβαρά προβλήματα όπως απώλειες μνήμης και μη φυσιολογικές συμπεριφορές στους ασθενείς, χτυπώντας στο κεντρικό νευρολογικό σύστημα του ανθρώπου [23]. Η βάση παρουσιάζεται στην βιβλιογραφία. [23]

Η προσπάθεια ξεκίνησε το 2003. Τρία κλινικά κέντρα της κατά πλάκα σκλήρυνσης (MS) συνεργάστηκαν για να πάρουν βιολογικά δείγματα από ασθενείς. Τα δύο ήταν

στην Ευρώπη και το ένα στην Αμερική. Κατά την μελέτη αυτή χρησιμοποιήθηκαν ασθενείς με ιστορική καταγωγή από την βόρεια Ευρώπη με προτίμηση να έχουν μια αρχική υποτροπή MS, αν και τα άτομα με όλες τις κλινικές υποκατηγορίες της ασθένειας συμμετείχαν, συμπεριλαμβανομένου του κλινικά απομονωμένου συνδρόμου (CIS), της επιστρέφοντας υποτροπής MS (RRMS), τα δευτεροβάθμια προοδευτικά MS (SPMS), τα αρχικά προοδευτικά MS (PPMS) και την προοδευτική υποτροπή MS (PRMS) [23].

Το αποτέλεσμα της συλλογικής αυτής προσπάθειας, κατέληξε στην δημιουργία της βάσης που χρησιμοποιήθηκε στα πλαίσια της μελέτης αυτής και περιέχει 1853 άτομα, ασθενείς και μη ασθενείς και μελέτη 327 σημειακών πολυμορφισμών. [23]

3.1.1.2 Τύποι Αρχείων που χρησιμοποιήθηκαν

Τα .ped αρχεία περιέχουν τα διάφορα γενετικά σημεία διαφόρων ατόμων που συλλέχτηκαν σε διάφορα βιολογικά εργαστήρια. Κάθε γραμμή αντιστοιχεί σε ένα άτομο. Οι στήλες τους διαχωρίζονται με white – space(space ή tab) και πάντα οι πρώτες έξι στήλες είναι υποχρεωτικές και αποτελούνται από τις στήλες

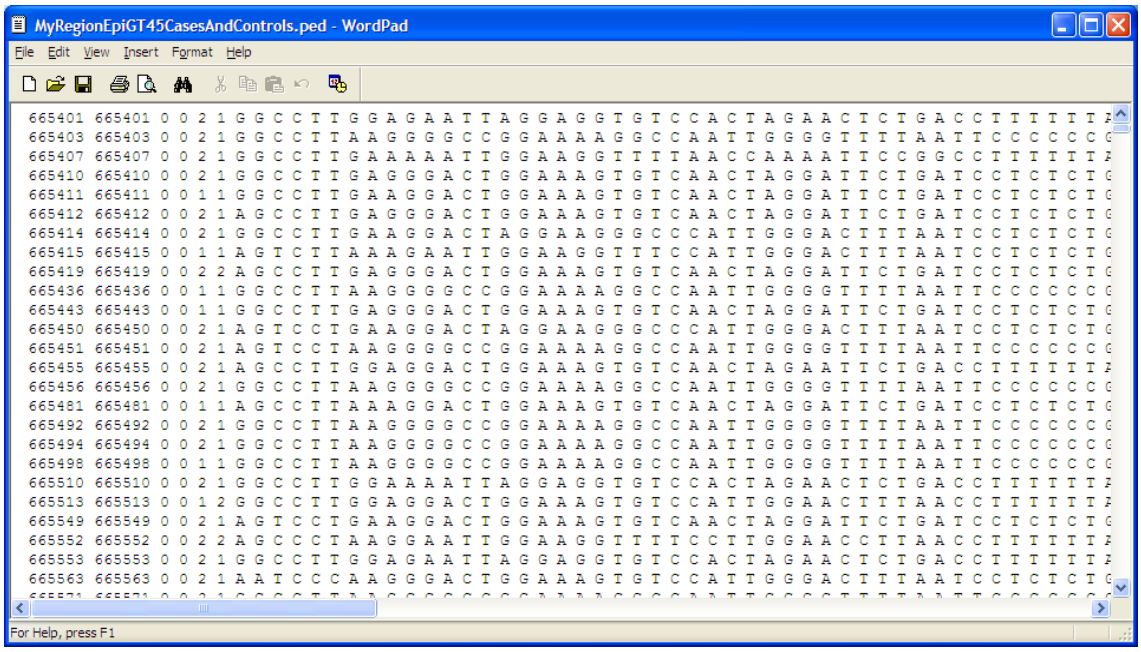
- Family ID
- Individual ID
- Paternal ID
- Maternal ID
- Sex (1=male; 2=female; other=unknown)
- Phenotype

Οι τιμές αυτών στην στηλών είναι αριθμοί, και ο συνδυασμός μεταξύ family ID και individual ID καθορίζουν μοναδικά ένα άτομο. Ο φαινότυπος καθορίζεται στην τελευταία στήλη δηλαδή κατά πόσο κάποιο άτομο είναι ασθενής ή υγιής. [8]

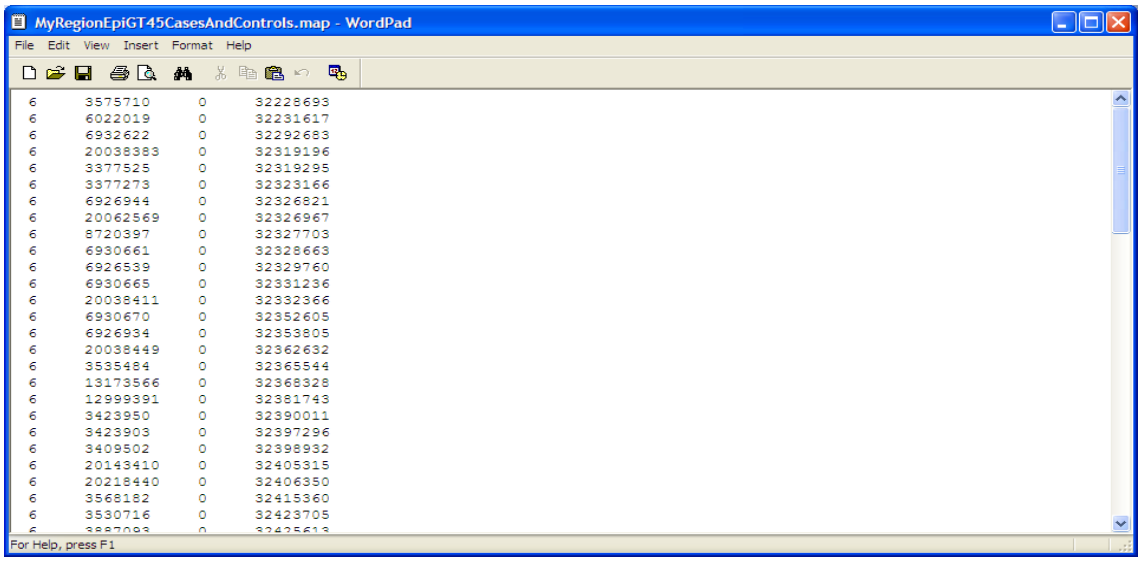
Τα .map αρχεία αποτελούν ουσιαστικά περιγραφές του κάθε γενετικού σημείου που υπάρχει σε ένα αντίστοιχο .ped αρχείο. Δηλαδή η τιμή κάθε στήλης (εκτός από τις 6 απαραίτητες) του .ped αρχείου χαρακτηρίζεται από μια γραμμή του .map αρχείου και συνήθως περιλαμβάνουν τις ακόλουθες πληροφορίες σε μία γραμμή.

- chromosome (1-22, X, Y or 0 if unplaced)
- rs# or snp identifier
- Genetic distance (morgans)
- Base-pair position (bp units)

Όπως και στα .ped αρχεία, οι στήλες διαχωρίζονται μεταξύ τους με white – space. [8]



Εικόνα 3.1 Παράδειγμα ped αρχείου



Εικόνα 3.2 Παράδειγμα map αρχείου

3.2 Εφαρμογές που Χρησιμοποιηθήκαν

3.2.1 Εφαρμογή PLINK για την επεξεργασία βιολογικών δεδομένων

Για την εξαγωγή διαφόρων γενετικών σημείων με σημαντική πληροφορία δηλαδή με μεγάλο βαθμό συσχετίσεων μεταξύ τους (ψηλό βαθμό επίστασης) χρησιμοποιήθηκε η εφαρμογή PLINK. Αποτελεί εφαρμογή ανοικτού κώδικα και προσφέρει μια πλειάδα από διάφορες βιολογικές και στατιστικές αναλύσεις σε γενετικά δεδομένα. Η χρήση του PLINK ήταν ιδιαίτερα σημαντική καθώς παρείχε την δυνατότητα χρήσης μόνο των γενετικών σημείων, τα οποία θεωρούνται ιδιαίτερα σημαντικά για την επίδραση μιας ασθένειας. [8]

Στο στάδιο της προεπεξεργασίας, έγινε χρήση των διαφόρων φίλτρων που παρέχονται μέσω της εφαρμογής αυτής για να παραχθούν αρχεία με συγκεκριμένο αριθμό σημειακών νουκλεοτιδικών πολυμορφισμών, όπως επίσης και για την δημιουργία ξεχωριστών αρχείων που περιέχουν είτε μόνο άτομα που φέρουν την ασθένεια (cases) είτε μόνο υγιείς άτομα (controls).

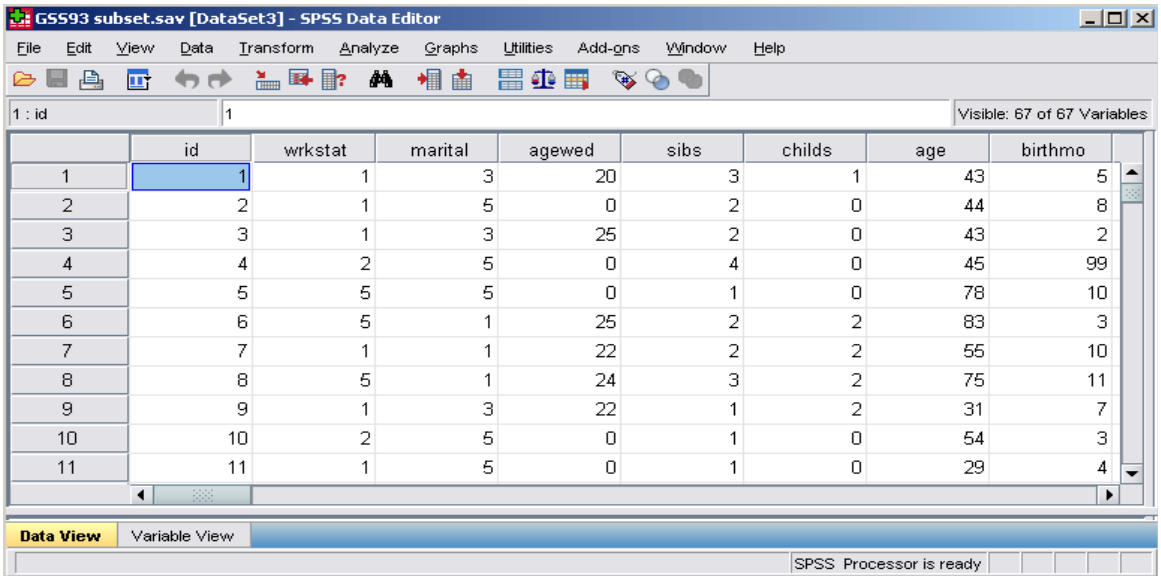
3.2.2 Εφαρμογή SPSS

Αποτελεί μια ομάδα εφαρμογών που προσφέρουν στατιστικές αναλύσεις. Προσφέρει στο τελικό χρήστη ένα πολύ φιλικό – προς – το – χρήστη περιβάλλον, καθώς και πάρα πολλές δυνατότητες στον χρήστη. Μπορεί να διαβάσει δεδομένα από σχεδόν όλων των τύπων τα αρχεία, και να τα επεξεργαστεί παράγοντας αναφορές, γραφικές παραστάσεις, και πάρα πολλές άλλες στατιστικές αναλύσεις.

Στο στάδιο της προεπεξεργασίας, έγινε χρήση των διαφόρων φίλτρων που παρέχονται μέσω της εφαρμογής αυτής για να παραχθούν αρχεία με συγκεκριμένο αριθμό σημειακών νουκλεοτιδικών πολυμορφισμών, όπως επίσης και για την απόρριψη ατόμων που δεν ικανοποιούσαν συγκεκριμένες βιολογικές ανάγκες.

Επιπρόσθετα, το SPSS χρησιμοποιήθηκε και κατά το στάδιο της μεταεπεξεργασίας για την παραγωγή διάφορων γραφικών παραστάσεων για απεικόνιση των αποτελεσμάτων

από τις εφαρμογές που έχουν αναπτυχθεί στα πλαίσια της μελέτης αυτής. Ο λόγος χρησιμοποίησης του SPSS αντί οποιασδήποτε άλλης εφαρμογής που δημιουργεί γραφικές παραστάσεις – εικόνες ήταν ότι είναι μια ιδιαίτερα φιλική προς τον χρήστη εφαρμογή. Επιπλέον, αφού υπήρχε ήδη μια εξοικείωση με τον περιβάλλον της από την προηγούμενη χρήση της, επιλέγηκε για την δημιουργία των απεικονίσεων μας. [16]



Εικόνα 3.3 Εισαγωγή Δεδομένων στο SPSS από 'Statistical Computing Group, University of Pennsylvania' (http://www.ssc.upenn.edu/scg/spss/verybasicSPSS.pdf)

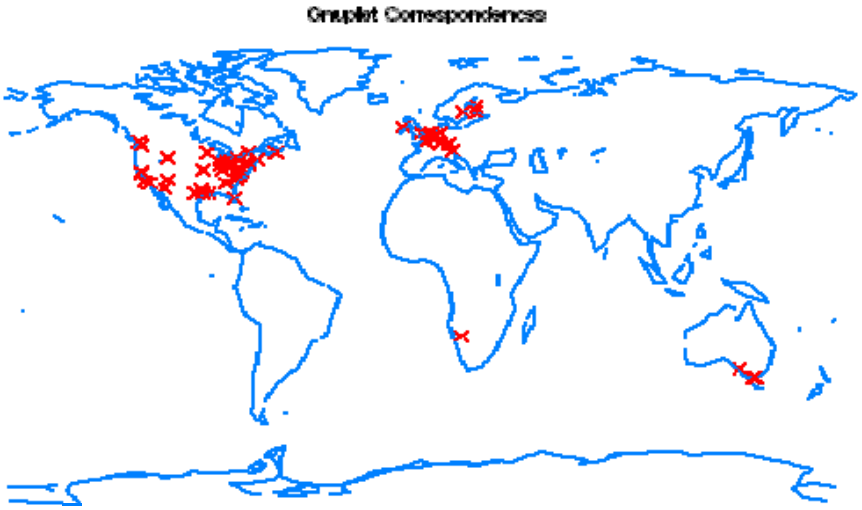
3.2.3 Εφαρμογή Gnuplot

Η εφαρμογή Gnuplot είναι εφαρμογή ανοικτού κώδικα η οποία μπορεί να παράξει δύο και τριών διαστάσεων γραφικές παραστάσεις. Είναι command – line πρόγραμμα δηλαδή δεν παρέχει γραφικό περιβάλλον – αντίθετα τα πάντα γίνεται με εντολές σε διάφορα τερματικά. Είναι φορητή (portable) εφαρμογή – τρέχει σε Unix – Linux, MS Windows, MAC OS και σε αρκετές άλλες πλατφόρμες. Έχει αναπτυχθεί στην γλώσσα C. Αν και είναι εφαρμογή ανοικτού κώδικα, συνεπώς ο πηγαίος κώδικας (source code) παρέχεται δωρεάν, εντούτοις δεν επιτρέπεται στους χρήστες – σε περίπτωση που τροποποιήσουν τον κώδικα – να τον διανέμουν ελευθέρα. [20]

Σήμερα το Gnuplot, απευθύνεται κυρίως σε επιστήμονες που απαιτούν εξειδικευμένες γραφικές παραστάσεις, αν και το 1986 όταν πρωτοεμφανίστηκε στον κόσμο, στόχευε να βοηθήσει κυρίως μαθητές και φοιτητές για απεικόνιση μαθηματικών συναρτήσεων

και δεδομένων. Επιπλέον, λόγω της ιδιαίτερα αυξημένης χρήσης και αναγνώρισης που απολαμβάνει έχουν αναπτυχθεί διάφορες βιβλιοθήκες σε πολλές γλώσσες προγραμματισμού που επιτρέπουν την χρήση του Gnuplot μέσα από άλλα προγράμματα. [20]

Στην μελέτη αυτή, χρησιμοποιήθηκε το Gnuplot, για την δημιουργία γραφικών παραστάσεων, ιδιαίτερα παραστάσεων που αφορούσαν τεράστιο όγκο δεδομένων.



Εικόνα 3.4 Παράδειγμα Εκτέλεσης στο Gnuplot
(http://www.gnuplot.info/demo/world.html)

3.3 Προεπεξεργασία Δεδομένων

Ένα από τα πιο σημαντικά στάδια για την εκπόνηση της μελέτης αυτής ήταν το στάδιο της προεπεξεργασίας των δεδομένων. Η μεγαλύτερη πρόκληση στο στάδιο αυτό ήταν ο περιορισμός του μεγέθους των δεδομένων, καθώς υπήρχαν πάρα πολλά γενετικά δεδομένα που δεν ήταν απαραίτητα. Τα δεδομένα αυτά πέρασαν από διάφορα φίλτρα, μέσω των οποίο αφαιρέθηκαν τα αχρείαστα δεδομένα. Όμως ακόμα και μετά το φιλτράρισμα των δεδομένων, ο όγκος των σημαντικών δεδομένων εξακολουθούσε να είναι τεράστιος. Έτσι, αποφασίστηκε η ανάπτυξη ενός αλγορίθμου συμπίεσης εξειδικευμένο για βιολογικά δεδομένα που θα μείωνε το μέγεθος των αρχείων στο μέγιστο δυνατό βαθμό, ο οποίος προσφέρεται σαν μοντέλο ανοικτού κώδικα.

3.3.1 Φιλτράρισμα Δεδομένων

Τα «ωμά δεδομένα» έπρεπε πρώτα να φιλτραριστούν έτσι ώστε να εξαχθούν τα δεδομένα τα οποία δεν ήταν χρήσιμα για τους σκοπούς της εφαρμογής. Έγιναν συγκεκριμένα 6 φιλτραρίσματα.

Συγκεκριμένα, πρώτα από όλα, δημιουργήθηκαν 3 διαφορετικά αρχεία δεδομένων που περιείχαν μόνο ασθενείς(cases), μόνο υγιείς (controls), και συνδυασμό τους. Οι διαχωρισμοί αυτοί έγιναν έτσι ώστε να εξαχθούν διάφορα συμπεράσματα όσο αφορά συσχετίσεις με τις ασθένειες καθώς μπορούν να ανακαλυφθούν χαρακτηριστικά που μπορεί να οφείλονται για την ασθένεια και παρουσιάζονται μόνο στους ασθενείς, ή χαρακτηριστικά στα οποία οφείλεται η καταστολή της ασθένειας και παρουσιάζονται μόνο στους υγιείς, ή χαρακτηριστικά τα οποία δεν ευθύνονται τα οποία είναι επίσης σημαντικά, για τον αποκλεισμό διάφορων δεδομένων.

Οι 3 υπόλοιποι διαχωρισμοί που έγιναν αφορούσαν και πάλι την ίδια κατηγοριοποίηση (cases, controls, cases and controls), όμως πρώτα από όλα επιλέγηκαν τα άτομα για τα οποία τα γενετικά τους σημεία παρουσιάζουν επίσταση μεγαλύτερη της τιμής 45. Ο λόγος επιλογής των ατόμων αυτών, οφείλεται στο γεγονός ότι τα γενετικά τους σημεία παρουσιάζουν, με βάση την στατιστική, ιδιαίτερη σημασία καθώς μπορεί να οφείλονται για την παρουσία ή όχι μιας συγκεκριμένης ασθένειας, καθώς κάποια SNPs επικαλύπτουν ή επηρεάζουν την δράση κάποιων άλλων σε μεγάλο βαθμό.

3.3.2 Δυαδικό Πρότυπο/Format για γενετικά δεδομένα – Συμπίεση

Όπως έχει προαναφερθεί, ακόμα και μετά την εφαρμογή φίλτρων στα δεδομένα, ο αποθηκευτικός χώρος που απαιτείτο για την διατήρηση των γενετικών δεδομένων, ήταν ιδιαίτερα μεγάλος σε σημείο που αποφασίστηκε η ανάπτυξη ενός αλγορίθμου ανοικτού κώδικα που θα ειδικεύεται στην συμπίεση των δεδομένων. Πιο συγκεκριμένα, χρησιμοποιήθηκαν οι δυνατότητες binary αποθήκευσης που προσφέρει η γλώσσα προγραμματισμού C++. Ο κώδικας σήμερα μπορεί να χρησιμοποιηθεί από όλους και έχει ήδη γραφτεί ένα conference paper που περιγράφει το πρωτόκολλο που δημιουργήθηκε για το σκοπό αυτό και δημοσιεύτηκε τον Οκτώβριο του 2008. [17]

Είναι σημαντικό να αναφερθεί ότι η δυνατότητα αυτή προσφέρεται μέσω της εφαρμογής PLINK [8] όμως η κωδικοποίηση που εφαρμόζει το PLINK αν και επιτυγχάνει την μείωση του μεγέθους των αρχείων στο μισό από ότι επιτυγχάνει ο προτεινόμενος αλγόριθμος, εντούτοις υπάρχουν κάποια σημαντικά στοιχεία τα οποία δεν μπορεί να διαχωρίσει και έτσι έχουμε απώλεια των αρχικών δεδομένων, σε αντίθεση με τον προτεινόμενο αλγόριθμο που μπορεί να αναγνωρίσει οποιοδήποτε γενετικό σημείο. [17]

3.3.2.1 Περιγραφή Αλγορίθμου

Η κωδικοποίηση στηρίχτηκε στον περιορισμένο αριθμό καταστάσεων στον οποίο μπορεί να βρίσκονται τα alleles. Πιο συγκεκριμένα, κάθε allele (ετερόζυγου σημειακού πολυμορφισμού) μπορεί να βρίσκεται σε μία από τρεις πιθανές καταστάσεις, που περιλαμβάνουν τα δύο πιθανά νουκλεοτίδια που μπορεί να αποτελείται καθώς και την κατάσταση να μην υπάρχουν δεδομένα για εκείνο το σημείο (ορίζεται ως missing data). Συνεπώς για την κωδικοποίηση των τριών αυτών καταστάσεων απαιτούνται 2 bits. Στο προτεινόμενο πρωτόκολλο προτείνεται όπως χρησιμοποιηθεί και η τέταρτη περίπτωση που μας δίνουν τα 2 bits έτσι ώστε να διαχωρίζονται οι διαγραφόμενοι (deleted) δείκτες (markers) από τα missing data, στοιχείο το οποίο είναι ιδιαίτερα σημαντικό σε κάποιες ειδικές αναλύσεις όπως οι Copy Number Variants (CNV) αναλύσεις. Επιπρόσθετα υπάρχει και το επιπλέον πλεονέκτημα ότι η κωδικοποίηση ετερόζυγων alleles διαχωρίζεται, κάτι που δεν ξεχωρίζει σε άλλους αλγόριθμους, και είναι σημαντική για αναλύσεις που στηρίζονται στην ακολουθία (sequence) των alleles [17] Η τελική κωδικοποίηση παρουσιάζεται στους πίνακες 3.1 και 3.2.

.

Με βάση τα πιο πάνω, υλοποιήθηκε ο αλγόριθμος συμπίεσης, χρησιμοποιώντας τις δυνατότητες της γλώσσας προγραμματισμού C++ και πιο συγκεκριμένα της Standard Template Library που παρέχει διάφορες κλάσεις για την ανάπτυξη εφαρμογών. [18] Συγκεκριμένα χρησιμοποιήθηκε η κλάση bitset που απεικονίζει πίνακες από bits παρέχοντας την δυνατότητα στο προγραμματιστή να αποθηκεύει bits αν και σε όλες τις γλώσσες προγραμματισμού η ελάχιστη μονάδα αποθήκευσης που παρέχεται στον προγραμματιστή είναι το byte.

Ουσιαστικά, κάθε 8 bits, αποθηκεύονταν σε ένα αρχείο στην μορφή χαρακτήρα καθώς ένα 1 byte αντιστοιχεί σε 8 bits και το byte, όπως έχει προαναφερθεί αποτελεί την ελάχιστη μονάδα αποθήκευσης. Έτσι γινόταν μια μετατροπή από bits σε bytes σαν χαρακτήρας, και το αντίθετο.

Επιπλέον πρέπει να αναφερθεί ότι ο συγκεκριμένος αλγόριθμος λόγω της αρκετής εργασίας που απαιτείτο, έπρεπε να μοιραστεί σε δύο εργασίες. Αφού αποφασίστηκε η γενική ιδέα για το πού θα στηριζόμαστε με την συνεργασία όλης της ομάδας, ο κύριος Άριστος Αριστοδήμου ανέλαβε την αποκρυπτογράφηση και διάβασμα των κωδικοποιημένων δεδομένων και ο εγώ ,ο υποφαινόμενος, ανέλαβα την κωδικοποίηση των δεδομένων.

PROPOSED ALLELE ENCODING

Allele	Encoding
Unknown	00
A	01
a	10
Deleted	11

Πίνακας 3.1 Η προτεινόμενη κωδικοποίηση παρθέν από το [17]

PROPOSED ENCODING FOR BI-ALLELIC MARKERS

Allele 1	Allele 2	Encoding
Unknown	Unknown	0000
Unknown	A	0001
Unknown	a	0010
Unknown	Deleted	0011
A	Unknown	0100
A	A	0101
A	a	0110
A	Deleted	0111
a	Unknown	1000
a	A	1001
a	a	1010
a	Deleted	1011
Deleted	Unknown	1100
Deleted	A	1101
Deleted	a	1110
Deleted	Deleted	1111

Πίνακας 3.2 Η προτεινόμενη κωδικοποίηση για οργανισμούς με δύο alleles παρθέν από το [17]

3.4 Περιγραφή Αλγορίθμων Επεξεργασίας

3.4.1 Categorical Δεδομένα για Χάρτη αυτό – οργάνωσης

Ένα από τα πιο σημαντικά τμήματα της μελέτης αυτής ήταν η αναζήτηση τρόπων χρήσης categorical δεδομένων σε χάρτες αυτό – οργάνωσης. Ως categorical δεδομένα, ορίζονται δεδομένα που αποτελούν ένα πεπερασμένο σύνολο δεδομένων (συνήθως μικρό μέγεθος δεδομένων), το οποίο το κάθε δεδομένο ανταποκρίνεται σε μια συγκεκριμένη κατηγορία ή όνομα. Δηλαδή δεν είναι δεδομένα που αναπαριστούνται με αριθμητικές τιμές. Για παράδειγμα, σε μια βάση δεδομένων που περιέχει μαθητές, το χαρακτηριστικό «τμήμα» είναι categorical δεδομένο καθώς περιέχει συγκεκριμένα ονόματα πανεπιστημίων. Ένας απλός ορισμός τέτοιων δεδομένων είναι να μπορεί να δοθεί μια 1 προς 1 αντιστοιχία μεταξύ των δεδομένων αυτών και ενός πεπερασμένου υποσυνόλου των φυσικών αριθμών. [24,25,26]

Οι χάρτες αυτοί όπως έχουν προταθεί [15], μπορούσε να χρησιμοποιηθεί μόνο για αριθμητικά δεδομένα. Όποτε, έπρεπε να βρεθεί τρόπος για λύση του προβλήματος αυτού. Μέσα την βιβλιογραφική επισκόπηση και γενικότερη μελέτη, χρησιμοποιήθηκε η δυαδική προσέγγιση – αναπαράσταση των δεδομένων

Έχουν προταθεί πολλοί άλλοι τρόποι για λύση του προβλήματος των categorical δεδομένων σε χάρτες αυτό – οργάνωσης, όπως για παράδειγμα χρήση πινάκων γράφων που απεικονίζουν την σχέση μεταξύ των δεδομένων κτλ. [26]. Ειδικότερα αυτή η προσέγγιση δεν μπορούσε στην προκειμένη περίπτωση να εφαρμοστεί, αν και παρουσιάζει θετικά αποτελέσματα, για δύο λόγους. Πρώτο απαιτεί σημαντική ποσότητα μνήμης για δημιουργία των γράφων και δεύτερο και το πιο σημαντικό είναι ότι με τα δεδομένα τα οποία υπήρχαν, δεν υπήρχε τρόπος να «βαθμολογηθεί» η σχέση μεταξύ των δεδομένων δηλαδή να δοθούν βάρη στους πολυμορφισμούς. Έτσι ακολουθήθηκε η προσέγγιση που ακολουθείται πιο συχνά για τέτοιες περιπτώσεις όπως αναφέρεται στην βιβλιογραφία [24,25,26] που είναι η απεικόνιση των δεδομένων σε δυαδική μορφή.

Η προτεινόμενη κωδικοποίηση - μετατροπή παρουσιάζεται στον πίνακα 3.3. Ακολουθήθηκε μια allelic προσέγγιση. Δηλαδή ανάλογα με τα ποια είναι τα alleles σε ένα σημειακό πολυμορφισμό καθορίζεται η ανάλογη αναπαράσταση, αφού το κάθε allele σε ένα SNP μπορεί να βρίσκεται σε 3 διαφορετικές καταστάσεις, είτε missing είτε ένα από τα δύο νουκλεοτίδια που ένα SNP μπορεί να έχει. Ο αλγόριθμος «βλέπει» το κάθε allele ξεχωριστά. Αγνοεί τα missing δεδομένα τα οποία εξηγούνται πιο κάτω το πώς τα διαχειρίζεται ο αλγόριθμος. Συνεπώς μένουν μόνο τα δύο πιθανά νουκλεοτίδια. Συνεπώς η παρουσία του ενός νουκλεοτίδιου στο allele αναπαριστάται από την τιμή 1 και η παρουσία του άλλου νουκλεοτιδίου αναπαριστάται από την τιμή 0. Για παράδειγμα αν τα νουκλεοτίδια ενός SNP είναι AT και ένα bi-allelic SNP έχει την τιμή AA τότε η δυαδική αναπαράσταση είναι 1 0.

Ο αλγόριθμος αναγνωρίζει πρώτα από όλα το ζεύγος των πιθανών νουκλεοτιδίων για το κάθε SNP, μέσα από τα δεδομένα. Ανάλογα με το ζεύγος που προκύπτει γίνεται η κωδικοποίηση όπως παρουσιάζεται στον πίνακα 3.3. Ένα παράδειγμα εκτέλεσης θα ήταν αναγνώριση των νουκλεοτιδίων T G. Τότε για ένα άτομο που έχει G G για το συγκεκριμένο SNP, η κωδικοποίηση που θα γίνει θα είναι 0 1 και για κάποιο άλλο άτομο που έχει G T ως τιμές των alleles, η κωδικοποίηση που θα γίνει θα είναι 0 0.

Με αυτό τον τρόπο, γίνεται η μετατροπή των categorical δεδομένων σε δυαδική μορφή και δίνονται ως είσοδος στο δίκτυο.

Όσο αφορά τον σειριακό αλγόριθμο με γραμμικά δεδομένα (linear data), εργασία για την οποία ήταν υπεύθυνος ο κύριος Άριστος Αριστόδημου, λόγω της μελέτης που έγινε πιο πάνω, αποφασίστηκε να ακολουθηθεί η ίδια προσέγγιση δηλαδή μετατροπή των categorical δεδομένων σε άλλη μορφή. Η διαφορά έγκειται στο ότι μετατρέπονται τα δεδομένα σε γραμμικά, αντί να χρησιμοποιείται δυαδική απεικόνιση. Για κάθε SNP, αναπαριστώνται 3 καταστάσεις ανάλογα με τις τιμές των alleles τους. AA μετατρέπεται σε 1, Aa ή aA σε 2 και aa σε 3. Η βασική διαφορά μεταξύ των δύο αναπαραστάσεων αφορά τον διαχωρισμό μεταξύ aA και Aa, αφού στην υλοποίηση με τα γραμμικά δεδομένα δεν διαχωρίζεται – θεωρείται σαν ίδια περίπτωση – ενώ στην δυαδική αναπαράσταση το δίκτυο μπορεί να τα αναγνωρίσει έως ξεχωριστές παραστάσεις.

Allele	Encoding
AA	01
aA	11
Aa	00
aa	10

Based on that the alleles found by the algorithm are aA

Πίνακας 3.3 Η προτεινόμενη κωδικοποίηση που χρησιμοποιήθηκε στα πλαίσια της μελέτης αυτής

3.4.2 Χειρισμός Missing Δεδομένων σε Γενετικά Δεδομένα

Σε αναλύσεις γενετικών δεδομένων, μια από τις πιο σημαντικές αποφάσεις που αναλαμβάνει κάποιος αναλυτής, αφορά το πώς θα διαχειριστεί τα missing δεδομένα που παρουσιάζονται σε διάφορες βάσεις γενετικών δεδομένων. Τα δεδομένα αυτά προκύπτουν είτε μέσα από σφάλματα ακρίβειας των μηχανών αναγνώρισης του DNA δηλαδή κατά την φάση του genotyping - της ανάγνωσης του γονότυπου -, είτε γιατί ένα συγκεκριμένο άτομο δεν έχει ένα σημειακό νουκλεοτιδικό πολυμορφισμό, καθώς μπορεί να βρίσκεται σε μια διαγραμμένη περιοχή. Τα δεδομένα αυτά, αν και δεν εμφανίζονται σε μεγάλο βαθμό, υπάρχουν περιπτώσεις κατά τις οποίες μπορεί να προκαλέσουν πολλά προβλήματα και συνεπώς λανθασμένα αποτελέσματα σε διάφορες αναλύσεις.

Στον αλγόριθμο που παρουσιάζεται στα πλαίσια της μελέτης αυτής, η παρουσία των missing δεδομένων, έπρεπε να ληφθεί υπόψη καθώς θα επηρέαζε τα αποτελέσματα σε μεγάλο βαθμό τα αποτελέσματα. Θεωρητικά, όσο αυξάνεται ο αριθμός των missing δεδομένων, τόσο θα αυξάνεται και ο βαθμός που επηρεάζονται τα αποτελέσματα δηλαδή το πόσο προκατειλημμένα (biased) θα είναι όσο προς τα missing δεδομένα. [27,28] Για το λόγο αυτό, έχει αποφασιστεί μια διαφοροποίηση στο αλγόριθμο.

Αυτό που γίνεται ουσιαστικά, είναι μια προσπάθεια να μην λαμβάνονται υπόψη σε κανένα υπολογισμό σημειακοί νουκλεοτιδικοί πολυμορφισμοί που έχουν missing

δεδομένα. Συγκεκριμένα, όταν ο αλγόριθμος αναγνωρίσει missing δεδομένα σε ένα SNP, τότε δεν λαμβάνεται υπόψη κατά τον υπολογισμό των αποστάσεων για την εύρεση του BMU. Έτσι, η επιλογή δεν επηρεάζεται καθόλου από τα δεδομένα αυτά. Επιπρόσθετα, στους κόμβους στους οποίους με βάση τον BMU, πρέπει να τροποποιήσουν τα βάρη τους, βάρη τα οποία αντιστοιχούν σε SNP για τα οποία για το συγκεκριμένο άτομο, παρουσιάζουν missing δεδομένα, δεν τροποποιούνται τα βάρη τους. Για την αναγνώριση, αν είναι missing δεδομένα, υπάρχει μια δομή, η οποία κατά την διαδικασία διαβάσματος των δεδομένων, με βάση το αν υπάρχουν 0 0 ένδειξη στα δεδομένα εισόδου, αποθηκεύει την πληροφορία αυτή. Έτσι, μπορεί το σύστημα ανά πάσα στιγμή να αναγνωρίσει σε ποια SNP υπάρχουν missing δεδομένα και να τα αγνοήσει.

3.4.3 Σειριακός Αλγόριθμος χάρτη αυτό-οργάνωσης

Στόχος της μελέτης αυτής είναι η ανάπτυξη ενός παράλληλου αλγορίθμου αυτό – οργάνωσης και χρήση του για εξαγωγή συμπερασμάτων με την χρήση βιολογικών δεδομένων όπως έχει προαναφερθεί. Μέσω της βιβλιογραφικής επισκόπησης που έγινε ανακαλύφθηκε και το μειονέκτημα του σειριακού αλγορίθμου, που είναι ο μεγάλος χρόνος εκπαίδευσης. Όμως, επειδή ο χρόνος αυτός εξαρτάται από τον αριθμό των δεδομένων εισόδου, αποφασίστηκε αρχικά η υλοποίηση του σειριακού έτσι ώστε να προκύψουν διάφορα συμπεράσματα κατά πόσο θα ήταν επαρκής η υλοποίηση αυτή για τους σκοπούς της μελέτης αυτής.

Συγκεκριμένα αναπτύχθηκε ο αλγόριθμός όπως ακριβώς περιγράφεται στο υποκεφάλαιο 2.2.4. Αρχικά ο αλγόριθμος δημιουργεί ένα δισδιάστατο πλέγμα που αποτελείται από νευρώνες και ανάλογα με το μέγεθος του διανύσματος εισόδου, σε κάθε νευρώνα δημιουργούνται τα διανύσματα των βαρών (έχουν το ίδιο μέγεθος με τα διανύσματα εισόδου) και αρχικοποιούνται με τιμές από 0 – 1. Στη συνέχεια ορίζονται ρυθμός μάθησης και η ακτίνα του δικτύου από τον χρήστη. Ακολούθως, αρχίζει η διαδικασία μάθησης. Παρουσιάζεται ένα διάνυσμα εισόδου, ένα κάθε φορά ωσότου περάσουν όλα τα διανύσματα κάτι που σημαίνει την ολοκλήρωσης μιας εποχής. Για κάθε διάνυσμα εισόδου, όλοι οι νευρώνες υπολογίζουν την ευκλείδεια απόσταση μεταξύ των βαρών τους και του διανύσματος εισόδου, και επιλέγεται ο νευρώνας με

την μικρότερη απόσταση. Ακολούθως, όλοι οι νευρώνες υπολογίζουν ξανά την ευκλείδεια απόσταση μεταξύ αυτή την φορά, των αποστάσεων τους στο πλέγμα. Οι νευρώνες για τους οποίους η ακτίνα του δικτύου, είναι μικρότερη από την τιμή που έχουν υπολογίσει, προσαρμόζουν τα βάρη τους, με τον τρόπο που έχει εξηγηθεί στο 2.2.4. Όταν ολοκληρωθεί μια εποχή, μειώνεται η ακτίνα και ο ρυθμός μάθησης. Η διαδικασία ολοκληρώνεται όταν ολοκληρωθεί ο αριθμός των εποχών που δίνεται από τον χρήστη.

Τα αποτελέσματα επιβεβαίωσαν τον μεγάλο χρόνο εκτέλεσης που απαιτείτο για την ολοκλήρωση - εκπαίδευση του αλγορίθμου, που οδήγησε στην ανάπτυξη του παράλληλου αλγόριθμου.

Επιπρόσθετα, πρέπει να σημειωθεί ότι ο αλγόριθμος αυτός, χρησιμοποιήθηκε κυρίως για categorical δεδομένα με τον τρόπο που επεξηγήθηκε πιο πάνω, όπως και για γραμμικά δεδομένα (linear data), εργασία για την οποία ήταν υπεύθυνος ο Άριστος Αριστοδήμου.

3.4.4 Παράλληλος Αλγόριθμος χάρτη αυτό-οργάνωσης

3.4.4.1 Εισαγωγή – Ανάπτυξη Παράλληλου Αλγόριθμου χάρτη αυτό-οργάνωσης

Στα πλαίσια της επεξεργασίας των δεδομένων έχει αναπτυχθεί ένας παράλληλος αλγόριθμος μάθησης που στηρίζεται στο σειριακό αλγόριθμο που προτάθηκε από τον Teuvo Kohonen δηλαδή στον αλγόριθμο Kohonen που στηρίζεται σε ένα δισδιάστατο πλέγμα. Ο αλγόριθμος αυτός έχει υλοποιηθεί με βασικό στόχο να μειωθεί ο χρόνος που χρειάζεται για την εκπαίδευση ενός χάρτη αυτό – οργάνωσης Kohonen. Ο κλασσικός αλγόριθμος όπως έχει προταθεί από τον Teuvo Kohonen [15] αν και οι δυνατότητες που προσφέρει είναι τεράστιες για την κατηγοριοποίηση δεδομένων, εντούτοις ο χρόνος που απαιτείται για την εκμάθηση του δικτύου αποτελεί μειονέκτημα. Συνδυάζοντας τις δυνατότητες που προσφέρει η παράλληλη επεξεργασία, ο χρόνος αυτός μπορεί να μειωθεί , δημιουργώντας ένα γρήγορο και ακριβή αλγόριθμο κατηγοριοποίησης.

Ακολούθως, αφού αναπτύχθηκε ο αλγόριθμος με τυχαία δεδομένα, χρησιμοποιήθηκε για την εισαγωγή βιολογικών δεδομένων. Η ιδέα πίσω από την χρήση του χαρτών αυτό-οργάνωσης Kohonen για τα συγκεκριμένα δεδομένα είναι η εξής: έχοντας δεδομένα από διάφορα άτομα, όπως έχει ήδη προαναφερθεί στην εισαγωγή, δηλαδή γενετικά σημεία της περιοχής HLA για την ασθένεια της κατά πλάκα σκλήρυνσης, και γνωρίζοντας ήδη από προηγούμενες βιολογικές έρευνες ότι προκύπτουν πολλαπλές αλληλεπιδράσεις συσχετιζόμενες με την κατά πλάκα σκλήρυνση μεταξύ των γενετικών σημείων στην περιοχή αυτή, θα γίνει μια προσπάθεια κατηγοριοποίησης των γενετικών σημείων αυτών, με την ελπίδα ότι θα δημιουργηθούν διάφορες ομάδες οι οποίες θα μας διαχωρίσουν τους ασθενείς με τους υγιείς. Αν όντως ο αλγόριθμος κατηγοριοποίησης δημιουργήσει τον προαναφερθέν διαχωρισμό, και δεδομένου ότι ο οι χάρτες αυτό – οργάνωσης Kohonen κατηγοριοποιούν, αναγνωρίζοντας κοινά χαρακτηριστικά και στην προκειμένη περίπτωση, κοινά γενετικά σημεία θα γίνει δυνατή η εύρεση των κοινών αυτών γενετικών σημείων που πολύ πιθανόν να οφείλονται για την παρουσία της ασθένειας αυτής.

3.4.4.2 Προγραμματιστικό Περιβάλλον για την ανάπτυξη του Αλγόριθμου

Όπως έχει προαναφερθεί, μετά από συζήτηση, πάρθηκε η απόφαση ο αλγόριθμος να αναπτυχθεί σε ένα περιβάλλον που υπάρχει κάποια μορφή κοινής μνήμης. Μέσω μελέτης των διαφόρων διαθέσιμων μοντέλων για ανάπτυξη παράλληλων εφαρμογών που εκπονήθηκε στην αρχή της μελέτης αυτής με σκοπό την αξιολόγηση τους, αποφασίστηκε η χρήση των POSIX threads ή όπως είναι ευρέως γνωστά ως pthreads. [19]

Τα pthreads αποτελούν ένα πρότυπο που δίνει την δυνατότητα χρήσης threads. Ένα thread είναι η βασική μονάδα στην οποία το λειτουργικό σύστημα εκχωρεί χρόνο επεξεργασίας. Μια εφαρμογή αποτελεί από πολλά processes(διαδικασίες). Ένα process(διαδικασία) αποτελεί ουσιαστικά ένα εκτελέσιμο πρόγραμμα. Ένα thread μπορεί να εκτελέσει οποιοδήποτε μέρος του κώδικα μιας διαδικασίας και συνήθως ζουν στο χώρο μνήμης των διαδικασιών. Τα σημερινά λειτουργικά συστήματα υποστηρίζουν multi-threading δίνοντας την δυνατότητα στους προγραμματιστές να εκτελούν

προγράμματα που δημιουργούν πολλά threads τα οποία εκτελούνται σε διάφορους επεξεργαστές οδηγώντας στην δημιουργία παράλληλων εφαρμογών.

Ο αλγόριθμος έχει αναπτυχθεί στην αντικειμενοστραφή γλώσσα προγραμματισμού C++ χρησιμοποιώντας ιδιαίτερα την STL (Standard Template Library) που προσφέρει πάρα πολλές δυνατότητες στους προγραμματιστές όπως την δημιουργία δυναμικών πινάκων, διάφορων αλγορίθμων κτλ. [18]

3.4.4.3 Ανάγκες – Απαιτήσεις Αλγορίθμου

Ο αλγόριθμος που έχει αναπτυχθεί, έχει κάποιες απαιτήσεις, περιορισμούς. Η υλοποίηση του αλγορίθμου προϋποθέτει την ύπαρξη μιας κοινής μνήμης (shared memory) στην μηχανή στην οποία θα τρέξει η εφαρμογή. Αν και σίγουρα αποτελεί περιορισμό όσο αφορά τον αλγόριθμο, εντούτοις με βάση τις μηχανές που υπήρχαν διαθέσιμες, όταν παίρναμε την απόφαση για τον αλγόριθμο, καταλήξαμε στο συμπέρασμα ότι η μηχανή στην οποία θα τρέχαμε και θα είχαμε τα καλύτερα αποτελέσματα (όσο αφορά κυρίως την πρόσβαση σε αυτή), ήταν μηχανή με αρχιτεκτονική κοινής μνήμης, οπότε αποφασίστηκε η εκμετάλλευση της δυνατότητας αυτής για την παράγωγη των καλύτερων δυνατών αποτελεσμάτων.

3.4.4.4 Περιγραφή Αλγορίθμου – Λογικό Διάγραμμα Αλγορίθμου

Ο αλγόριθμος έχει αναπτυχθεί με βάση το μοντέλο Master Slave (SPMD), όπου κάθε thread εκτελεί τον ίδιο κώδικα αλλά σε διαφορετικά δεδομένα. Η γενική ιδέα είναι ότι το κάθε thread αναλαμβάνει ένα μέρος του πλέγματος του χάρτη, ακολούθως με συνεργασία βρίσκουν τον best matching νευρώνα και ανανεώνουν τα βάρη των νευρώνων που αναλαμβάνει το κάθε thread.

Αρχικά, λοιπόν, ανάλογα με τον αριθμό των threads και των διαστάσεων του πλέγματος του χάρτη τα οποία ορίζονται από τον χρήστη, γίνεται η ανάθεση των νευρώνων που θα διαχειρίζεται το κάθε thread. Ο υπολογισμός του αριθμού των νευρώνων που ανατίθεται σε κάθε thread γίνεται με βάση με τις διαστάσεις του πλέγματος. Η ανάθεση γίνεται πάντα με τέτοιο τρόπο έτσι ώστε να είναι μοιρασμένοι οι νευρώνες με μεγάλη

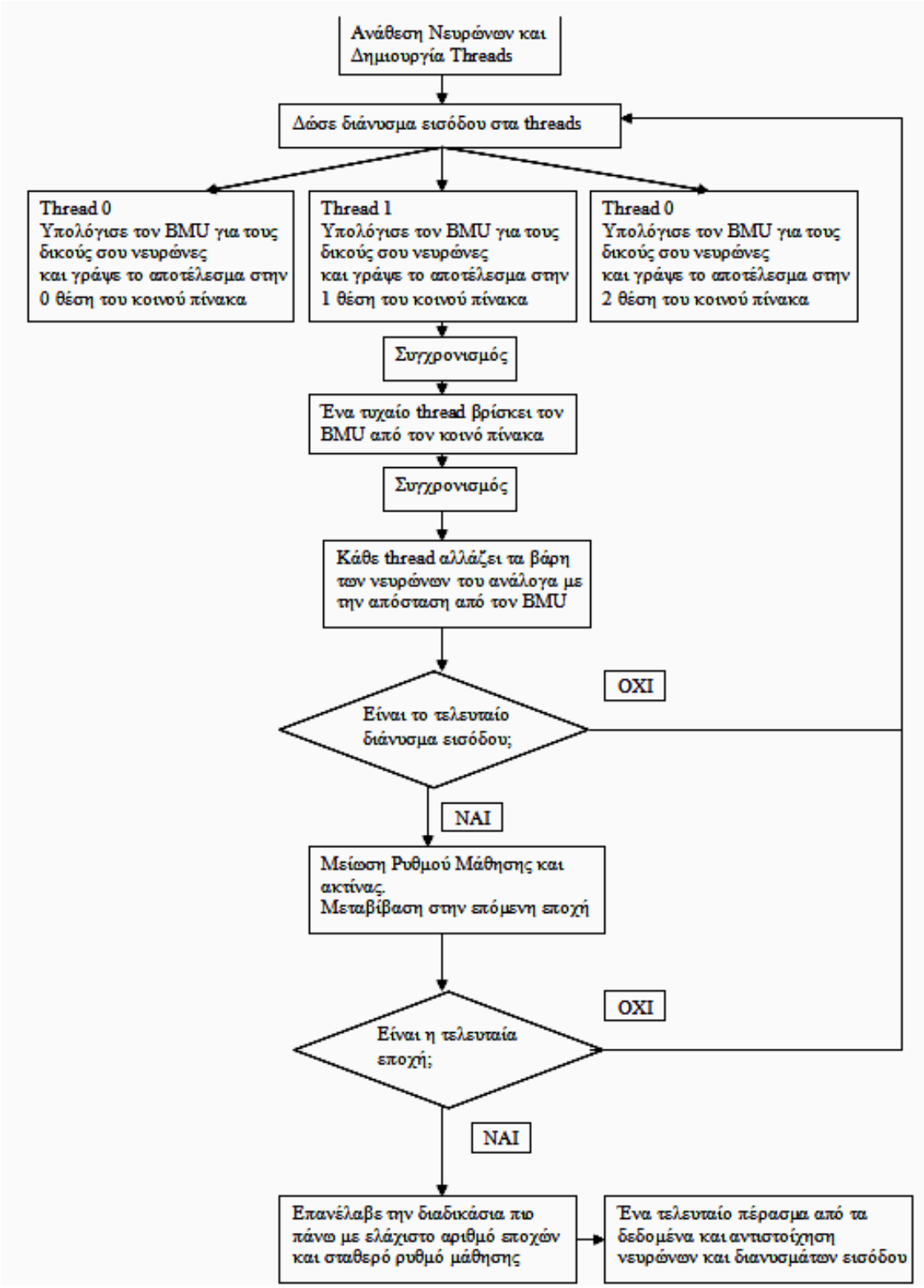
ακρίβεια έτσι ώστε να παρατηρείται ισορροπημένος φόρτος (load) σε κάθε thread. Το μειονέκτημα του αλγορίθμου εδώ είναι ότι η ανάθεση των νευρώνων γίνεται ουσιαστικά στατικά, οπότε δεν λαμβάνονται υπόψη πιθανά προβλήματα κακής χρονοδρομολόγησης από το λειτουργικό, πιθανής διαφορετικότητας των επεξεργαστών και γενικά προβλήματα που παρουσιάζονται κατά την εκτέλεση του αλγορίθμου.

Στη συνέχεια, σε κάθε thread δίνεται το ίδιο δεδομένο εισόδου, και αναλαμβάνει για τους νευρώνες τους οποίους διαχειρίζεται να υπολογίσει το νευρώνα με τα βάρη που έχουν την ελάχιστη απόσταση από το διάνυσμα εισόδου, γράφοντας το αποτέλεσμα σε ένα κοινό πίνακα (κάθε θέση του πίνακα αντιστοιχεί σε ένα thread). Ακολούθως ένα thread αναλαμβάνει να βρει με βάση τις τιμές του πίνακα τον νευρώνα με την ελάχιστη απόσταση. Μετά το βήμα αυτό είναι απαραίτητος ο συγχρονισμός καθώς όλα πρέπει να γράψουν την δική τους τιμή στο πίνακα ώστε να υπολογιστεί ο best matching νευρώνας.

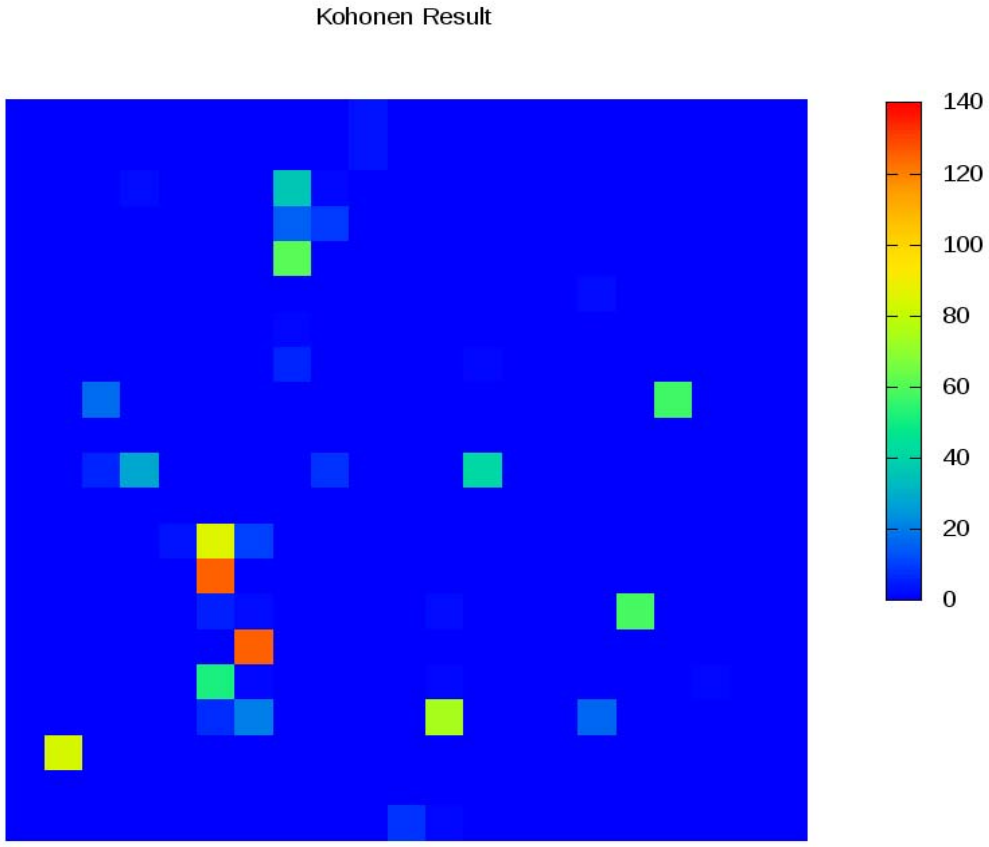
Με βάση το νευρώνα που αποτελεί τον best matching unit, το κάθε thread αναλαμβάνει να αλλάξει τα βάρη των νευρώνων των οποίων διαχειρίζεται ανάλογα με την απόσταση τους από τον best matching unit. Ακολούθως, όλα τα threads προχωρούν στο επόμενο διάνυσμα εισόδου, ωσότου ολοκληρωθεί η εποχή (εποχή στην υλοποίηση που έγινε ορίστηκε ένα πέρασμα από τα δεδομένα για να μην υπάρχει περίπτωση τα αποτελέσματα – κατηγοριοποίηση που θα γίνει να είναι επηρεασμένη από τις πόσες φορές παρουσιάστηκε στο νευρωνικό δίκτυο συγκεκριμένο διάνυσμα εισόδου). Όταν ολοκληρωθεί μια εποχή, έχουμε την μείωση του ρυθμού μάθησης όπως επίσης και της ακτίνας του δικτύου.

Αφού ολοκληρωθούν οι εποχές που έχει ορίσει ο χρήστης επαναλαμβάνεται ο αλγόριθμος για ένα ακόμα μικρό αριθμό εποχών με σταθερή ακτίνα 0.99 και ρυθμό μάθησης 0.01 έτσι ώστε ο κάθε νευρώνας να προσαρμοστεί ακόμα καλύτερα σε κάθε διάνυσμα εισόδου. Το πρώτο στάδιο στόχο έχει να καθοριστούν οι γειτονιές και το δεύτερο στάδιο να ξεκαθαρίσουν καλύτερα οι νευρώνες μέσα στις γειτονιές. Κάπου εδώ τελειώνουν και τα threads ολοκληρώνουν τις εργασίες τους.

Τέλος σαν τελευταίο στάδιο, έχουμε ένα τελευταίο πέρασμα από τα δεδομένα (ουσιαστικά ακόμα μια εποχή) όπου εδώ διατηρούνται πληροφορίες όσο αφορά το ποιος νευρώνας αντιστοιχεί στο κάθε διάνυσμα εισόδου (δηλαδή πόσες φορές ένας νευρώνας αντικατοπτρίζει ένα διάνυσμα εισόδου), όπως επίσης και εξειδικευμένες πληροφορίες όσο αφορά τα συγκεκριμένα βιολογικά δεδομένα που χρησιμοποιήθηκαν, όπως για παράδειγμα ποια διανύσματα εισόδου αντιστοιχούν σε ασθενείς και σε υγιείς, συγκεκριμένη κωδικοποίηση των δεδομένων κτλ.



Εικόνα 3.5 Λογικό Διάγραμμα Παράλληλου Αλγορίθμου



Εικόνα 3.6 Αποτέλεσμα Εκτέλεσης Αλγορίθμου

3.4.4.5 Επιλογή-Δικαιολόγηση Παραμέτρων Δικτύου Παράλληλου Αλγορίθμου

Πριν γίνει η παρουσίαση των αποτελεσμάτων του αλγορίθμου, πρέπει να δικαιολογηθούν οι διάφορες επιλογές που έγιναν όσο αφορά τις παραμέτρους του δικτύου. Η επιλογή των παραμέτρων παίζει πολύ καθοριστικό ρόλο στην απόδοση ενός δικτύου δηλαδή τον βαθμό κατηγοριοποίησης που επιτυχαίνει στην προκειμένη περίπτωση.

3.4.4.5.1 Ρυθμός Μάθησης

Ως ρυθμός μάθησης έχει επιλεγεί η τιμή 0.1. Όπως έχει προαναφερθεί, η τιμή αυτή καθορίζει το μέγεθος των αλλαγών που γίνονται στα βάρη. Συνεπώς μεγάλη τιμή αρχικά συνεπάγεται μεγάλες αλλαγές στα βάρη. Όμως πολύ μεγάλες τιμές συνεπάγεται ταυτόχρονα μεγαλύτερος χρόνος σύγκλισης του δικτύου. Οπότε έχει επιλεχθεί μια μέση επιλογή με βάση εμπειρικές δοκιμές. Μετά από διάφορες δοκιμές με πολλές διαφορετικές τιμές, έχει παρατηρηθεί ότι τα καλύτερα αποτελέσματα - κατηγοριοποίηση παρατηρείται όταν ο αρχικός ρυθμός μάθησης είναι 0.1.

Η πιο πάνω θέση επιβεβαιώνεται και από την βιβλιογραφία [15] πιο συγκεκριμένα από τον T. Kohonen στον οποίο ανήκει η ιδέα των χαρών αυτό-οργάνωσης. Όπως αναφέρει συγκεκριμένα, η επιλογή είναι καθαρά εμπειρική. Όσο αφορά την συνάρτηση μείωσης του ρυθμού μάθησης, αν και ο T. Kohonen προτείνει μια συνάρτηση για βέλτιστη μείωση, εντούτοις αναφέρει ότι τα αποτελέσματα δεν είναι σίγουρα και συνήθως παρατηρούνται ικανοποιητικά αποτελέσματα όταν χρησιμοποιηθεί μια συνάρτηση της μορφής $a(t) = \frac{A}{t + B}$ η οποία έχει εφαρμοστεί κατά την ανάπτυξη του αλγορίθμου όπως έχει προαναφερθεί στο κεφάλαιο 3.

3.4.4.5.2 Ακτίνα Δικτύου

Για την επιλογή της αρχικής τιμής της ακτίνας επιλέγεται η τιμή του μισού της διαμέτρου του δικτύου. Για παράδειγμα ένα 50x50 πλέγμα, η αρχική ακτίνα θα είναι 25. Η επιλογή έγινε με στόχο στα αρχικά βήματα του αλγορίθμου να έχουμε αρκετούς νευρώνες στις γειτονιές δημιουργώντας μεγάλες γειτονιές έτσι ώστε να ομαδοποιούνται παρόμοια στοιχεία σε γειτονικούς νευρώνες. Όσο πιο μικρή είναι η αρχική γειτονιά οι αλλαγές που θα γίνονται ακολούθως, θα είναι τοπικές σε σημείο που θα δημιουργηθούν κομμάτια εντελώς ανεξάρτητα που δεν είναι αντικατοπτρίζουν στην πλειοψηφία των περιπτώσεων τα κοινά χαρακτηριστικά αφού συνήθως υπάρχουν δεδομένα που ανήκουν σε περισσότερες από μία ομάδες. Αυτά τα δεδομένα θα κατηγοριοποιηθούν σε μία από τις ομάδες χωρίς να μπορεί να διακριθεί η σχέση τους με τις άλλες ομάδες. Όποτε αρχικά, ορίζεται μεγάλη γειτονιά έτσι ώστε οι αλλαγές να γίνουν σχεδόν σε όλο

το χάρτη και σιγά σιγά να βρεθούν τα δεδομένα με την μεγαλύτερη σχέση μειώνοντας σταθερά την γειτονιά. [15]

Για την επιλογή της συνάρτησης μείωσης της γειτονιάς, όπως προτείνεται στην βιβλιογραφία, η κάθε υλοποίηση μπορεί να χρησιμοποιεί διαφορετική συνάρτηση. Η πιο ευρέως διαδεδομένη συνάρτηση αποτελεί η Gaussian function όπως έχει προαναφερθεί στο κεφάλαιο 3, γι' αυτό έχει επιλεγεί και για την ανάπτυξη του αλγορίθμου .

3.4.4.5.3 Αριθμός Εποχών

Όσο αφορά τον αριθμό των εποχών, αν και συνήθως σε άλλες εφαρμογές που χρησιμοποιήθηκαν χάρτες αυτό-οργάνωσης, η επιλογή γίνεται με καθαρά εμπειρικό τρόπο, για πιο ακριβή αποτελέσματα και για επιβεβαίωση ότι γίνεται σωστή κατηγοριοποίηση – clustering και αλλαγή των βαρών, έχει χρησιμοποιηθεί η μετρική αλλαγής βαρών που παρουσιάζεται στο 3.6.2.

Τα αποτελέσματα που παρουσιάζονται στο επόμενο κεφάλαιο δικαιολογούν πλήρως την επιλογή των 1000 εποχών και θα εξηγηθούν στο κεφάλαιο της συζήτησης

3.4.4.5.4 Διαστάσεις Δικτύου

Η επιλογή των διαστάσεων του δικτύου, όπως και σε όλες τις εφαρμογές που χρησιμοποιούν χάρτες αυτό – οργάνωσης, γίνεται εμπειρικά. Δυστυχώς δεν υπάρχουν τρόποι να ανακαλύψει κάποιος τις βέλτιστες διαστάσεις χωρίς να δοκιμάσει πολλές εκτελέσεις. Είναι ένα από τα σοβαρότερα μειονεκτήματα του αλγορίθμου, καθώς χρειάζεται αρκετός χρόνος και πολλές δοκιμές για να καταλήξει κάποιος στην βέλτιστη και σωστότερη κατηγοριοποίηση.

Στα πλαίσια της συγκεκριμένης μελέτης έχουν δοκιμαστεί διαφόρων ειδών μεγέθους πλέγματα (η υλοποίηση έγινε με την υπόθεση ότι το πλέγμα είναι ορθογώνιο). 5x5, 10x10, 20x20, 50x50, 100x100. Μέσα από τις δοκιμές αυτές, παρατηρήθηκε ότι τα καλύτερα αποτελέσματα προήλθαν κυρίως από την κατηγοριοποίηση στο πλέγμα

50x50 και 20x20 καθώς ήταν εντελώς ξεκάθαρα τα αποτελέσματα. Ο λόγος επιλογής μεγάλων πλεγμάτων ήταν ότι αναζητήθηκε όσο το δυνατό μεγαλύτερη ανεξαρτητοποίηση των νευρώνων, έτσι ώστε να «ξεφύγει» ο αλγόριθμος από το επίπεδο των γειτονιών και να κατηγοριοποιήσει τα δεδομένα εισόδου σε επίπεδο νευρώνων, ώστε κάθε νευρώνας να γίνει αντιπρόσωπος συγκεκριμένων νευρώνων. Σκοπός ήταν να βρεθούν τα άτομα με τον μεγαλύτερο βαθμό ομοιότητας γι' αυτό προτιμήθηκε η προσέγγιση αυτή δηλαδή η επιλογή μεγάλων πλεγμάτων.

3.4.4.5.5 Διάνυσμα εισόδου

Τα διανύσματα εισόδου ορίζονται από τον χρήστη ανάλογα με την αριθμό των δεδομένων που υπάρχουν στα αρχεία εισόδου. Επειδή η εφαρμογή στηρίζεται σε βιολογικά δεδομένα και η κωδικοποίηση που έγινε και στηρίζεται στα alleles, είναι δυαδική (binary) τα δεδομένα εισόδου αποτελούνται από 0 και 1 ανάλογα με τα είδη των alleles. Απλά ο χρήστης πρέπει να ορίσει το μέγεθος των διανυσμάτων.

Γενικότερα, σε όλες τις δοκιμές, που έγιναν χρησιμοποιήθηκαν 3 διαφορετικά σύνολα δεδομένων (datasets). Το $1^{ο}$ αποτελείτο από μόνο ασθενείς, το $2^{ο}$ από μόνο μη ασθενείς και το τρίτο συνδυασμός του $1^{ου}$ και του $2^{ου}$. Το κοινό χαρακτηριστικό τους ήταν η τιμή της επίστασης που ήταν 45 (η τιμή της επίστασης επιλέγηκε ως 45 καθώς το αντίστοιχο p-value ήταν 10^{-9}, κάτι που σημαίνει στατιστικά σημαντική πληροφορία). Επιπρόσθετα έχει δοκιμαστεί και το δίκτυο στα «ωμά» δεδομένα (raw data) δηλαδή δεδομένα που δεν έτυχαν κάποιας επεξεργασίας ή δεν παρουσιάζουν κάποια κοινή ιδιότητα. Τα αποτελέσματα που θα παρουσιαστούν πιο κάτω θα αποτελούνται κυρίως από τα αποτελέσματα του $1^{ου}$ συνόλου δεδομένου και κατά δεύτερο λόγο από του $3^{ου}$, καθώς ήταν τα αποτελέσματα με την πιο σημαντική πληροφορία.

3.4.4.6 Βελτιστοποιήσεις Παράλληλου Αλγορίθμου

Μέσα στα πλαίσια ανάπτυξης του αλγορίθμου, έγιναν πολλές προσπάθειες για βελτιστοποίηση της διαδικασίας μάθησης. Αν και έγινε προσπάθεια μελέτης και βιβλιογραφικής επισκόπησης, δεν προέκυψε οποιαδήποτε βοήθεια. Έτσι, προσπάθησε η ομάδα να φτάσει σε μερικές βελτιστοποιήσεις με βάση καθαρά εισηγήσεις των μελών

της. Υπεύθυνος για το τμήμα αυτό ήταν ο κύριος Άριστος Αριστοδήμου, στον οποίο ανήκουν σε μεγάλο βαθμό οι ιδέες βελτιστοποιήσεων.

Ίσως η πιο σημαντική βελτιστοποίηση, αποτελεί ο προϋπολογισμός των αποστάσεων μεταξύ των κόμβων του δικτύου. Όπως έχει προαναφερθεί, όταν βρεθεί ο BMU, υπολογίζονται οι αποστάσεις του κάθε νευρώνα από τον BMU. Αυτές οι αποστάσεις δεν χρειάζεται να υπολογίζονται σε κάθε βήμα, κάτι που κοστίζει πάρα πολύ στην συνολική επίδοση του αλγορίθμου. Η πρώτη βελτιστοποίηση συνεπώς αφορούσε τον προϋπολογισμό των αποστάσεων μεταξύ των νευρώνων και αποθήκευση τους σε πίνακες, κάτι που βελτιώνει πάρα πολύ την επίδοση του συστήματος.

Η δεύτερη βελτιστοποίηση αφορά στον τρόπο που γίνονται οι δύο συγχρονισμοί. Αντί να γίνονται δύο ξεχωριστοί συγχρονισμοί, έχει αναπτυχθεί ένας μικρός κώδικας που συνενώνει τους δύο συγχρονισμούς και τον υπολογισμό του γενικού BMU, κάτι που οδηγεί σε σημαντική βελτιστοποίηση του σταδίου του συγχρονισμού καθώς αντί να εκτελείται δύο φορές ο ίδιος αλγόριθμος συγχρονισμού, εκτελείται μία και μεταξύ του κώδικα αυτού, υπολογίζεται και ο γενικός BMU.

3.4.4.7 Χρήση Κύριας Μνήμης στον Παράλληλο Αλγόριθμο

Ένας από τους βασικούς στόχους του παράλληλου αλγορίθμου ήταν η μείωση όσο το δυνατό της μνήμης που χρειάζεται για να εκτελεστεί. Μελετώντας τον αλγόριθμο των χαρτών αυτό – οργάνωσης, το πρώτο πράγμα που μπορεί να αντιληφθεί κάποιος είναι τις μεγάλες απαιτήσεις του σε μνήμη – μοντελοποίηση νευρώνων και των βαρών τους, αποθήκευση δεδομένων εισόδου, double precision πράξεις και αποθήκευση double αριθμών - . Για την σωστή και αποδοτική εκτέλεση του αλγόριθμου, δεν μπορούν να αποφευχθεί η έντονη χρήση της μνήμης.

Η σκέψη που έγινε εδώ, ήταν αφού μοντελοποιήθηκαν τα categorical δεδομένα με δυαδική αναπαράσταση, να γίνει εκμετάλλευση του γεγονότος αυτού, αφού το μόνο που απαιτείτο είναι η αναπαράσταση των δεδομένων binary. Γι’ αυτό τον σκοπό χρησιμοποιήθηκε η κλάση bitset [18] που είναι πάρα πολύ χρήσιμη για τέτοιες περιπτώσεις.

Ο αλγόριθμος θα συγκριθεί με τον αλγόριθμο που έχει αναπτύξει ο κύριος Άριστος Αριστοδήμου με τις βελτιστοποιήσεις για τις οποίες ήταν υπεύθυνος, οι οποίες απαιτούν αρκετή χρήση της κύριας μνήμης για εξαγωγή συμπερασμάτων όσο αφορά την αποδοτική χρήση της.

Σημαντική διαφορά, επιπλέον εκτός από την χρήση bits, αποτελεί και το διαχωρισμός του πλέγματος στα threads, καθώς η δυαδική υλοποίηση ορίζει τους κόμβους του πλέγματος να είναι κοινός (shared) σε όλα τα threads, ενώ στην γραμμική υλοποίηση το κάθε thread έχει στο private address space του τους κόμβους για τους οποίους είναι υπεύθυνο.

3.5 Μεταεπεξεργασία

3.5.1 Περιγραφή Διαδικασίας – Βημάτων στο στάδιο Μεταεπεξεργασίας

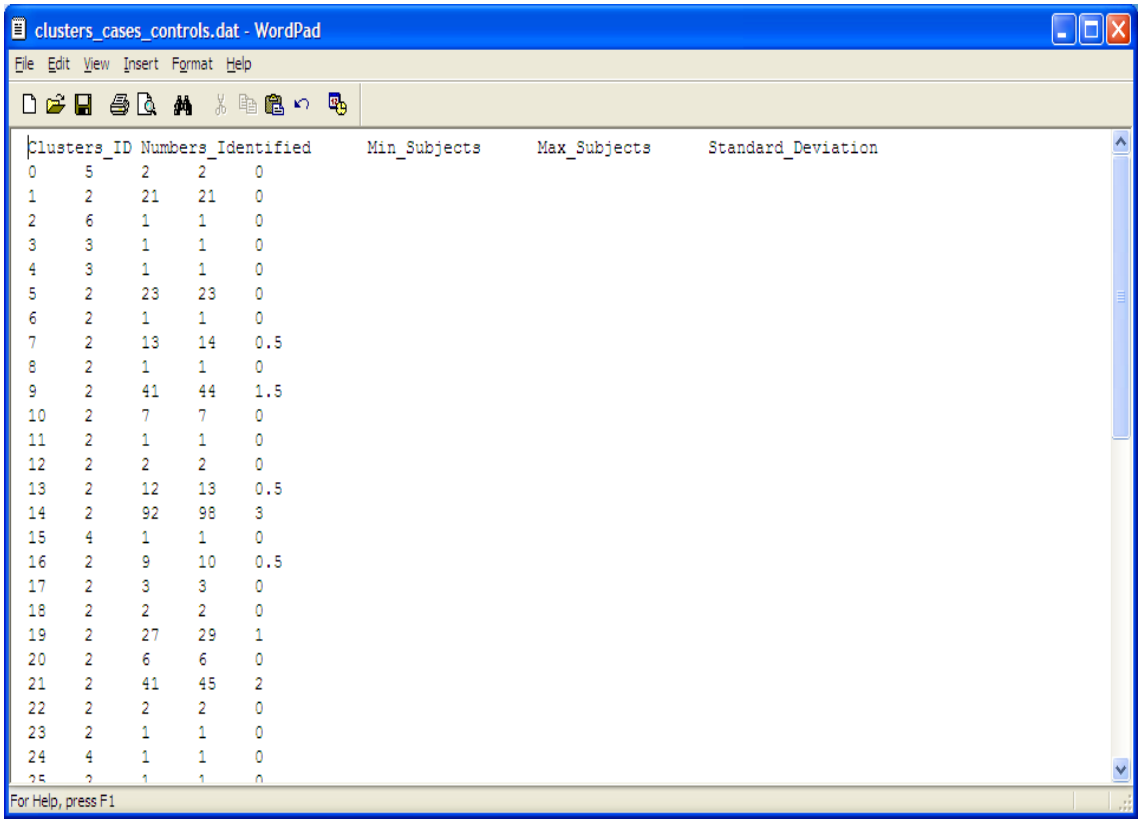
Ως μέρος της μετά επεξεργασίας των αποτελεσμάτων ήταν αρχικά, σαν πρώτο βήμα, η δημιουργία των διάφορων γραφικών παραστάσεων και εικόνων με την χρήση διαφόρων εφαρμογών και συγκεκριμένα των εφαρμογών Gnuplot και SPSS. Ειδικά για την χρήση της εφαρμογής Gnuplot, έχει γραφτεί αρκετός κώδικας με την μορφή script, ο οποίος αναλαμβάνει να παράξει τα διάφορα αποτελέσματα που παράχθηκαν από τον αλγόριθμο.

Το δεύτερο βήμα στο στάδιο αυτό, αποτελούσε η συγχώνευση των διαφόρων διαφορετικών runs που έγιναν. Συγκεκριμένα, για να είναι συνεπή τα αποτελέσματα, και να αποφευχθούν οποιαδήποτε τυχαία αποτελέσματα, αφού στον αλγόριθμο που έχει αναπτυχθεί υπάρχει μια τυχαιότητα (randomness) στην αρχική ανάθεση των βαρών του κάθε νευρώνα και συνεπώς, οποιοσδήποτε θα μπορούσε να ισχυριστεί ότι με ένα μόνο run του αλγορίθμου, τα αποτελέσματα θα ήταν τυχαία. Οπότε, αποφασίστηκε να τρέξει ο αλγόριθμος αρχικά 10 φορές και στην συνέχεια 50 και σαν μεταεπεξεργασία, από τις διάφορες ομάδες που έχουν ανακαλυφθεί σε κάθε run, να βρεθούν οι ομάδες που έχουν πολύ μεγάλη ομοιότητα σε όλα τα runs.

3.5.2 Αλγόριθμος Εύρεσης Κοινών Ομάδων – Clusters

Όπως έχει προαναφερθεί, το δεύτερο στάδιο στο βήμα αυτό αφορούσε την ανάπτυξη ενός αλγορίθμου για την συγχώνευση των αποτελεσμάτων των διαφόρων runs. Πιο συγκεκριμένα, στόχος του αλγορίθμου αυτού ήταν να διαβάζει όλες τις ομάδες – clusters που δημιουργεί το κάθε run και να βρίσκει στην συνέχεια πόσες φορές εμφανίζεται η κάθε ομάδα σε κάθε run ανάλογα με ένα ποσοστό ομοιότητας που ορίζει ο χρήστης από πριν.

Ο αλγόριθμος αυτός ακολουθεί μια σχετικά απλή μεθοδολογία. Αρχικά μελετά κάθε ομάδα με όλες τις άλλες. Αν βρίσκει ομοιότητα με κάποια από τις άλλες, αποθηκεύονται τα στοιχεία της ομάδας αυτής, αυξάνεται ένας μετρητής ο οποίος χρησιμοποιείται για την αποθήκευση των φορών που παρατηρήθηκε η συγκεκριμένη ομάδα και η δεύτερη ομάδα η οποία έχει βρεθεί να είναι όμοια με την πρώτη διαγράφεται, αφού δεν υπάρχει περίπτωση να ανήκει σε κάποια άλλη μετέπειτα. Αυτή η διαδικασία συνεχίζεται για όλες τις ομάδες που έχουν διαβαστεί. Τα αποτελέσματα εμφανίζονται σε ένα αρχείο με την μορφή που φαίνεται στην εικόνα 6.3



Clusters_ID	Numbers_Identified	Min_Subjects	Max_Subjects	Standard_Deviation
0	5	2	2	0
1	2	21	21	0
2	6	1	1	0
3	3	1	1	0
4	3	1	1	0
5	2	23	23	0
6	2	1	1	0
7	2	13	14	0.5
8	2	1	1	0
9	2	41	44	1.5
10	2	7	7	0
11	2	1	1	0
12	2	2	2	0
13	2	12	13	0.5
14	2	92	98	3
15	4	1	1	0
16	2	9	10	0.5
17	2	3	3	0
18	2	2	2	0
19	2	27	29	1
20	2	6	6	0
21	2	41	45	2
22	2	2	2	0
23	2	1	1	0
24	4	1	1	0

Εικόνα 3.7 Αποτέλεσμα του αλγορίθμου εύρεσης κοινών ομάδων - clusters

3.6 Τρόποι Σύγκρισης και Αξιολόγησης Αλγορίθμων

Για την αξιολόγηση του αλγορίθμου, θα χρησιμοποιηθούν διάφορες μετρικές όπως η επιτάχυνση που χρησιμοποιείται για την αξιολόγηση του παραλληλισμού, οι χρόνοι ολοκλήρωσης του αλγορίθμου και το ποσοστό σφάλματος. Για την αποφυγή οποιοδήποτε σφαλμάτων στις μετρήσεις και για συνέπεια των αποτελεσμάτων, έχουν παρθεί διάφορες μετρήσεις στην κάθε μετρική και παρουσιάζονται οι μέσες τιμές των αποτελεσμάτων.

3.6.1 Επιτάχυνση (Speed Up)

Η επιτάχυνση, στο παράλληλο υπολογισμό, ορίζει το πόσο γρηγορότερος είναι ο παράλληλος αλγόριθμος σε σχέση με τον αντίστοιχο σειριακό. Με απλά λόγια, καθορίζει το πόσο επεκτάσιμος είναι ένας αλγόριθμος (scalability), ανάλογα με τον αριθμό των επεξεργαστών που υπάρχουν διαθέσιμα σε μια μηχανή. Είναι, ίσως η πιο διαδεδομένη μετρική όσο αφορά τον παράλληλο υπολογισμό.

Η επιτάχυνση έχει υπολογιστεί με βάση τον ορισμό της επιτάχυνσης δηλαδή

$$\text{Επιτάχυνση} = \frac{\text{Χρόνος Σειριακού Αλγορίθμου}}{\text{Χρόνος Παράλληλου Αλγορίθμου}}.$$

3.6.2 Αλλαγή Βαρών (Weight Adjustment)

Η μετρική αυτή, μετρά τις αλλαγές που γίνονται στα βάρη ολόκληρου του δικτύου. Πιο συγκεκριμένα, για κάθε διάνυσμα εισόδου αθροίζονται οι αλλαγές στα βάρη του κάθε νευρώνα και ακολούθως όλων των νευρώνων. Στο τέλος κάθε εποχής το άθροισμα αυτό διαιρείται με τον αριθμό των νευρώνων του δικτύου δίνοντας μια μέτρηση – ένδειξη για την αλλαγή των βαρών.

3.6.3 Σφάλμα Κβαντισμού

Το σφάλμα κβαντισμού (quantization error) είναι μια μετρική που καθορίζει τον βαθμό μάθησης που παρατηρείται σε ένα χάρτη αυτό – οργάνωσης. Είναι ένας τρόπος, με απλά λόγια να ελέγξει κάποιος αν το δίκτυο που έχει δημιουργήσει μαθαίνει μέσω της διαδικασίας μάθησης. [21] Είναι ευρέως διαδεδομένη μετρική (χρησιμοποιείται το σφάλμα κβαντισμού συνήθως και σε μερικές περιπτώσεις το τοπογραφικό σφάλμα) και ουσιαστικά αξιολογεί την ποιότητα της μάθησης. Ο ορισμός του σφάλματος κβαντισμού ορίζεται ως εξής:

$$Quantization\_Error = \frac{1}{N}\sum_{i=1}^{N}\left\|x(i) - w_{BMU}(i)\right\|$$

Όπου x είναι το διάνυσμα εισόδου, το w το αντίστοιχο διάνυσμα των βαρών του BMU και το N το μέγεθος των διανυσμάτων x και w. Συνεπώς το σφάλμα αυτό υπολογίζεται για κάθε διάνυσμα εισόδου. Σ' αυτό το σημείο πρέπει να αναφερθεί ότι στην βιβλιογραφία μια εποχή θεωρείται η παρουσίαση ενός δεδομένου στο δίκτυο. [15] Για τον συγκεκριμένο αλγόριθμο, μια εποχή θεωρείται ένα πέρασμα από όλα τα δεδομένα καθώς, αν υιοθετείτο η προαναφερθείσα προσέγγιση το αποτέλεσμα της κατηγοριοποίησης θα ήταν προκατειλημμένο (bias) ως προς τα δεδομένα που παρουσιάζονται πιο συχνά στο δίκτυο καθώς η επιλογή θα γινόταν τυχαία από το input space. Επειδή, στόχος ήταν η κατηγοριοποίηση όλων των δεδομένων, ως εποχή ορίστηκε ένα πέρασμα από όλα τα δεδομένα. Γι' αυτό για τον υπολογισμό του σφάλματος κβαντισμού, υπολογίζεται με μια μικρή διαφορά.

$$Quantization\_Error = \frac{1}{M}\sum_{j=1}^{M}\ \frac{1}{N}\sum_{i=1}^{N}\left\|x(i) - w_{BMU}(i)\right\|$$

Όπου M ο αριθμός των διανυσμάτων εισόδου. Συνεπώς, η διαφορά είναι ότι αθροίζεται το σφάλμα κβαντισμού για κάθε διάνυσμα εισόδου και στο τέλος κάθε εποχής, το άθροισμα διαιρείται με τον αριθμό των διανυσμάτων εισόδου.

Κεφάλαιο 4

Αποτελέσματα

4.1 Αποτελέσματα Προεπεξεργασίας

4.1.1 Φιλτράρισμα Δεδομένων

Μέσω του φιλτραρίσματος και συζήτησης που έγινε στα μέλη της ομάδας, αποφασίστηκε να χρησιμοποιηθούν τα τρία φιλτραρίσματα που αφορούσαν SNPs με επίσταση 45. Το αποτέλεσμα και στις 3 περιπτώσεις, ήταν να χρησιμοποιηθούν 84 SNPs (168 alleles). Όσο αφορά τους ασθενείς, μετά το φιλτράρισμα, δημιουργήθηκε ένα αρχείο με 972 ασθενείς και μέγεθος 340KB και για τους υγιείς, 881 άτομα με 309KB αντίστοιχα. Τέλος το αρχείο που περιείχε και τις δύο ομάδες, περιείχε συνολικά 1853 άτομα και μέγεθος 650KB.

4.1.2 Αποτελέσματα Δυαδικού Πρότυπου για γενετικά δεδομένα – συμπίεση

Όπως έχει προαναφερθεί, ο αλγόριθμος συμπίεσης μειώνει σημαντικά το μέγεθος που απαιτείται για αποθήκευση των γενετικών δεδομένων. Ο πίνακας 4.1 παρέχει λεπτομέρειες για του λόγου του αληθές. Αρχεία βάσης δεδομένων συνολικού μεγέθους 3.62 GB μετατρέπονται σε ένα μόνο αρχείο της τάξεως των 496MB. Συνεπώς, παρατηρείται ότι ο προτεινόμενος αλγόριθμος να μειώνει περίπου 8 φορές το μέγεθος

αρχείων με pedigree format. Αν και ο αλγόριθμος του PLINK, μειώνει περίπου 16 φορές το μέγεθος των αρχείων, εντούτοις πολλές σημαντικές πληροφορίες χάνονται όπως έχει εξηγηθεί πιο πάνω.

TABLE IV
COMPARISON OF FILE FORMATS

QTDR Merlin	Proposed Protocol	Plink 's protocol
Information Loss for bi-allelic markers		
Used as Reference	None	strand location of heterozygote alleles Aa,aA Missing one of the two alleles A0, 0A, a0,0a
Added Information capability		
tri or quad allelic markers	Alleles deleted from a specific strand	None
Size*		
Pedigree File 3.6 GB Map File 12.6 MB **Total:** 3.612 GB	One binary file 496 MB	.bed file : 229.6 MB .fam file : 30 KB .bim file : 14.1 MB **Total** : 243.7 MB

Πίνακας 4.1 Αποτελέσματα του αλγορίθμου συμπίεσης παρθέν από το [17]

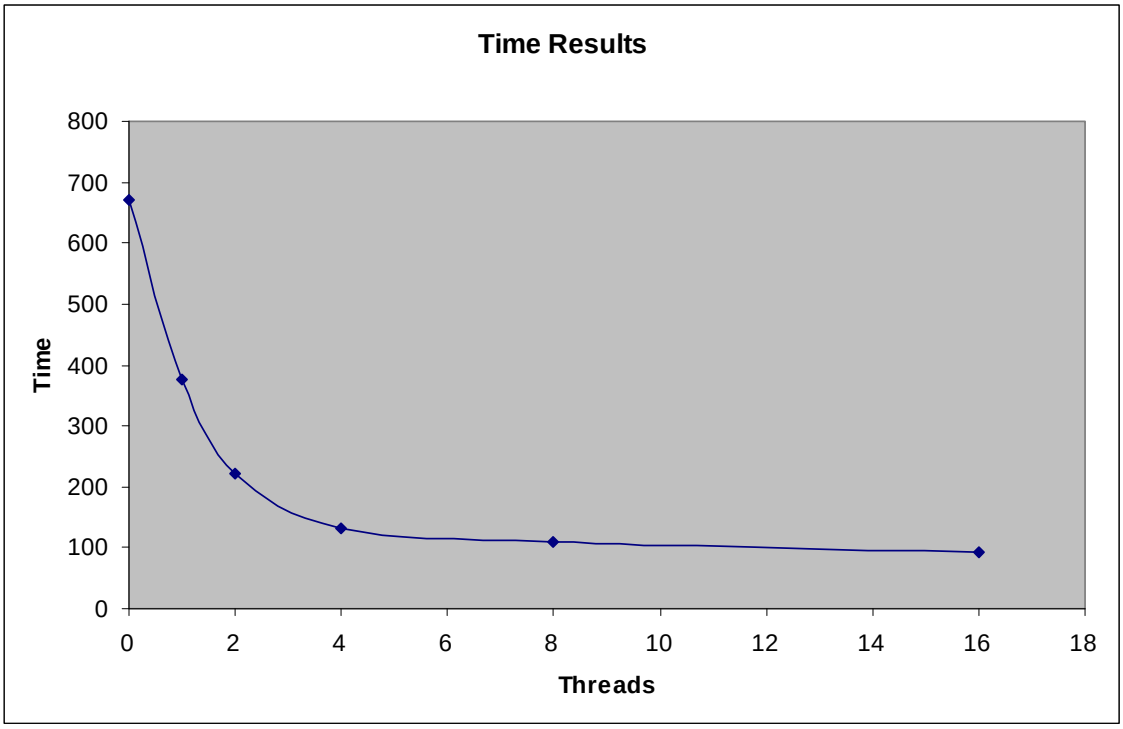
4.2 Αποτελέσματα Παράλληλου Αλγορίθμου

4.2.1 Αποτελέσματα με διάφορα threads

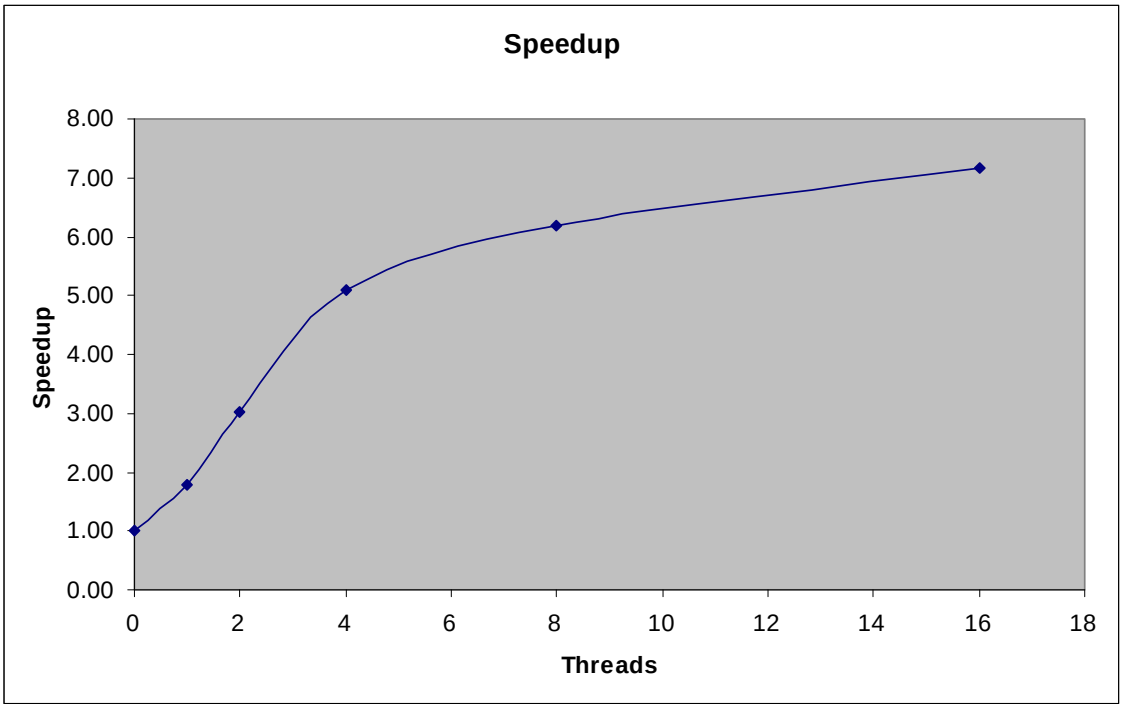
Όπως έχει προαναφερθεί, μια από τις βασικές αιτίες για να παρθεί η απόφαση να αναπτυχθεί ο παράλληλος αλγόριθμος εκμάθησης χαρτών αυτό-οργάνωσης αποτελούσε ο υπερβολικός χρόνος που χρειαζόταν ο σειριακός αλγόριθμος για να ολοκληρώσει την εκτέλεση του. Συνεπώς, πρέπει να παρουσιαστούν αποτελέσματα όσο αφορά επιτάχυνση (speedup) για να εξεταστεί κατά πόσο άξιζε η υλοποίηση και ανάπτυξη του αλγορίθμου αυτού. Η εικόνα 4.1 απεικονίζει τους χρόνους που χρειάστηκε ο αλγόριθμος σε συνάρτηση με τον αριθμό των threads που χρησιμοποιήθηκαν σε κάθε εκτέλεση του αλγορίθμου. Συγκεκριμένα έχουν γίνει πολλές δοκιμές χρησιμοποιώντας

1 – 16 threads (δεν ήταν δυνατό να χρησιμοποιηθούν περισσότερα threads καθώς η μηχανή η οποία ήταν διαθέσιμη για τις εκτελέσεις χρησιμοποιεί 8 επεξεργαστές, αφού το κόστος από την εναλλαγή των threads σε ένα επεξεργαστή αν ο αριθμός των threads είναι μεγάλος είναι μεγάλο προκαλώντας σοβαρή καθυστέρηση στο συνολικό χρόνο της εφαρμογής) και φαίνεται ξεκάθαρα ότι ο χρόνος μειώνεται αισθητά.

Παρόμοια αποτελέσματα παρουσιάζει και η επιτάχυνση (speedup) η οποία παρουσιάζεται στην εικόνα 4.2. Συγκεκριμένα, η εικόνα αυτή, παρουσιάζει στον άξονα Y την επιτάχυνση και στον άξονα X των αριθμών των threads. Η επιτάχυνση έχει υπολογιστεί με βάση τον ορισμό της επιτάχυνσης. Τα αποτελέσματα είναι ανάλογα με τα αποτελέσματα της εικόνας 4.1 καθώς παρατηρείται σοβαρή επιτάχυνση για διάφορους λόγους που θα εξηγηθούν πιο κάτω αλλά και το φαινόμενα να μειώνεται σταδιακά (θα εξηγηθεί και αυτή η αιτία στο επόμενο κεφάλαιο), αν και επιθυμία ήταν η γραμμική επιτάχυνση. Όμως ο τρόπος με τον οποίο εμφανίζεται η επιτάχυνση είναι ο αναμενόμενος λόγω των ιδιαιτεροτήτων που παρουσιάζει ο αλγόριθμος, ιδιαιτερότητες που θα αναφερθούν στο επόμενο κεφάλαιο.



Εικόνα 4.1 Χρόνοι που χρειάστηκαν συγκεκριμένες εκτελέσεις του αλγόριθμου.



Εικόνα 4.2 Επιτάχυνση σε σύγκριση με διάφορες εκτελέσεις.

4.2.2 Αποτελέσματα Τριών Συγκεκριμένων Εκτελέσεων

Ο αλγόριθμος εκτελέστηκε πάρα πολλές φορές με τα αποτελέσματα να είναι πολύ παρόμοια σε όλες σχεδόν τις εκτελέσεις, κάτι που αποδεικνύεται στο 4.3.1 μέσω του αλγορίθμου συνέπειας που έχει αναπτυχθεί. Γι ‘ αυτό το λόγο θα παρουσιαστούν τα αποτελέσματα από τρεις συγκεκριμένες εκτελέσεις ως αντιπροσωπευτικές για όλες τις εκτελέσεις, αφού δεν είναι δυνατόν να παρουσιαστούν όλα τα αποτελέσματα του αλγορίθμου από όλες τις εκτελέσεις.

Τα αποτελέσματα που θα παρουσιαστούν στα τμήματα 4.2.2.1, 4.2.2.2 και 4.2.2.3 θα αφορούν τρεις εκτελέσεις κατά τις οποίες χρησιμοποιήθηκαν πλέγματα 50 x 50, 20 x 20, 10 x 10 με αριθμό εποχών εκπαίδευσης 1000, με αρχικό ρυθμό μάθησης 0,1 και ακτίνα 25, 10 και 5 αντίστοιχα χρησιμοποιώντας 4 threads. Οι λόγοι επιλογής των παραμέτρων αυτών έχουν εξηγηθεί πιο πάνω. Όσο αφορά τα δεδομένα, χρησιμοποιήθηκαν τα 3 αρχεία που παράχθηκαν στο στάδιο της προεπεξεργασίας (ασθενείς, υγιείς και συνδυασμός τους)

Όσο αφορά τον τρόπο με τον οποίο δημιουργείται η ομαδοποίηση, όταν εκπαιδευτεί το δίκτυο, ο αλγόριθμος περνά ακόμα μια φορά από όλα τα δεδομένα εισόδου. Για κάθε διάνυσμα εισόδου, ο αλγόριθμος βρίσκει το BMU, και αναθέτει στον νευρώνα αυτό το διάνυσμα εισόδου αυξάνοντας ταυτόχρονα ένα μετρητή που διατηρεί τον αριθμό των δεδομένων που «κτύπησαν» σε ένα συγκεκριμένο νευρώνα. Τα αποτελέσματα εξάγονται σε αρχεία τα οποία μέσω scripts μετατρέπονται σε γραφικές παραστάσεις μέσω του εργαλείου gnuplot.

Στην εικόνα 4.4, παρουσιάζεται η κατηγοριοποίηση που εξάγεται από τον αλγόριθμο σε ASCII μορφή. Είναι τα δεδομένα που χρησιμοποιούνται για την εξαγωγή των απεικονίσεων του δικτύου που παρουσιάζονται πιο κάτω. Κάθε τιμή αναπαριστά ένα νευρώνα (η τιμή αναπαριστά τον μετρητή) και διαχωρίζονται μεταξύ τους με tab. Λόγω του μεγέθους του πλέγματος δεν ήταν δυνατό να παρουσιαστούν όλα τα δεδομένα.

```
7	1	1	0	0	0	1	0	0	0	1	0	1	0	1	1	0	2	0
0	1	0	2	0	1	0	1	0	2	0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0
0	1	0	2	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0
0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	0	0	1
1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	2	0	0	0
0	1	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0
1	0	0	0	0	0	1	0	0	0	1	0	1	0	2	0	0	0	0
0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1
0	0	1	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0	0
0	1	0	0	1	0	0	0	1	0	0	18	0	1	1	0	0	0	0
1	0	2	0	0	1	0	0	0	1	0	0	0	0	0	1	0	7	0
0	5	0	1	1	0	1	0	1	0	0	1	0	1	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	1	0	0	0	1	0	2	0	0	2	0	1	0	0
0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	1	0	0	2	0	1	0	1	0	1	0
9	0	0	9	0	0	2	0	0	1	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	4	0	0	0
0	0	2	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0
2	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0
0	0	3	0	1	0	1	0	1	0	0	1	0	0	1	0	0	0	0
2	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	1	0	1
0	0	1	0	1	0	1	0	2	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	0
0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0
0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1
1	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0
0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	1	0	0
1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
0	0	1	0	1	0	0	0	5	0	0	0	1	0	0	0	0	1	0
2	0	0	0	0	0	1	0	0	0	1	0	0	1	0	1	0	0	0
0	1	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0
0	0	0	1	0	2	0	0	0	0	0	1	0	0	0	0	0	1	0
4	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
```

Εικόνα 4.4 Παράδειγμα Εξόδου Κατηγοριοποιήσης Πλέγματος (clusteringASCII.dat)

Στην εικόνα 4.5, παρουσιάζεται η κατηγοριοποίηση των ατόμων στον κάθε νευρώνα με την πρόσθετη πληροφορία του διαχωρισμού σε ασθενείς (cases) και υγιείς (controls). Η μορφή (format) το οποίο ακολουθείται στο αρχείο αυτό αποτελείται από 4 γραμμές σε κάθε νευρώνα. Η πρώτη γραμμή αναπαριστά τους δείκτες (indexes) του κάθε νευρώνα στο πλέγμα. Για παράδειγμα [1][2] αναπαριστά τον νευρώνα 1 , 2 στο πλέγμα. Στην επόμενη γραμμή περιέχονται οι δείκτες των ασθενών (cases) που «κτύπησαν» στον νευρώνα αυτό με βάση την θέση τους στο αρχείο .ped που δίνεται στην είσοδο. Αντίστοιχα, αναπαριστώνται οι υγιείς (controls) στην επόμενη γραμμή. Η τελευταία γραμμή περιέχει τα άτομα στα οποία δεν υπήρχε πληροφορία στο αρχείο εισόδου κατά πόσο ήταν ασθενείς ή όχι. Με άλλα λόγια, ο φαινότυπος τους ήταν άγνωστος κατά την συλλογή των δεδομένων. Δεν ήταν δυνατό να δοθούν όλα τα δεδομένα στην εικόνα 4.5, λόγω του μεγέθους του πλέγματος στο παράδειγμα.

```
[0][40]
695

[0][41]

[0][42]

[0][43]
39    117   119   257   338   767   821

[0][44]

[0][45]
663

[0][46]

[0][47]
115

[0][48]

[0][49]
23    57    63    100   125   148   149   181   195   209   332   370   375   378   416   418   434   469   487
```

Εικόνα 4.5 Παράδειγμα Εξόδου Κατηγοριοποιήσης Πλέγματος (clusteringCases_ControlsIndex.dat)

Τέλος, στην εικόνα 4.6 παρουσιάζεται το αποτέλεσμα των βαρών όλων των νευρώνων. Το αρχείο που παράγεται χρησιμοποιείται ως είσοδος σε script για την παραγωγή των εικόνων που παρουσιάζουν τα βάρη νευρώνων και θα παρουσιαστούν πιο κάτω. Η μορφή (format) που ακολουθείται ορίζεται ως εξής: αρχικά ορίζονται οι δείκτες που ορίζουν τον νευρώνα, όπως χρησιμοποιείται και στο clusteringCases_ControlsIndex.dat αρχείο. Ακολούθως στις επόμενες γράμμες (μέχρι τον επόμενο δείκτη) αναγράφονται τα βάρη του κάθε νευρώνα (κάθε γραμμή ένα βάρος).

```
  Weights
==================
Neuron: [0][0]
Weights:
1
1
1.28862e-53
2.47033e-323
0.019985
6.27213e-16
1
5.13723e-131
4.76453e-131
1.8141e-74
1.44264e-158
1
1.44264e-158
1
1.90195e-83
2.49464e-204
1.90195e-83
2.49464e-204
1.44264e-158
1
1
1
1
1
1.90195e-83
2.49464e-204
1
1
1
1
2.96439e-323
1
1.90195e-83
2.49464e-204
1
1
1
```

Εικόνα 4.6 Αναπαράσταση Βαρών Πλέγματος 50x50 (weights.dat)

4.2.2.1 Κατηγοριοποίηση Εκτέλεσης 50x50 πλέγματος

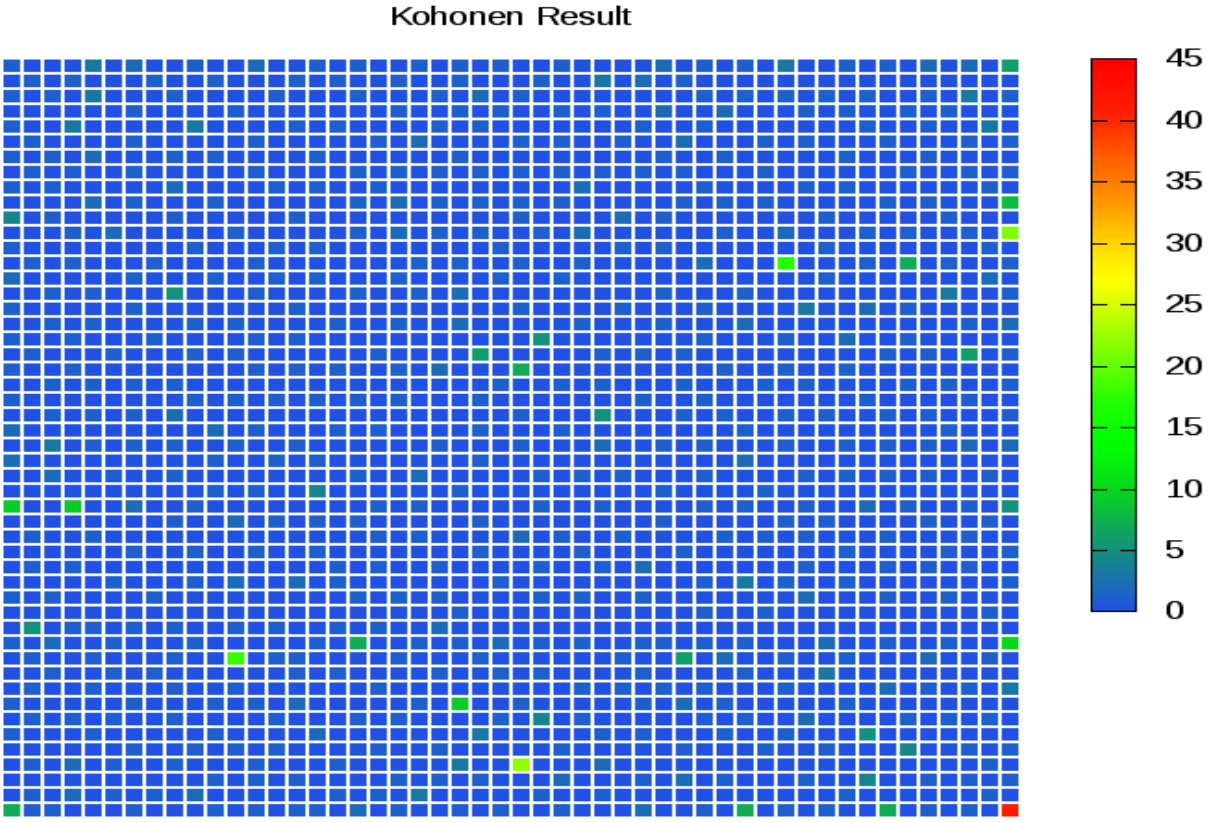
Στα τμήματα 4.2.2.1.1, 4.2.2.1.2 και 4.2.2.1.3 παρουσιάζονται τα αποτελέσματα του αλγορίθμου σε ένα 50 x 50 πλέγμα. Κάθε τετράγωνο αναπαριστά ένα νευρώνα. Το χρώμα του κάθε νευρώνα αναπαριστά τον μετρητή. Μέσω της αναπαράστασης αυτής είναι πολύ εύκολο να εξαχθούν διάφορα συμπεράσματα όσο αφορά το ποιες ομάδες δημιουργούνται, όπως επίσης συμπεράσματα για το ποιοι νευρώνες χρησιμοποιήθηκαν περισσότερο και αναμένεται να έχουν σημαντική βιολογική αξία, με βάση την θεωρία στην οποία στηρίζεται η μελέτη αυτή, δηλαδή στην εύρεση κοινών χαρακτηριστικών στα γενετικά δεδομένα διαφόρων ατόμων.

Επιπλέον σε κάθε τμήμα, παρουσιάζονται και οι αναπαραστάσεις των βαρών κάποιων συγκεκριμένων νευρώνων. Η επιλογή των νευρώνων για αναπαράσταση των βαρών τους δεν ήταν τυχαία καθώς οι νευρώνες αυτοί παρουσιάζουν την μεγαλύτερη κίνηση, κάτι που σημαίνει ότι η γενετική ακολουθία των ατόμων που απεικονίζουν παρουσιάζει μεγάλη ομοιότητα. Συνεπώς, πρέπει να παρατηρείται σημαντική βιολογική σημασία στα βάρη του νευρώνα αυτού, που αντικατοπτρίζουν ουσιαστικά τα κοινά χαρακτηριστικά των ατόμων αυτών. Όπως έχει ήδη προαναφερθεί, τα δεδομένα που ήταν διαθέσιμα, προέρχονται από την γενετική περιοχή HLA που είναι ιδιαίτερα πολύπλοκή και είναι σίγουρο ότι ευθύνεται για πολλές ασθένειες αλλά δεν έχει ξεκαθαρίσει ποια γονίδια ή τμήματα της περιοχής αυτής ευθύνονται για την κάθε ασθένεια. Συνεπώς μια τέτοια συσχέτιση μπορεί να βοηθήσει πάρα πολύ γι' αυτό και απεικονίζονται ξεχωριστά οι νευρώνες αυτοί.

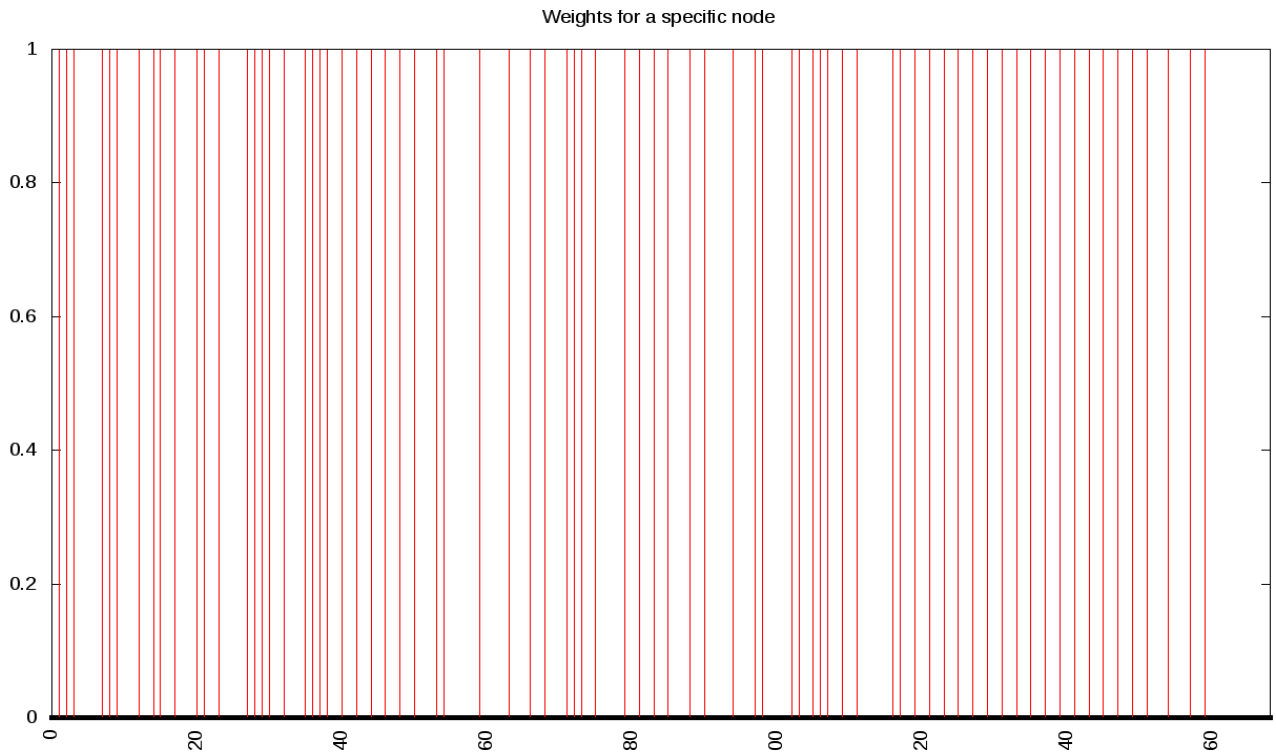
Πιο συγκεκριμένα, στις εικόνες αυτές στον άξονα X βρίσκονται οι δείκτες των βαρών (π.χ η τιμή 20 αντικατοπτρίζει το 20 βάρος) και ουσιαστικά αποτελούν τα alleles (οπότε ανά δυο αντιστοιχούν σε ένα σημειακό νουκλεοτιδικό πολυμορφισμό) και στον άξονα Y οι αντίστοιχες τιμές των βαρών. Το σημαντικό είναι ότι παρατηρούνται μόνο τιμές 0 ή 1 κάτι που σημαίνει ότι οι τιμές αυτές βρίσκονται στην κωδικοποιημένη μορφή όλων των ατόμων που αναπαριστώνται από τους συγκεκριμένους νευρώνες.

4.2.2.1.1 Αποτελέσματα Ασθενών

Στην εικόνα 4.7, παρουσιάζεται ο αριθμός των ατόμων που αντιπροσωπεύει ο κάθε νευρώνας. Οι αριθμοί αναπαριστώνται με βάση το χρώμα. Η πιο σημαντική ομάδα παρατηρείται στο νευρώνα [0][49] που αντιπροσωπεύει 41 άτομα. Για το λόγο αυτό, αναπαριστώνται στην εικόνα 4.8, τα βάρη του νευρώνα αυτού. Αυτό που αξίζει να σημειωθεί είναι ότι δεν παρατηρούνται καθόλου γειτονιές, και αυτός ήταν ο στόχος να καταφέρει το δίκτυο να κατηγοριοποιήσει βάση ανεξάρτητων νευρώνων και όχι βάση γειτονιών. Από την απεικόνιση των βαρών φαίνεται ξεκάθαρα ότι υπάρχει μεγάλη ομοιότητα στα άτομα που αντιπροσωπεύονται από τον νευρώνα αυτό.



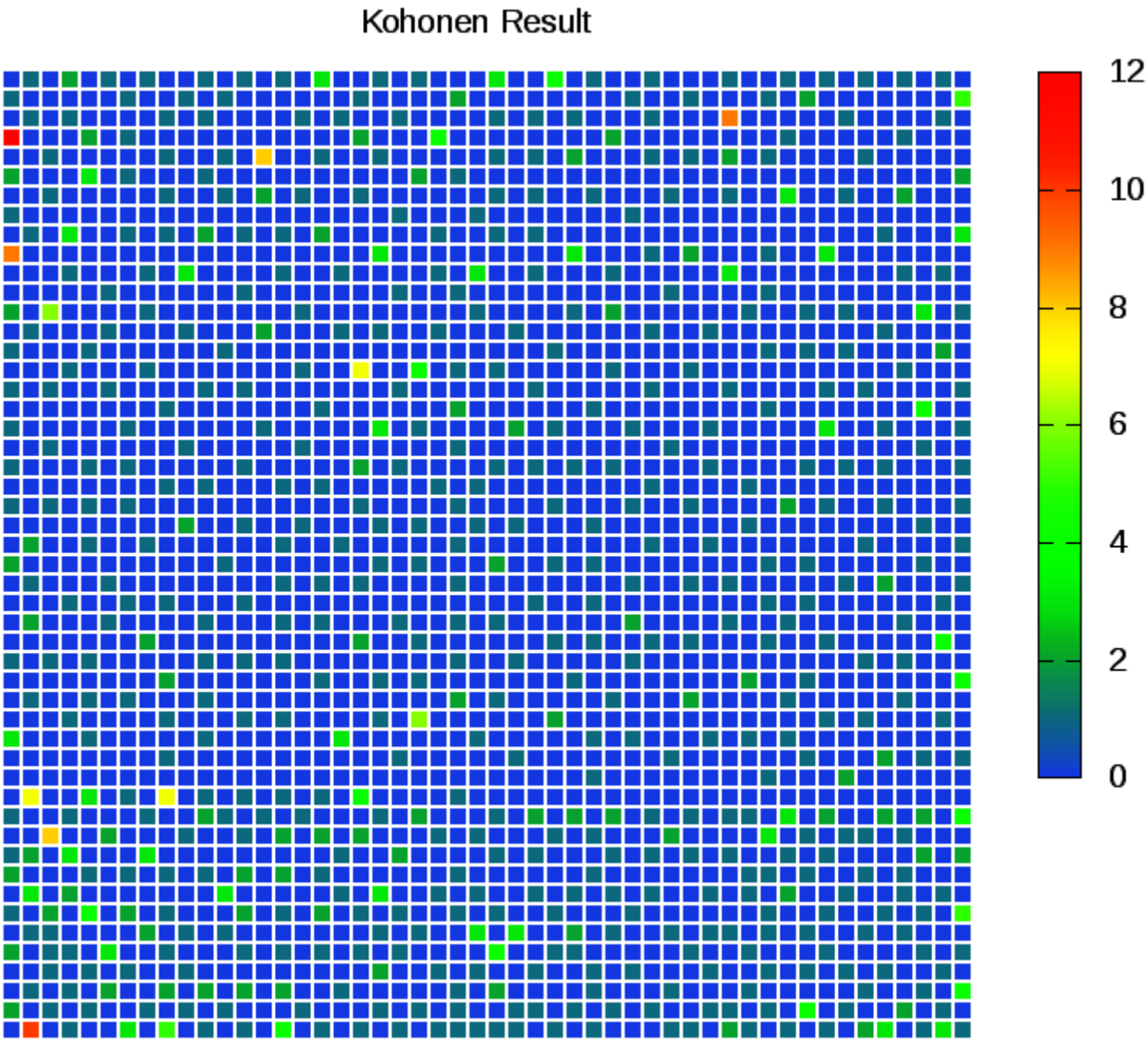
Εικόνα 4.7 Απεικόνιση Κατηγοριοποιήσης 50x50 Ασθενών



Εικόνα 4.8 Αναπαράσταση Βαρών Νευρώνα [40] [49] Πλέγματος 50x50 – Ασθενείς

4.2.2.1.2 Αποτελέσματα Υγιών

Στην εικόνα 4.9, παρουσιάζεται ο αριθμός των ατόμων που αντιπροσωπεύει ο κάθε νευρώνας. Οι αριθμοί αναπαριστώνται με βάση το χρώμα. Δεν θα παρουσιαστεί η απεικόνιση βαρών συγκεκριμένου κόμβου καθώς δεν έχει παρουσιαστεί κάποιος νευρώνας που να ξεχωρίζει. Γενικά παρατηρείται μια πολύ συγκεχυμένη κατηγοριοποίηση καθώς έχουν χρησιμοποιηθεί σχεδόν όλοι οι νευρώνες σε μικρό βαθμό, κάτι που οδηγεί στο συμπέρασμα ότι είτε δεν υπάρχουν πολύ μεγάλες ομοιότητες μεταξύ των μη ασθενών, είτε ότι είναι πολύ μεγάλο το πλέγμα για το συγκεκριμένο σύνολο δεδομένων.

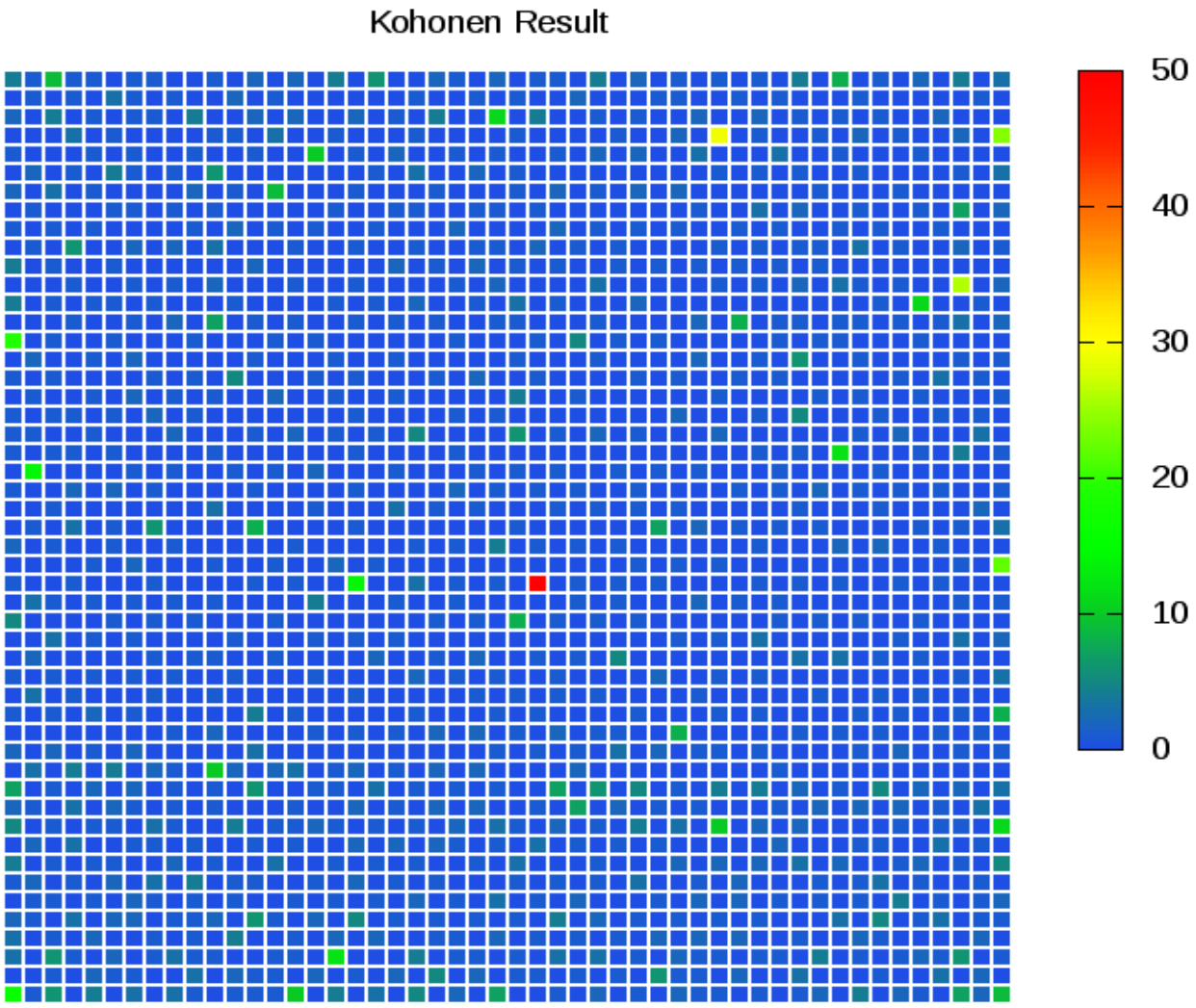


Εικόνα 4.9 Απεικόνιση Κατηγοριοποιήσης 50x50 Υγιείων

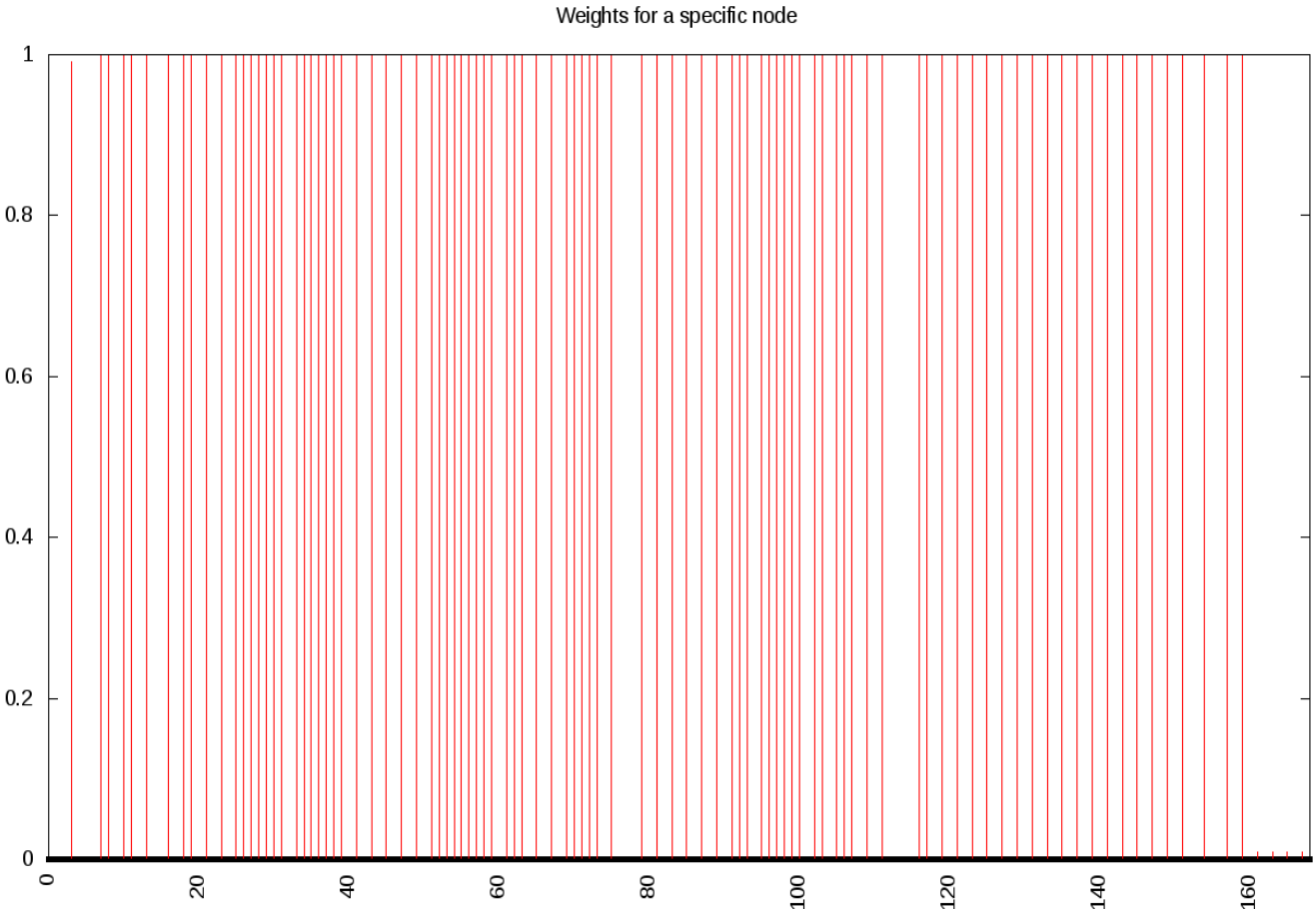
4.2.2.1.3 Αποτελέσματα Συνδυασμού Ασθενών - Μη Ασθενών

Στην εικόνα 4.10, παρουσιάζεται ο αριθμός των ατόμων που αντιπροσωπεύει ο κάθε νευρώνας. Οι αριθμοί αναπαριστώνται με βάση το χρώμα. Η πιο σημαντική ομάδα παρατηρείται στο νευρώνα [22][25] που αντιπροσωπεύει 50 άτομα. Για το λόγο αυτό, αναπαριστώνται στην εικόνα 4.11, τα βάρη του νευρώνα αυτού. Σαν πρώτη παρατήρηση, επιβεβαιώνει η εικόνα 4.10 τα αποτελέσματα που παρουσιάζονται στην εικόνα 4.7 καθώς παρουσιάζονται περίπου οι ίδιες ομάδες, με κάποια περισσότερα άτομα, άτομα τα οποία είναι κυρίως μη ασθενείς, και έχουν ομαδοποιηθεί μαζί με τις

ομάδες των ασθενών που παρουσιάστηκαν στην εικόνα 4.7. Δεν παρατηρούνται καθόλου γειτονιές, κάτι το οποίο ήταν στόχος, όπως έχει προαναφερθεί. Επιπλέον τα βάρη επιβεβαιώνουν την ομοιότητα των ατόμων που αντιπροσωπεύονται από το συγκεκριμένο νευρώνα.



Εικόνα 4.10 Απεικόνιση Κατηγοριοποιήσης 50x50 Ασθενών – Μη ασθενών



Εικόνα 4.11 Αναπαράσταση Βαρών Νευρώνα [22] [25] Πλέγματος 50x50 – Ασθενείς και Υγιείς

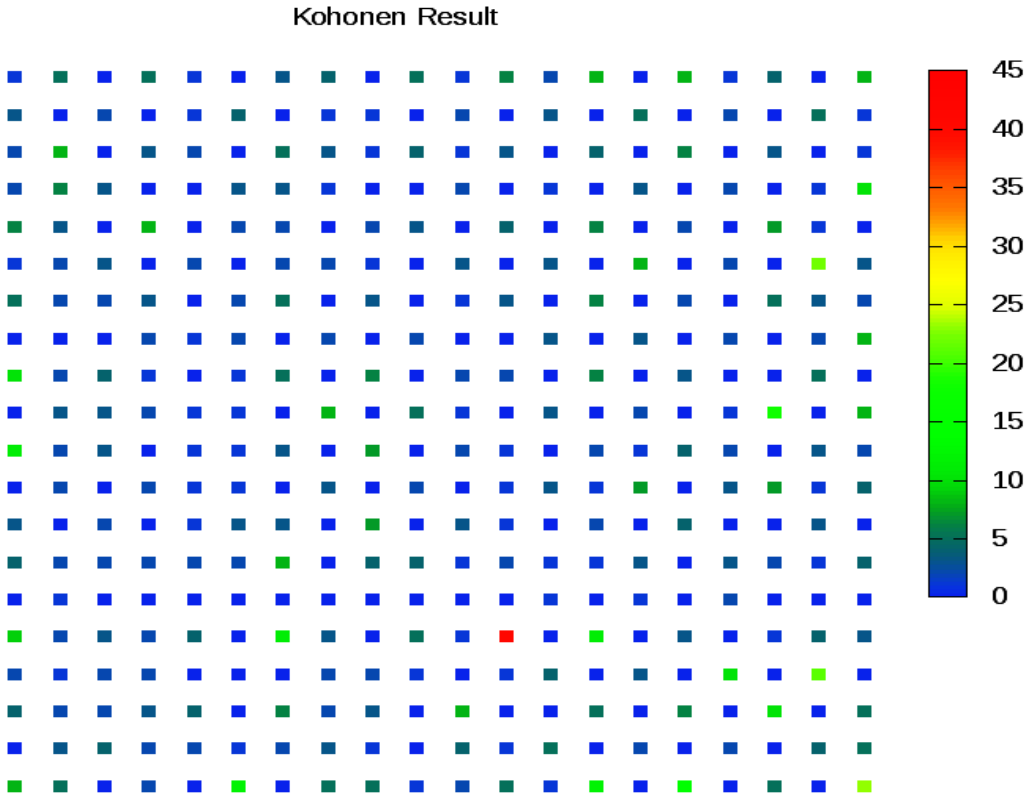
4.2.2.2 Κατηγοριοποίηση Εκτέλεσης 20x20 πλέγματος

Στα τμήματα 4.2.2.2.1, 4.2.2.2.2, 4.2.2.2.3 παρουσιάζονται τα αποτελέσματα του αλγορίθμου σε ένα 20 x 20 πλέγμα. Θα παρουσιαστεί η κατηγοριοποίηση όπως προκύπτει σε εικονική μορφή, όπως επίσης και τα βάρη των πιο σημαντικών νευρώνων. Τα αποτελέσματα που προκύπτουν είναι παρόμοια με τις αντίστοιχες κατηγοριοποιήσεις στο 50x50 πλέγμα.

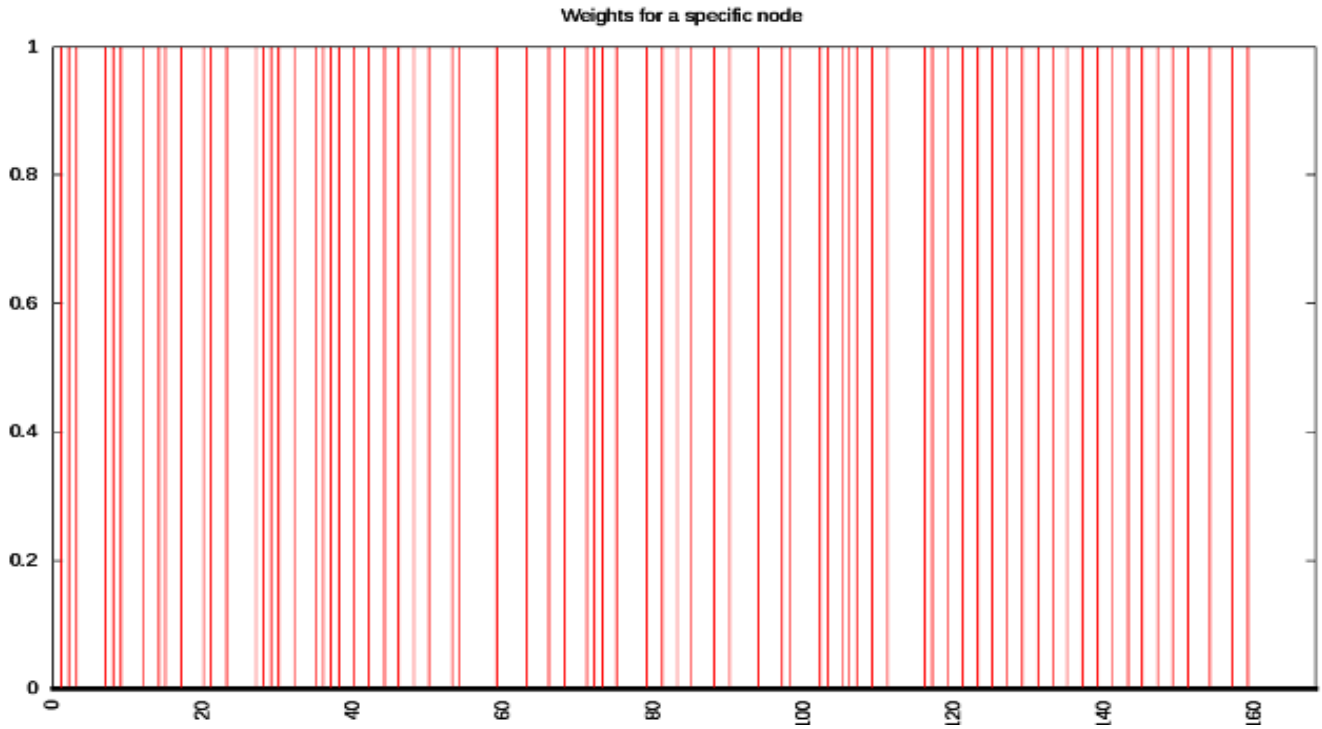
4.2.2.2.1 Αποτελέσματα Ασθενών

Ο νευρώνας [4][11] έχει ιδιαίτερη σημασία, καθώς αντιπροσωπεύει 41 ασθενείς. Συνεπώς, είναι ιδιαίτερα σημαντικό να απεικονιστούν τα βάρη του συγκεκριμένου

νευρώνα έτσι ώστε να προκύψουν διάφορα συμπεράσματα και να βρεθούν τα κοινά χαρακτηριστικά αυτών των 41 ασθενών. Αυτό που πρέπει να παρατηρηθεί είναι μια μικρή εμφάνιση κάποιων γειτονιών σε αντίθεση με το 50x50 πλέγμα, όπου η κατηγοριοποίηση ήταν σε επίπεδο νευρώνων και όχι γειτονιών, που ήταν και ο αρχικός στόχος της μελέτης αυτής, εξού και ο αλγόριθμος συνέπειας αναπτύχθηκε με την υπόθεση των ανεξάρτητων νευρώνων. Στην απεικόνιση των βαρών παρατηρούνται μόνο 0 και 1 τιμές κάτι που σημαίνει ότι για τα συγκεκριμένα άτομα που αναπαριστά ο συγκεκριμένος νευρώνας υπήρχε πολύ μεγάλη ομοιότητα.



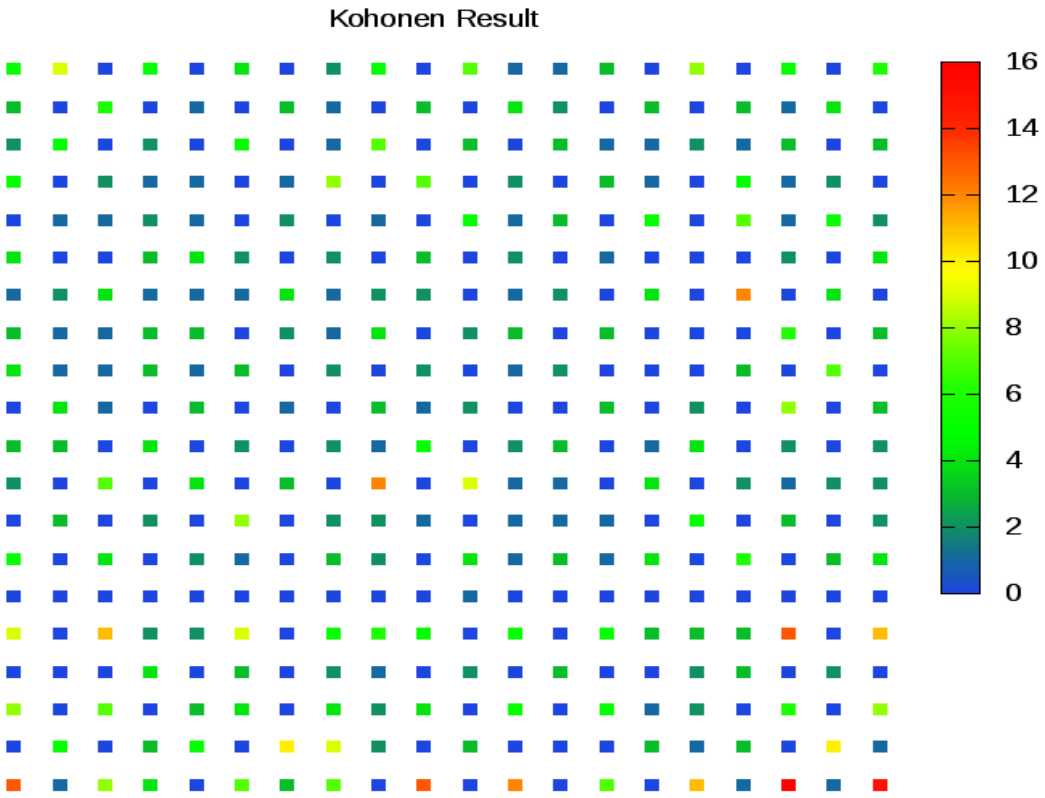
Εικόνα 4.12 Απεικόνιση Κατηγοριοποιήσης 20x20 Ασθενών



Εικόνα 4.13 Αναπαράσταση Βαρών Νευρώνα [4] [11] Πλέγματος 20x20 – Ασθενείς

4.2.2.2.2 Αποτελέσματα Υγιών

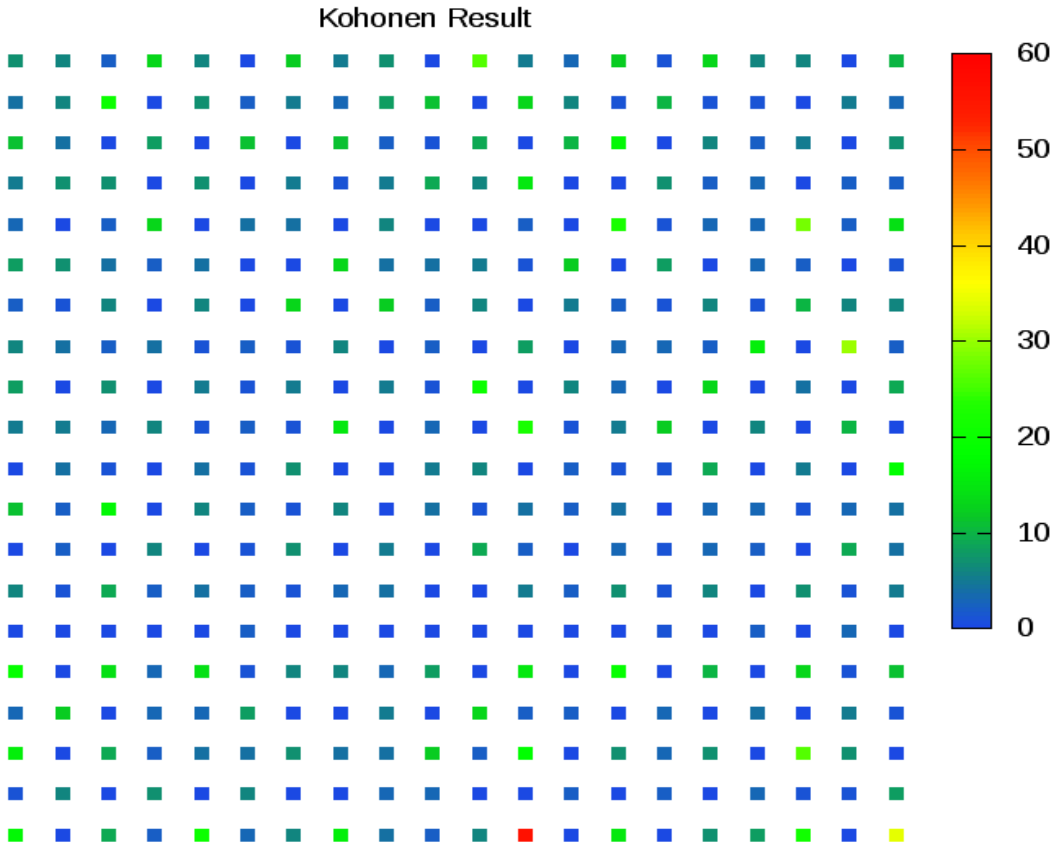
Στην εικόνα 4.14, παρουσιάζεται ο αριθμός των ατόμων που αντιπροσωπεύει ο κάθε νευρώνας. Οι αριθμοί αναπαριστώνται με βάση το χρώμα. Δεν θα παρουσιαστεί η απεικόνιση βαρών συγκεκριμένου κόμβου καθώς δεν έχει παρουσιαστεί κάποιος νευρώνας που να ξεχωρίζει. Παρατηρείται έντονα το φαινόμενο των γειτονιών, αλλά δεν δίνουν ιδιαίτερη πληροφορία καθώς ο κάθε νευρώνας αναπαριστά πολύ μικρό αριθμό ατόμων (μέγιστος 15) και γενικότερα τα άτομα έχουν διασκορπιστεί σε όλο το πλέγμα, κάτι που δείχνει δεν υπάρχει μεγάλη ομοιότητα μεταξύ των μη ασθενών.



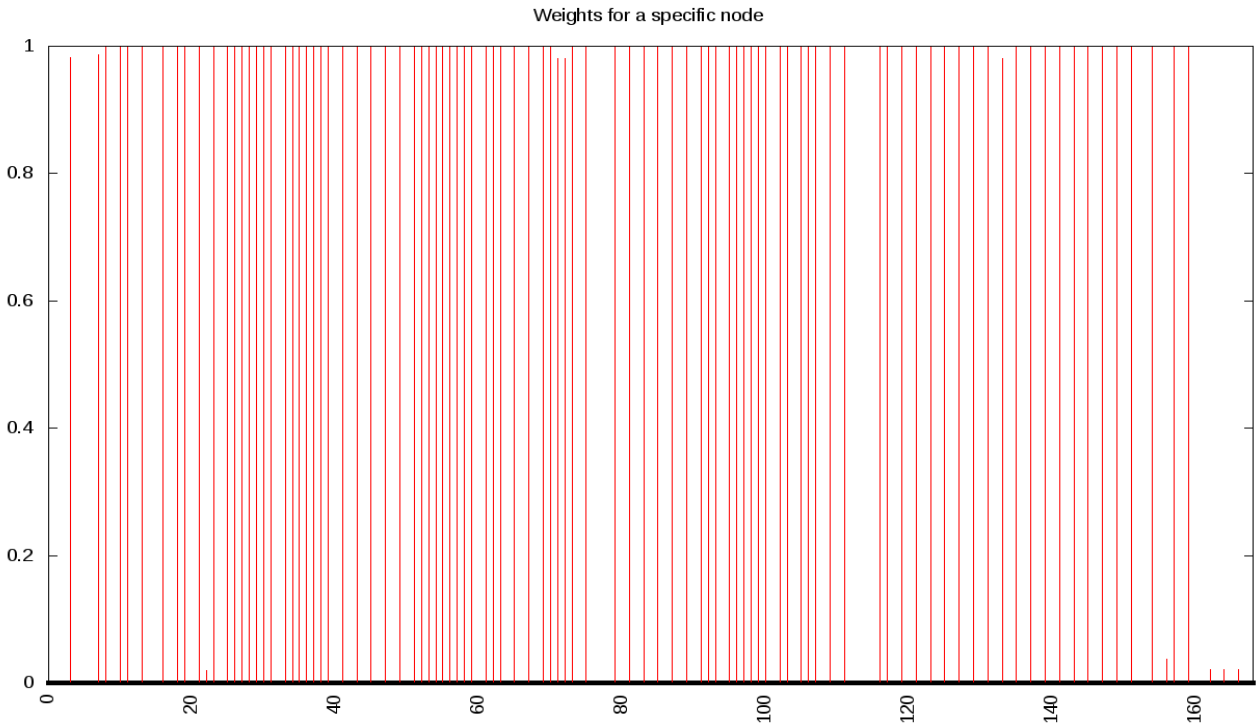
Εικόνα 4.14 Απεικόνιση Κατηγοριοποιήσης 20x20 Υγιών

4.2.2.2.3 Αποτελέσματα Συνδυασμού Ασθενών - Μη Ασθενών

Στην εικόνα 4.15, παρουσιάζεται ο αριθμός των ατόμων που αντιπροσωπεύει ο κάθε νευρώνας. Οι αριθμοί αναπαριστώνται με βάση το χρώμα. Η πιο σημαντική ομάδα παρατηρείται στο νευρώνα [0][11] που αντιπροσωπεύει 56 άτομα. Για το λόγο αυτό, αναπαριστώνται στην εικόνα 4.16, τα βάρη του νευρώνα αυτού. Και σε αυτή εικόνα αρχίζουν να διαφαίνονται οι γειτονιές, ένδειξη του ότι αρχίζει ο αλγόριθμος να ξεφεύγει από την προσέγγιση του εντελώς ανεξάρτητου νευρώνα και προσπαθεί να κατηγοριοποιήσει δίνοντας περισσότερη έμφαση στις γειτονιές.



Εικόνα 4.15 Απεικόνιση Κατηγοριοποιήσης 20x20 – Ασθενείς και Υγιείς



Εικόνα 4.16 Αναπαράσταση Βαρών Νευρώνα [4] [11] 20x20 – Ασθενείς και Υγιείς

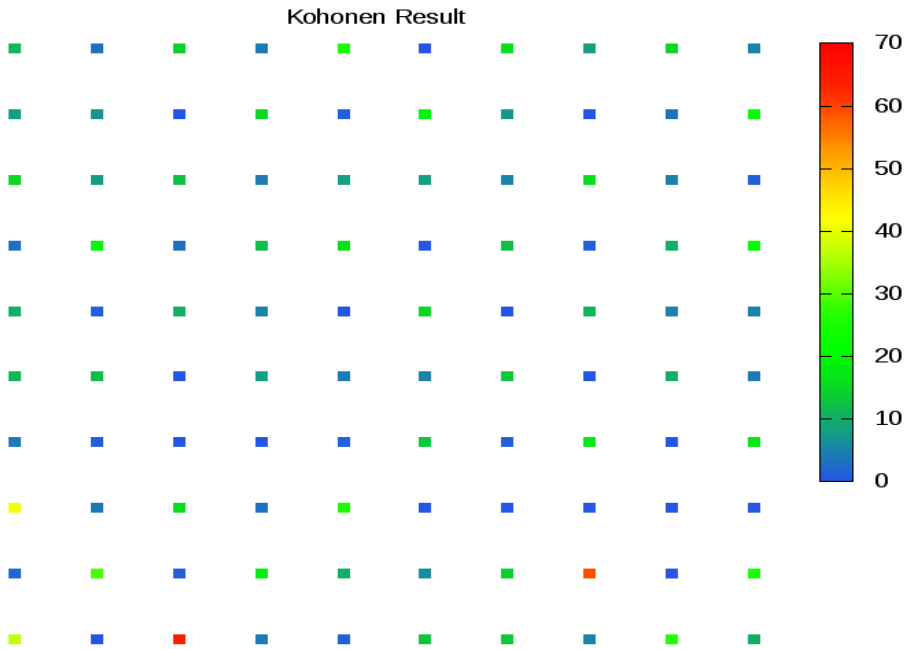
4.2.2.3 Κατηγοριοποίηση Εκτέλεσης 10x10 πλέγματος

Στις ακόλουθες εικόνες, παρουσιάζεται το αποτέλεσμα του αλγορίθμου σε ένα 10 x 10 πλέγμα. Θα παρουσιαστεί η κατηγοριοποίηση όπως προκύπτει σε εικονική μορφή.

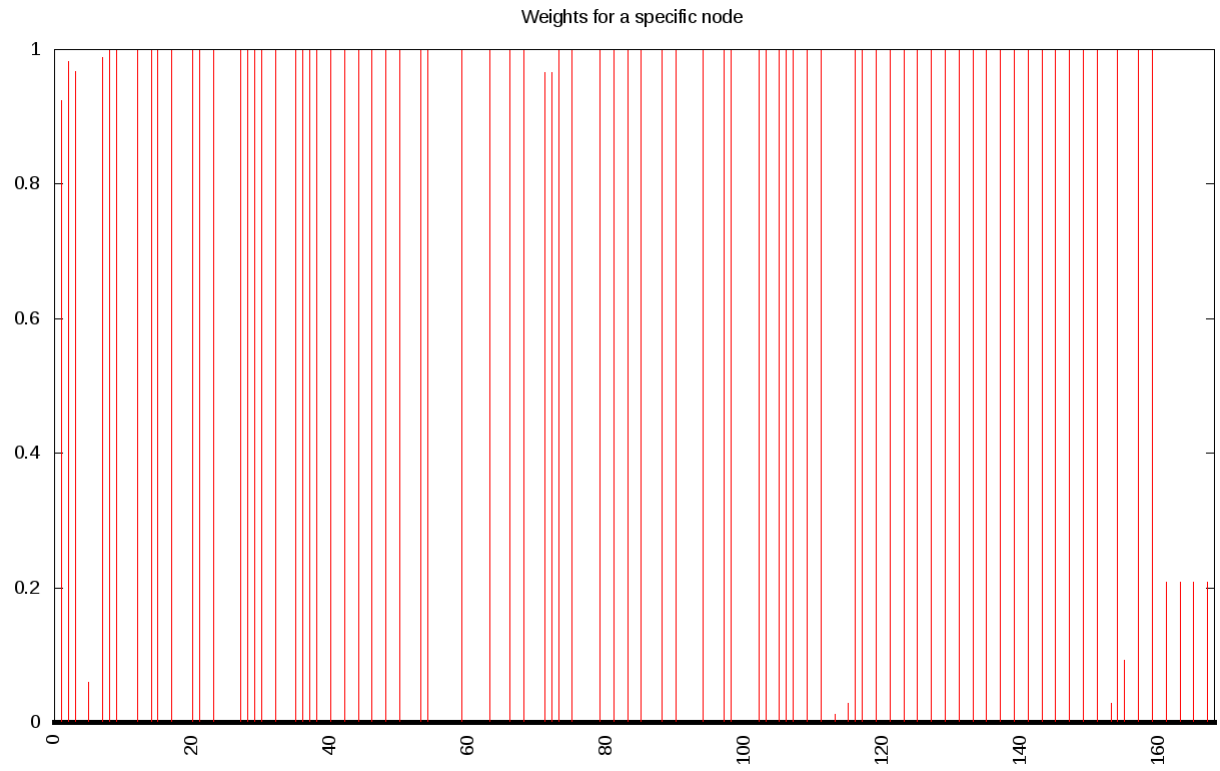
Τα αποτελέσματα που προκύπτουν εδώ διαφέρουν σε σύγκριση με τις άλλες κατηγοριοποιήσεις και αυτό οφείλεται στο ότι είναι μικρότερο το πλέγμα.

4.2.2.3.1 Αποτελέσματα Ασθενών

Στην εικόνα 4.17, παρουσιάζεται ο αριθμός των ατόμων που αντιπροσωπεύει ο κάθε νευρώνας. Οι αριθμοί αναπαριστώνται με βάση το χρώμα. Η πιο σημαντική ομάδα παρατηρείται στο νευρώνα [0][2] που αντιπροσωπεύει 64 άτομα. Για το λόγο αυτό, αναπαριστώνται στην εικόνα 4.18, τα βάρη του νευρώνα αυτού. Η σημαντική διαφορά με τις προηγούμενες αναπαραστάσεις βαρών, έγκειται στο ότι παρουσιάζονται και τιμές που δεν είναι ούτε 0 ούτε 1, κάτι που συνεπάγεται ότι λόγω της μείωσης του πλέγματος κάποια άτομα που κατηγοριοποιούνται μαζί, στα συγκεκριμένα SNPs δεν παρουσίαζαν όλοι τις ίδιες τιμές, κάτι που δεν συνεπάγεται πλήρης ομοιότητα. Επιπλέον έχουν την έντονη εμφάνιση γειτονιών, φαινόμενο που οφείλεται κυρίως στο ότι μικραίνει το πλέγμα και ξεφεύγουμε από την προσέγγιση του ανεξάρτητου νευρώνα.



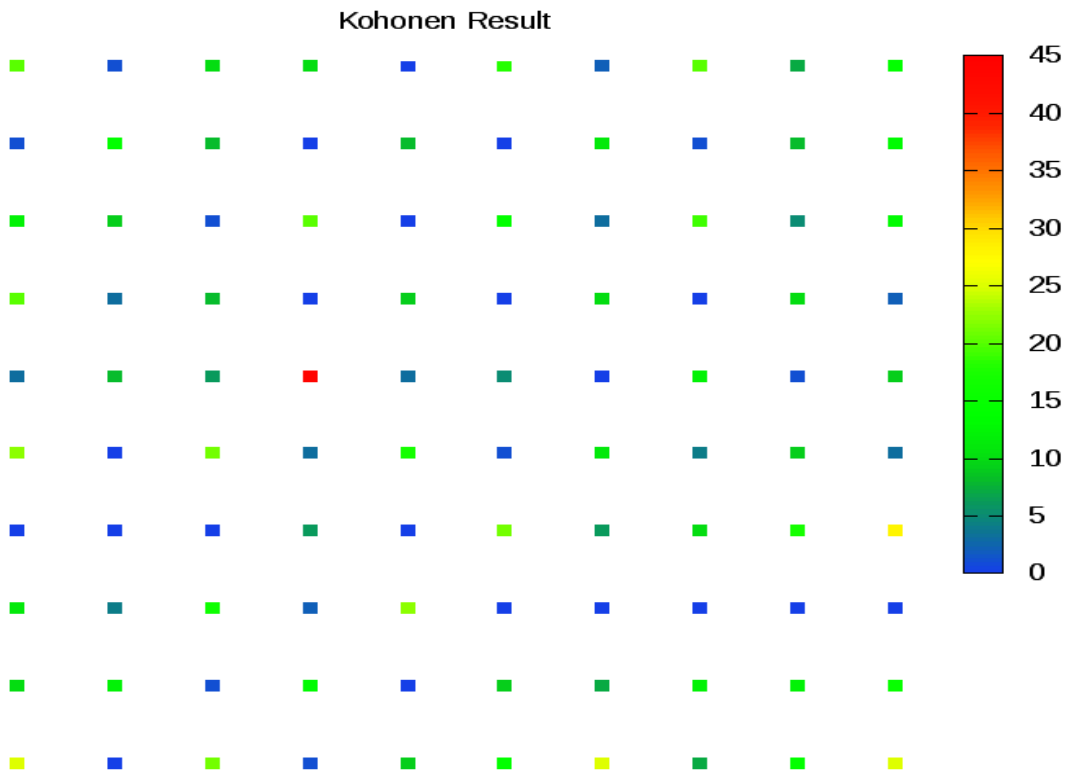
Εικόνα 4.17 Απεικόνιση Κατηγοριοποιήσης 10x10 Ασθενών



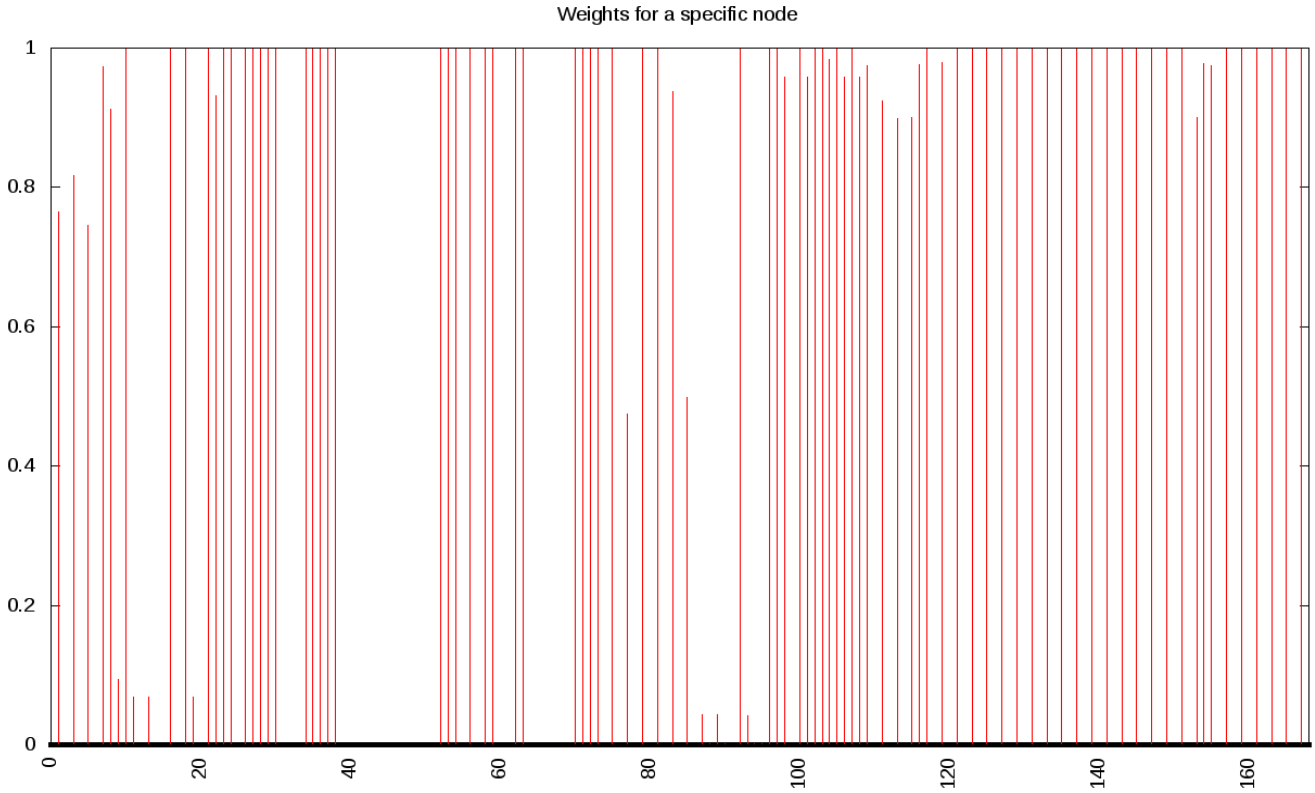
Εικόνα 4.18 Αναπαράσταση Βαρών Νευρώνα [0] [2] 10x10 – Ασθενείς

4.2.2.3.2 Αποτελέσματα Υγιών

Στην εικόνα 4.19, παρουσιάζεται ο αριθμός των ατόμων που αντιπροσωπεύει ο κάθε νευρώνας. Οι αριθμοί αναπαριστώνται με βάση το χρώμα. Η πιο σημαντική ομάδα παρατηρείται στο νευρώνα [5][3] που αντιπροσωπεύει 64 άτομα. Για το λόγο αυτό, αναπαριστώνται στην εικόνα 4.20, τα βάρη του νευρώνα αυτού. Το ίδιο φαινόμενο που παρουσιάστηκε στα βάρη του σημαντικότερου νευρώνα στα αποτελέσματα των ασθενών παρουσιάζεται και στους μη ασθενείς, δηλαδή της εμφάνισης τιμών που δεν είναι ούτε 0 ούτε 1 αλλά κάπου ενδιάμεσα. Επιπλέον έχουν και το φαινόμενο των γειτονιών, που παρατηρήθηκε και στους ασθενείς.



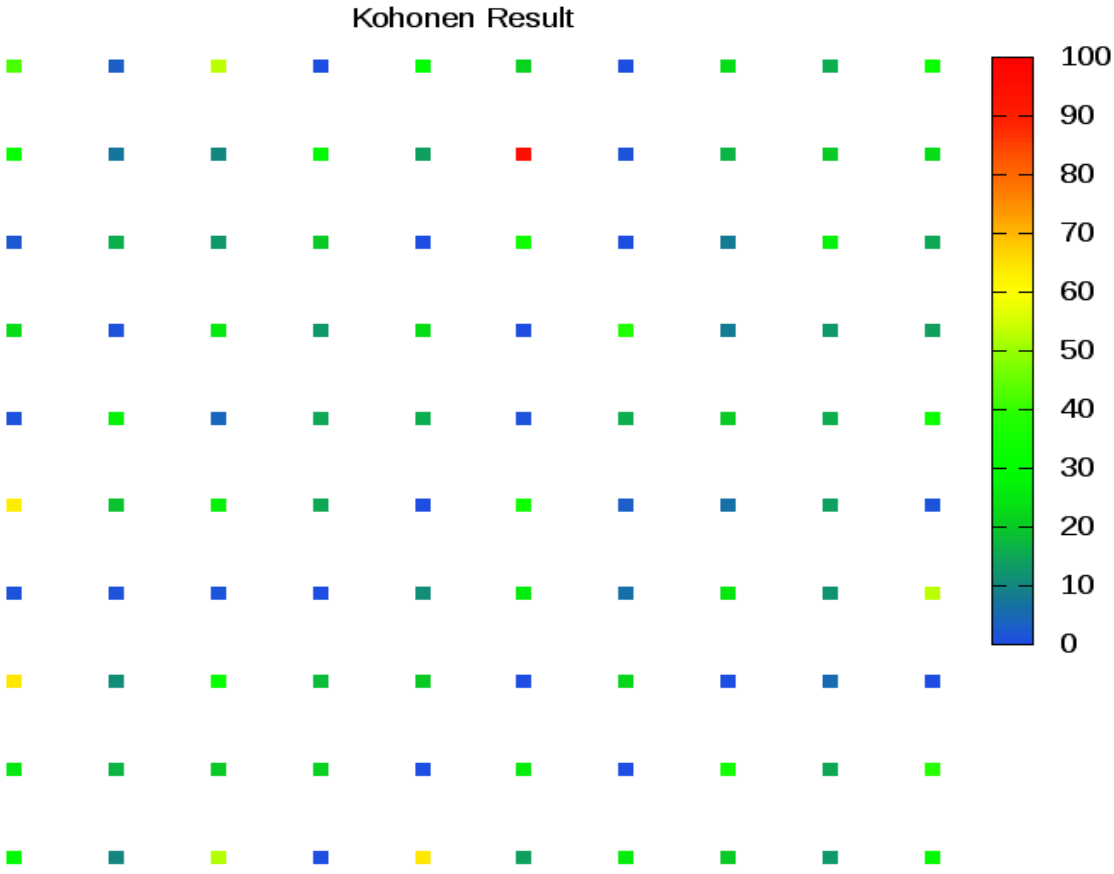
Εικόνα 4.19 Απεικόνιση Κατηγοριοποιήσης 10x10 Ασθενών



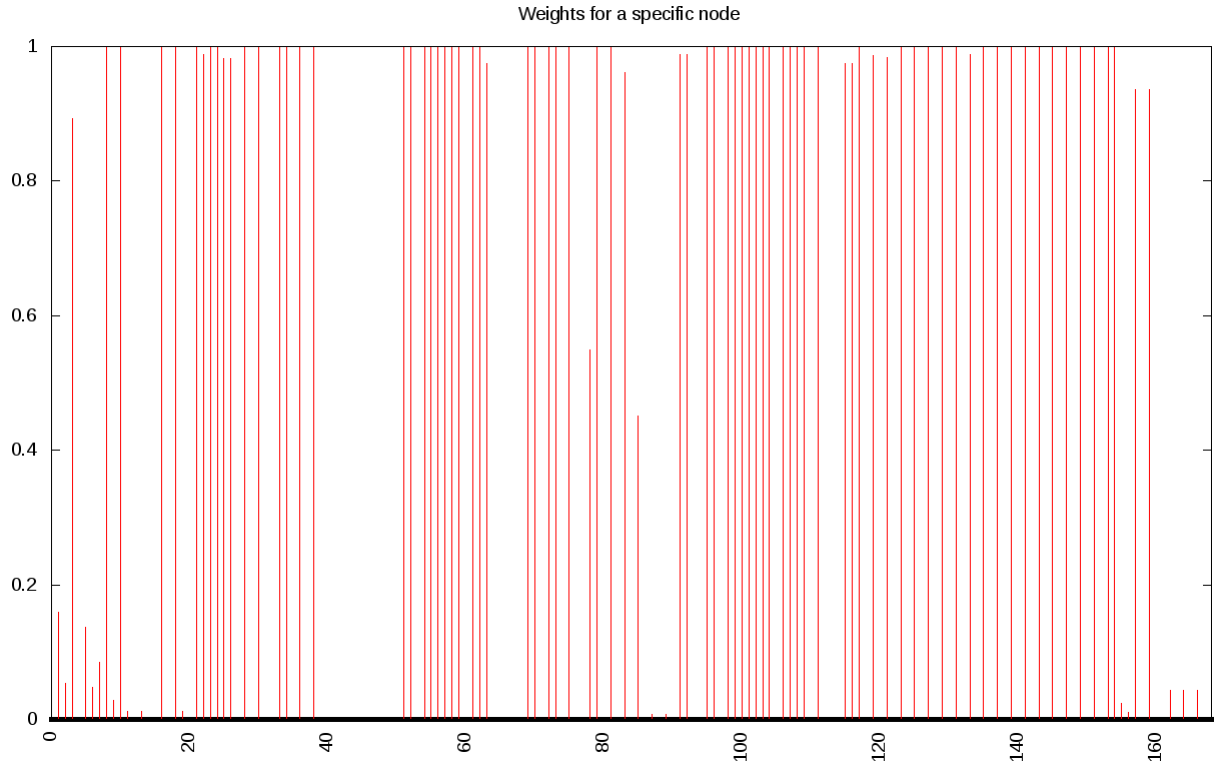
Εικόνα 4.20 Αναπαράσταση Βαρών Νευρώνα [5] [3] 10x10 – Υγιείς

4.2.2.3.3 Αποτελέσματα Συνδυασμού Ασθενών - Μη Ασθενών

Στην εικόνα 4.21, παρουσιάζεται ο αριθμός των ατόμων που αντιπροσωπεύει ο κάθε νευρώνας. Οι αριθμοί αναπαριστώνται με βάση το χρώμα. Η πιο σημαντική ομάδα παρατηρείται στο νευρώνα [8][5] που αντιπροσωπεύει 64 άτομα. Για το λόγο αυτό, αναπαριστώνται στην εικόνα 4.22, τα βάρη του νευρώνα αυτού. Το ίδιο φαινόμενο που παρουσιάστηκε στα βάρη του σημαντικότερου νευρώνα στα αποτελέσματα των ασθενών και μη ασθενών παρουσιάζεται και στους μη ασθενείς, δηλαδή της εμφάνισης τιμών που δεν είναι ούτε 0 ούτε 1 αλλά κάπου ενδιάμεσα. Επιπλέον έχουν και το φαινόμενο των γειτονιών, που παρατηρήθηκε και στους ασθενείς και μη ασθενείς. Γενικότερα τα αποτελέσματα εδώ αποτελούν ένα μίγμα των δύο προηγούμενων αποτελεσμάτων με την ένωση ομάδων που προέκυψαν στις δύο προηγούμενες περιπτώσεις.



Εικόνα 4.21 Απεικόνιση Κατηγοριοποιήσης 10x10 Ασθενών - Υγιών

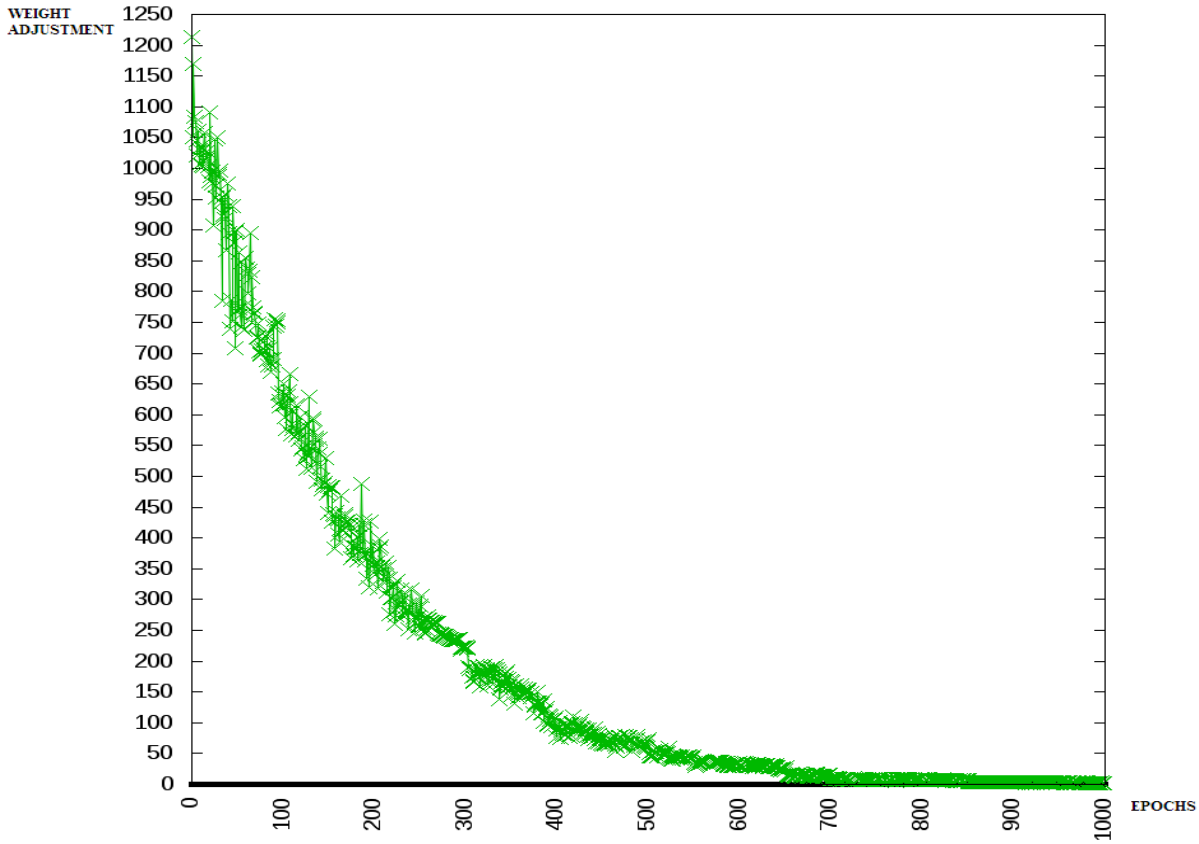


Εικόνα 4.20 Αναπαράσταση Βαρών Νευρώνα [8] [5] 10x10 Ασθενών - Υγιών

4.2.3 Δικαιολόγηση Παραμέτρων

4.2.3.1 Αριθμός Εποχών

Τα αποτελέσματα αλλαγής βαρών παρουσιάζονται στην εικόνα 4.21 που απεικονίζει τις τιμές αυτές στις διάφορες εποχές, με μέγιστο αριθμό εποχών τις 1000. Ο άξονας Χ αντικατοπτρίζει τον αριθμό των εποχών και ο άξονας Υ την μέση αλλαγή βαρών που παρατηρούνται στο δίκτυο. Ο τρόπος υπολογισμού των τιμών αυτών έχει ήδη εξηγηθεί στο κεφάλαιο 3.6.2 και 3.4.4.5.3



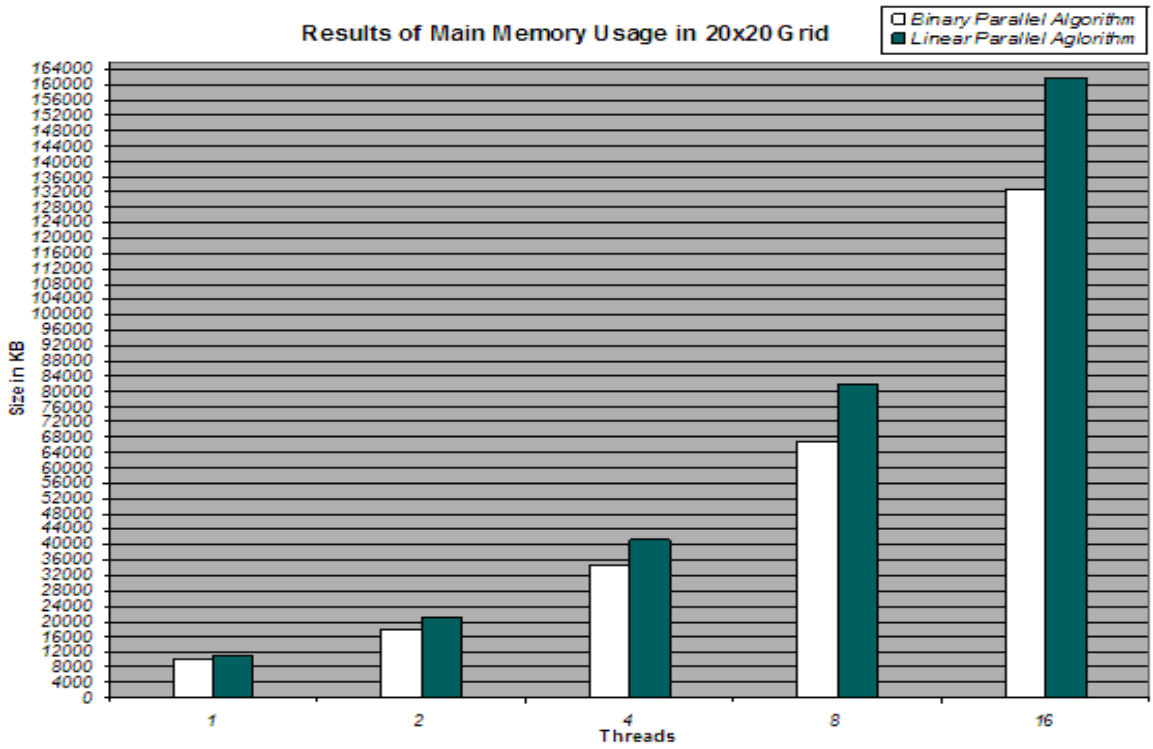
Εικόνα 4.21 Weight Adjustment από συγκεκριμένο run του αλγορίθμου.

4.2.4 Χρήση Κύριας Μνήμης στον Παράλληλο Αλγόριθμο

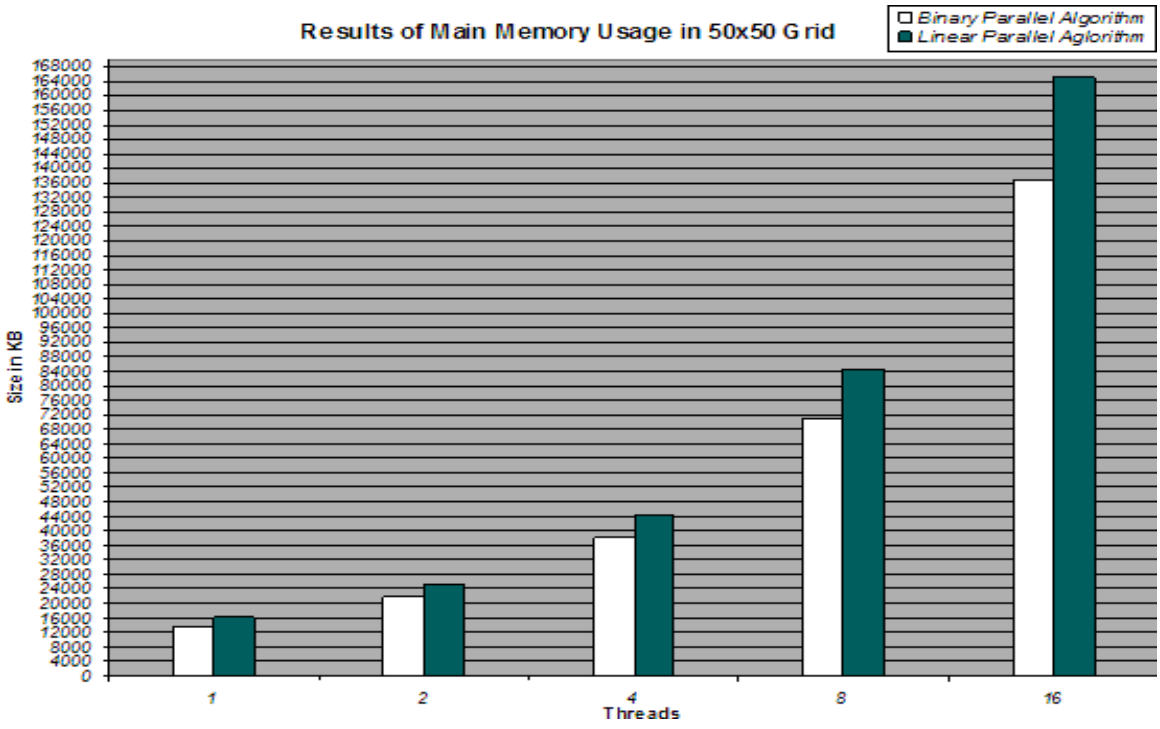
Στις εικόνες 4.22 και 4.23 παρουσιάζονται τα αποτελέσματα που προέκυψαν κατά την εκτέλεση των δύο παράλληλων αλγορίθμων δυαδικής μορφής και γραμμικής με 20x20 και 50x50 πλέγμα. Ο άξονας Χ αναπαριστά τις εκτελέσεις που έγιναν με βάση τον αριθμό των threads σε κάθε εκτέλεση και ο άξονας Υ το μέγεθος της μνήμης που χρειάστηκε μια συγκεκριμένη εκτέλεση σε KB. Με άσπρο χρώμα αναπαριστώνται οι εκτελέσεις του δυαδικού παράλληλου αλγόριθμου ενώ με πράσινο χρώμα αναπαριστώνται οι αντίστοιχες εκτελέσεις του γραμμικού παράλληλου αλγόριθμου.

Για την συλλογή των αποτελεσμάτων, χρησιμοποιήθηκε η εντολή pmap των Unix/Linux, η οποία με βάση τον μοναδικό process id, υπολογίζει το μέγεθος της μνήμης που χρησιμοποιεί η συγκεκριμένη διαδικασία.

Τα αποτελέσματα πρόεκυψαν με βάση το αρχείο των ασθενών με επίσταση 45 (μέγεθος 340KB)



Εικόνα 4.22 Σύγκριση Χρήσης Κύριας Μνήμης Δυαδικού και Γραμμικού Αλγόριθμου σε 20x20 Πλέγμα



Εικόνα 4.23 Σύγκριση Χρήσης Κύριας Μνήμης Δυαδικού και Γραμμικού Αλγόριθμου σε 50x50 Πλέγμα

4.3 Μεταεπεξεργασία

4.3.1 Αποτελέσματα των ομάδων – clusters και Συνέπεια Αλγορίθμου

Όπως έχει προαναφερθεί στο κεφάλαιο 3, αναπτύχθηκε ένας αλγόριθμος για την επιβεβαίωση της συνέπειας του αλγόριθμου, ο οποίος βρίσκει τις ομάδες που παρουσιάζονται πιο συχνά. Θα παρουσιαστούν αποτελέσματα από ασθενείς, υγιείς και συνδυασμού ασθενών και μη ασθενών από τις τρεις κατηγορίες πλεγμάτων που παρουσιάστηκαν πιο πάνω.

Ο αλγόριθμος συνέπειας παρουσιάζει πόσο συχνά εμφανίζονται συγκεκριμένες ομάδες – clusters σε κάποιες εκτελέσεις του παράλληλου αλγορίθμου. Οι έξοδοι του αλγορίθμου συνέπειας παρουσιάζονται στην εικόνα 4.24 και 4.25. Στην εικόνα 4.24 η δομή που ακολουθείται είναι η ακόλουθη: στην πρώτη στήλη ορίζεται ένα μοναδικό χαρακτηριστικό για την κάθε ομάδα. Η επόμενη στήλη παρουσιάζει τον αριθμό των φορών που εμφανίζεται η κάθε ομάδα. Οι επόμενες δύο στήλες περιέχουν τον αριθμό των ελάχιστων και μέγιστων ατόμων αντίστοιχα που βρέθηκαν να συμμετέχουν στις διάφορες εκτελέσεις. Τέλος, η τελευταία στήλη υπολογίζει την τυπική απόκλιση μεταξύ των δύο προαναφερθεισών τιμών. Η εικόνα 4.25 αποτελεί συνέχεια της εικόνας 4.24. Ουσιαστικά δείχνει ποια άτομα συμμετέχουν σε κάθε ομάδα.

Clusters_ID	Numbers_Identified	Min_Subjects	Max_Subjects	Standard_Deviation
0	6	5	5	0
1	7	2	2	0
2	3	12	12	0
3	7	1	1	0
4	4	2	2	0
5	5	5	5	0
6	8	1	1	0
7	6	12	12	0
8	3	14	14	0
9	9	5	5	0
10	3	23	23	0
11	2	3	3	0
12	3	4	4	0
13	2	2	2	0
14	8	2	2	0
15	5	1	1	0
16	5	2	2	0
17	4	3	3	0
18	3	1	1	0
19	3	4	4	0
20	7	1	1	0
21	10	5	5	0
22	2	4	4	0
23	4	5	5	0
24	5	2	2	0
25	8	4	4	0
26	6	6	6	0
27	7	2	2	0
28	7	3	3	0
29	5	8	8	0
30	4	5	5	0
31	2	6	6	0
32	5	5	5	0
33	2	2	2	0
34	5	1	1	0
35	4	2	2	0

Εικόνα 4.24 Παράδειγμα Εξόδου Αποτελέσματων Αλγορίθμου Συνέπειας

```
Cluster 0
363   411   536   753   835
Cluster 1
326   453
Cluster 2
41    45    321   391   465   537   613   670   736   768   837   938
Cluster 3
863
Cluster 4
87    695
Cluster 5
249   268   372   672   966
Cluster 6
47
Cluster 7
98    154   206   295   322   345   346   468   602   617   786   911
Cluster 8
215   220   354   374   410   501   528   616   661   729   755   879   900   905
Cluster 9
83    543   589   740   878
Cluster 10
152   175   196   250   288   302   317   337   349   437   473   477   490   508   576   652   686   689   693
Cluster 11
168   398   413
Cluster 12
299   497   644   747
Cluster 13
170   351
Cluster 14
492   950
Cluster 15
573
Cluster 16
147   875
Cluster 17
113   207   417
Cluster 18
381
Cluster 19
118   226   793   968
Cluster 20
```

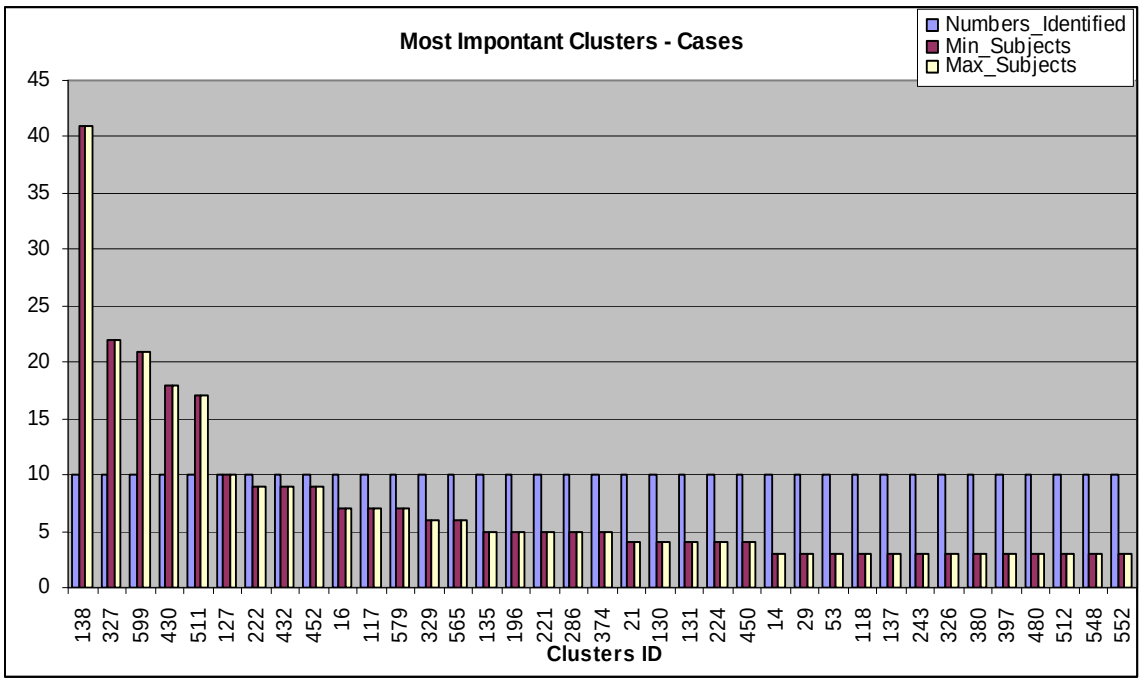
Εικόνα 4.25 Παράδειγμα Εξόδου Αποτελέσματων Αλγορίθμου Συνέπειας

4.3.1.1 Αποτελέσματα Κατηγοριοποίησης 50x50

Στις εικόνες 4.26, 4.27, 4.28 παρουσιάζονται οι ομάδες που παρουσιάστηκαν πιο συχνά στις εκτελέσεις που έγιναν σε 50x50 πλέγμα. Στον άξονα Χ, παρουσιάζονται οι μοναδικές τιμές που χαρακτηρίζουν κάθε ομάδα. Με μπλε χρώμα απεικονίζονται ο αριθμός των φορών που παρουσιάζονται, με κόκκινο ο ελάχιστος αριθμός ατόμων που παρατηρούνται στην αντίστοιχη ομάδα και με κίτρινο ο μέγιστος αριθμός. Τα ιστογράμματα αντικατοπτρίζουν τον αριθμό των φορών που εμφανίστηκαν σε 10 εκτελέσεις, τον ελάχιστο και μέγιστο αριθμό των ατόμων που παρουσιάστηκαν σε κάθε ομάδα αντίστοιχα. Το ποσοστό ομοιότητας ορίστηκε στο 95% στο πλέγμα αυτό καθώς λόγω του μεγάλου αριθμού κόμβων, δημιουργήθηκε μεγάλη ανεξαρτησία μεταξύ των νευρώνων και έτσι αναμένεται ότι τα άτομα που ανήκουν σε κάποιο νευρώνα πρέπει να μοιάζουν πάρα πολύ.

4.3.1.1.1 Αποτελέσματα Ασθενών

Στην εικόνα 4.26, παρουσιάζονται τα αποτελέσματα των ασθενών στο πλέγμα 50x50. Το πρώτο που πρέπει να σημειωθεί είναι ότι οι ομάδες που παρατηρούνται, εμφανίζονται σε όλες τις εκτελέσεις κάτι που υποδεικνύει η μπλε στήλη. Παρατηρούνται επιπλέον, κάποιες σημαντικές ομάδες όπως η ομάδα 138 που περιέχει 41 άτομα και παρουσιάζεται σχεδόν σε όλες τις μετέπειτα εκτελέσεις (με περισσότερα άτομα) και οι ομάδες 327 και 599 με 22 και 21 άτομα αντίστοιχα. Αυτές οι ομάδες, φαίνεται να παρουσιάζουν ιδιαίτερη αξία και πρέπει να μελετηθούν ξεχωριστά τα άτομα που τις αποτελούν και οι νευρώνες που τους αντιπροσωπεύουν για την εξαγωγή των ομοιοτήτων τους και περαιτέρω ανάλυση τους.

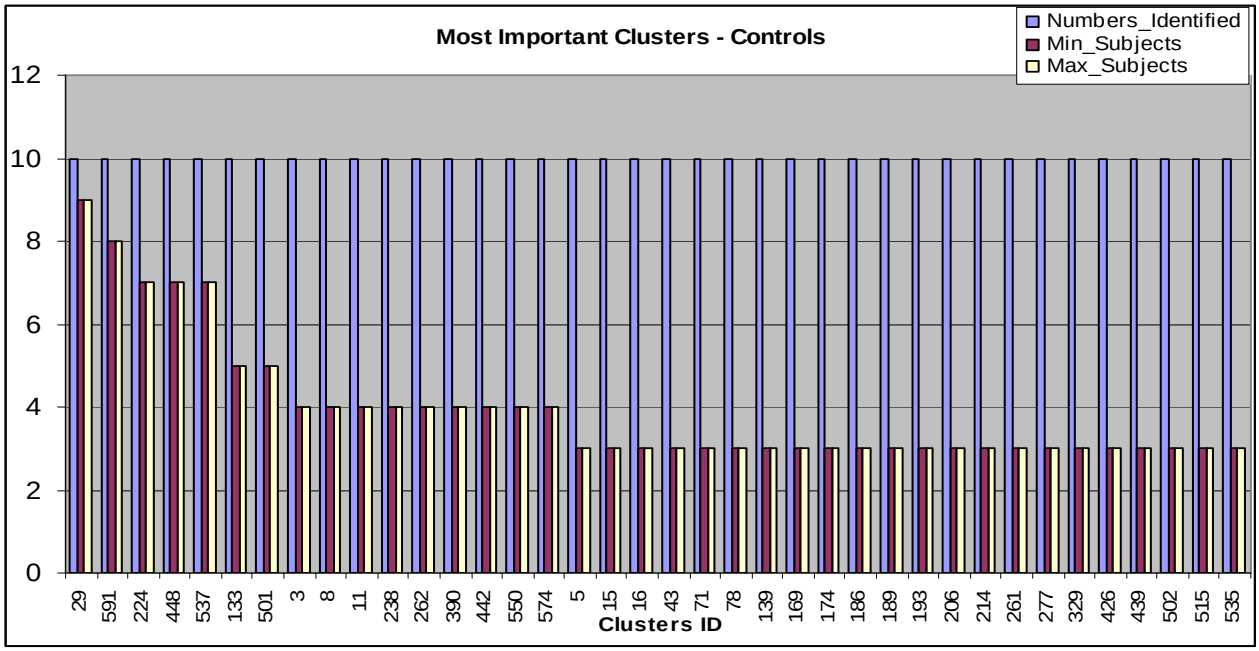


Εικόνα 4.26 Αναπαράσταση Ομάδων Ασθενών που παρουσιάζονται πιο συχνά 50x50

4.3.1.1.2 Αποτελέσματα Υγιών

Σε αντίθεση με τους ασθενείς, τα αποτελέσματα των μη ασθενών, που παρουσιάζονται στην εικόνα 4.27, δεν είναι καθόλου ενθαρρυντικά, καθώς, αν και οι ομάδες που παρατηρούνται εμφανίζονται σχεδόν σε όλες τις εκτελέσεις, εντούτοις είναι πολύ μικρές οι ομάδες αυτές (η μέγιστη ομάδα περιέχει μόλις 9 άτομα). Αποτέλεσμα το οποίο δεν μπορεί να χρησιμοποιηθεί για εξαγωγή συμπερασμάτων και οδηγεί στο

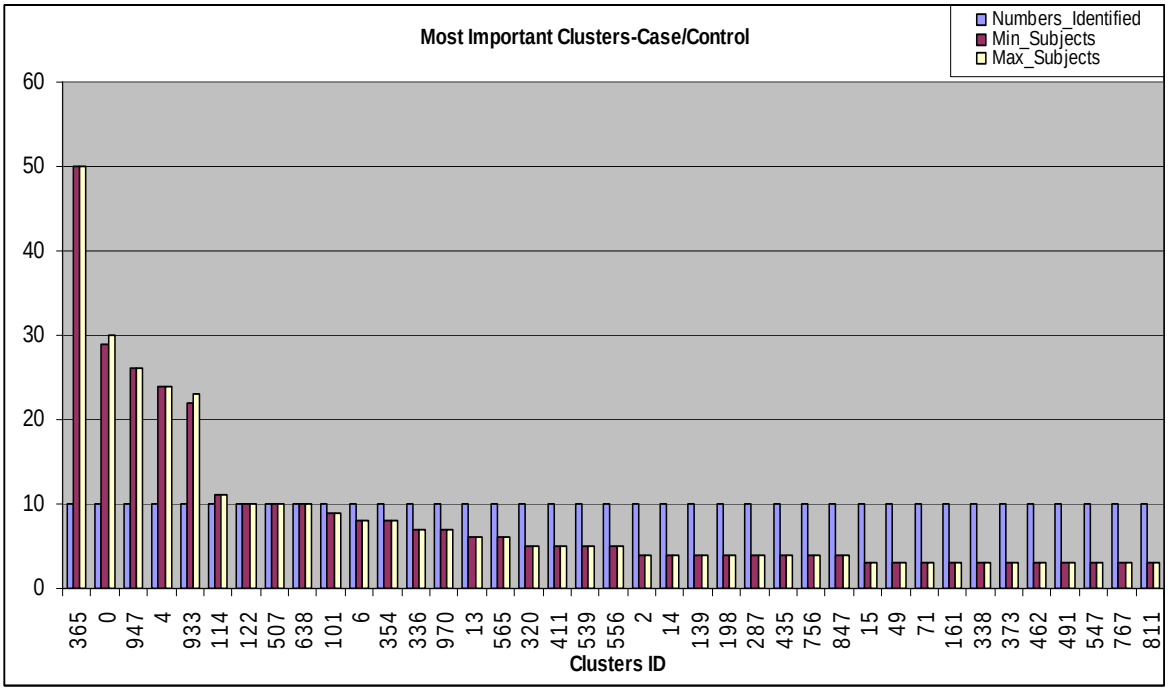
συμπέρασμα ότι στους μη ασθενείς δεν υπάρχει κάτι που να είναι ιδιαίτερα κοινό σε κάποια άτομα.



Εικόνα 4.27 Αναπαράσταση Ομάδων Υγιών που παρουσιάζονται πιο συχνά 50x50

4.3.1.1.3 Αποτελέσματα Συνδυασμού Ασθενών - Μη Ασθενών

Σημαντικά αποτελέσματα παρατηρούνται στον συνδυασμό των ασθενών και μη ασθενών. Αποτελέσματα τα οποία συμπίπτουν πάρα πολύ με τους ασθενείς, κάτι που οδηγεί στο συμπέρασμα ότι τα αποτελέσματα αυτά οφείλονται κυρίως στις συσχετίσεις μεταξύ των ασθενών, συμπέρασμα πολύ σημαντικό. Συγκεκριμένα παρατηρούνται οι ίδιες ομάδες που παρατηρούνται και στους ασθενείς, με την διαφορά ότι περιέχονται και κάποια επιπλέον άτομα σε σύγκριση με τις ομάδες που παρατηρήθηκαν στους ασθενείς, όπως για παράδειγμα οι ομάδες 365 και 0, οι οποίες είναι αντίστοιχες 138 και 327 των ασθενών. Συνεπώς η εικόνα 4.28 αποτελεί μια επιβεβαίωση των αποτελεσμάτων που παρουσιάστηκαν στην εικόνα 4.26 και επιβεβαιώνουν την ύπαρξη σημαντικής ομοιότητας ανάμεσα στους ασθενείς.



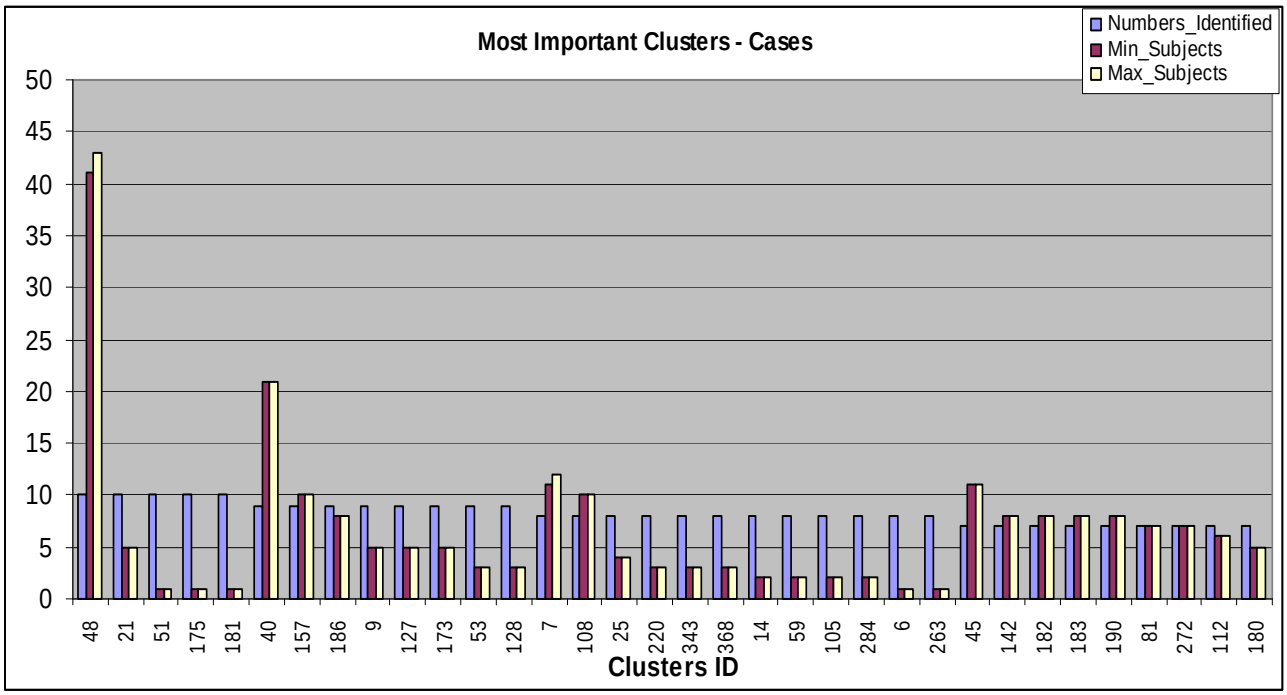
Εικόνα 4.28 Αναπαράσταση Ομάδων Ασθενών - Υγιών που παρουσιάζονται πιο συχνά 50x50

4.3.1.2 Αποτελέσματα Κατηγοριοποίησης 20x20

Στις εικόνες 4.29, 4.30, 4.31 παρουσιάζονται οι ομάδες που παρουσιάστηκαν πιο συχνά στις εκτελέσεις που έγιναν σε πλέγμα 50x50. Στον άξονα Χ, παρουσιάζονται οι μοναδικές τιμές που χαρακτηρίζουν κάθε ομάδα. Με μπλε χρώμα απεικονίζονται ο αριθμός των φορών που παρουσιάζονται, με κόκκινο ο ελάχιστος αριθμός ατόμων που παρατηρούνται στην αντίστοιχη ομάδα και με κίτρινο ο μέγιστος αριθμός. Τα ιστογράμματα αντικατοπτρίζουν τον αριθμό των φορών που εμφανίστηκαν σε 10 εκτελέσεις, τον ελάχιστο και μέγιστο αριθμό των ατόμων που παρουσιάστηκαν σε κάθε ομάδα αντίστοιχα. Το ποσοστό ομοιότητας ορίστηκε στο 90%, σε αντίθεση με το πλέγμα 50x50, καθώς μικραίνει αρκετά το πλέγμα και αναμένεται ότι κάποια άτομα τα οποία θα ομαδοποιούνταν σε διαφορετικό νευρώνα με μεγαλύτερο πλέγμα, θα μπουν σε άλλους νευρώνες που αντιπροσωπεύουν άτομα με πολύ μεγάλη ομοιότητα σε πιο μικρό πλέγμα, γι' αυτό μειώνεται και το ποσοστό ομοιότητας έτσι ώστε ο αλγόριθμος συνέπειας θα αγνοήσει αυτά τα άτομα και να βρει τις όμοιες ομάδες.

4.3.1.2.1 Αποτελέσματα Ασθενών

Στην εικόνα 4.29, παρουσιάζονται οι ομάδες που παρουσιάστηκαν πιο συχνά στις εκτελέσεις που έγιναν χρησιμοποιώντας το αρχείο των ασθενών. Παρατηρούνται κάποιες ενδιαφέροντες ομάδες όμως η ομάδα με μοναδικό χαρακτηριστικό το 48 που εμφανίζεται σε όλες τις εκτελέσεις και περιέχει 41 – 43 άτομα, η ομάδα 40 με 21 άτομα που εμφανίζεται 9 φορές και οι ομάδες 7 και 108 με 10-11 και 10 άτομα αντίστοιχα, οι οποίες εμφανίζονται 7 φορές. Τα άτομα των ομάδων αυτών συνεπάγεται ότι παρουσιάζουν μεγάλη ομοιότητα μεταξύ τους

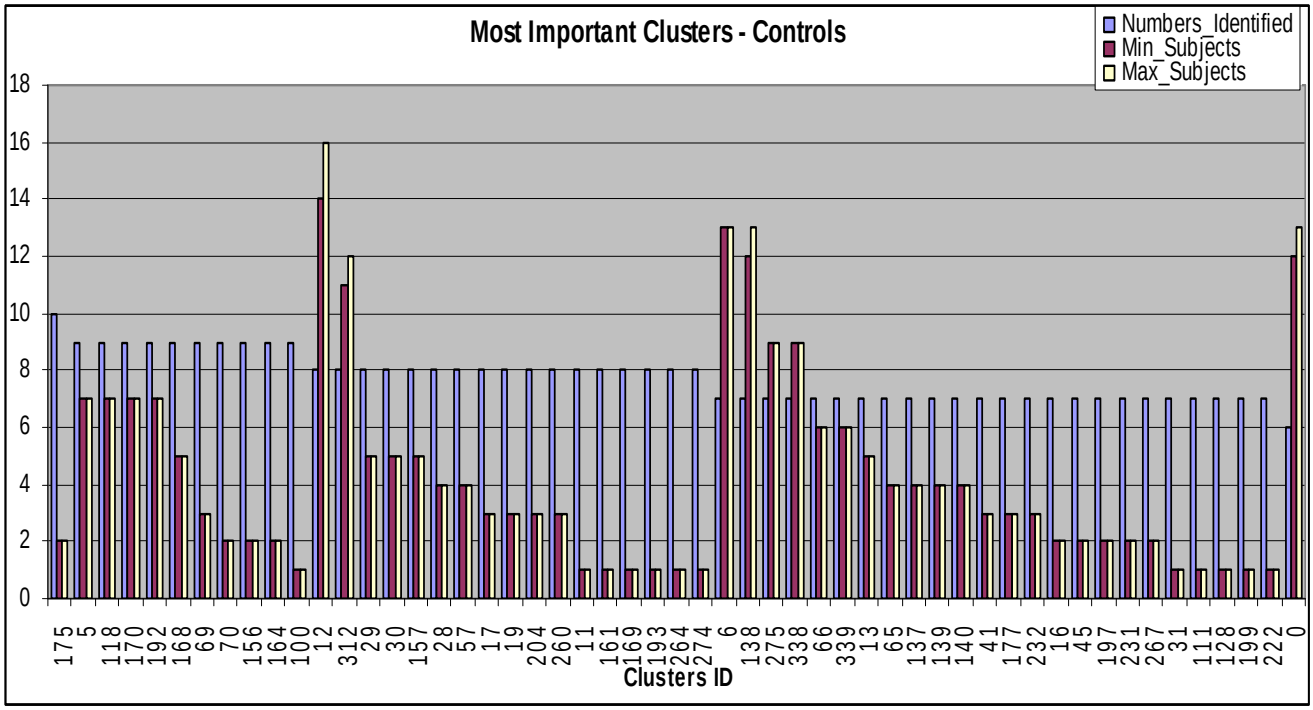


Εικόνα 4.29 Αναπαράσταση Ομάδων Ασθενών που παρουσιάζονται πιο συχνά 20x20

4.3.1.2.2 Αποτελέσματα Υγιών

Στην εικόνα 4.29, παρουσιάζονται οι ομάδες που παρουσιάστηκαν πιο συχνά στις εκτελέσεις που έγιναν χρησιμοποιώντας το αρχείο των ασθενών. Είναι η αντίστοιχη γραφική παράσταση της 4.28 με άτομα τα οποία δεν φέρουν την ασθένεια. Εδώ δεν παρουσιάζονται ενδιαφέροντες ομάδες. Ο μέγιστος αριθμός ατόμων σε μια ομάδα φτάνει μόλις τα 16 άτομα, κάτι που σίγουρα δεν είναι ικανοποιητικό. Επιπλέον, πολύ λίγες σημαντικές ομάδες παρουσιάζονται συχνά, καθώς η μεγαλύτερη ομάδα με 16 άτομα εμφανίζεται 8 φορές και από εκεί και πέρα η επόμενη μεγάλη ομάδα με 13 άτομα

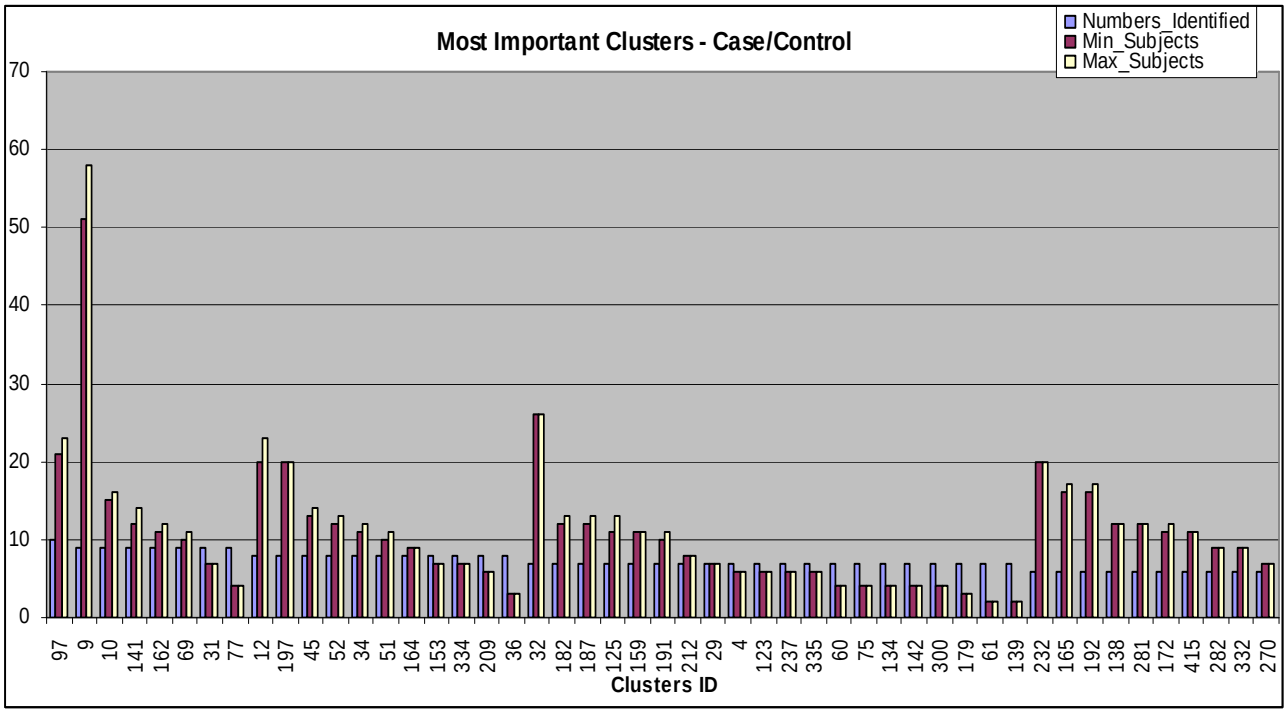
εμφανίζεται μόλις 7 φορές. Υπάρχουν κάποιες ομάδες που παρουσιάζονται αρκετά συχνά όμως τα άτομα που αντιπροσωπεύουν είναι ελάχιστα, κάτι που δεν δίνει σημαντικές πληροφορίες και συμπεράσματα για την μελέτη αυτή.



Εικόνα 4.30 Αναπαράσταση Ομάδων Υγιών που παρουσιάζονται πιο συχνά 20x20

4.3.1.2.3 Αποτελέσματα Συνδυασμού Ασθενών - Μη Ασθενών

Παρόμοια αποτελέσματα παρουσιάζονται δεδομένα ασθενών – μη ασθενών (εικόνα 4.31) συγκρινόμενα με την ομάδα των ασθενών. Υπάρχουν κάποιες καλές ομάδες που εμφανίζονται αρκετά συχνά όπως η ομάδα 9 που περιέχει 51 – 60 άτομα και εμφανίζεται 9 στις 10 εκτελέσεις Κάτι που σίγουρα είναι σημαντικό καθώς δείχνει ότι οι ασθενείς αυτής της ομάδας έχουν κάποια κοινά χαρακτηριστικά που μπορεί να οδηγήσουν σε σημαντικά ευρήματα. Υπάρχει επίσης η ομάδα 97 που περιέχει 21 – 23 και εμφανίζεται σε όλες τις εκτελέσεις, στοιχείο πολύ σημαντικό, όπως επίσης και η ομάδα 32 που περιέχει 26 άτομα και εμφανίζεται 8 φορές στις 10. Τέλος, υπάρχουν και οι ομάδες 232, 165,192 που εμφανίζονται 7 φορές και περιέχουν 20,16,16 άτομα αντίστοιχα, που επίσης είναι σημαντικές ομάδες. Γενικά, τα αποτελέσματα στην εικόνα 4.31 είναι ιδιαίτερα ενθαρρυντικά με τις ομάδες αυτές να χρήζουν περαιτέρω μελέτης.



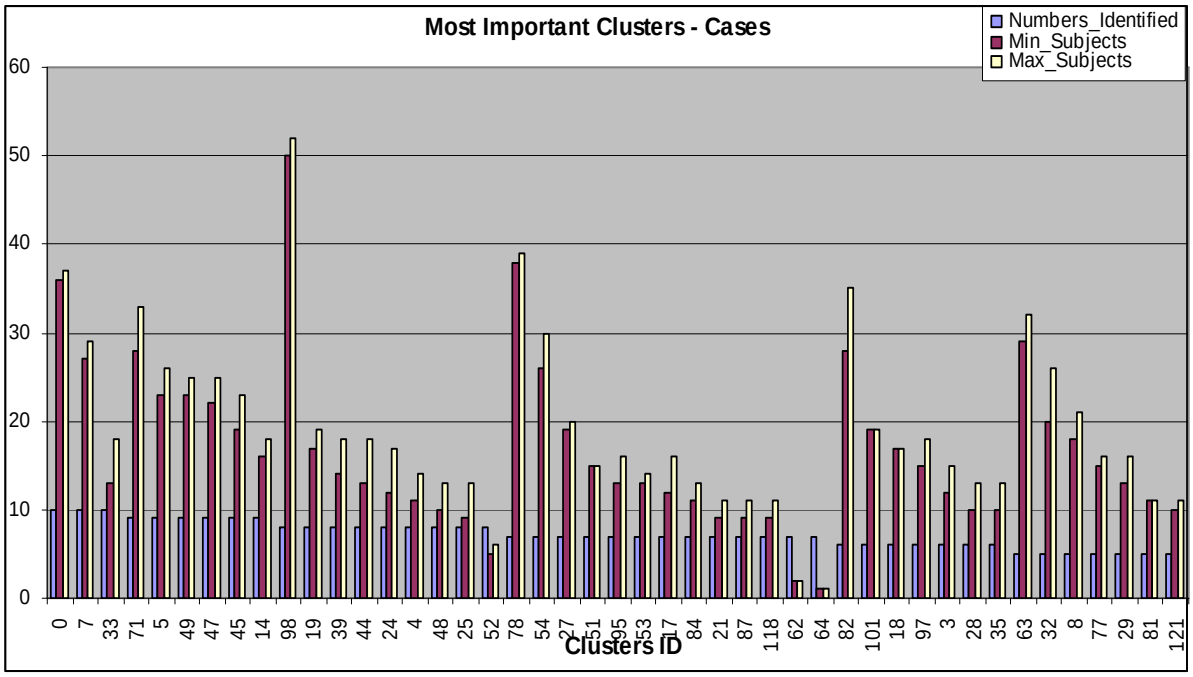
Εικόνα 4.31 Αναπαράσταση Ομάδων Ασθενών και Μη Ασθενών που παρουσιάζονται πιο συχνά 20x20

4.3.1.3 Αποτελέσματα Κατηγοριοποίησης 10x10

Στις εικόνες 4.29, 4.30, 4.31 παρουσιάζονται οι ομάδες που παρουσιάστηκαν πιο συχνά στις εκτελέσεις που έγιναν σε πλέγμα 10x10. Στον άξονα Χ, παρουσιάζονται οι μοναδικές τιμές που χαρακτηρίζουν κάθε ομάδα. Με μπλε χρώμα απεικονίζονται ο αριθμός των φορών που παρουσιάζονται, με κόκκινο ο ελάχιστος αριθμός ατόμων που παρατηρούνται στην αντίστοιχη ομάδα και με κίτρινο ο μέγιστος αριθμός. Τα ιστογράμματα αντικατοπτρίζουν τον αριθμό των φορών που εμφανίστηκαν σε 10 εκτελέσεις, τον ελάχιστο και μέγιστο αριθμό των ατόμων που παρουσιάστηκαν σε κάθε ομάδα αντίστοιχα. Το ποσοστό ομοιότητας ορίστηκε στο 80%, σε αντίθεση με το πλέγμα 50x50 (95%) και 20x20 (90%) , με την ίδια λογική ότι μειώνεται αρκετά το μέγεθος του πλέγματος και κάποια άτομα θα μεταφερθούν σε νευρώνες που αντιπροσωπεύουν άτομα με πολύ μεγάλη ομοιότητα στις διάφορες εκτελέσεις και ο αλγόριθμος δεν θα μπορέσει να αναγνωρίσει τις ομάδες στις οποίες έχουν προστεθεί άτομα λόγω του ότι είναι μικρό το πλέγμα, αν υπάρχει μεγάλο ποσοστό ομοιότητας.

4.3.1.3.1 Αποτελέσματα Ασθενών

Στην εικόνα 4.32, παρουσιάζονται τα αποτελέσματα των ασθενών σε 10x10 πλέγμα. Προκύπτουν αρκετές ομάδες, ειδικά η ομάδα 0 που παρουσιάζεται και στις κατηγοριοποιήσεις με άλλα πλέγματα, κάτι που αποδεικνύει την μεγάλη ομοιότητα των μελών αυτής της ομάδας. Επιπρόσθετα, λόγω και του μικρού μεγέθους του πλέγματος για το συγκεκριμένα δεδομένα, παρατηρούνται αρκετές ομάδες, οι οποίες χρήζουν περισσότερης ανάλυσης, όπως για παράδειγμα η ομάδα 7 με 28 άτομα και εμφανίζεται σε όλες τις εκτελέσεις, και η ομάδα 98 που αποτελεί την μεγαλύτερη ομάδα στις εκτελέσεις αυτές με 50 άτομα και εμφανίζεται 8 στις 10 φορές.

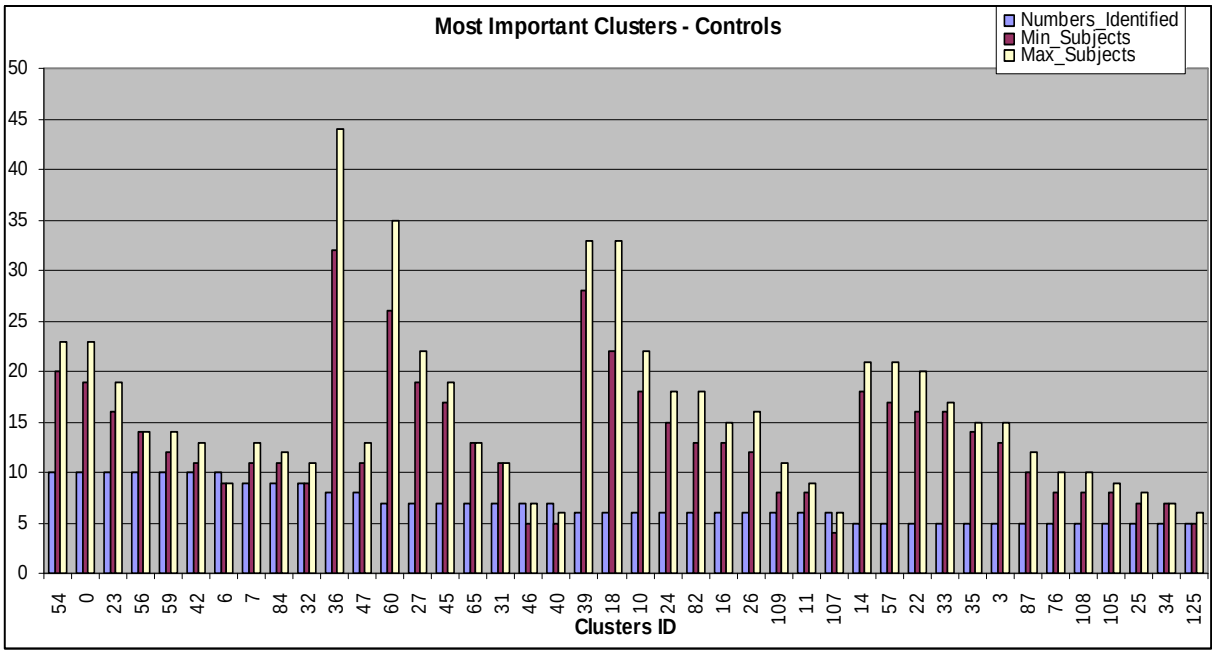


Εικόνα 4.32 Αναπαράσταση Ομάδων Ασθενών που παρουσιάζονται πιο συχνά 10x10

4.3.1.3.2 Αποτελέσματα Υγιών

Στην εικόνα 4.33, παρουσιάζονται τα αποτελέσματα των μη ασθενών σε 10x10 πλέγμα. Τα αποτελέσματα συμπίπτουν γενικά με τα αποτελέσματα που παρουσιάστηκαν και στα προηγούμενα κατηγοριοποιήσεις αν εξαιρεθούν οι 2 μεγάλες ομάδες που

παρουσιάζονται οι οποίες πολύ πιθανόν να οφείλονται στην μείωση του πλέγματος και του ποσοστού ομοιότητας. Προκύπτουν αρκετές ομάδες, ειδικά η ομάδα 0 που παρουσιάζεται και στις κατηγοριοποιήσεις με άλλα πλέγματα, κάτι που αποδεικνύει την μεγάλη ομοιότητα των μελών αυτής της ομάδας. Επιπρόσθετα, λόγω και του μικρού μεγέθους του πλέγματος για το συγκεκριμένα δεδομένα, παρατηρούνται αρκετές ομάδες, οι οποίες χρήζουν περισσότερης ανάλυσης, όπως για παράδειγμα η ομάδα 7 με 28 άτομα και εμφανίζεται σε όλες τις εκτελέσεις, και η ομάδα 98 που αποτελεί την μεγαλύτερη ομάδα στις εκτελέσεις αυτές με 50 άτομα και εμφανίζεται 8 στις 10 φορές.

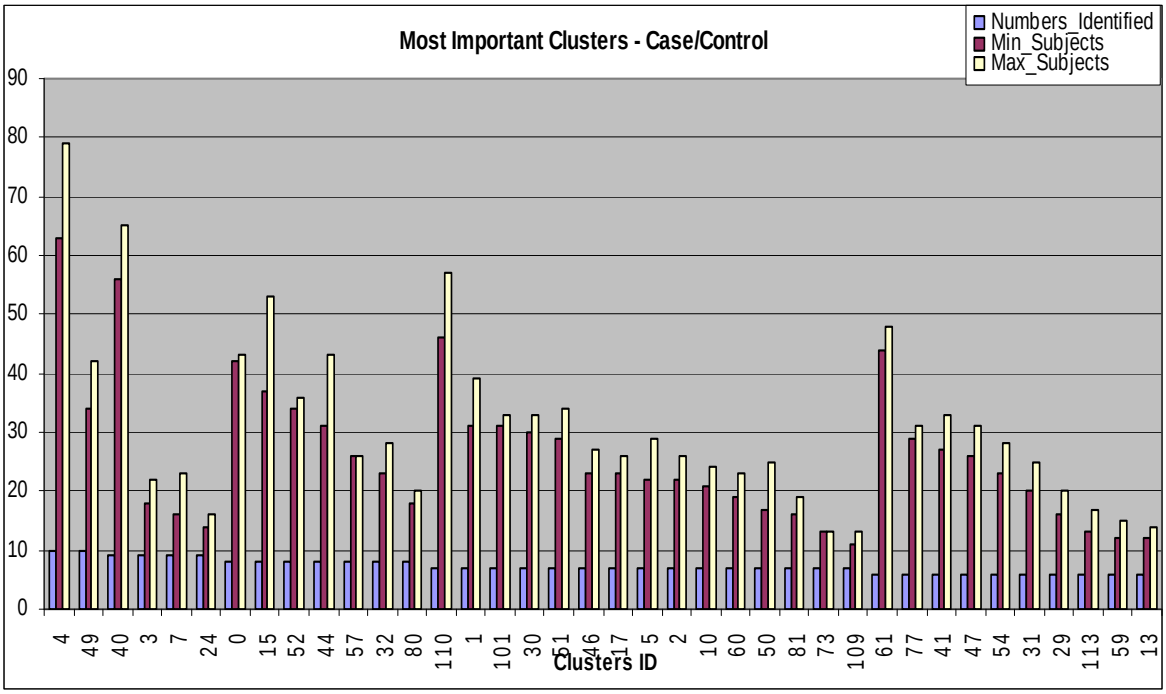


Εικόνα 4.33 Αναπαράσταση Ομάδων Μη Ασθενών που παρουσιάζονται πιο συχνά 10x10

4.3.1.3.3 Αποτελέσματα Συνδυασμού Ασθενών - Μη Ασθενών

Μεγάλη σημασία παρουσιάζουν τα αποτελέσματα του συνδυασμού ασθενών και μη ασθενών στην εικόνα 4.34. Παρουσιάζονται αρκετές ομάδες, σχεδόν σε όλες τις εκτελέσεις, με αρκετά άτομα π.χ. η ομάδα 4 με άτομα 63-79 και η ομάδα 40 με άτομα 55-65. Σίγουρα τα αποτελέσματα αυτά επηρεάζονται και από το μικρό πλέγμα και από την μείωση του ποσοστού ομοιότητας. Το σημαντικό είναι ότι οι ομάδες που παρατηρούνται και στις άλλες εκτελέσεις αποτελούν μέρη των ομάδων που έχουν

παρουσιαστεί και στο μικρότερο πλέγμα, κάτι που συνεπάγεται την ύπαρξη σημαντικής ομοιότητας των ατόμων που τις αποτελούν.



Εικόνα 4.34 Αναπαράσταση Ομάδων Ασθενών και Μη Ασθενών που παρουσιάζονται πιο συχνά 10x10

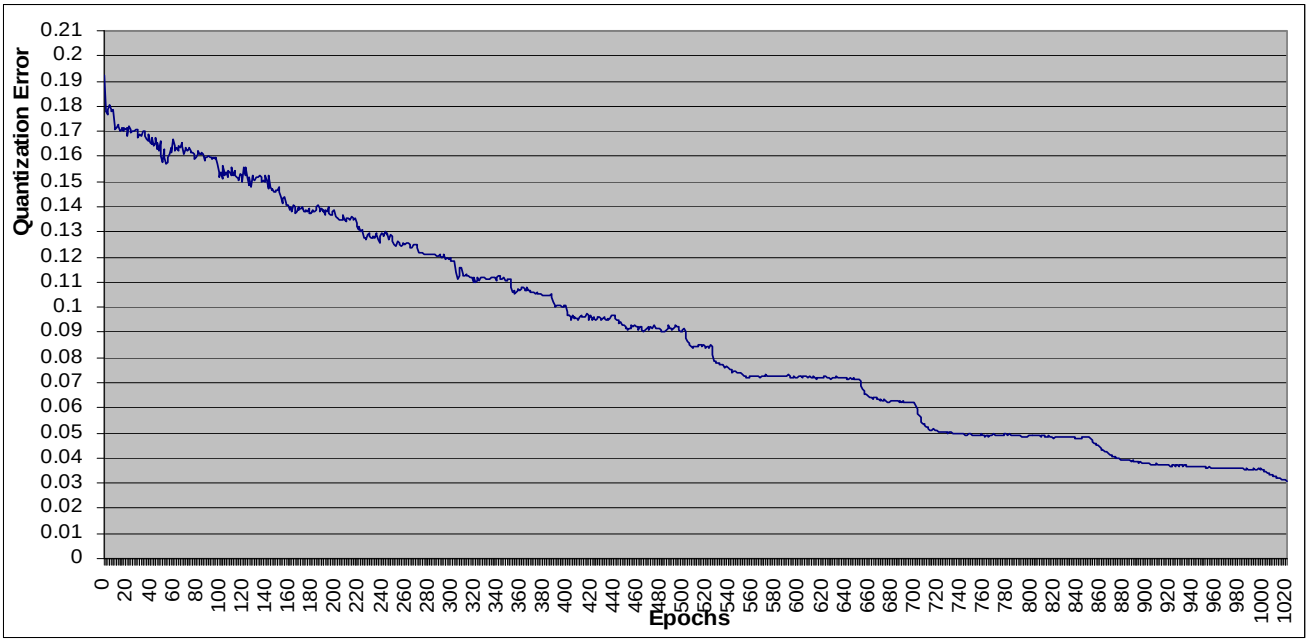
4.4 Τρόποι Σύγκρισης και Αξιολόγησης Αλγορίθμων

4.4.1 Αποτελέσματα Σφάλματος Κβαντισμού (Quantization Error)

Τα αποτελέσματα που θα παρουσιαστούν θα αφορούν την εκτέλεση του αλγορίθμου με 20x20 πλέγμα. Δεν υπάρχει λόγος να παρουσιαστούν και τα αποτελέσματα του σφάλματος για το 50x50 ή το 10x10 πλέγμα καθώς το σφάλμα μειώνεται ανάλογα, οπότε δεν δίνει κάποιο σημαντικό στοιχείο που αξίζει να παρατηρηθεί και να σχολιαστεί.

Τα αποτελέσματα παρουσιάζονται στην εικόνα 4.35. Στον άξονα Χ παρουσιάζεται ο αριθμός των εποχών και στο άξονα Υ το αντίστοιχο σφάλμα κβαντισμού. Παρατηρείται ξεκάθαρα ότι το σφάλμα μειώνεται σταδιακά με την αύξηση των εποχών. Τα αποτελέσματα δικαιολογούν πλήρως και την επιλογή του αριθμού των εποχών καθώς

μετά από τις 850 εποχές περίπου παρατηρείται μια σταθεροποίηση του σφάλματος, αν και συνεχίζει να μειώνεται. Απλά η μείωση είναι αρκετά μικρή σε σημείο που οδηγεί στο συμπέρασμα ότι δεν πρόκειται να μειωθεί αισθητά με μεγαλύτερο αριθμό εποχών. Οπότε, έγινε η επιλογή των 1000 εποχών έτσι ώστε να μπορέσει το δίκτυο να σταθεροποιήσει περίπου το σφάλμα και να φτάσει έτσι στην κατώτερη δυνατή τιμή, που όπως φαίνεται είναι πολύ μικρό της τάξεως 0.01 και είναι εξαιρετικά ικανοποιητικά τα αποτελέσματα.



Εικόνα 4.35 Αναπαράσταση Σφάλματος Κβαντισμού 20x20 πλέγματος

Κεφάλαιο 5

Συζήτηση

5.1 Συζήτηση Αποτελεσμάτων Προεπεξεργασίας

5.1.1 Συζήτηση Αποτελεσμάτων Δυαδικού Προτύπου για γενετικά δεδομένα – Συμπίεση

Όπως φαίνεται από τον πίνακα 4.1, η μείωση του μεγέθους των δεδομένων είναι σημαντική καθώς μειώνει περίπου 8 φορές το μέγεθος των βάσεων δεδομένων. Το αποτέλεσμα αυτό χρησιμοποιήθηκε στα πλαίσια της μελέτης αυτής για την εισαγωγή στο παράλληλο αλγόριθμο που αναπτύχθηκε πολύ μεγαλύτερου μεγέθους δεδομένα. Επιπρόσθετα, ο αλγόριθμος αυτός, έχει χρησιμοποιηθεί και για άλλους σκοπούς, καθώς ακολουθεί το μοντέλο του ανοικτού κώδικα και συνεπώς μπορεί να χρησιμοποιηθεί και

να τροποποιηθεί από οποιοδήποτε κάτω από ορισμένη άδεια χρήσης. Ήδη έχει χρησιμοποιηθεί από διάφορα άτομα. Όπως γίνεται αντιληπτό, ο αλγόριθμος και το πρωτόκολλο που έχει προταθεί [17], παρουσίασε καλά αποτελέσματα και χρησιμοποιήθηκε ιδιαίτερα, τόσο στα πλαίσια της μελέτης αυτής, όσο και σε εφαρμογές που δεν ανήκουν στο φάσμα της μελέτης αυτής.

5.2 Συζήτηση Αποτελεσμάτων Παράλληλου Αλγόριθμου

5.2.1 Συζήτηση Αποτελεσμάτων με διάφορα threads

Μελετώντας τα αποτελέσματα χρόνου, παρατηρείται ιδιαίτερη μείωση του απαιτούμενου συνολικού χρόνου εκτέλεσης. Τα σημαντικό γεγονός όμως είναι ότι δεν έχει επιτευχθεί γραμμική μείωση του χρόνου και ο λόγος είναι οι συχνοί συγχρονισμού που απαιτούνται από τον αλγόριθμο, που δεν μπορούσαν να αποφευχθούν. Συγκεκριμένα, σε κάθε επεξεργασία ενός διανύσματος εισόδου απαιτούνται 2 τουλάχιστον συγχρονισμοί (συγχρονισμός στα πλαίσια της μελέτης αυτής ορίζεται ένα φράγμα – barrier όπου όλα τα threads όταν φτάσουν σε ένα σημείο στον κώδικα, πρέπει να περιμένουν όλα τα άλλα threads να φτάσουν στο ίδιο σημείο, ώστε να μπορέσουν να συνεχίσουν την εκτέλεση τους). Ο πρώτος συγχρονισμός απαιτείται κατά το σημείο όπου όλοι θα υπολογίσουν τον BMU των νευρώνων για τους οποίους είναι υπεύθυνοι (το αποτέλεσμα του κάθε thread αποθηκεύεται σε μια θέση ενός πίνακα) καθώς πρέπει όλα τα threads να γεμίσουν τον πίνακα αυτό για να μπορέσει να υπολογιστεί ο BMU ολόκληρου του δικτύου. Ο δεύτερος συγχρονισμός απαιτείται μετά τον υπολογισμό του γενικού BMU που αναλαμβάνει ένα thread (αλγόριθμος ελαχίστου σε ένα πίνακα). Τα threads εδώ πρέπει να συγχρονιστούν περιμένοντας το thread που αναλαμβάνει τον υπολογισμό, να ολοκληρώσει την εργασία του έτσι ώστε να έχουν σωστή ενημέρωση για το ποιος νευρώνας είναι ο BMU για να αλλάξουν αναλόγως τα βάρη τους. Συνεπώς δεν ήταν δυνατό να αποφευχθούν οι 2 συγχρονισμοί αυτοί, αν και προκαλούν μεγάλο υπολογιστικό κόστος (αύξηση χρόνου εκτέλεσης) στον παράλληλο αλγόριθμο. Γι' αυτό, δεν επιτυγχάνεται γραμμική μείωση του χρόνου εκτέλεσης.

Το προαναφερθέν πρόβλημα εμφανίζεται και στα αποτελέσματα της επιτάχυνσης. Δεν επιτυγχάνεται γραμμική επιτάχυνση λόγων των συγχρονισμών. Αυτό που πρέπει να

προαναφερθεί εδώ είναι οι επιταχύνσεις που παρουσιάζονται κατά την χρήση 1 και 2 thread αντίστοιχα. Επιταχύνσεις που ξεπερνούν τον αριθμό των threads κάτι που συνήθως σημαίνει σουπερ γραμμική επιτάχυνση. Όμως το φαινόμενο αυτό δεν συνεχίζεται και στις επόμενες εκτελέσεις και οφείλονται κυρίως σε βελτιστοποιήσεις που έγιναν στον κώδικα κατά την μετατροπή του σειριακού σε παράλληλου έτσι ώστε να επιτευχθεί όσο τον δυνατό μεγαλύτερη χρήση των επεξεργαστών, μειώνοντας έτσι τον συνολικό χρόνο εκτέλεσης.

Όπως φαίνεται ξεκάθαρα υπάρχει σοβαρή μείωση του χρόνου αν ληφθεί υπόψη ότι χρησιμοποιήθηκε μηχανή με οκτώ επεξεργαστές. Συνεπώς με περισσότερο αριθμό επεξεργαστών σίγουρα αναμένονται καλύτερα αποτελέσματα.

Επιπλέον, υπάρχει περίπτωση να παρατηρηθεί στον αλγόριθμο ανισορροπία στην ανάθεση φόρτου (work load) στα threads καθώς στα threads ανατίθενται στατικά οι νευρώνες για τους οποίους είναι υπεύθυνα, και έτσι πολύ πιθανόν, κάποια threads να πρέπει να εκτελέσουν πολύ περισσότερες πράξεις, ειδικότερα στο σημείο που αποφασίζεται ποιοι νευρώνες πρέπει να προσαρμόσουν τα βάρη τους. Πρόβλημα που μπορεί να λυθεί μόνο με δυναμική ανάθεση φόρτου, κάτι που θα συνεπάγεται επιπλέον κόστος (overhead) κατά την φάση της ανάθεσης εργασιών στα threads. Συνεπώς παρατηρείται ένα δίλημμα (trade off) μεταξύ στατικής και δυναμικής ανάθεσης, που όμως στην περίπτωση αυτή, αναμένεται ότι το κόστος δυναμικής ανάθεσης θα είναι πολύ μικρό σε σύγκριση με τη σωστή ανάθεση φόρτου.

5.2.2 Συζήτηση Αποτελεσμάτων Τριών Συγκεκριμένων Εκτελέσεων

Σχολιάζοντας τα αποτελέσματα που προκύπτουν από τις κατηγοριοποιήσεις που παρουσιάστηκαν στο προηγούμενο κεφάλαιο, μπορούν να εξαχθούν πολλά συμπεράσματα.

Αρχικά, η πρώτη παρατήρηση που προκύπτει από τις τρεις εκτελέσεις, είναι όσο μικραίνει το μέγεθος του πλέγματος, τόσο πρέπει να μικραίνει το ποσοστό ομοιότητας. Η παρατήρηση αυτή οφείλεται στο γεγονός ότι όσο μικραίνει το πλέγμα, κάποια άτομα, τα οποία θα κατηγοριοποιούνταν σε ξεχωριστούς νευρώνες σε μεγάλα πλέγματα, θα

μεταφερθούν αναγκαστικά σε άλλους νευρώνες ώστε να «χωρέσουν» στο πλέγμα και ουσιαστικά θα δημιουργηθούν ομάδες που δεν έχουν τόση μεγάλη ομοιότητα όσο θα είχαν σε ένα πιο μεγάλο πλέγμα. Συνεπώς ο αλγόριθμος συνέπειας , πρέπει να «αγνοήσει» με κάποιο τρόπο αυτά τα επιπλέον άτομα, τα οποία κατηγοριοποιήθηκαν σε κάποιες ομάδες εξ ανάγκης, γι' αυτό αναγκαστικά πρέπει να μειωθεί το ποσοστό ομοιότητας έτσι ώστε να αναγνωριστούν οι πραγματικές ομάδες και να μην αγνοηθούν ομάδες που έχουν μεγαλώσει.

Επιπλέον, σαν γενική παρατήρηση, τα πιο σημαντικά αποτελέσματα παρατηρούνται στις ομάδες των ασθενών και στις τρεις εκτελέσεις. Παρατηρούνται κάποιες ομάδες ατόμων που είναι κοινές σε όλες τις εκτελέσεις των ασθενών (αν και σε πιο μικρά πλέγματα περιέχονται και άλλα άτομα στις ομάδες αυτές), αν και σχετικά με τον αριθμό των ασθενών, αναμενόταν η δημιουργία μεγαλύτερων ομάδων. Το ίδιο ισχύει και για τα αποτελέσματα του συνδυασμού των ασθενών και μη ασθενών, καθώς παρατηρήθηκε ότι παρουσιάζονται πολύ παρόμοια αποτελέσματα με αυτά των ασθενών σε όλα τα πλέγματα (παρουσιάζονται μεγαλύτερες ομάδες καθώς κάποιοι μη ασθενείς ομαδοποιούνται με τους ασθενείς), κάτι που ενισχύει την πεποίθηση ότι, παρά τις σχετικά μικρές ομάδες, υπάρχει σημαντική ομοιότητα μεταξύ των ατόμων που αποτελούν τις ομάδες αυτές και σίγουρα είναι πολύ σημαντικό εύρημα, που χρήζει περαιτέρω μελέτης και ανάλυσης. Όσο αφορά τους μη ασθενείς, εδώ γενικά και στις τρεις εκτελέσεις δεν έχουν προκύψει σημαντικά αποτελέσματα καθώς οι ομάδες που δημιουργήθηκαν είναι πάρα πολύ μικρές της τάξεως των 10 ατόμων, πληροφορία που δεν είναι ικανοποιητική. Βέβαια, τα αποτελέσματα για τους μη ασθενείς, ήταν αναμενόμενα καθώς η ιδέα στην οποία στηρίχτηκε η μελέτη αυτή είναι ότι υπάρχουν κάποια σημεία που είναι κοινά σε όλους τους ασθενείς και μπορεί να προκαλούν την εμφάνιση διαφόρων γενετικών ασθενειών. Οπότε αυτά τα όμοια σημεία έπρεπε να απουσιάζουν από τους μη ασθενείς, κάτι που εν μέρει επιβεβαιώνεται από τα αποτελέσματα.

Επιπρόσθετα, παρατηρούνται κάποιες σημαντικές ομάδες και στις εκτελέσεις που παρουσιάστηκαν αλλά και γενικότερα στην μεγάλη πλειοψηφία των εκτελέσεων που έγιναν στα πλαίσια της μελέτης αυτής, ομάδες που επιβεβαιώνονται από τον αλγόριθμο συνέπειας. Ειδικά, υπάρχει μια ομάδα που εμφανίζεται συνεχώς, κάτι που δείχνει ότι

υπάρχει μεγάλη ομοιότητα στα άτομα της ομάδας αυτής. Αν και δεν έχουν προκύψει πολλές μεγάλες ομάδες (και ίσως αυτό να οφείλεται και στο μέγεθος του δικτύου που επιλέγηκε που «αναγκάζει το δίκτυο» να βρει άτομα με πολύ μεγάλη ομοιότητα δημιουργώντας όσο το δυνατό πιο ανεξάρτητες ομάδες σε επίπεδο ατομικών νευρώνων και επιπλέον να ευθύνεται και ο τρόπος με τον οποίο επιλέγονται οι ομάδες καθώς έχει στοχοθετηθεί η μελλοντική μεταεπεξεργασία των αποτελεσμάτων έτσι ώστε να μπορεί το δίκτυο να ομαδοποιήσει τα άτομα ανάλογα με την επιθυμητή) σε σύγκριση πάντα με την αναλογία των ατόμων (η μεγαλύτερη ομάδα περιείχε 41 άτομα στους 972), εντούτοις τα αποτελέσματα μπορεί να είναι πολύ σημαντικά, δεδομένου ότι υπάρχουν βλέψεις για αξιολόγηση των ομάδων που προκύπτουν. Σίγουρα, αναμένονταν πολύ μεγαλύτερες ομάδες. Όμως, τα αποτελέσματα αυτά μπορούν να αποτελέσουν σημαντική βάση για περαιτέρω έρευνα. Υπάρχει σκέψη για περαιτέρω μεταεπεξεργασία των ομάδων αυτών που προκύπτουν, με ανάπτυξη νέων αλγορίθμων. Πιστεύεται ότι με γενικότερη αξιολόγηση των ομάδων που προκύπτουν και περαιτέρω ανάλυση των ατόμων που αποτελούν τις ομάδες αυτές, τότε είναι πολύ πιθανόν να προκύψουν πολύ σημαντικότερες ομάδες και συνεπώς, ανακάλυψη συσχετίσεων ανάμεσα στα SNPs.

Είναι σημαντικό να αναφερθεί, επίσης, ότι, ειδικότερα μετά την μελλοντική μεταεπεξεργασία, τα αποτελέσματα που προκύπτουν, θα μπορέσουν να απλοποιήσουν αναλύσεις – ειδικότερα αναλύσεις που απαιτούν πολύ χρόνο επεξεργασίας – καθώς θα μπορούν πλέον οι αναλύσεις αυτές να εφαρμόζονται σε επίπεδο ομάδων, μειώνοντας κατά πολύ τον αριθμό των στοιχείων, που θα απαιτείτο για ανάλυση όλων των ατόμων.

Όσο αφορά τους νέους αλγόριθμους, για να μπορέσουν να αξιοποιήσουν τα αποτελέσματα που προκύπτουν από τις κατηγοριοποιήσεις, θα χρησιμοποιήσουν τα βάρη των νευρώνων που παρουσιάζουν μεγάλη «κίνηση». Και ο λόγος είναι ότι μέσω των βαρών, μπορούν να ανακαλυφθούν τα SNPs τα οποία είναι όμοια σε όλα τα άτομα που αντιπροσωπεύει ένας νευρώνας. Βάρη με τιμές 0 ή 1 ή με τιμές που πλησιάζουν με πολύ μεγάλο βαθμό τις τιμές αυτές, δίνουν σαφή πληροφορία ότι υπάρχει ομοιότητα, καθώς με βάση τον αλγόριθμο μάθησης τα βάρη θα καταλήξουν να «μοιάζουν» με τα δεδομένα εισόδου. Αν προκύψουν τέτοιες τιμές σημαίνει ότι η μεγάλη πλειοψηφία (αν όχι όλα) των SNPS, είχαν τις ίδιες τιμές. Αν προκύπτουν τιμές μεταξύ 0.3-0.7 (κατά προσέγγιση), τότε σημαίνει ότι δεν υπήρχε σοβαρή ομοιότητα μεταξύ των δεδομένων

εισόδου. Με απλά λόγια η τιμή του κάθε βάρους δείχνει στο περίπου το ποσοστό των 0 και 1 στα δεδομένα, αν αγνοηθεί και ο παράγοντας τυχαιότητας (randomness) κατά την αρχικοποιήση των βαρών. Για παράδειγμα τιμή 0.3, αντιστοιχεί περίπου σε 70% 0 και 30% 1. Συνεπώς, μπορεί να αντιληφθεί ο οποιοσδήποτε την σημαντικότητα απεικόνισης των βαρών και τον λόγο για τον οποίο παρουσιάζονται στα αποτελέσματα.

Με την βοήθεια λοιπόν, των νέων αλγορίθμων μεταεπεξεργασίας, που θα δουλεύουν σε γενικότερο επίπεδο και θα ψάχνουν για ομοιότητες μεταξύ των ομάδων ξεχωριστά (και όχι άτομο προς άτομο), θα αναγνωριστούν και άλλες ομάδες και τα αποτελέσματα που παρουσιάζονται στα πλαίσια της μελέτης αυτής, μαζί με τα νέα αποτελέσματα, θα δοθούν στους επιστήμονες πληροφορικής και βιολόγους, για ανάλυση τόσο σε επίπεδο ομάδων και ανάλυση ατόμων που ανήκουν σε συγκεκριμένες ομάδες.

5.2.3 Συζήτηση Δικαιολόγησης Παραμέτρων

5.2.3.1 Συζήτηση Αποτελεσμάτων Αλλαγής Βαρών - Αριθμός Εποχών

Όσο αφορά τα αποτελέσματα που προκύπτουν από την μέση αλλαγή των βαρών, οι αλλαγές που παρατηρούνται στα βάρη είναι ελάχιστες μετά από περίπου 700 – 800 εποχές συνεπώς το δίκτυο έχει φτάσει σε μια σταθερή σχεδόν κατάσταση, αφού δεν χρειάζονται μεγάλες αλλαγές, έτσι έγινε η επιλογή για να τρέξει ο αλγόριθμος για 1000 εποχές.

Επιπρόσθετα, επειδή οποιοσδήποτε μπορεί να ισχυριστεί ότι το αποτέλεσμα αυτό οφείλεται και σε μεγάλο βαθμό και στην μείωση της ακτίνας αλλά κυρίως στην σταθερή μείωση του ρυθμού μάθησης, πρέπει να σημειωθεί ότι υπολογίζεται και το σφάλμα κβαντισμού, που επιβεβαιώνει τα πιο πάνω αποτελέσματα και θα εξηγηθεί πιο κάτω.

5.2.4 Συζήτηση Αποτελεσμάτων για την Χρήση Κύριας Μνήμης

Με βάση τα αποτελέσματα που παρουσιάστηκαν στο 4.2.4, είναι εμφανής η διαφορά μεταξύ των δύο εκτελέσεων όσο αφορά το ποσοστό μνήμης που χρησιμοποιούν. Αν

αναλογιστεί κανείς ότι το μέγεθος του αρχείου που χρησιμοποιήθηκε ήταν 340KB, είναι εμφανής η διαφορά που παρουσιάζεται σε όλες τις εκτελέσεις, οφείλεται στο ότι χρησιμοποιήθηκε για αποθήκευση των δεδομένων εισόδου, bits αντί αριθμοί. Πολύ σημαντική παρατήρηση καθώς, σε μεγαλύτερο δεδομένα θα είναι εμφανής η διαφορά στην χρήση των δύο αλγορίθμων καθώς μειώνεται περίπου στο ¼ ο χώρος που απαιτείται για την αποθήκευση των δεδομένων για επεξεργασία, κάτι που δίνει την δυνατότητα χρήσης της δυαδικής αναπαράστασης σε πολύ μεγάλα σύνολα δεδομένων.

5.2.5 Μειονεκτήματα Αλγορίθμου

Ο αλγόριθμος παρουσιάζει και κάποια μειονεκτήματα. Πρώτο μειονέκτημα του αλγόριθμου, με τον τρόπο τον οποίο έχει αναπτυχθεί, είναι ο πολύ συχνός συγχρονισμός που απαιτείται να γίνεται μεταξύ των διαφορών threads που συνεργάζονται. Δεν ήταν δυνατό να αποφευχθεί ο συγχρονισμός λόγω ότι σε κάθε επανάληψη του αλγορίθμου, πρέπει όλοι οι νευρώνες του δικτύου να γνωρίζουν τον best matching unit (BMU) από όλους τους νευρώνες, αφού με βάση αυτό θα συνεχίσουν για να αλλάξουν τα βάρη τους. Όποτε σε κάθε επανάληψη του αλγορίθμου, είναι απαραίτητος ο συγχρονισμός των threads.

Επιπρόσθετα, ο αλγόριθμος δεν παρέχει δυνατότητες portability (φορητότητα) καθώς κάνει χρήση των pthreads δηλαδή του POSIX standard για threads, υλοποιήσεις του οποίου υπάρχουν κυρίως σε Unix – Linux συστήματα. Έτσι ο αλγόριθμος αυτός λειτουργεί μόνο σε τέτοια περιβάλλοντα, αν και πολύ σύντομα θα υπάρχουν υλοποιήσεις και σε windows βασισμένα συστήματα, οπότε θα μπορεί εύκολα ο αλγόριθμος να λειτουργήσει και σε windows χωρίς την παραμικρή αλλαγή.

Επίσης, ένα επιπρόσθετο μειονέκτημα, είναι ότι ο χρήστης, λόγω του ότι χρησιμοποιήθηκε η κλάση bitset για βελτιστοποίηση της χρήσης της μνήμης, πρέπει να ορίσει πριν την εκτέλεση του αλγορίθμου το αριθμό των SNPs και alleles που υπάρχουν στο αρχείο εισόδου. Ήταν κάτι που δεν μπορούσε να αποφευχθεί, καθώς η κλάση bitset δεν επιτρέπει στον προγραμματιστή να επεκτείνει δυναμικά το μέγεθος του αντίστοιχου αντικειμένου που χρησιμοποιεί στο πρόγραμμα του για σκοπούς καθαρά βελτιστοποίησης. [18]

Πρόβλημα παρατηρείται επίσης και στην ανάθεση φόρτου εργασίας στον παράλληλο αλγόριθμο. Λόγω του ότι στα threads ανατίθενται στατικά οι νευρώνες για τους οποίους είναι υπεύθυνα, πολύ πιθανόν κάποια threads να πρέπει να εκτελέσουν πολύ περισσότερες πράξεις, ειδικότερα στο σημείο που αποφασίζεται ποιοι νευρώνες πρέπει να προσαρμόσουν τα βάρη τους. Υπάρχει περίπτωση να πρέπει ένα thread να αλλάξει όλα τα βάρη των νευρώνων του, ενώ ένα αντίστοιχο thread να μην χρειαστεί να εκτελέσει οποιαδήποτε αλλαγή στα βάρη. Είναι πρόβλημα ανάθεσης φόρτου και δίνει και μεγαλύτερη σημασία στο συγχρονισμό καθώς κάποια thread περιμένουν τα άλλα και έτσι το thread με τον μεγαλύτερο φόρτο γίνεται αυτομάτως bottleneck στον αλγόριθμο.

5.3 Συζήτηση Μεταεπεξεργασίας

5.3.1 Συζήτηση Αποτελεσμάτων των ομάδων – clusters και Συνέπεια Αλγορίθμου

Ο αλγόριθμος που αναπτύχθηκε στο στάδιο αυτό, έχει αναπτυχθεί για να αποδείξει την συνέπεια του αλγορίθμου, όπως έχει προαναφερθεί. Μέσω των αποτελεσμάτων που παρουσιάστηκαν στο προηγούμενο κεφάλαιο αλλά και διάφορων άλλων εκτελέσεων που έγιναν στα πλαίσια των δοκιμών, αποδείχθηκε ότι ο αλγόριθμος είναι συνεπής. Σε όλες σχεδόν τις περιπτώσεις, παρουσιάζονται ομάδες που εμφανίζονται σε όλες τις εκτελέσεις (εκτελέσεις με ίδιες παραμέτρους) και ομάδες που παρουσιάζονται περίπου 80% με 90% των εκτελέσεων. Συνεπώς, μόνο τυχαίες δεν είναι οι ομάδες αυτές, και σίγουρα τα άτομα τα οποία ανήκουν στις ομάδες αυτές, έχουν , αν μη τι άλλο, κάποιες πολύ σημαντικές ομοιότητες. Ομοιότητες οι οποίες μπορούν να εξαχθούν με βάση την απεικόνιση των βαρών του νευρώνα από τον οποίο αναπαριστώνται.

Επιπρόσθετα, πρέπει να σημειωθεί ότι ο αλγόριθμος αυτός έπρεπε να αναπτυχθεί, καθώς η μελέτη αυτή θα δημοσιευθεί και, συνεπώς, δεν μπορούσαν τα αποτελέσματα να πιστοποιηθούν για την ορθότητα και συνέπεια τους. Πόσο μάλλον, όταν ένας από τους βασικούς στόχους της μελέτης αυτής, είναι να δοθεί στους επιστήμονες και να χρησιμοποιηθούν τα συμπεράσματα ως βάση για περαιτέρω έρευνα, όχι μόνο για την

κατά πλάκα σκλήρυνση, αλλά και να χρησιμοποιηθεί η προτεινόμενη μεθοδολογία και για άλλες ασθένειες.

Όπως έχει προαναφερθεί, αν και δεν ήταν αναμενόμενα τα αποτελέσματα, καθώς τουλάχιστον θεωρητικά αναμένοντο ομάδες με περισσότερα άτομα, εντούτοις με επέκταση του αλγορίθμου συνέπειας και περαιτέρω μεταεπεξεργασία, θα ανακαλυφθούν περισσότερες και σημαντικότερες ομάδες.

5.4 Τρόποι Σύγκρισης και Αξιολόγησης Αλγορίθμων

5.4.1 Συζήτηση Αποτελεσμάτων Σφάλματος Κβαντισμού (Quantization Error)

Όσο αφορά τα αποτελέσματα του σφάλματος κβαντισμού, είναι προφανές ότι το δίκτυο μαθαίνει, αφού το σφάλμα μειώνεται σε πολύ μεγάλο βαθμό, σε σημείο που σχεδόν μπορεί να θεωρηθεί αμελητέο.

Το γεγονός ότι παρουσιάζεται ένα τόσο μικρό σφάλμα κβαντισμού έχει το πλεονέκτημα ότι επιβεβαιώνει την ορθότητα του αλγορίθμου σε συνδυασμό με τα αποτελέσματα από του αλγορίθμου για την συνέπεια του αλγορίθμου. Δίνει την δυνατότητα της σιγουριάς όσο αφορά την κατηγοριοποίηση που γίνεται.

Παρ’ όλα αυτά, υπάρχει και ένα μειονέκτημα το οποίο δυστυχώς εδώ δεν μπορεί να προβλεφθούν προβλήματα υπέρ – εκπαίδευσης. Συνήθως τέτοια προβλήματα για να παρατηρηθούν πρέπει να υπάρχουν δεδομένα εκπαίδευσης (training data) και δεδομένα ελέγχου (test data). Όμως για την συγκεκριμένη εφαρμογή, δεν υπάρχει λόγος διαχωρισμού αφού στόχος δεν ήταν να γενικεύει το δίκτυο αλλά να ομαδοποιεί μόνο τα συγκεκριμένα δεδομένα με τα οποία εκπαιδεύεται. Στα πλαίσια της μελέτης αυτής, στόχος ήταν να δούμε τα κοινά χαρακτηριστικά των γενετικών δεδομένων διαφόρων ατόμων. Δεν ήταν στόχος να εκπαιδευτεί το δίκτυο έτσι ώστε να μπορεί να αναγνωρίζει ασθενείς ή υγιείς, ούτε να δέχεται νέες εισόδους για πρόβλεψη. Τα δεδομένα ήταν σταθερά και δεν επρόκειτο να αλλάξουν. Η αλλαγή, μπορεί να γίνει πολύ εύκολα βέβαια, αν και είναι εκτός του πλαισίου της μελέτης αυτής, οπότε μπορεί να θεωρηθεί

ως μελλοντική εργασία σε περίπτωση που αλλάξουν οι στόχοι ή σε περίπτωση που θα χρησιμοποιηθεί για άλλους σκοπούς.

Κεφάλαιο 6

Συμπεράσματα

6.1 Γενικά Συμπεράσματα

Στόχος της μελέτης αυτής ήταν να γίνει προσπάθεια κατηγοριοποίησης γενετικών δεδομένων στην περιοχή HLA, με ασθενείς που παρουσιάζουν στον φαινότυπο τους την ασθένειας της κατά πλάκα σκλήρυνσης. Αποφασίστηκε η χρήση των χαρτών αυτό-οργάνωσης με αλγόριθμο μάθησης όπως προτάθηκε από τον T. Kohonen. [15]. Τα αποτελέσματα που προέκυψαν, δεν επιβεβαίωσαν στον αναμενόμενο βαθμό, τουλάχιστον, τις αρχικές εκτιμήσεις. Όμως, έχουν εξαχθεί σημαντικά συμπεράσματα στα πλαίσια της μελέτης αυτής.

Το πρώτο συμπέρασμα αφορά τον αλγόριθμο συμπίεσης για γενετικά δεδομένα, που επιτυγχάνει να μειώσει το μέγεθος των αρχικών δεδομένων κατά 8 φορές περίπου και μπορεί να χρησιμοποιηθεί για όλες τις επιστημονικές αναλύσεις καθώς διατηρεί όλες τις γενετικές πληροφορίες. Άλλο σημαντικό συμπέρασμα αποτελεί η μη γραμμική επιτάχυνση που επιτυγχάνει ο παράλληλος αλγόριθμος, αν και στόχος ήταν η γραμμική επιτάχυνση, εντούτοις λόγων των συγχρονισμών, δεν ήταν δυνατή η περαιτέρω βελτιστοποίηση του αλγορίθμου, κάτι που έχει οριοθετηθεί ως μελλοντική εργασία με τροποποίηση του αλγόριθμου μάθησης. Οι βελτιστοποιήσεις που έχουν γίνει στον προτεινόμενο αλγόριθμο, βοήθησαν όμως σίγουρα το πιο σημαντικό πρόβλημα αποτελεί ο συνεχής συγχρονισμός, στο οποίο συμβάλει και η στατική ανάθεση νευρώνων στα threads. Όσο αφορά τις κατηγορίες που προκύπτουν, αν και οι ομάδες δεν ήταν οι αναμενόμενες, εντούτοις τα αποτελέσματα θα τύχουν περισσότερης και πιο γενικής μεταεπεξεργασίας και θα δοθούν στους επιστήμονες για περαιτέρω ανάλυση.

Τα πλεονεκτήματα είναι πάρα πολλά, καθώς μπορούν πλέον να γίνουν περίπλοκες αναλύσεις σε επίπεδο ομάδων πλέον, μειώνοντας το χρόνο που χρειάζονταν προηγουμένως, ή να γίνουν περισσότερες αναλύσεις στα άτομα που εμφανίζονται σε συγκεκριμένες ομάδες, που σύμφωνα με τον αλγόριθμο παρουσιάζουν μεγάλη ομοιότητα. Ομοιότητα που παρουσιάζεται μέσω των απεικονίσεων των βαρών. Τέλος, οι διάφορες μετρικές που παρουσιάστηκαν και ο αλγόριθμος συνέπειας αποδεικνύουν ότι ο αλγόριθμος είναι συνεπής, καθώς υπάρχουν ομάδες που παρουσιάζονται σχεδόν κάθε φορά στις εκτελέσεις, και εκπαιδεύεται σωστά καθώς το σφάλμα κβαντισμού, που αποτελεί την πιο σημαντική μετρική αξιολόγησης, μειώνεται σταθερά σε κάθε εποχή.

Ολοκληρώνοντας, πρέπει να επισημανθεί ότι στον αλγόριθμο υπάρχουν μεγάλες δυνατότητες βελτίωσης, εργασία που θα γίνει στο εγγύς μέλλον, και τα αποτελέσματα, με περαιτέρω μεταεπεξεργασία, θα μελετηθούν και θα αξιολογηθούν από εμπειρογνώμονες βιολόγους με τους οποίους συνεργάζεται η ομάδα, και με την πολύτιμη συνεισφορά τους, θα καθοριστούν τα επόμενα βήματα στην μεγάλη αυτή προσπάθεια για την ανακάλυψη συσχετίσεων μεταξύ γενετικών σημείων. Το σημαντικότερο ίσως σημείο που πρέπει να σημειωθεί είναι ότι, αν η σημασία των αποτελεσμάτων επιβεβαιωθεί και από τους επιστήμονες, η μελέτη αυτή μπορεί να αποτελέσει βάση για νέες προσεγγίσεις σε πολλές χρόνιες ασθένειες που οφείλονται στην παρουσία ή απουσία γενετικών πληροφοριών, δηλαδή χρήση παρόμοιων αλγορίθμων κατηγοριοποίησης για ομαδοποίηση της γενετικής πληροφορίας με βάση τις ομοιότητες που παρουσιάζει στα διάφορα άτομα και εξαγωγή περιοχών που υπάρχουν συχνά και πιθανόν να οφείλονται για την εμφάνιση μιας ασθένειας.

6.2 Μελλοντική Εργασία

Στο πεδίο της συγκεκριμένης μελέτης υπάρχουν αρκετά θέματα τα οποία προσφέρονται για μελλοντική εργασία.

Όσο αφορά τον παράλληλο αλγόριθμο που έχει αναπτυχθεί, όπως έχει προαναφερθεί, τα δύο μειονεκτήματα του ήταν οι συχνοί συγχρονισμοί που ήταν απαραίτητοι για την λειτουργία του, καθώς και η στατική ανάθεση νευρώνων στα threads. Το πρώτο πρόβλημα θα λυθεί αρχικά με την ανάπτυξη διαφορετικού αλγορίθμου συγχρονισμού,

δηλαδή υλοποίηση του συγχρονισμού με διαφορετικό τρόπο, και μετέπειτα με την υλοποίηση του batch αλγόριθμου [15], ο οποίος αλγόριθμος αποτελεί μια παραλλαγή του online αλγορίθμου που υλοποιήθηκε στα πλαίσια της μελέτης αυτής. Ουσιαστικά η διαφορά έγκειται στο ότι αλλάζονται τα βάρη αφού παρουσιαστούν όλα τα δεδομένα μια φορά στο δίκτυο, κάτι που επιτρέπει την αποφυγή συγχρονισμού μεταξύ των διανυσμάτων εισόδου. Όσο αφορά την στατική ανάθεση των νευρώνων στα threads, υπάρχει η σκέψη τα threads αναλαμβάνουν δυναμικά τους νευρώνες για τους οποίους θα είναι υπεύθυνοι, ανάλογα με το ποιοι νευρώνες χρειάζεται να αλλάξουν τα βάρη τους. Η σκέψη αυτή είναι στο στάδιο της συζήτησης με τα υπόλοιπα μέλη της ομάδας και θα υπάρξει μεγαλύτερη επικοινωνία και συνεργασία για το καλύτερο δυνατό αποτέλεσμα.

Στο θέμα της μεταεπεξεργασίας, υπάρχει πρόθεση για ανάπτυξη ενός γενικότερου αλγορίθμου απόδειξης της συνέπειας. Ο προτεινόμενος αλγόριθμος στηρίχτηκε στο γεγονότος ότι κάθε νευρώνας αποτελεί μια ομάδα, γι' αυτό και άλλωστε επιλέγηκαν μεγάλα πλέγματα, έτσι ώστε να ανεξαρτητοποιηθούν όσο περισσότερο γίνεται οι ομάδες και να φτάσουν σε επίπεδο νευρώνων. Πιστεύεται ότι αν ο αλγόριθμος επεκταθεί με την ικανότητα να ελέγχει σε γειτονικό επίπεδο, αντί σε επίπεδο νευρώνων θα προκύψουν καλύτερα αποτελέσματα. Επιπλέον, υπάρχουν βλέψεις για καλύτερη μεταεπεξεργασία όσο αφορά τα άτομα κάθε ομάδας. Υπάρχουν δύο στόχοι. Να γίνει έλεγχος στις όμοιες περιοχές των ατόμων που ανήκουν σε ομάδες και να ελεχθεί η πιθανότητα κάποιες από τις περιοχές αυτές να οφείλονται για την εμφάνιση της κατά πλάκα σκλήρυνσης. Ο δεύτερος στόχος είναι να γίνει έλεγχος σε επίπεδο ομάδων. Δεδομένου ότι είναι γνωστές πλέον οι ομάδες που αντιπροσωπεύουν τα διάφορα άτομα, μπορούν να γίνουν διάφορες αναλύσεις με τους αντιπροσώπους τους για εξαγωγή περισσότερων συμπερασμάτων, όπως ποιοι αντιπρόσωποι παρουσιάζουν μεγαλύτερη ομοιότητα και τον βαθμό ομοιότητας.

Συνεπώς, υπάρχει πολύ προοπτική για αξιολόγηση των αποτελεσμάτων που προέκυψαν στα πλαίσια της μελέτης αυτής. Σε συνεργασία με ειδικούς βιολόγους, υπάρχει η πεποίθηση ότι θα προκύψουν πολύ σημαντικότερα αποτελέσματα που θα βοηθήσουν στην αναγνώριση των αιτίων που προκαλούν την εμφάνιση πολλών γενετικών ασθενειών.

Βιβλιογραφία

[1] Scott A. Lesley, “High-Throughput Proteomics: Protein Expression and Purification in the Postgenomic World”, Genomics Institute, Novartis Research Foundation, 3115 Merryfield Row, San Diego, California 92121, June 14, 2001.

[2] J. Ott, "Brief Research Communication - Neural networks and disease association studies," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics,* vol. 105 Issue 1, pp. 60 - 61, Jan. 2001.

[3] D. Mount, Bioinformatics Sequence and Genome Analysis, Second Edition. Cold Spring Harbor Laboratory Press, 2004.

[4] T. Strachan and A. Read. Molecular Genetics 2, 2nd ed. United States: BIOS Scientific Publishers Ltd, 1999 [E-book] Available: NCBI Bookshelf

[5] P. Yue and J. Moult, "Identification and analysis of deleterious human SNPs," Journal of molecular biology, Volume 356 Issue 5, pp. 1263–1274, Mar. 2006.

[6] L. M. Havill, T. D. Dyer, D. K. Richardson, M. C. Mahaney and J. Blangero, " The quantitative trait linkage disequilibrium test: a more powerful alternative to the quantitative transmission disequilibrium test for use in the absence of population stratification," BMC Genet, vol. 6(Suppl. 1), Dec. 2005

[7] H. J. Cordell. "Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans," *Human Molecular Genetics*, vol. 11, no. 20, pp. 2463 - 2468, July 2002.

[8] S. Purcell, B. Neale, K. T. Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M.J. Daly and P. C. Sham, PLINK: a toolset for whole-genome association and population-based linkage analysis. [Open

Source]. Massachusetts General Hospital: American Journal of Human Genetics vol. 81, 2007.

[9] Mat Buckland, "Kohonen's Self Organizing Feature Maps," ai-junkie.com, 2002. [Online]. Available: http://www.ai-junkie.com/ann/som/som1.html [Accessed: July. 06, 2009].

[10] Jeremy Wyatt, "Kohonen Networks," cs.bham.ac.uk, July 1999. [Online]. Available:http://www.cs.bham.ac.uk/~jlw/sem2a2/Web/Kohonen.htm[Accessed: July. 07, 2009].

[11] Bashir Magomedov, "Self-Organizing Feature Maps (Kohonen maps)," codeproject.com, 07 November 2006. [Online]. Available: http://www.codeproject.com/KB/recipes/sofm.aspx [Accessed: July. 08, 2009].

[12] STATSOFT, "SOFM Networks," statsoft.com, [Online]. Available: http://www.statsoft.com/textbook/stneunet.html#sofm [Accessed: July. 09, 2009].

[13] David J. Haglin, " Kohonen Neural Networks for Unsupervised Clustering," grb.mnsu.edu. [Online]. Available: http://grb.mnsu.edu/grbts2/doc/manual/node20.html [Accessed: July. 10, 2009].

[14] Casey Chesnut, "Self Organizing Map AI for Pictures," generation5.org, 03 November 2004. [Online]. Available: http://www.generation5.org/content/2004/aisompic.asp?Print=1 [Accessed: July. 10, 2009].

[15] Teuvo Kohonen, *Self-Organizing Maps*, 3rd Edition, New York: Springer, 2000.

[16] SPSS Inc, SPSS Base 17.0 for Windows User's Guide. Chicago IL: SPSS Inc, 2008.

[17] A. Antoniades, L. Loizou, A. Aristodimou, C.S. Pattichis, A binary format for genetic data designed for large whole genome studies that enable both marker and strand based analyses, in CD-ROM Proceedings of the 8th IEEE International Conference on BioInformatics and BioEngineering (BIBE 2008), October 8-10, Athens, Greece, 6 pages, 2008.

[18] Nicolai M. Josuttis, *The C++ Standard Library - A Tutorial and Reference,*U.S.A: Addison-Wesley , 1999.

[19] U. Drepper and I. Molnar, "The Native POSIX Thread Library for Linux," Red Hat Linux NTFL, February 21, 2005.

[20] P. K. Janert, *Gnuplot in Action*. U.S.A.: Manning, 2009

[21] T. Kohonen, I. T. Nieminen, and T. Honkela, " On the Quantization Error in SOM vs. VQ: A Critical and Systematic Study.," presented at 7th International Workshop on Advances in Self-Organizing Maps, Florida, USA, 2009.

[22] G. R. Abecasis, S. S. Cherny, W. O. Cookson and L. R. Cardon, "Merlin-rapid analysis of dense genetic maps using sparse gene flow trees," Nature Genetics vol. 30, pp. 97-101, 2002.

[23] S. E. Baranzinin, J. Wang, et al, "Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis," *Human Molecular Genetics.*, vol. 18, no. 4 pp. 767-778, November 2008.

[24] F. Lourenço, V. Lobo, and F. Bação, "Binary-based similarity measures for categorical data and their application in self- organizing maps," April 2004. [Online]. Available: http://www.isegi.unl.pt/docentes/vlobo/Publicacoes/lobo042.pdf

[25] P. Kudova, H. Rezankova, D. Husek, V. Snasel, "Categorical Data Clustering Using Statistical Methods and Neural Networks," Jun. 2, 2009 [Online]. http://syrcodis.citforum.ru/2006/kudova.pdf [10 December 2009].

[26] C.-C. Hsu, "Generalizing Self-Organizing Map for Categorical Data," IEEE Transactions on Neural Networks, vol. 17, no. 2, pp. 294 - 304, Mar. 2006.

[27] P. A. McCaskie, K. W. Carter, S. R. McCaskie and L. J. Palmer, Eds., *The effect of missing data on linkage disequilibrium mapping and haplotype association analysis in the GAW14 simulated datasets: Genetic Analysis Workshop 14: Microsatellite and single-nucleotide polymorphism, July 07-10, 2004, Noordwijkerhout, The Netherlands*. Australia: Laboratory for Genetic Epidemiology, Western Australian Institute for Medical Research, UWA Centre for Medical Research, University of Western Australia, 2005.

[28] A. Passaro, F. Baronti and V. Maggini, Eds., *Exploring relationships between genotype and oral cancer development through XCS: Genetic And Evolutionary Computation Conference, Washington, D.C., U.S.A.,2005.* U.S.A. N.Y.: ACM, 2005, pp. 147 - 151

Παράρτημα Α

Στο παράρτημα αυτό, θα γίνει μια αναφορά στις δημοσιεύσεις που προέκυψαν από την μελέτη αυτή. Έχει ήδη εκδοθεί ο αλγόριθμος συμπίεσης στο στάδιο της προεπεξεργασίας [17]. Η επόμενη δημοσίευση θα αφορά τα αποτελέσματα των ομάδων που προέκυψαν με την χρήση του δυαδικού παράλληλου αλγόριθμου, σε συνδυασμό με τις επιδόσεις που εμφανίζει ο αλγόριθμος αυτός. Η τρίτη δημοσίευση θα αφορά τις βελτιστοποιήσεις που έγιναν στον παράλληλο αλγόριθμο, βελτιστοποιήσεις που επί τω πλείστον, μπορούν να εφαρμοστούν και στον απλό σειριακό αλγόριθμο μάθησης χαρτών αυτό-οργάνωσης. Τέλος, η επόμενη δημοσίευση θα αφορά τον τρόπο με τον όποιο διαχειρίζεται η μελέτη αυτή τα categorical δεδομένα για να χρησιμοποιηθούν σε ένα αλγόριθμο που επεξεργάζεται γραμμικά δεδομένα, και πώς διαχειρίζεται ο συγκεκριμένος παράλληλος αλγόριθμος τα missing δεδομένα.

Ακολουθεί η δημοσίευση που έχει ήδη εκδοθεί.

A binary format for genetic data designed for large whole genome studies that enable both marker and strand based analyses

Athos Antoniades, Loizos Loizou, Aristos Aristodimou,
Constantinos S. Pattichis, *Senior Member, IEEE*

***Abstract*— Recent advances in genotyping technology have enabled large studies with data from thousands of subjects to contain half a million or more of single nucleotide polymorphisms (SNPs) marker per subject. This rapid increase in the size of data has generated the need to compress the data in order to reduce the storage capacity requirements and the memory required at run time to perform analysis on the data. The availability of so many markers across the whole genome has created opportunities for new methodologies to be implemented that take advantage of the relatively high density of the markers to perform analyses that take into account the Linkage Disequilibrium (LD), an effect where some combinations of genetic markers are non-randomly associated. Classical techniques for transforming genotypic data into a binary format are already in use by several applications however we demonstrate in this paper that the traditional transformations are not adequate for certain types of analyses as some information key to new methodologies of analyses is lost. We propose a new protocol for formatting binary genotypic data that can be used in all types of analyses while still offering a very high compression rate.**

I. INTRODUCTION

Until recently in genetic studies, due to limitations in genotyping technology, only a small subset of the genome could be analyzed. The introduction of affordable high throughput genotyping technologies allowed the assay of more than half a million loci variants (SNPs) per subject across the whole genome. Genetic association studies applying such technology have now become common as they allow investigation of the majority of common loci variants in the genome; such studies are typically called genome wide association scans(GWAS). In this paper we present a non lossy format of encoding the data in the datasets generated from GWAS to a binary form.

The substance that encodes the genetic instructions of living organisms is Deoxyribonucleic acid (DNA). DNA consists of two long units called strands having a shape of a double helix [7], [8]. The genetic code is specified by the four nucleotide "letters" A (adenine), C (cytosine), T (thymine), and G (guanine). There are multiple different types of variations that can occur on the DNA strands, however traditionally the genetic analyses seem to focus on analyzing Single Nucleotide Polymorphisms (SNPs). SNP variation occurs when a single nucleotide, such as an A, replaces one of the other three nucleotide letters—C, G, or T [9]. For a variation to be considered a SNP it also needs to be present in at least 1% of the population. Due to the Linkage Disequilibrium effect (LD), SNPs serve as biological markers for pinpointing a region on the genome associated with a phenotypic trait even thought the SNP itself is not necessarily the variation responsible for the association. Rather it's one or more of the variations that are in LD with the SNP that is causing the association.

The need to reduce the size of the data is owed to the 100 fold increase in the genotyping capacity available today combined with the massive reduction in cost. Today's technology has the ability to analyze datasets up to 550,000 or even 1 million SNPs per subject with >99% accuracy, at a rate of >100 K genotypes per day

Manuscript received August 8, 2008.

Athos Antoniades is with the Department of Computer Science, University of Cyprus, Nicosia, CY (corresponding author: +357-22-892685; e-mail: athos@cs.ucy.ac.cy).

Loizos Loizou is with the Department of Computer Science, University of Cyprus, Nicosia, CY. (e-mail: cs06ll1@cs.ucy.ac.cy).

Aristos Aristodimou is with the Department of Computer Science, University of Cyprus, Nicosia, CY

and at a cost of around 20–30 cents per genotype [2], [6], [10].

Traditionally the genetic data in these studies is stored in the QTDT format introduced in the program Merlin [1]. Input files describe relationships between individuals in a dataset, store marker genotypes, disease status and quantitative traits and provide information on marker locations and allele frequencies.

There is already a technique for compressing this data by utilizing a binary encoding [3]. However, that technique was focused at encoding SNPs as markers to be used in analyses rather than strand based information. Today as more GWAS datasets became available researchers are developing innovative new methodologies to analyze them. Some of these methodologies are not relying so much on the SNPs as markers; rather they look at the sequence of genotyped alleles on each strand of DNA separately. The methodology used in plink to encode the data looses the information of which strand holds each allele's genotype for heterozygote SNPs, therefore making it impossible to run analyses that use strand information using the binary input format for GWAS. Also, recent technological advances in genotyping are enabling the detection of deletion regions. This may result in future datasets to incorporate markers in them that may have an allele missing not due to a genotyping error but due to a deletion on one of the two strands. The encoding format proposed in this paper will offer an efficient lossles compression that is also capable of encoding deletions separately from errors in genotyping or quality control removed markers.

The structure of the paper is as follows: Section II Presents the traditional methodology of formatting GWAS data as well as the current alternative to compressing the data and the format proposed in this paper. Section III Offers a comparison of the three formats identifying advantages and disadvantages of each. Section IV concludes describing the situations under which each format is optimal.

II. METHODS AND MATERIAL

A. THE COMMONLY USED FORMAT QTDT MERLIN[1]

The QTDT file format described in Merlin is the one traditionally used for this type of data. It's split into three files; Pedigree files contain phenotypes for discrete and quantitative traits and marker genotypes for a specific number of subjects. They are usually white-space delimited files. The first (usually 6) columns contains information about the subject (Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype). The combination of the information of each subject must be unique. The next columns contain biallelic markers; typically SNPs. Marker genotypes are encoded as two consecutive integers, one for each allele, or using the letters "A", "C", "T" and "G". To denote missing alleles a sentinel value is used, typically "0".

Map files contain information for each single nucleotide polymorphism. They are used to analyze genetic markers into the equivalent pedigree file. Each line per marker usually contains 3, 4 or 5 columns (chromosome, SNP identifier, morgans or centimorgans and base-pair position). Each column is separated by white space.

Dat files describe the pedigree file. They include one row per data item in the pedigree file, indicating the data type providing a one-word label for each item.

B. PLINK'S METHOD FOR BINARY PED FILES [3]

Plink is an excellent, open source program offering a comprehensive range of basic large-scale whole genome association analysis methodologies. It has been widely adopted since high dimensionality GWAS have become available as it enables researchers to efficiently analyze these large datasets in a computationally efficient manner.

In plink there is an encoding format for transforming QTDT Merlin data into binary formatted files. The approach used in plink uses 2 bits for encoding biallelic markers with 4 possible states. Plink uses the encoding for each genotypegiven in Table I. [3]

TABLE I
PLINK BINARY PED FILE ENCODING [3]

Allele On Strand +	Allele On Strand -	Marker Encoding
A	A	00
A a	a A	01
A	a	11
Missing Data	Missing Data	10

Testing on plink binary format showed that the exported binary file was 15 times smaller than the original file. The drawback however is that encoding of the heterozygote allele is the same regardless of what strand it's actually derived from. Therefore any analyses that rely on the sequence of the alleles on the strand will be missing this information.

One analyses technique that needs the lost information is imputation. Imputation analysis is the practice of 'filling in' missing data with plausible values. It is a method for uncovering the genetic basis of human disease and it is used for inferring genotypes at observed or unobserved SNPs that can detect causal variants that have not been directly genotyped [5]. It is in essence an in-silico approach to discover the probability of the existence of a specific genotype for loci that haven't being directly genotyped but are known to be in LD with genotyped markers.

C. PROPOSED METHOD

Since imputation analyses require knowledge of which strand the alleles of heterozygote SNPs exist on, we need to encode each allele on each strand separately for all cases. There are a minimum of three states each allele can be in, these include the two possible nucleotides (commonly denoted as A and a) as well as the possibility of missing data at that location. The smallest number of bits that can encode the 3 states of an allele is 2, however with 2 bits we can actually encode a fourth state. In many existing studies this may not be used, even though it will have no impact on the capacity requirements the databases will have for storage. However, we propose that the fourth state is set to denote markers that are in deleted regions. This utilizes the extra available coding identifying a deleted allele from a missing allele due to quality control concerns, or genotyping error.

An analytical technique that enables the detection of these deletions that has started being applied is copy number variants (CNV) analyses. It refers to the genetic trait of differences in the number of copies of a particular region (for example a gene) present in the genome of an individual [4, 10]. To perform CNV analyses most algorithms rely on raw data from the genotyping platform. CNV algorithms are able to detect deletion regions as well as regions that are duplicated that may exist in some individual's strands.

However it should be clearly noted that CNV incorporates more information than just deleted regions. It can also detect regions that exist in more than one copy per strand. That information is not reflected in the proposed protocol; therefore methodologies that use that information would still rely on an external file with CNV regions.

In the proposed format the data is structured as a two dimensional vector of alleles. The first dimension's size is equal to the number of strands the subject has. Typically in humans all chromosomes have two strands with the exception of X and Y chromosomes in males that each have 1 strand. Even in these cases a second strand can exist listing the alleles of the second strand on males as missing.

Each element in the vector of each strand will encode an allele. The allele will be 2 bits long enabling encoding of a total of 4 states per allele, missing data, nucleotide 1, nucleotide 2 or deleted.

Table II presents the 4 different states that can be coded per allele. The term "Unknown" is used rather than the more typical "missing" to denote alleles that it's unclear what their genotypes are or if they are deleted so as not to confuse it with the deleted state that defines alleles that do not exist on that strand.

Table III shows how two alleles are encoded and create a biallelic allele such as SNPs, while still enabling the identification of deleted alleles from missing values if that information is available. By comparing two alleles together and using the encoding as proposed above the resulted encoding is shown in Table III.

The two strands in each Subject's vector need to be perfectly aligned, that is, the *i* th element of each vector will point to the same Marker's alleles one for each strand. This makes it easy to access information in the way it was traditionally accessed. To access the *i*'th marker's alleles the two bits at position *i* in each strand will carry a total of 4 bits, using the encoding column of Table III the genotype of Marker *i* can then be identified.

III. EXPERIMENTAL RESULTS

To compare the different formats we will consider the size of the resulting files in relation to the original QTDT Merlin format, the amount of information lost through the encoding of the original QTDT Merlin format and the ability of each format to retain different types of information available today. Table IV provides a summary of the comparison.

A. INFORMATION LOSS

On one hand loosing information due to encoding is undesired; however, if the encoding is used to simply speed up analyses and reduce scratch space then it's not an issue so long as the original raw file is kept for future analyses. However if you are developing an algorithm that needs the information of which strand each heterozygote marker's alleles are on , or if there are deletion regions overlapping the markers in either strand, then utilizing the encoding methodology proposed in this paper will in a single table or file encode all of that information efficiently. Another issue is the actual storage of the data for long term use or for transferring over the internet. Utilizing a non-lossy approach to compressing the data that incorporates all genetic information into a single file can reduce the resources required.

B. ADDED INFORMATION CAPABILITY

In this paper we focused on bi-allelic markers as they are the ones that current high throughput genotyping technologies are able to genotype. However it should be noted that the format of QTDT Markers enables encoding of tri or quad allelic markers as well since each allele is encoded as an ASCII character. Neither plink's binary ped file nor the format proposed could handle tri or quad allelic markers. Large datasets with tri-allelic markers are not existing today (to the knowledge of the authors) while information on deleted markers is available through CNV analyses and other methodologies available to the various genotyping platforms. Therefore tri-allelic and quad allelic markers were ignored in this encoding until high throughput technologies genotyping them are available making it necessary to generate a new protocol for that.

C. SIZE OF ENCODED DATA

The compression rate can easily be estimated

TABLE II
PROPOSED ALLELE ENCODING

Allele	Encoding
Unknown	00
A	01
a	10
Deleted	11

TABLE IV
COMPARISON OF FILE FORMATS

QTDR Merlin	Proposed Protocol	Plink 's protocol
Information Loss for bi-allelic markers		
Used as Reference	None	strand location of heterozygote alleles Aa,aA Missing one of the two alleles A0, 0A, a0,0a
Added Information capability		
tri or quad allelic markers	Alleles deleted from a specific strand	None
Size*		
Pedigree File 3.6 GB Map File 12.6 MB **Total:** 3.612 GB	One binary file 496 MB	.bed file : 229.6 MB .fam file : 30 .bim file : 14.1 MB **Total** : 243.7 MB

TABLE III
PROPOSED ENCODING FOR BI-ALLELIC MARKERS

Allele 1	Allele 2	Encoding
Unknown	Unknown	0000
Unknown	A	0001
Unknown	A	0010
Unknown	Deleted	0011
A	Unknown	0100
A	A	0101
A	a	0110
A	Deleted	0111
a	Unknown	1000
a	A	1001
a	a	1010
a	Deleted	1011
Deleted	Unknown	1100
Deleted	A	1101
Deleted	a	1110
Deleted	Deleted	1111

since both encodings are deterministic, however we also provide a as an example an actual test we performed using an average dataset size of 1806 subjects and 532,579 markers. The plink binary ped file was able to compress the file to $1/15^{th}$ of it's original size while the proposed method compressed the file to exactly double the size of that achieving just $1/7.5^{th}$ of the original size. However, both compressions produce enough of a compression to overcome the issue of the large data since the data can now fit on the average computer's physical memory (RAM) as today's typical computer has at least 1 GB.

IV. CONLUSION

The proposed format encodes genotypic data into a binary form in order to compress it and at the same time preserve all the information relating to the bi-allelic markers and the strand location of each allele. However it does double in size the resulting datasets from existing encoding methodologies that are lossy, but loose information only necessary to certain type of analytical approaches. The analytical approaches that would benefit from the proposed format of encoding are primarily the ones that take into account the strand on which heterozygote alleles are based on, the existence of the marker on just one of two possible strands, identifying if the second wasn't available due to genotyping errors or a deletion over the marker on that strand. Due to the need to use two bits per allele for encoding while only needing 3 states for each allele we were left with an available fourth state. That fourth state we proposed that is used to denote markers that were deleted as that information is becoming commonly available from genotyping platforms available already; however, future researchers may choose to use the fourth state to code a different state an allele can be in. Also in cases where compression of the data in a non-lossy way for storage, backup or data transfer is used the methodology proposed would be ideal.

REFERENCES

[1] G. R. Abecasis, S. S. Cherny, W. O. Cookson and L. R. Cardon, "Merlin-rapid analysis of dense genetic maps using sparse gene flow trees," Nature Genetics vol. 30, pp. 97-101, 2002.

[2] S. Jenkins and N. Gibson, "High-Throughput SNP Genotyping," Comparative and Functional Genomics, vol. 3, no. 1, pp. 57-66, 2002.

[3] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P.I.W De Bakker, Daly MJ and Sham PC, "PLINK: a toolset for whole-genome association and population-based linkage analysis," American Journal of Human Genetics, vol. 81, Sep. 2007, pp 559-75.

[4] J. Redon, "Global variation in copy number in the human genome," Nature, vol. 444, No. 7118. Nov. 2006, pp. 444-454.

[5] R.E. Fay, "Valid inference from imputed survey data," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 41-48, 1993.

[6] P.Y. Kwok, "High-throughput genotyping assay approaches," Pharmacogenomics, vol. 1, Feb 2000, pp. 95–100.

[7] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walters, *Molecular Biology of the Cell* 4th ed. New York and London: Garland Science, 2002, ch. 1 .

[8] Butler, John M. (2001). Forensic DNA Typing. Elsevier. ISBN 978-0-12-147951-0. pp. 14–15.

[9] Weiner MP; Hudson TJ (2002) Introduction to SNPs: Discovery of markers for disease

[10] Freeman, JL, et al (2006). "Copy number variation: New insights into genome diversity". Genome Research 16: 949–61

[11] Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95–1