

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

**ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΒΙΟΔΕΙΚΤΕΣ:
ΕΦΑΡΜΟΓΕΣ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ**



Καραλής Γεώργιος

Καραλής Γέωργιος

**ΒΙΟΠΛΗΡΟΦΟΡΙΚΗ ΚΑΙ ΒΙΟΔΕΙΚΤΕΣ:
ΕΦΑΡΜΟΓΕΣ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ**

Επιβλέπων Καθηγητής: Βλάμος Παναγιώτης
Μέλη Επιτροπής: Αυλωνίτης Μάρκος, Μυλωνάς Φοίβος

Η εκπόνηση αυτή της εργασίας θα ήταν αδύνατη χωρίς την βοήθεια της
Αβραμούλη Αντιγόνης

Περιεχόμενα

0.0 Σκοπός πτυχιακής εργασίας

0.1 Εισαγωγή

0.2 Πληροφορική & Ιατρική

0.3 Κατηγορίες σύμπραξης πληροφορικής και ιατρικής

0.4 Εισαγωγή στη Βιοπληροφορική

1.1 Τι είναι οι βιοδείκτες;

1.2 Τελικές ενδείξεις και βιοδείκτες

1.3 Πρωτεϊνικοί Βιοδείκτες

2 Βιοδείκτες, Ανάγκη Ιατρικής προσέγγισης

3.1 Επεξεργασία και ανάλυση δεδομένων Βιολογικών Συστημάτων

3.1.2 Εύρεση Βιοδείκτη

3.2 Επεξεργασία και Ανάλυση

4 Μηχανική Μάθηση

4.1 Εκπαίδευση

4.2 Τύποι Μηχανικής Μάθησης

4.3 Μοντέλο Γραμμικής Παλινδρόμησης

4.4 Νευρωνικά Δίκτυα

4.5 Στρώματα Νευρωνικών δικτύων

4.6 Πολυστρωματικά Τροφοδοτούμενα προς τα εμπρός Δίκτυα

4.7 K-MEANS (γνώση πλήθους ομάδων)

4.8 Τεχνικές Οπτικοποίησης

4.9 Self Organizing Maps

4.10 3D Printing & U*Matrix

4.11 Principal Component Analysis

4.12 SNP Ανάλυση

4.13 Significance Analysis of Microarrays

4.14 SUPPORT VECTOR MACHINE

4.15 Naïve Bayes

5.1 Βιολογικές Βάσεις Δεδομένων

5.2 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΝΟΥΚΛΕΟΤΙΔΙΚΩΝ ΑΛΛΗΛΟΥΧΙΩΝ

5.3 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΠΡΩΤΕΪΝΙΚΩΝ ΑΛΛΗΛΟΥΧΙΩΝ

5.4 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΓΕΝΕΤΙΚΗΣ ΠΟΙΚΙΛΟΜΟΡΦΙΑΣ

5.5 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

5.6 ΔΕΥΕΡΟΓΕΝΕΙΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

5.7 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΟΙΚΟΓΕΝΕΙΩΝ

5.8 ΕΞΕΙΔΙΚΕΥΜΕΝΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

6.1 Decision Trees

6.2 Ταξινόμηση με Δέντρα Απόφασης

6.2.1 Single Decision Tree Classifiers

6.3 Decision Forests

6.4 Data Mining (Εξόρυξη Δεδομένων)

7 Omics and Translational Research Platforms

8.1 Ο ανθρώπινος εγκέφαλος (δομή)

8.1.1 Εγκεφαλικός φλοιός

8.2 Κυτταρική Οργάνωση Του Εγκεφάλου

8.3 Νευρωνικές Συνάψεις

9.1 Εισαγωγή στο Αλτσχάιμερ

9.2 Ήπια Γνωστική Εξασθένηση (Mild Cognitive Impairment)

9.3 Παράγοντες Κινδύνου

9.4 Βιοχημεία

9.5 Νευροπαθολογία

9.6 Βιοδείκτες

9.7 Συμπτώματα

9.8 Διάγνωση Αλτσχάιμερ

10 Συμπέρασμα

BIBΛΙΟΓΡΑΦΙΑ

0.0 Σκοπός πτυχιακής εργασίας

Η εκπόνηση της συγκεκριμένης πτυχιακής εργασίας, έχει ως στόχο να μελετήσει την σύμπραξη των επιστημών της Ιατρικής και της Πληροφορικής. Λεπτομερώς θα ερευνηθούν οι εφαρμογές και τα εργαλεία που χρησιμοποιούνται στο πεδίο της Βιοπληροφορικής, οι αλγόριθμοι που εφαρμόζονται, οι ασθένειες στις οποίες μπορούν να εφαρμοστούν και η συνεισφορά τους, στην διαδικασία αντιμετώπισης τους.

0.1 Εισαγωγή

Η εξέλιξη της ιατρικής ανά τους αιώνες συνδυάζεται πάντα με τις αντίστοιχες τεχνολογικές και κοινωνικές εξελίξεις. Η ραγδαία εξέλιξη της πληροφορικής, από μια ιδέα στη φαντασία των οραματιστών πριν από έναν αιώνα μέχρι τη σημερινή της μορφή, ως συνδετικός ιστός όλων των ανθρώπινων δραστηριοτήτων, με την παράλληλη ανάπτυξη των τηλεπικοινωνιακών δικτύων και την επανάσταση της πληροφορίας, αποτέλεσε και αποτελεί το γόνιμο έδαφος πάνω στο οποίο πολλοί τομείς της ανθρώπινης δραστηριότητας εξελίσσονται σε νέα επίπεδα. Ο τομέας της ιατρικής και γενικότερα της υγείας αποτελεί ίσως τον σημαντικότερο τομέα έρευνας και ανάπτυξης, καθώς έχει ως επίκεντρό του τον άνθρωπο. Η πληροφορική της υγείας (Health informatics) είναι το αποτέλεσμα αυτής της σύμπραξης μεταξύ τηλεπικοινωνιών και πληροφοριακών εφαρμογών με την ιατρική και όλων των συναφών με αυτή λειτουργιών. Η έννοια της πληροφορικής της υγείας είναι ιδιαίτερα ευρεία και περιλαμβάνει όλες τις εφαρμογές πληροφορικής και τηλεπικοινωνιών στον χώρο της υγείας, συμπεριλαμβανομένων συνεπώς όλων των υπολοίπων συναφών εννοιών, όπως η ιατρική πληροφορική (medical informatics), τα πληροφοριακά συστήματα υγείας (Health Information Systems), ο ηλεκτρονικός ιατρικός φάκελος (electronic medical record), η τηλεϊατρική (telemedicine), η κινητή υγεία (mhealth) και φυσικά η Βιοπληροφορική - Bioinformatics. [1]

0.2 Πληροφορική & Ιατρική

Η πληροφορική είναι η επιστήμη που ασχολείται με την έννοια της πληροφορίας και των δεδομένων, καθώς και όλων των σχετικών με αυτά ενεργειών, όπως η απόκτηση, η καταγραφή, η διαχείριση, η επεξεργασία, η κωδικοποίηση, η μετάδοση, η ανάλυση, η κρυπτογράφηση και η συμπίεση. Η πληροφορική, που είναι ίσως η νεότερη από τις σύγχρονες επιστήμες, γνώρισε αλματώδη εξέλιξη με την ανάπτυξη των ηλεκτρονικών υπολογιστών και των υπολογιστικών συστημάτων, καθώς μέσω αυτών μπορούν να υλοποιηθούν με αυτοματοποιημένο τρόπο τεχνικές επεξεργασίας μεγάλου όγκου δεδομένων. Τα πληροφοριακά συστήματα είναι συστήματα που είναι σχεδιασμένα να υλοποιούν μια ή περισσότερες από τις παραπάνω λειτουργίες σε ένα σύνολο δεδομένων ή πληροφοριών. Επίσης, η ανάπτυξη της πληροφορικής έχει μεγάλο αντίκτυπο σε άλλα πεδία ανθρώπινης δραστηριότητας. Η ιατρική είναι η επιστήμη που ασχολείται με την έρευνα και την εφαρμογή τεχνικών και μεθόδων για τη διάγνωση και τη θεραπεία των ασθενειών του ανθρώπου. Θέτοντας στο επίκεντρό της τον άνθρωπο και την υγεία του, η ιατρική αποτελεί την αρχαιότερη των επιστημών με την εξέλιξή της να είναι συνυφασμένη με αυτή του ανθρώπου, έχοντας τις απαρχές της στις πρώτες ανθρώπινες κοινωνίες. Βγαίνοντας από το χώρο της τέχνης και της μαγείας, όπου ανήκε μέχρι πριν από λίγες εκατοντάδες χρόνια, η ιατρική απέκτησε επιστημονική υπόσταση και πλέον αποτελεί τον ερευνητικό κλάδο, ο οποίος στο σύγχρονο ερευνητικό τοπίο συγκεντρώνει το μεγαλύτερο ενδιαφέρον. Ο όρος «ιατρική πληροφορική» αναφέρεται κυρίως στον κλάδο που σχετίζεται με τις εφαρμογές της πληροφορικής στην ιατρική. Συνδυάζοντας την αρχαιότερη με μια από τις νεότερες επιστήμες, η ιατρική πληροφορική είναι ένας χώρος σύνθεσης ιδεών, με την πληροφορική να δρα επικουρικά ως προς την

ιατρική αλλά και ταυτόχρονα να μεγεθύνει σε υπερθετικό βαθμό τα αποτελέσματα αυτής στην κοινωνία. Παράλληλα, μέσω της ιατρικής πληροφορικής ανοίγονται και νέοι δρόμοι παροχής υπηρεσιών υγείας, όπως στην τηλεϊατρική ή την κινητή υγεία, οι οποίοι παρόλο που δεν αποτελούν νέους κλάδους της ιατρικής επιστήμης, επιτρέπουν την εφαρμογή της σε περιοχές που ήταν απρόσιτες με τις έως τώρα κοινά εφαρμοζόμενες πρακτικές. Τυπικά, η ιατρική πληροφορική είναι υποσύνολο της πληροφορικής της υγείας, καθώς αφορά μόνο ένα από τα υποσύνολα της υγείας, αυτό της ιατρικής και δεν καλύπτει άλλα σημαντικά τμήματα, όπως η νοσηλευτική και η φαρμακευτική. απ' όλα αυτά όμως, ο όρος «ιατρική πληροφορική» χρησιμοποιείται ως συνώνυμο αυτού της «πληροφορικής της Υγείας», όπως προκύπτει και από ορισμούς που έχουν παρουσιαστεί στη βιβλιογραφία, όπως: «ιατρική πληροφορική είναι ο διεπιστημονικός κλάδος που ασχολείται με τη μελέτη, τον σχεδιασμό, την ανάπτυξη, την υιοθέτηση και την εφαρμογή της πληροφορικής στον τομέα των υπηρεσιών υγειονομικής περίθαλψης, διαχείρισης και σχεδιασμού», που έχει δοθεί από την Εθνική ιατρική Βιβλιοθήκη των Η.Π.Α. [1]

0.3 Κατηγορίες σύμπραξης πληροφορικής και ιατρικής

Η πληροφορική της υγείας περιλαμβάνει τα ακόλουθα πεδία (όπως αναφέρονται στο Ηλεκτρονικό Ακαδημαϊκό σύγγραμμα “Ιατρική Πληροφορική Ψηφιακή Επεξεργασία Βιοσημάτων” των Μάρκος Γ. Τσιπούρας Αλέξανδρος Τ. Τζάλλας Ευάγγελος Χ. Καρβούνης Νικόλαος Γιαννακέας) :

- Ιατρική πληροφορική (medical informatics) με επιμέρους διεπιστημονικά πεδία την κλινική πληροφορική (clinical informatics), την πληροφορική της παθολογίας (pathology informatics) και την πληροφορική ιατρικής εκπαίδευσης (medical education informatics)
- Νοσηλευτική πληροφορική (nursing informatics)
- Φαρμακευτική πληροφορική (pharmacy informatics)
- Πληροφορική δημόσιας υγείας (public health informatics)
- Πληροφορική για έρευνα στην υγεία (informatic for health research)
- Πληροφορική για εκπαίδευση στην υγεία (healthcare education informatics)
- Πληροφορική καταναλωτών υγείας (consumer health informatics)
- Βιοπληροφορική (bioinformatics) και κλινική βιοπληροφορική (clinical bioinformatics)

0.4 Εισαγωγή στη Βιοπληροφορική

Ο όρος βιοπληροφορική περιλαμβάνει την σύμπραξη της μοριακής βιολογίας με την πληροφορική, τη στατιστική και τα μαθηματικά και μελετά νέους τρόπους για προσέγγιση των βιολογικών προβλημάτων, καθώς και την αντίληψη βασικών αρχών της βιολογίας. Ως ανεξάρτητη επιστήμη πρωτοεμφανίστηκε την δεκαετία του 1990, αλλά χρωστάει την ύπαρξη του στις απαρχές της σύγχρονης μοριακής βιολογίας (δεκαετίες 1950 - 1960) και σε συγκεκριμένα πειράματα που προσδιόρισαν την ύπαρξη κάποιας μορφής ψηφιακής πληροφορίας στις βιολογικές αλληλουχίες. Η έκρηξη της μοριακής βιολογίας εκείνες τις δεκαετίες, εισήγαγε τα θεμέλια της βιοπληροφορικής με τον προσδιορισμό της τρισδιάστατης δομής του DNA και την εύρεση του RNA. Το DNA και το RNA αποτελούν πρωτεΐνες ως ψηφιακή πληροφορία και είναι βιολογικά μοκρομόρια, ουσιαστικά στον κλάδο της πληροφορικής θεωρούνται συμβολοσειρές. Το DNA μπορεί θεωρείται και ως μια ακολουθία χιλιάδων νουκλεοτιδίων, τα οποία αποτελούνται από τέσσερα είδη και τα σύμβολα τους, την αδενίνη

(A), την θυμίνη (T), την γουανίνη (G) και την κυτοσίνη (C). Κάνοντας χρήση των συμβόλων των νουκλεοτιδίων κατασκευάζονται οι συμβολοσειρές. Αντίστοιχα με παρόμοιο τρόπο, οι πρωτεΐνες έχουν την δυνατότητα περιγραφής ως ακολουθίες αμινοξέων καιέτοιμο είδους αναπαραστάσεις είναι ιδανικές για επεξεργασία σε ηλεκτρονικό υπολογιστή με σκοπό την εξαγωγή χρήσιμων συμπερασμάτων.

Μια από τις μεγαλύτερες προκλήσεις που προέκυψαν ως μέρος των πρόσφατων προόδων στον τομέα της βιοπληροφορικής, είναι το τεράστιο ποσό των δεδομένων που παράγεται με έναν συνεχώς αυξανόμενο ρυθμό. Χρησιμοποιώντας τεχνικές όπως η επόμενη γενιά RNA και αλληλουχία του DNA, οι ερευνητές έχουν την δυνατότητα πρόσβασης σε εξαιρετικά ακριβείς πληροφορίες σχετικά με ολόκληρα γονιδιώματα. Αυτός ο τεράστιος όγκος δεδομένων δημιούργησε ένα πρόβλημα πολυπλοκότητας, γεγονός που καθιστά αδύνατη την εύρεση απαντήσεων στον επιθυμητό βαθμό ακρίβειας. Η μηχανική μάθηση είναι ένας διεπιστημονικός τομέας της βιοπληροφορικής που περιλαμβάνει μια κατηγορία δεδομένων αλγορίθμων που αναζητούν λύσεις σε ένα δεδομένο πρόβλημα, με την εφαρμογή μελέτης και ανάλυσης προτύπων σε σύνολα δεδομένων βάσει γνωστών παραγόντων όπως η έκφραση γονιδίων και οι κλινικές πληροφορίες σε ένα πλήθος από περιπτώσεις. Αυτές οι προσεγγίσεις έχουν χρησιμοποιηθεί ευρέως και με επιτυχία στην βιολογία, ιδιαίτερα στις μελέτες εκείνες που έχουν ως στόχο την ανακάλυψη βιοδεικτών, λόγω της πολυμορφίας τους και της ισχύς που παρέχουν, η μελέτη αυτών έχει οδηγήσει στην δυνατότητα χρήσης αλγορίθμων μηχανικής μάθησης και μεθοδολογίας. Με την συγκεκριμένη μέθοδο, υπάρχει η δυνατότητα εφαρμογής και ανάπτυξης ενός Τεχνητού Νευρωνικού Δικτύου με σκοπό να ανακαλύψουν, να αναλύσουν και να επικυρώσουν νέους βιοδείκτες στις διάφορες ασθένειες. [1,2]

1.1 Τι είναι οι βιοδείκτες;

Ο όρος «βιοδείκτης» εμφανίστηκε στο προσκήνιο των επιστημών της ιατρικής και της βιοπληροφορικής το 1998. Ο ορισμός που δόθηκε τότε στην έννοια των βιοδεικτών από το National Institutes of Health Biomarkers Definitions Working Group((εθνικά ινστιτούτα υγείας και ομάδας εργασίας ορισμού βιοδεικτών) ήταν ο εξής: «Ένα χαρακτηριστικό το οποίο είναι αντικειμενικά μετρήσιμο και αξιολογήσιμο ως ένδειξη βιολογικών/παθογόνων διαδικασιών ή φαρμακολογικών αντιδράσεων σε μια θεραπευτική παρέμβαση». Με άλλα λόγια σύμφωνα με τον παραπάνω ορισμό, η έννοια βιοδείκτης αποτελεί μια υποκατηγορία μετρήσιμων ιατρικών ενδείξεων του ασθενή που παρατηρούνται εξωτερικά αυτού και βοηθούν στην διάγνωση μιας πάθησης-αρρώστιας καιτην παρακολούθηση της εξέλιξης της. Αναλύοντας την εξήγηση αυτή, υπό μια οπτική μπορεί να τα ταυτίσει κανείς με τα συμπτώματα ενός ασθενή, παρ' όλα αυτά δεν σχετίζονται άμεσα. Οι βιοδείκτες ποικίλουν, από τις απλούστερες μορφές τους, τις οποίες μπορεί κανείς να μετρήσει με καθημερινά «εργαλεία» μέτρησης, όπως ένα πιεσόμετρο που μπορεί να μετρήσει την αρτηριακή πίεση του αίματος, όπως ένας παλμογράφος που καταγράφει τους παλμούς, ακόμα κι ένα απλό θερμόμετρο που χρησιμοποιείται για τον έλεγχο της θερμοκρασίας του σώματος, μέχρι και πολυσύνθετα εργαστηριακά πειράματα που απαιτούν τελείως διαφορετικά εργαλεία από τα απλά καθημερινά και έχουν μεγαλύτερη ακρίβεια σε πολύ πιο εξειδικευμένες μετρήσεις, τα οποία εφαρμόζονται στο αίμα και σε άλλους ιστούς. Τέτοιες περιπτώσεις των βιοδεικτών αποτελούν πειράματα των οποίων οι μετρήσεις αναφέρονται σε αντικειμενικά ποσοτικοποιημένα χαρακτηριστικά βιολογικών διεργασιών της κλινικής κατάστασης του ασθενούς, που αποτελεί και το κύριο αντικείμενο των περισσότερων ερευνών, αν όχι όλων.

Στα κλινικά εργαστηριακά πειράματα που πραγματοποιήθηκαν με χρήση βιοδεικτών. Οι μετρήσεις αναφέρονται σε ουσίες, δομές ή διαδικασίες που προέρχονται από το σώμα (το ίδιο το σώμα ή παράγωγα του), οι οποίες μπορούν να προβλέψουν σε τι κλινική κατάσταση-περίπτωση βρίσκεται ο ασθενής, ουσιαστικά να διαγνώσουν την ύπαρξη μίας πιθανής αρρώστιας, δηλαδή στην φάση της διάγνωσης. Η μέτρηση αυτών μπορεί να υπολογιστεί βιοχημικά σε κυτταρικό επίπεδο ή μέσω μοριακών αντιδράσεων. Με άλλα λόγια οι βιοδείκτες είναι βιολογικά χαρακτηριστικά τα οποία μετριοούνται αντικειμενικά και αξιολογούνται ως δείκτες κανονικών βιολογικών διεργασιών, παθογόνων διεργασιών ή φαρμακολογικών αποκρίσεων σε θεραπευτική παρέμβαση, δηλαδή δεν συμβάλουν μόνο στην διάγνωση. Αν και η συνεισφορά τους μπορεί να αποδειχθεί ιδιαίτερα ωφέλιμη καθώς εντοπίζει την πιθανή ασθένεια που θα πλήξει το αντικείμενο ακόμα και πριν αυτή κάνει την εμφάνιση της. Κυρίως σε θανάσιμες ασθένειες, η διάγνωση τους διαδραματίζει τεράστιο ρόλο στην βιωσιμότητα του αντικειμένου, επομένως μπορούν να αποτελέσουν έναν τεράστιο εφόδιο. Οι βιοδείκτες επίσης μπορούν να χρησιμοποιηθούν για τον προσδιορισμό της εμφάνισης της νόσου, της εξέλιξης, της αποτελεσματικότητας της θεραπείας με φάρμακα, της ευαισθησίας του ασθενούς για την ανάπτυξη ενός συγκεκριμένου τύπου ασθένειας ή την πρόβλεψη της αποτελεσματικότητας της θεραπείας σε ένα συγκεκριμένο στάδιο της νόσου. Οι πρωτεϊνικοί μοριακοί βιοδείκτες είναι ιδιαίτερα δημοφιλείς λόγω της διαθεσιμότητας ενός μεγάλου εύρους αναλυτικών οργάνων-εργαλείων, τα οποία μπορούν να αναγνωρίσουν και να ποσοτικοποιήσουν πρωτεΐνες από σύνθετα βιολογικά δείγματα.

Οι πρωτεϊνικοί μοριακοί βιοδείκτες αναφέρονται σε πρωτεΐνες η οποίες είναι βασικές ενώσεις στη βιοσύνθεση, στη σηματοδότηση των κυττάρων, των ιστών και των οργάνων και παρέχουν δομική σταθερότητα κυττάρων και ιστών σε ζωντανούς οργανισμούς. Οι πρωτογενείς πρωτεϊνικές αλληλουχίες κωδικοποιούνται στο γονιδίωμα. Ωστόσο, οι σύνθετες μετά-μεταφραστικές τροποποιήσεις (PTM) και οι τρισδιάστατες δομές είναι αρκετά απρόβλεπτες από τις πληροφορίες γονιδιώματος με αποτέλεσμα να είναι δύσκολη η απευθείας εξαγωγή συμπερασμάτων, ειδικά όταν η δομή τους είναι πολυδιάστατη. [3]

1.2 Τελικές ενδείξεις και βιοδείκτες

Στην περίπτωση διάγνωσης μιας νόσου και κατά την παρακολούθηση της, το αντικείμενο μελέτης (ασθενής) εξετάζεται συνήθως ως προς κάποιες ενδείξεις που μπορεί πολλές φορές ακόμα και ο ίδιος να διαπιστώσει και να προσδιορίσει έμμεσα ή άμεσα, σύμφωνα με το πως νιώθει, λειτουργεί, ή επιβιώνει ο ίδιος. Αυτές οι ενδείξεις αναφέρονται ως τελικές. Δεδομένου ότι όλες οι φαρμακευτικές αγωγές που παρέχονται μπορούν να ελλοχεύουν κινδύνους ως προς την υγεία του, ένας ασθενής θα επέλεγε να ακολουθήσει μια τέτοια αγωγή μόνο αν ήταν βέβαιος ότι τα φάρμακα που του συνταγογραφήθηκαν θα έχουν ως αποτέλεσμα ένα όφελος που μπορεί να το διακρίνει ο ίδιος, ή αν μειώνει τον κίνδυνο εμφάνισης μιας ανεπιθύμητης πάθησης ή επιπλοκής της νόσου από την οποία πλήττεται ήδη. Ουσιαστικά πρόκειται για μεταβλητές που αναπαριστούν το αποτέλεσμα της έρευνας της υγείας του αντικειμένου, από την οπτική του ίδιου. Συνεπώς, οι πρωτεύοντες τελικές ενδείξεις πρέπει να είναι ένα άμεσο μέτρο αυτών των ανεπιθύμητων συνθηκών ή επιπλοκών της νόσου. Υπάρχουν περιπτώσεις νόσων που κατά την μελέτη τους σε ασθενείς, οι παραπάνω τελικές ενδείξεις δεν μπορούν να είναι αντιπροσωπευτικά δείγματα μετρήσεων της νόσου υπό μελέτη και της πιθανής εξέλιξης της ή της βιολογική αντίδρασης του ασθενή κατά την συνταγογράφηση των φαρμάκων. Σε τέτοιου είδους κλινικές δοκιμές θεωρείται επιτακτική η ανάγκη προσδιορισμού της πορείας της νόσου στο αντικείμενο μελέτης με ακριβείς μετρήσεις.

Η χρήση των βιοδεικτών έχει την ιδιότητα αναπληρωματικών τελικών ενδείξεων, δηλαδή διαδραματίζουν τον ρόλο υποκατάστατων στις δεύτερες. Η μορφή αυτών είναι σε μικρά υποσύνολα που προσδιορίζουν συγκεκριμένες ιατρικές μετρήσεις όπως προαναφέρθηκε. Οι βιοδείκτες μπορούν να χρησιμοποιηθούν για διαφορετικούς σκοπούς κατά την διάρκεια μιας μελέτης, όπως η διαστρωμάτωση του κινδύνου από την ασθένεια, η πρόληψη, η διαλογή, η διάγνωση, η ταξινόμηση και η πρόγνωση αυτής. Ωστόσο, παρόλο που πολλοί βιοδείκτες μπορούν να συσχετιστούν με μια ασθένεια και να έχουν ένα ευρύ φάσμα χρήσεων, μόνο μερικοί δυνητικά μπορούν να χαρακτηριστούν ως υποκατάστατα των τελικών ενδείξεων που θα χρησιμοποιηθούν σε κλινικές δοκιμές. Για να χαρακτηριστεί ως υποκατάστατη τελική ένδειξη, ένας βιοδείκτης πρέπει να επιδείξει σημαντική ικανότητα στην πρόβλεψη ενός κλινικά σχετικού και ακριβή αποτελέσματος. [2] Η διαφορά έγκειται στην λεπτομέρεια που παρέχουν οι βιοδείκτες, καθώς αποτελούν πολύ συγκεκριμένες μετρήσεις και οι ασθενείς δεν μπορεί να τις εντοπίσει ή να νιώσει την αλλαγή τους άμεσα όπως θα μπορούσε να συμβεί στην αλλαγή της θερμοκρασίας του για παράδειγμα.

Τα δεδομένα που παρέχουν οι βιοδείκτες λόγω της ακρίβειας τους δίνουν στους ερευνητές την ευκαιρία να διακόψουν τις παρεμβάσεις που μακροπρόθεσμα θεωρούνται επιβλαβή για τα αντικείμενα, προτού τα κλινικά δεδομένα είναι διαθέσιμα. Σε άλλες περιπτώσεις, οι βιοδείκτες μπορεί να επιτρέπουν στους ερευνητές να σχεδιάσουν μικρότερες και αποδοτικότερες μελέτες, μειώνοντας το πλήθος των αντικειμένων που εκτίθενται σε πειραματικές θεραπείες. Η μείωση του χρόνου έγκρισης μιας θεραπείας είναι όσο θετικό ακούγεται, καθώς οδηγεί στην εφαρμογή περισσότερων αποδοτικών δοκιμών και έτσι επιταχύνεται η διαδικασία ανάπτυξης των κατάλληλων για τον ασθενή φαρμάκων, επιτρέποντας στις επιβεβαιωμένες αποδοτικές πλέον θεραπείες να επηρεάσουν μεγαλύτερες ομάδες ασθενών και μάλιστα νωρίτερα. Γεγονός που οδηγεί στο θετικό παρεπόμενο, εξοικονόμησης και υλικών και ανθρώπινων πόρων για άλλα ερευνητικά πρότζεκτ ακόμα και για άλλες κλινικές δοκιμές. Η αντικατάσταση των πρωταρχικών ενδείξεων με βιοδείκτες μπορεί να αποδειχτεί προσοδοφόρα σε συγκεκριμένες περιπτώσεις νόσων. Ειδικότερα σε αυτές που η πορεία θεραπείας του ασθενή είναι δύσκολο να παρακολουθηθεί με την χρήση των έμμεσων ενδείξεων, αλλά απαιτείται ειδικότερη μελέτη της βιολογικής υπόστασης του αντικειμένου. Τέτοιες βιολογικές μετρήσεις σε κλινικά πειράματα πραγματοποιούνται σε μοριακό και κυτταρικό επίπεδο, για αυτό βέβαια η εφαρμογή των βιοδεικτών δεν βρίσκεται ακόμα σε καθολικό επίπεδο, αν και τα θετικά επακόλουθα της εφαρμογής τους θα είναι πολλαπλά.

1.3 Πρωτεϊνικοί Βιοδείκτες

Ο ρόλος της βιοπληροφορικής στην έρευνα των βιολογικών δεικτών είναι η εξαγωγή πρωτεϊνών και πεπτιδίων - κυρίως πρωταρχικής ακολουθίας συμπεριλαμβανομένων ενδεχόμενων τροποποίησης και ποσοτικοποίησης πληροφοριών αυτόματα από σύνθετα και μεγάλου όγκου δεδομένα που αποκτούν οι αναλυτικές πλατφόρμες. Επιπρόσθετα, η βιοπληροφορική συμβάλλει στην επιλογή πρωτεϊνικών ή πεπτιδικών βιοδεικτών με την υποστήριξη στατιστικών μεθόδων στη φάση ανακάλυψης και για την καθιέρωση της στατιστικής ισχύος του υποψηφίου σε ανεξάρτητο μεγάλο δείγμα στη φάση επικύρωσης. Οι αλγόριθμοι βιοπληροφορικής είναι προσαρμοσμένοι για συγκεκριμένα δεδομένα και ως εκ τούτου απαιτούν διαφορετικούς αλγόριθμους για επεξεργασία δεδομένων πρωτεϊνών, 2DE ή LC-MSn. Γενικά, οι ροές εργασίας βιοπληροφορικής είναι πολύπλοκες και επηρεάζονται από τον πειραματικό σχεδιασμό. Για παράδειγμα, μια ροή εργασίας για την ανάλυση των δεδομένων της ανακάλυψης πρωτεϊνικών LC-MSn απαιτεί ένα τμήμα προ επεξεργασίας με μονάδες προσδιορισμού και ποσοτικοποίησης πεπτιδίων / πρωτεϊνών, μια ενότητα για την ενσωμάτωση πληροφοριών ποσοτικού προσδιορισμού και ταυτοποίησης, μια ενότητα για στατιστική ανάλυση και μια ενότητα για διερεύνηση της πιθανής εμπλοκής των υποψηφίων βιοδεικτών σε βιοσύνθεση,

ρυθμιστικές ή σηματοδοτικές οδούς. Οι ενότητες ποσοτικοποίησης μπορούν να περιλαμβάνουν μια ποικιλία στοιχείων, όπως φιλτράρισμα για τον θόρυβο, ταυτοποίηση και ποσοτικοποίηση κορυφών, ευθυγράμμισης του χρόνου, κανονικοποίηση, βαθμονόμηση μάζας, αντιστοίχιση των ίδιων κορυφών σε πολλαπλά χρωματογραφήματα, με σκοπό την παροχή ενός μη εξαντλητικού καταλόγου εργασιών. Η τρέχουσα πρόκληση που αντιμετωπίζουν οι μελετητές είναι να εκτιμηθεί με ακρίβεια η απόδοση αυτών των πολύπλοκων ροών εργασίας, γεγονός που υποστηρίζεται από την αυξανόμενη διαθεσιμότητα τυποποιημένων δεδομένων με γνωστές ποσοτικές πληροφορίες και πληροφορίες ταυτότητας για ορισμένα ή όλα τα πρωτεϊνικά συστατικά. [40] Οι τρέχουσες λύσεις βιοπληροφορικής είναι ανεπαρκείς όσον αφορά την ακριβή εξαγωγή ποσοτικών και πρωτεϊνικών πληροφοριών αναγνώρισης και οι δημοσιευμένες νέες εξελίξεις συχνά δεν περιλαμβάνονται σε ένα περιβάλλον λογισμικού που επιτρέπει στους μη ειδικούς από τα «wet laboratories» να εκτελούν την επεξεργασία δεδομένων. Ένα καλό παράδειγμα είναι η συνήθως χρησιμοποιούμενη εμπορική database-centric μέθοδος προσδιορισμού πεπτιδίων για δεδομένα LC-MSn, η οποία θεωρείται ανεπαρκή στην ταυτοποίηση απροσδόκητων PTMs. Εντούτοις, αναδύεται μια νέα σειρά εναλλακτικών μεθόδων αναγνώρισης όπως οι φασματικές βιβλιοθήκες ή τα φασματικά αρχεία, η ανάλυση φασματικών δικτύων ή η αναζήτηση ανοιχτών τροποποιήσεων καθώς και οι αλγόριθμοι de novo προσδιορισμού αλληλουχίας, οι οποίοι είναι πολύ καλύτεροι στην ανίχνευση απλών μη αναμενόμενων PTM που εμφανίζονται ως μετατοπίσεις μάζας στα φάσματα. Υπάρχει επείγουσα ανάγκη για περαιτέρω εξελίξεις, για παράδειγμα για τον προσδιορισμό της πρωταρχικής δομής των γλυκοπεπτιδίων με αυτοματοποιημένο τρόπο ή για την ανάπτυξη αποτελεσματικών προσεγγίσεων μεταφοράς κορυφών για την αύξηση του εύρους δυναμικής συγκέντρωσης των σημειωμένων κορυφών σε μονοφασικά δεδομένα LC-MSn.

2 Βιοδείκτες, Ανάγκη Ιατρικής προσέγγισης

Οι περισσότερες από τις νόσους που πλήττουν τον ανθρώπινο οργανισμό είναι σύνθετες ασθένειες, οι οποίες θα μπορούσαν να προκληθούν από πολλές γενετικές οδούς. Αυτό σημαίνει ότι για ένα δοσμένο φαινότυπο, (δηλ. Όλα εκείνα τα χαρακτηριστικά που εκδηλώνει ένας οργανισμός και αποτελούν το μέρος του γονότυπου του οργανισμού το οποίο είναι εφικτό έμμεσα ή άμεσα να παρατηρηθεί.) υπάρχουν πολλαπλά πιθανά γονίδια που θα μπορούσαν να είναι γενωμικά ή επιγενετικά (δηλαδή μεταλλάξεις, παραλλαγές αριθμού αντιγράφων, επιγενετικές τροποποιήσεις κ.ο.κ.). Ως εκ τούτου, είναι κατανοητό ότι διαφορετικά άτομα που μοιράζονται τον ίδιο φαινότυπο / ασθένειες μπορεί να έχουν διαφορετικά αιτιώδη γονίδια και έτσι μπορεί να χρειάζονται φάρμακα διαφορετικού τύπου, γεγονός που περιπλέκει την ασθένεια και δεν βοηθά στην άμεση αντιμετώπιση της μέσω της εμπειρικής γνώσης. Για παράδειγμα, τα μεταλλαγμένα γονίδια είναι σπάνια μεταξύ των καρκινοπαθών του ίδιου τύπου καρκίνου [4]. Επιπρόσθετα, για ένα δεδομένο φάρμακο, μόνο το 10% - 30% των ασθενών του ίδιου τύπου καρκίνου ανταποκρίνονται σε αυτό. Γεγονός που υποδηλώνει ότι τα γονιδιωμικά και άλλα χαρακτηριστικά γνωρίσματα ή / και το περιβάλλον και φυσικά ο τρόπος ζωής, θα μπορούσαν να συμβάλουν στην διαφορετική ανταπόκριση των φαρμάκων. [5] Είναι λοιπόν σαφές ότι θα πρέπει να χορηγηθεί το «σωστό φάρμακο» στον «σωστό ασθενή» τη σωστή στιγμή. Ένας από τους σκοπούς των προγραμμάτων δημιουργίας φαρμάκων που βρίσκονται σε εξέλιξη είναι να επιτευχθεί αυτός ο στόχος χρησιμοποιώντας omic (δηλ. γονιδιωματική, πρωτεϊνωματική, επιγενετική και ούτω καθεξής) και / ή περιβαλλοντικούς παράγοντες και παράγοντες του τρόπου ζωής των ατόμων και όλα αυτά για κάθε ασθενή ξεχωριστά.

Οι τεχνολογίες υψηλής απόδοσης οδηγούν στην εξέλιξη των πεδίων της βιολογίας και της ιατρικής. Για να επιτευχθεί ιατρική ακρίβεια, είναι απαραίτητος ο εντοπισμός βιολογικών δεικτών που χρησιμοποιούν είτε omic δεδομένα μόνα τους είτε με περιβαλλοντικούς παράγοντες και τους παράγοντες του τρόπου ζωής των ασθενών. Η πρόκληση είναι το πως θα μετατραπούν τα δεδομένα σε βιοδείκτες που θα μπορούσαν να προβλέψουν τα κλινικά αποτελέσματα, την ανταπόκριση των ασθενών στην φαρμακευτική αγωγή που τους παρέχεται ή σε άλλα καίρια χαρακτηριστικά που θα συμβάλουν στην διάγνωση, στην παρακολούθηση καθώς και στην θεραπεία μίας νόσου. Σε γενικές γραμμές, θεωρείται δύσκολη η ταυτοποίηση υψηλού επιπέδου βιοδεικτών, δηλαδή με υψηλή ακρίβεια και ευρωστία, με τη χρήση omic δεδομένων όπως τα πρωτεϊνωματικά δεδομένα κ.ο.κ. Για παράδειγμα, πολλοί omic-based καρκινικοί βιοδείκτες δεν χαρακτηρίζονται ως εύρωστοι, πράγμα που σημαίνει ότι ο βιοδείκτης που προσδιορίζεται από μια ομάδα ασθενών χάνει την δύναμη της πρόβλεψης σε σχέση με άλλες ομάδες ασθενών που πλήττονται από τον ίδιο τύπο καρκίνου [3 & 5]. Έχουν καταβληθεί προσπάθειες για την ανάπτυξη νέων αλγορίθμων με σκοπό την λύση του παραπάνω προβλήματος. Για παράδειγμα, ο πολλαπλός έλεγχος επιβίωσης (MSS) και Ανάλυση Σημαντικών Προγνωστικών Υπογραφών (SAPS) έχουν αναπτυχθεί για τον εντοπισμό ισχυρού καρκίνου. [4,6,7].

Στην εποχή που ακολούθησε μετά τον προσδιορισμό της αλληλουχίας του γονιδιώματος, ο προσδιορισμός αυτής όπως είναι λογικό γίνεται όλο και φθηνότερος, πράγμα που καθιστά την αλληλουχία του γονιδιώματος να γίνει οικονομικά προσιτή και προσβάσιμη σε μια ευρύτερη κλινική κλίμακα. Ως εκ τούτου, ο εντοπισμός βιοδεικτών χρησιμοποιώντας το σύνολο γενώματος δεδομένων ή το σύνολο exome δεδομένων, είναι πολύ πιο ελκυστικό. Παρ' όλα αυτά, γνωρίζοντας τα προαναφερθέντα χαρακτηριστικά των σύνθετων ασθενειών, θεωρείται πρόκληση η κατασκευή προγνωστικών μοντέλων (μέσω βιοδεικτών) χρησιμοποιώντας την αλληλουχία ολόκληρου του

γονιδιώματος ή του συνόλου αλληλουχίας exome δεδομένων. Επειδή οι αλληλεπιδράσεις πολλαπλών γονιδίων διέπουν τους υποκείμενους μοριακούς μηχανισμούς των σύνθετων ασθενειών, η προσέγγιση του γραμμικού μοντέλου δεν αποτελεί επιλογή για την αναγνώριση βιοδεικτών που χρησιμοποιούν τα δεδομένα αλληλουχίας γονιδιώματος. Μία δικτυοκρατική και μη γραμμική προσέγγιση θα μπορούσε να δώσει την επίλυσή του προβλήματος. Για παράδειγμα, το Cancer Hallmark Network Framework (CHNF) [9] παρέχει μια λύση στο πρόβλημα. Πρόσφατα, έχει αναπτυχθεί ένας αλγόριθμος που βασίζεται σε CHNF και προβλέπει με ακρίβεια την επανεμφάνιση του καρκίνου του μαστού χρησιμοποιώντας το σύνολο δεδομένων αλληλουχίας των exome των όγκων (Milanese et al., unpub-δεδομένα

Ένα κύτταρο χαρακτηρισμού omic και χωρίς DNA (δηλ. Υγρή βιοψία) έχει οδηγήσει στον εντοπισμό μη επεμβατικών βιοδεικτών, οι οποίοι είναι εξαιρετικά χρήσιμοι στις κλινικές μελέτες. Σύμφωνα με τους συγγραφείς του “Disease Biomarkers for Precision Medicine: Challenges and Future Opportunities” Edwin Wang, Willian C.S. Cho, S.C. Cesar Wong και Siqi Liu, θα καταβληθούν πολύ περισσότερες προσπάθειες προς αυτή την κατεύθυνση στο εγγύς μέλλον. Επιπλέον, ή πρόσφατη ανάπτυξη της μονοκυτταρικής τεχνολογίας omic θα μπορούσε να φέρει νέες δυνατότητες αναγνώρισης βιοδεικτών. Τέλος, σχεδόν όλες οι προσπάθειες που καταβλήθηκαν έχουν επικεντρωθεί στον εντοπισμό των βιοδεικτών χρησιμοποιώντας omic δεδομένα στο παρελθόν. Ωστόσο, οι περισσότερες ασθένειες προκαλούνται από την αλληλεπίδραση των γενετικών και περιβαλλοντικών παραγόντων ή των παραγόντων του τρόπου ζωής. Ως εκ τούτου, για την ακριβή πρόβλεψη των κλινικών χαρακτηριστικών των σύνθετων ασθενειών, οι μελλοντικές εργασίες πρέπει να επικεντρωθούν στον εντοπισμό βιολογικών δεικτών με την ενσωμάτωση του συνδυασμού των omic δεδομένων και των παραγόντων του περιβάλλοντος και του τρόπου ζωής. [6]

3.1 Επεξεργασία και ανάλυση δεδομένων Βιολογικών Συστημάτων

Τι είναι ΒΙΟΛΟΓΙΑ ΣΥΣΤΗΜΑΤΩΝ;

Ως βιολογία Συστημάτων ορίζεται η μελέτη των γενετικών, πρωτεϊνικών, μεταβολικών και βιοσυνθετικών δράσεων, δηλαδή των ενεργειών που αλληλεξαρτούνται και συνθέτουν ένα ολοκληρωμένο βιολογικό σύστημα.

Οι εφαρμογές της βιολογίας συστημάτων ποικίλουν και αποτελούνται από τις εξής: Κλινική διάγνωση και πρόγνωση ασθενειών, Εξατομικευμένη ιατρική (Personalized medicine), Έλεγχος Τοξικότητας, Βιοτεχνολογικές μελέτες σε φυτά, Μελέτες στις επιστήμες διατροφής και τέλος Μελέτες της επίδρασης δραστηριοτήτων και συνηθειών του ανθρώπου. [9] Όλα τα παραπάνω πεδία εφαρμογών της Βιολογίας συστημάτων έχουν ως στόχο την εξαγωγή ενός συνόλου δεδομένων, με την μορφή βιοδεικτών και δεδομένων “omics”. Τα παραπάνω δεδομένα χάρη στην πρόσφατη ανάπτυξη των φασματοσκοπικών τεχνικών ανάλυσης και της υπολογιστικής ισχύος, μπορούν να “περάσουν” από ολιστική ανάλυση, δηλαδή την ανάλυση δειγμάτων με στόχο τη λήψη του ολικού «αποτυπώματος» του δείγματος· μια «ολική χαρτογράφηση» των συστατικών του. Ουσιαστικά πρόκειται για μια επιλεγόμενη μέθοδο η οποία μπορεί να συλλέξει δεδομένα, πλούσια σε πληροφορίες. Τα δεδομένα αναλύονται με προηγμένα υπολογιστικά εργαλεία, όπως η στατιστική ανάλυση πολλαπλών μεταβλητών και τεχνικές όπως Principal Component Analysis (PCA), Partial Least Squares-Discriminant Analysis (PLS-DA) καθώς και άλλες. Οι κύριες αναλυτικές τεχνικές οι οποίες εφαρμόζονται στην ολιστική ανάλυση είναι οι τεχνικές συστοιχιών-πινάκων (array), η φασματοσκοπία πυρηνικού συντονισμού (NMR) και η φασματομετρία μαζών σε διάφορες μορφές της είτε μόνη της είτε σε συνδυασμό με υγρή χρωματογραφία (LC-MS), αέρια χρωματογραφία (GC-MS), ή τριχοειδή ηλεκτροφόρηση (CE-MS). [9]

3.1.2 Εύρεση Βιοδείκτη

Σκεπτόμενος το πρόβλημα της εύρεσης του βιοδείκτη για μια ασθένεια, η σταθερότητα αυξάνεται ως μια σημαντική ιδιότητα του αλγορίθμου γιατί θα πρέπει να βρει το ίδιο σύνολο χαρακτηριστικών σε διαφορετικές διαδρομές. Οι Fortino, Kinarct, Fyhrquist, Alenius και Greco προτείνουν το 2014 μια μέθοδο επιλογής χαρακτηριστικών περιτυλίγματος που συνδυάζει ασαφή λογική και RF και είναι σε θέση να εγγυηθεί καλές επιδόσεις και υψηλή σταθερότητα. Ο αλγόριθμος αποτελείται από τρία μέρη: στο πρώτο βήμα, αποκαλούμενο discretization, τα γονιδιακά εκφραζόμενα δεδομένα εδώ μετασχηματίζονται σε ασαφή πρότυπα (FPs). Τα FPs παρέχουν πληροφορίες για τα πιο συναφή χαρακτηριστικά κάθε κατηγορίας. Στο δεύτερο βήμα, οι προηγούμενες γνώσεις σχετικά με τα FPs χρησιμοποιούνται για την κατάρτιση μιας RF και την ταξινόμηση δεδομένων. Στο τρίτο βήμα, τα επιλεγμένα χαρακτηριστικά ταξινομούνται από ένα μέτρο μεταβλητής σημασίας μετατόπισης. Η μέθοδος δοκιμάστηκε σε αρκετά δεδομένα γονιδιακής έκφρασης και σε σύγκριση με άλλες δύο μεθόδους επιλογής χαρακτηριστικών βάσει RF: varSelRF (Díaz-Uriarte & De Andres, 2006) και Borda (Kursa, Rudnicki, et al., 2010). Οι βαθμολογίες F και G αξιολογήθηκαν σε 30 επαναλήψεις για την εκτίμηση της ακρίβειας. Αυτά τα μέτρα είναι ιδιαίτερα κατάλληλα για τα μη ισορροπημένα προβλήματα πολλαπλών κλάσεων (Yu et al., 2013). Κατά τη διάρκεια των επαναλήψεων, επιλέχθηκαν χαρακτηριστικά συνάφειας. Η τελική μέτρηση σταθερότητας ορίστηκε ως ο λόγος μεταξύ του αριθμού των σταθερών χαρακτηριστικών και του συνολικού αριθμού των επιλεγμένων λειτουργιών.

3.2 Επεξεργασία και Ανάλυση

Μετά την επιτυχημένη μέτρηση των βιοδεικτών και των άλλων δεδομένων που μπορεί να προκύψουν από την μελέτη της Βάσης Δεδομένων, είναι απαραίτητη η Επεξεργασία και η Ανάλυση τους. Η ανάκτηση λοιπόν της πληροφορίας από τα δεδομένα ανάλυσης είναι περίπλοκη και πολλές φορές δύσκολη διαδικασία, που απαιτεί πόρους και χρόνο. Συγκριτικά με τις προηγούμενες δεκαετίες, στην τελευταία, η πρόοδος της αναλυτικής τεχνολογίας έχει πάρει τέτοια ανοδική πορεία που έχει εξοπλιστεί με αξιόπιστα όργανα τα οποία παράγουν τεράστιους όγκους δεδομένων υψηλού επιπέδου με ταχύτατους ρυθμούς. Οι τεράστιοι όγκοι δεδομένων και ο χειρισμός τους αποτελούν, πλέον μείζον ζήτημα για την επιστημονική κοινότητα τόσο στο παρόν όσο και για το μέλλον.

Η επεξεργασία των δεδομένων έχει αρκετές δυσκολίες ακόμα και για τα πιο απλά δισδιάστατα δεδομένα NMR. Υπάρχει σοβαρή διένεξη σε σχέση με την καλύτερη μέθοδο για την ανάκτηση των δεδομένων (data extraction). Χρησιμοποιούνται τεχνικές χωρισμού του φάσματος σε μικρές περιοχές (binning) οι οποίες στη συνέχεια εξετάζονται χωριστά αλλά παράλληλα εφαρμόζονται και τεχνικές αναγνώρισης και ευθυγράμμισης κορυφών σε όλο το φάσμα (peak-picking and alignment). Για την επεξεργασία πολυδιάστατων δεδομένων LC-MS, GC-MS, η κατάσταση είναι, σαφώς, πιο περίπλοκη.

Το βασικό πρόβλημα με τις μεθόδους ανάκτησης δεδομένων είναι η έλλειψη προτυποποίησης τους: όταν γίνεται ανάκτηση από τα ίδια δεδομένα, με n διαφορετικές μεθόδους, μπορεί να ληφθούν n διαφορετικά σετ δεδομένων (datasets) και πιθανότατα n διαφορετικά στατιστικά μοντέλα (εάν ακολουθηθεί ένα και μόνο είδος στατιστικής ανάλυσης). Σε εφαρμογές στις οποίες λαμβάνουν χώρα μεγάλης κλίμακας αλλαγές, όπως για παράδειγμα σε μοντέλα τοξικότητας, οι διαφορές στη μέθοδο ανάκτησης δεδομένων ίσως να μην είναι τόσο επιδραστικές ως προς το τελικό αποτέλεσμα και τα τελικά ευρήματα μπορεί να είναι πρακτικά τα ίδια. Σε άλλες περιπτώσεις, όμως, όπου οι αλλαγές στο σύνολο των μεταβολιτών είναι ελάχιστες (π.χ. επιδημιολογικές μελέτες περιβαλλοντικής έκθεσης ανθρώπων), ακόμη και οι μικρές διαφοροποιήσεις στην επεξεργασία των δεδομένων μπορεί να έχουν δραματικό αντίκτυπο στα αποτελέσματα μιας μελέτης. Γι' αυτόν το λόγο χρειάζεται προσοχή και μεθοδικότητα κατά την επεξεργασία των δεδομένων, η οποία πρέπει να γίνει με τις ίδιες βασικές παραμέτρους για όλα τα δείγματα, εκτός εάν συντρέχει κάποιος σοβαρός και ειδικός λόγος. Τεχνικές όπως η LC-MS, η GC-MS και κυρίως η GCxGC-MS, έχουν ως αποτέλεσμα πολύ πιο πολύπλοκα δεδομένα (τρισδιάστατα ή ακόμη και τετρ-αδιάστατα) σε σχέση με τα δεδομένα του NMR. Επίσης, οι φασματογράφοι μάζας έχουν ένα πλήθος λειτουργιών και μεθόδων καταγραφής (acquisition modes) καθώς και άλλων παραμέτρων, οι οποίες αυξάνουν την πολυπλοκότητα των δεδομένων. Για παράδειγμα, υπάρχουν οι εξής περιπτώσεις:[41]

1. Ανάκτηση δεδομένων σε centroid ή profile data acquisition mode.
2. Σταθερός χρόνος σάρωσης (όργανα τύπου TOF, Q) ή μεταβλητός (όργανα τύπου IT).
3. Θετική, αρνητική ή εναλλασσόμενη πολικότητα.
4. Σάρωση (full scan MS) ή επιλογή ιόντων (SIM MS).
5. Χαμηλή ή πολύ υψηλή διαχωριστική ικανότητα.
6. Την ικανότητα για συνδυασμό όλων των παραπάνω.
7. Τύποι δεδομένων και μέθοδοι καταγραφής μοναδικοί για το κάθε όργανο.

Τα διάφορα αυτοματοποιημένα εργαλεία εξαγωγής δεδομένων χρησιμοποιούν διαφορετικούς αλγόριθμους. Οι πιο συχνά απαντημένες προσεγγίσεις περιλαμβάνουν τις εξής διεργασίες:[41]

1. Ανίχνευση κορυφών και ολοκλήρωση, ακολουθούμενη από ευθυγράμμιση (alignment).
2. Στρέβλωση του επιπέδου (warping of the m/z –retention time plane (alignment at raw data level)), αφαίρεση σήματος υποβάθρου, ακολουθούμενη από εντοπισμό κορυφών και ολοκλήρωσή τους.
3. Αφαίρεση των ισοτοπικών κορυφών μετά τα βήματα 1 και 2 (de-isotoping of data).
4. Εξαγωγή μεταβολιτών με τη χρήση φασματοσκοπικών δεδομένων (deconvolution, φασματοσκοπικές βιβλιοθήκες, ακριβής μάζα, κλπ.).

Υπάρχουν περίπου 20 υπολογιστικά εργαλεία για την αυτόματη εξαγωγή δεδομένων, είτε εμπορικά είτε δωρεάν διαθέσιμα., τα οποία δίνονται στους παρακάτω πίνακες όπως τους παρουσιάζει ο Γεώργιος Θεοδωρίδης στο “Βιολογία Συστημάτων, Τεχνολογίες Omics”

ΟΝΟΜΑ	ΚΑΤΑΣΚΕΥΑΣΤΗΣ	ΕΦΑΡΜΟΓΕΣ
BlueFuse	BlueGnome, Cambridge UK	Metabolomics με δεδομένα MS και NMR
Chenomx NMR Suite	Chenomx, Edmonton, Canada	Metabolomic με δεδομένα NMR
Genedata Exoressionist	Genedata, Basel, Switzerland	Πλατφόρμα για γενική χρήση σε τεχνολογίες -omics. Το τμήμα των metabolomics δέχεται δεδομένα MS
LineUp	Infometrix	Χρωματογραφική ευθυγράμμιση
MarketLynx	Waters, Milford, MA, USA	Metabolomics με δεδομένα LC-MS
MasterView	Applied Biosystems, Foster Vity, CA	Metabolomics με δεδομένα LC-MS
MassHunter Profiling Software	Agilent Technologies, Santa Clara, CA	Proteomics με δεδομένα LC-MS
Metabolic Profiler	Bruker Daltonic & Bruker Biospin, Billerica, MA, USA	Metabolomics με δεδομένα MS και NMR
MS Resolver	Pattern Recognition Systems, Bergen, Norway	Δεδομένα GC-MS και LC-MS
Profile	Phenomenone Discoveries Sakatoon Canada	Metabolomics με δεδομένα MS και NMR
Rossetta Elucidation	Rosetta Biosoftware, Seattle, WA, USA	Proteomics με δεδομένα LC-MS
Sieve	Thermo Fisher Scientific, Waltham, MA, USA	Proteomics με δεδομένα LC-MS

ΟΝΟΜΑ	ΕΦΑΡΜΟΓΕΣ	ΠΛΑΤΦΟΡΜΑ	ΤΥΠΟΣ ΑΔΕΙΑΣ
Chrompare	Metabolomics με δεδομένα GC-FID	Microsoft Excel Visual Basic	Διαθέσιμο για download
COMSPARI	Metabolomics με δεδομένα LC-MS και GC-MS	Γραμμένο σε γλώσσα C, για κάθε πρόσφατη πλατφόρμα, όπως Linux και Windows	GNU General Public Licence
Continuous profile models	Proteomics με δεδομένα LC_MS	Toolbox για το Matlab	Δωρεάν για εκπαιδευτική και ερευνητική χρήση, διαθέσιμος πηγαίος κώδικας
HiRes	Metabolomics με δεδομένα NMR	Γραμμένο σε γλώσσα C++ για Windows	Δωρεάν για εκπαιδευτικούς και κλινικούς σκοπούς
LCMSWARP	Proteomics με δεδομένα LC-MS και LC-MS/MS	Γραμμένο σε γλώσσα C++	ΑΓΝΩΣΤΟ
MapQuant	Proteomics με δεδομένα LC-MS	Γραμμένο σε γλώσσα C, για Windows και Linux	Harvard University open-source compatible licence
MathDAMO	Metabolomics με δεδομένα LC-MS, GC-MS και CE-MS	Πακέτο για το Mathematica	Δωρεάν
metAlign	Για δεδομένα GC-MS και LC-MS	Windows	Δωρεάν
MET-IDEA	Metabolomics με δεδομένα LC-MS, GC-MS και CE-MS	Windows, πλατφόρμα .NET	Δωρεάν διαθέσιμο για ακαδημαϊκούς σκοπούς κατόπιν αίτησης
MSFACTs	Metabolomics με δεδομένα LC-MS και GC-MS	Γραμμένο σε Java	Δωρεάν διαθέσιμο για ακαδημαϊκούς σκοπούς κατόπιν αίτησης
MSight	Proteomics με δεδομένα LC-MS	Windows	Δωρεάν
msInspect	Proteomics με δεδομένα LC-MS	Γραμμένο σε Java, απαιτεί τη στατιστική γλώσσα R	Δωρεάν λογισμικό διαθέσιμο με άδεια Apache 2.0

MZmine	Metabolomics με δεδομένα LC-MS και GC-MS	Γραμμένο σε Java	GNU General Public Licence
SpecArray	Proteomics με δεδομένα LC-MS	Γραμμένο σε γλώσσα C, για Linux	GNU General Public Licence
Xalign	Proteomics με δεδομένα LC-MS	Γραμμένο σε γλώσσα C++	Διαθέσιμο μετά από αίτηση στον Προγραμματιστή
XCMS	Metabolomics με δεδομένα LC-MS και GC-MS ^{vvv}	Γραμμένο στη στατιστική γλώσσα R	GNU General Public Licence

[41]

Μετά την αρχική επεξεργασία δεδομένων, ακολουθεί η στατιστική ανάλυση με διάφορες πολυπαραμετρικές και μονοπαραμετρικές τεχνικές στατιστικής ανάλυσης. Για την ανάλυση αυτών των δεδομένων μικροσυστοιχιών υπάρχει διαθέσιμη για χρήση μια πληθώρα στατιστικών μεθόδων, αλγορίθμων ταξινόμησης και διάκρισης στοιχείων, τα οποία εντάσσονται στα πλαίσια της πολυπαραμετρικής ανάλυσης. Η πολυπαραμετρική στατιστική ανάλυση αφορά εκείνους τους αλγόριθμους που είναι ανεπτυγμένοι για την ανάλυση δεδομένων και που χρησιμοποιούν περισσότερες από μια μεταβλητές σε κάθε ένα από τα βήματα τους. Η στατιστική ανάλυση χωρίζεται σε δύο κατηγορίες, οι οποίες είναι οι επιβλεπόμενες και μη-επιβλεπόμενες αντίστοιχα, τεχνικές ανάλυσης. Στις τεχνικές με επίβλεψη είναι εφικτή η διάκριση γονιδίων, μεταβολιτών κ.ά. που διαφέρουν σε συγκεκριμένες καταστάσεις (όπως η SAM). Στην δεύτερη τεχνική ανάλυσης, τα δεδομένα αναλύονται ανεξάρτητα κάποιας δοσμένης πληροφορίας. Ουσιαστικά είναι μια “τυφλή” ανάλυση, που αποσκοπεί στην κατηγοριοποίηση των δεδομένων των πειραματικών μετρήσεων (παραδείγματα ή PCA και η HCL).

Οι αλγόριθμοι που εκτελούν την πολυπαραμετρική ανάλυση, αποτελούν αλγόριθμους ομαδοποίησης και ουσιαστικά ταξινομούν τα δεδομένα και τις ομάδες με βάση τον διαχωρισμό τους σύμφωνα με τα αποτελέσματα της μελέτης. Δηλαδή εφαρμόζονται τεχνικές ομαδοποίησης, σε μια διαδικασία που ονομάζεται ταξινόμηση και θα αναλυθεί παρακάτω λεπτομερώς. Αυτές οι ταξινομήσεις ή ομαδοποιήσεις αποτελούνται από δύο κατηγορίες, τις ιεραρχικές και τις μη ιεραρχικές. Οι ιεραρχικοί αλγόριθμοι ταξινόμησης πραγματοποιούν ταξινόμηση που προκύπτει με έναν αυξανόμενο αριθμό εμφωλευμένων κλάσεων και το αποτέλεσμα θυμίζει φυλογενετική ταξινόμηση, δηλαδή υπάρχει κάποιου είδους σχέσης μεταξύ των δεδομένων διαφορετικών ομάδων και επομένως υπάρχουν κοινά χαρακτηριστικά μεταξύ και των κλάσεων ταξινόμησης. Στους μη ιεραρχικούς αλγόριθμους από την άλλη, για παράδειγμα όπως ο k-Means, απλά διαχωρίζουν τα αντικείμενα σε διαφορετικές ομάδες χωρίς να επιδιώκουν να βρεθεί κάποια σχέση μεταξύ των ξεχωριστών αντικειμένων, που αποτελεί και το ζητούμενο των ιεραρχικών ομαδοποιήσεων.

Το κύριο πλεονέκτημα των ιεραρχικών αλγορίθμων ομαδοποίησης έγκειται στο γεγονός ότι η ομαδοποίηση είναι σχετικά απλή και μπορεί να γίνει εύκολα αντιληπτή μέσω γραφημάτων καθώς η οπτικοποίηση σε αυτήν την κατηγορία ταξινόμησης δεν είναι τόσο σύνθετη υπόθεση. Αποτελεί την πιο διαδιδόμενη μέθοδο ανάλυσης δεδομένων γονιδιακής έκφρασης. Επιπλέον θεωρείται μια συνενωτική μέθοδος κατά την οποία το κάθε δεδομένο εντάσσεται σε ομάδες οι οποίες ομαδοποιούνται και μεταξύ τους μέχρι να ολοκληρωθεί η διαδικασία η οποία θα δώσει ένα ιεραρχικό δένδρο. Η διαδικασία προχωρά με απλό τρόπο.

Αρχικά υπολογίζεται ένας πίνακας αποστάσεων για όλα τα γονίδια που πρέπει να ομαδοποιηθούν. Έπειτα παίρνουμε τον πίνακα αποστάσεων και γίνεται η εύρεση των δυο πιο όμοιων γονιδίων ή ομάδων γονιδίων. Κάθε ομάδα περιέχει αρχικά μόνο ένα γονίδιο, δηλαδή το πλήθος των ομάδων είναι ίσο με αυτό των γονιδίων. Σε περιπτώσεις που υπάρχουν περισσότερα του ενός ζεύγους γονιδίων με ίσες αποστάσεις, θα ήταν ορθότερο να έχει ληφθεί από πριν κανόνας απόφασης για το τι θα γίνει. Πλέον τα δύο γονίδια-ομάδες ενώνονται και παράγουν μια ομάδα με δύο μέλη το ελάχιστο. Έπειτα, επαναυπολογίζονται οι αποστάσεις μεταξύ της νέας ομάδας και των υπόλοιπων ομάδων. Θα ήταν αχρείαστο να υπολογιστούν από την αρχή όλες οι αποστάσεις, για αυτό συνηθίζεται η μέτρηση μόνο εκείνων που συμπεριλαμβάνουν την νέα ομάδα που σχηματίστηκε. Τέλος οι διαδικασίες σχηματισμού ομάδων επαναλαμβάνονται με βάση τις αποστάσεις μέχρι να ομαδοποιηθούν όλα τα δεδομένα σε μια και μόνο ομάδα.

Κατηγορίες Ταξινόμησης

Supervised classification

Γραμμική	Μη γραμμική
Linear Discriminant Analysis	Νευρωνικά δίκτυα



Support Vector Machine

Unsupervised Classification



Ιεραρχική	Μη-Ιεραρχική (επιμεριστική)
• Agglomerative • Disive	• k-means • SOM • Spectral

4 Μηχανική Μάθηση

Η ταξινόμηση των δεδομένων στην βιοπληροφορική διαδραματίζει καίριο ρόλο σε πολλά στάδια μελετών, με πιο δημοφιλές και σημαντικό αυτό της διάγνωσης της ασθένειας ενός αντικειμένου. Υπάρχουν αρκετοί αλγόριθμοι που θα μπορούσαν να χρησιμοποιηθούν για να επιτευχθεί ο στόχος της ταξινόμησης των δεδομένων σε κλάσεις. Συγκρίνοντας τους αλγορίθμους βρίσκει κανείς κάποια κοινά στοιχεία και κάποια διαφορετικά, αλλά οι περισσότεροι θα εντοπίσει κανείς ότι δουλεύουν με **Μηχανική Μάθηση**. [62]

Η Μηχανική Μάθηση (Machine Learning) αποτελεί μια νοήμον διαδικασία των υπολογιστών, από 'κει προκύπτει και ο όρος μάθηση. Ως πεδίο επιστήμης ανήκει σε αυτό της Τεχνητής Νοημοσύνης. Η Τεχνητή Νοημοσύνη αποτελεί σύμπτυξη πολλών επιστημών εκτός της πληροφορικής, όπως η ψυχολογία, η φιλοσοφία, η νευρολογία, η γλωσσολογία η επιστήμη μηχανικών, με κύριο στόχο την επιτυχία σύνθεσης συμπεριφοράς ενός υπολογιστή με στοιχεία συλλογιστικής μάθησης και προσαρμογής στο περιβάλλον. [63] Επομένως είναι πιο εύκολο να καταλάβουμε τον όρο της Μηχανικής μάθησης, η οποία περιγράφεται από τον Mitchell (1997) ως *“Ένα πρόγραμμα υπολογιστή που μαθαίνει από την εμπειρία E ως προς κάποια κλάση εργασιών T και μέτρο απόδοσης P , αν η απόδοση του σε εργασίες από το T , όπως μετριέται από το P , βελτιώνεται μέσω της εμπειρίας E ”* [62]

Ως κλάδος της Τεχνητής Νοημοσύνης, η Μηχανική μάθηση ασχολείται με τη μελέτη συγκεκριμένων αλγορίθμων, που αποσκοπούν στην βελτιώση της συμπεριφοράς της μηχανής-υπολογιστή που τους υποστηρίζει, σε κάποια εργασία που τους έχει ανατεθεί, χρησιμοποιώντας την εμπειρία που έχουν αποκτήσει, κατά την φάση της εκπαίδευσης. Η σχεδίαση των συστημάτων της Μηχανικής Μάθησης, για τα συστήματα που ανήκουν στην Τεχνητή Νοημοσύνη, η δυνατότητα μάθησης προσδιορίζεται ως ικανότητα απόκτησης επιπλέον γνώσης, που επιφέρει μεταβολές στην ήδη υπάρχουσα. Ο τομέας της Μηχανικής Μάθησης αναπτύσσει, επίσης, επιτυχώς την Εξελικτική Μάθηση (Evolutionary Learning), η οποία μιμείται διαδικασίες φυσικής αναπαραγωγικής σε έμβια όντα. Χρησιμοποιείται κυρίως σε προβλήματα βελτιστοποίησης. [62]

Εκτός της επιστήμης της Τεχνητής Νοημοσύνης, μεταξύ των επιστημονικών κλάδων που επωφελούνται από τη εφαρμογή της Μηχανικής Μάθησης συγκαταλέγονται οι: Πιθανότητες και Στατιστική, Θεωρία της Πολυπλοκότητας, Θεωρία της Πληροφορίας, Αριθμητική Βελτιστοποίηση, Θεωρία Ελέγχου (προσαρμοστική), Εξόρυξη Δεδομένων, Ψυχολογία (εξελικτική, γνωστική), Νευροβιολογία και Γλωσσολογία

4.1 Εκπαίδευση

Η δομή της εκπαίδευσης ενός αλγορίθμου Μηχανικής Μάθησης αποτελείται από 4 στοιχεία ή βήματα. Αρχικά έχουμε τα **Δεδομένα (Data)** που θέλουμε να διαχειριστούμε και να τα χρησιμοποιήσουμε για την εκπαίδευση του **Μοντέλου (Model)** το οποίο μπορεί να ποικίλει από μία απλή γραμμική παλινδρόμηση μέχρι ένα πολυσύνθετο νευρωνικό δίκτυο. με την εφαρμογή της **Συνάρτησης Στόχου (Objective function)**, υπολογίζεται πόσο κοντά στην επιθυμητή τιμή στόχο του μοντέλου είναι οι τιμές εξόδου που παράχθηκαν μέσω αυτού. Για παράδειγμα εάν η συνάρτηση που έχουμε υπολογίζει την πρόβλεψη λάθους του μοντέλου και όπως είναι λογικό η επιθυμία μας είναι να ελαχιστοποιήσουμε την περίπτωση σφάλματος δηλαδή να ελαχιστοποιήσουμε την τιμή της

συνάρτησης. Το τελευταίο στάδιο αποτελείται από τον **Αλγόριθμο Οπτικοποίησης (Optimization algorithm)**, ο οποίος περιλαμβάνει τους κατάλληλους μηχανισμούς μέσω των οποίων μεταβάλλουμε τις παραμέτρους του μοντέλου με απώτερο σκοπό την βελτιστοποίηση της συνάρτησης στόχου. Η διαδικασία αυτή είναι επαναληπτική. Παρακάτω υπάρχει μια εικονική αναπαράσταση των σταδίων.



Ένας αλγόριθμος μηχανικής μάθησης έχει την δυνατότητα να επιτρέπει σε ένα σύστημα να μαθαίνει βάσει των δικών του καταστάσεων, όταν ο ανθρώπινος παράγοντας αδυνατεί να προσφέρει ένα αυστηρό σύνολο κανόνων που θα ακολουθήσει ο υπολογιστής για να επιτύχει τον αρχικό στόχο. Ακόμα και αν είναι εφικτό ο άνθρωπος να μπορέσει να παράξει κανόνες τέτοιους ώστε να οδηγηθεί ο υπολογιστής στο επιθυμητό αποτέλεσμα, ένας αλγόριθμος μηχανικής μάθησης μπορεί να προσφέρει πολύ καλύτερο αποτέλεσμα καθώς έχει δοκιμάσει πιθανόν εκατομμύρια τρόπους για να οδηγηθεί στο τελικό αποτέλεσμα επιλέγοντας τον βέλτιστο.

4.2 Τύποι Μηχανικής Μάθησης

Μια υπολογιστική μηχανή μπορεί να εκπαιδευτεί μέσω Μηχανικής Μάθησης με τρεις διαφορετικούς τρόπους: **Μάθηση με Επίβλεψη (Supervised Learning)**, **Μάθηση χωρίς Επίβλεψη (unsupervised learning)**, **Ενισχυτική Μάθηση (Reinforcement)**.

- **Μάθηση με Επίβλεψη:** Εδώ παρέχονται στον αλγόριθμο με δεδομένα εισόδου τα οποία έχουν χαρακτηριστεί και τις επιθυμητές τιμές στόχο που θέλουμε ο αλγόριθμος να εξάγει. Η Μάθηση με επίβλεψη χωρίζεται σε δύο είδη εκ' των οποίων το ένα το είδαμε και προηγουμένως:
 1. **Ταξινόμηση:** εδώ τα δεδομένα διαχωρίζονται σε κατηγορίες.
 2. **Παλινδρόμηση:** τα δεδομένα σε αυτήν την υποκατηγορία είναι αριθμητικά

- **Μάθηση χωρίς Επίβλεψη:** Σε αυτήν την περίπτωση εισάγονται στον αλγόριθμο τα δεδομένα εισόδου χωρίς αυτήν την φορά να δοθεί κάποια επεξήγηση ή κάποια τιμή στόχο. Αντιθέτως ζητείται από τον αλγόριθμο να βρει κάποιου είδους λογική σχέση που περιγράφει τα δεδομένα που του δόθηκαν εξ αρχής.
- **Ενισχυτική Μάθηση:** Στην Ενισχυτική Μάθηση συμπεριφερόμαστε στο μοντέλο με τέτοιο τρόπο ώστε να ενεργεί μέσα σε ένα καθορισμένο περιβάλλον ανάλογα με τη επιβράβευση που του δίνεται.

4.3 Μοντέλο Γραμμικής Παλινδρόμησης

Η γραμμική παλινδρόμηση ασχολείται με το να περιγράψει και να αξιολογήσει τις σχέσεις μεταξύ μιας μεταβλητής (η οποία πολύ συχνά ονομάζεται εξαρτημένη) με μια ή περισσότερες μεταβλητές, οι οποίες ονομάζονται ανεξάρτητες. Στην Μηχανική Μάθηση όπως και στην στατιστική ένας τρόπος για να εκφράσουμε την Γραμμική Παλινδρόμηση είναι ο εξής:

1. Την τάση της εξαρτημένης μεταβλητής Y να σχετίζεται με την ανεξάρτητη μεταβλητή X συστηματικά.
2. Τον τρόπο με τον οποίο οι παρατηρήσεις είναι διασκορπισμένες γύρω από την καμπύλη της στατιστικής σχέσης.

Η γραμμική παλινδρόμηση υποθέτει ότι τα δεδομένα εισόδου είναι ανεξάρτητα μεταξύ τους. Αυτό σημαίνει ότι η τιμή μιας ανεξάρτητης μεταβλητής δε θα επηρεάζει την τιμή μιας άλλης. Αυτό φυσικά δεν είναι πάντα ο κανόνας. Πολλές φορές οι ανεξάρτητες μεταβλητές σε ένα πρόβλημα συσχετίζονται, γεγονός που επιφέρει αρνητικές συνέπειες στα τελικά αποτελέσματα. Με την ανάλυση συσχέτισης (correlation analysis) μπορούμε να μετρήσουμε και να ερμηνεύσουμε το κατά πόσο η γραμμική ή μη γραμμική σχέση μεταξύ δύο συνεχόμενων μεταβλητών είναι ισχυρή.

Η συνάρτηση στόχου ενός αλγορίθμου Μηχανικής Μάθησης που εφαρμόζεται σε μοντέλο Γραμμικής Παλινδρόμησης είναι η εξής:

$$f(x) = x * w + b$$

Ο στόχος του αλγορίθμου Μηχανικής Μάθησης είναι να βρει τις κατάλληλες τιμές για w και b , ώστε η έξοδος της συνάρτησης $f(x)$ να είναι όσο το δυνατόν πιο κοντά στην επιθυμητή τιμή.

Παρακάτω ακολουθεί κώδικας σε python, το παράδειγμα αναπαριστά μια απλή γραμμική παλινδρόμηση

Απλή Γραμμική Παλινδρόμηση

Βιβλιοθήκες python

```
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
```

Generate τυχαίες τιμές δεδομένων ως input για την εκπαίδευση

```
observations = 1000

xs = np.random.uniform(low = -10, high = 10, size = (observations, 1) )
zs = np.random.uniform(-10, 10, (observations, 1) )

inputs = np.column_stack((xs,zs))

print(inputs.shape)
```

(1000, 2)

Δημιουργούμε τις τιμές στόχους

```
noise = np.random.uniform(-1, 1, (observations, 1))

#η συνάρτηση είναι  $y = x*w + b$ 
targets = 2*xs - 3*zs + 5 + noise
# 2 and 3 are our weights and 5 is the bias

print(targets.shape)
```

(1000, 1)

Σχεδίαση των δεδομένων προς εκπαίδευση

```
#για να επιτευχθεί οπτικοποίηση 3διάστατου μοντέλου, απαιτείται μετασχηματισμός των δεδομένων
targets = targets.reshape(observations,)

fig = plt.figure()

#η μέθοδος που δημιουργεί το 3διάστατο μοντέλο
ax = fig.add_subplot(111, projection = '3d')

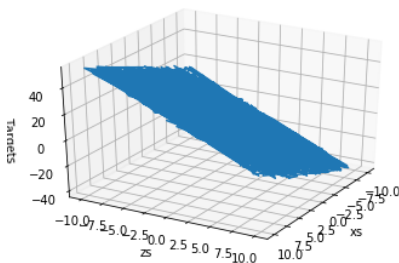
ax.plot(xs, zs, targets)

#άξονες
ax.set_xlabel('xs')
ax.set_ylabel('zs')
ax.set_zlabel('Targets')

#η μεταβλητή azim αναφέρεται στην οπτική από την οποία κοιτάμε το γράφημα
ax.view_init(azim = 30)

plt.show()

targets = targets.reshape(observations, 1)
```



Διάγραμμα 4.3.1

Αρχικοποίηση μεταβλητών

```
init_range = 0.1
weights = np.random.uniform(low = -init_range, high = init_range, size = (2, 1))
biases = np.random.uniform(low = -init_range, high = init_range, size = 1)
print(weights)
print(biases)
```

```
[[0.04206915]
 [0.05480915]]
[0.07618989]
```

Ορισμός ρυθμού μάθησης

```
learning_rate = 0.02
```

Εκπαίδευση μοντέλου

```
for i in range(100):
    #υπολογισμός τιμες εξόδου για τα αντίστοιχα weights και biases
    outputs = np.dot(inputs, weights) + biases
    deltas = outputs - targets

    #συνάρτηση υπολογισμού λάθους των τιμών εξόδου σε σχέση με τις τιμές στόχου
    loss = np.sum(deltas ** 2)/2/observations

    print(loss)

    #update των τιμών
    deltas_scaled = deltas / observations
    weights = weights - learning_rate * np.dot(inputs.T, deltas_scaled)
    biases = biases - learning_rate * np.sum(deltas_scaled)
```

```
0.38205597007473835
0.37347986038730774
0.3652429029670783
0.35733168564180134
0.3497333266392342
0.34243545361185357
0.3354261834910566
0.3286941031380605
0.32222825075997547
0.3160180980608049
...
0.17117125929105942
0.17093483162578313
0.17070775376847677
```

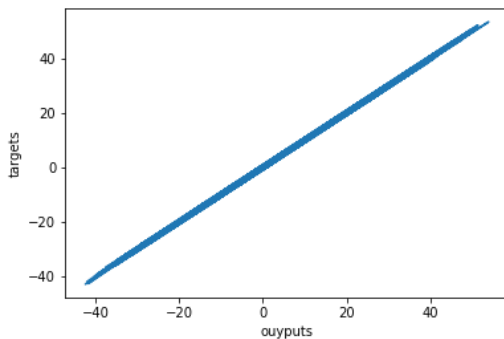
```
0.1704896559700718
0.1702801831036578
0.1700789940862323
0.16988576132331895
0.16970017017554662
0.1695219184463258
0.16935071588978354
0.16918628373815897
```

```
print(weights,biases)
```

```
[[ 2.00028049]
 [-3.00023516]] [4.93852219
```

Σχεδίαση τελικών τιμών εξόδου και τιμών στόχου

```
plt.plot(outputs, targets)
plt.xlabel('ouyputs')
plt.ylabel('targets')
plt.show()
```



Διάγραμμα 3.3.2.2

Στον αλγόριθμο που αναπτύχθηκε, εφαρμόστηκε Μηχανική Μάθηση. Αρχικά εισάχθηκαν οι απαραίτητες βιβλιοθήκες για την διεκπεραίωση του αλγορίθμου. Επιλέχθηκε η βιβλιοθήκη *NumPy* η οποία περιέχει όλες τις μαθηματικές πράξεις που χρειάστηκαν και επίσης είναι ιδιαίτερα γρήγορη, αυτοί είναι και οι κύριοι λόγοι που είναι τόσο δημοφιλής στην Επιστήμη των Δεδομένων (Data Science). Η δεύτερη βιβλιοθήκη, *matplotlib*, συμβάλει στην γραφική αναπαράσταση των δεδομένων. Είναι πολύ απλή, γεγονός που την καθιστά αρκετά εύχρηστη. Τέλος έγινε εισαγωγή του module *Axes3D* από *mpl_toolkits.mplot3d* το εργαλείο αυτό παρέχει στον αλγόριθμο την δυνατότητα να οπτικοποίηση τα δεδομένα σε τρισδιάστατη μορφή.

Στην συνέχεια παράχθηκαν τυχαία δεδομένα για την εκπαίδευση του μοντέλου. Στο τέλος γίνεται *print* των δεδομένων εισόδου και βλέπουμε ότι είναι ένας πίνακας μεγέθους 1000*2.

Δουλεύοντας σε ένα γραμμικό μοντέλο, θα πρέπει να χρησιμοποιηθεί η αντίστοιχη συνάρτηση. Με τα βάρη των μεταβλητών x και z να ισούνται με 2 και 3 και την *bias* να είναι 5 μας λείπει μόνο ο *θόρυβος*, όλα τα πραγματικά δεδομένα περιέχουν θόρυβο. Επομένως, θα πρέπει να προσθέσουμε *θόρυβο* στην συνάρτησή μας. Ο *θόρυβος* είναι από -1, μέχρι 1 και το μέγεθος του θα πρέπει να ισούται με αυτό των x και z . Πλέον έχουμε την συνάρτηση που περιγράφει τις τιμές στόχους του μοντέλου:

$$targets = f(xz) = 2 * x - 3 * z + 5 + noise$$

Πλέον όλα τα δεδομένα μας έχουν την μορφή πίνακα μεγέθους 1000*1 επομένως τυπώνοντας την μορφή που έχουν οι τιμές στόχοι του αλγορίθμου θα πάρουμε έναν τέτοιο πίνακα.

Έπειτα αυτό που θα συμβεί όπως και φαίνεται από το *Διάγραμμα 3.3.2.1* είναι να εμφανιστούν τα δεδομένα σε μια τρισδιάστατη μορφή πάνω σε ένα σύστημα αξόνων xs , $Targets$, zs όπου $Targets$ είναι οι τιμές στόχοι ενώ τα xs και zs είναι οι μεταβλητές μας. Η οπτική από την οποία κοιτάμε το τρισδιάστατο μοντέλο είναι ελεγχόμενη και από την εντολή που μας παρέχει το module *Axes3D* της βιβλιοθήκης *mpl_toolkits.mplot3d*: `ax.view_init(azim = 30)`.

Στην συνέχεια αρχικοποιούμε τις μεταβλητές. Ξεκινώντας από την *init_range* η οποία αντικατοπτρίζει το εύρος των αρχικών τιμών που θα πάρουν τα βάρη και τα *biases* και έπειτα θα παραχθούν οι τιμές των βαρών τυχαία. Το μέγεθος των πινάκων των βαρών θα είναι 2*1, λόγω του ότι κάθε φορά θα έχουμε δύο νέες μεταβλητές εισόδου και μία έξοδο. Αναλόγως παράγονται και οι τιμές που θα έχει κάθε φορά το *bias* που θα έχει πίνακα μεγέθους 1, καθώς για κάθε έξοδο της συνάρτησης έχουμε μόνο ένα. Τέλος τυπώνονται οι τιμές των βαρών και του *bias*.

Θα πρέπει να ορισθεί και ρυθμός μάθησης που στην μηχανική μάθηση συμβολίζεται με το ελληνικό γράμμα η και εκφράζει την ταχύτητα με την οποία εκπαιδεύεται ο αλγόριθμος. Το *learning rate* λοιπόν εκφράζεται ως εξής:

$$x_{i+1} = x_i - \eta * f'(x_i)$$

Το x_i είναι η φάση που βρίσκεται ο αλγόριθμος τώρα και το x_{i+1} είναι η επόμενη. Όπου το $f'(x_i) = 0$ τότε το $x_{i+1} = x_i$ και δεν υπάρχουν περαιτέρω βελτιώσεις και λέμε ότι έχουμε φτάσει στο ελάχιστο. Στους αλγορίθμους μηχανικής μάθησης, ο ρυθμός μάθησης είναι επιθυμητό να είναι αρκετά υψηλός ώστε να φτάσουμε στο συντομότερο ελάχιστο σε ένα λογικό χρονικό διάστημα. Αλλά και αντίστοιχα αρκετά χαμηλός ώστε να μην ταλαντευόμαστε κοντά στο ελάχιστο, χωρίς να το πετύχουμε.

Σειρά έχει η εκπαίδευση του μοντέλου, η οποία όπως είδαμε στην *Εικόνα 3.3.2.1* είναι επαναληπτική διαδικασία. Αυτό μας οδηγεί στο να χρησιμοποιήσουμε μια “λούπα” με την χρήση της εντολής *for* με πλήθος επαναλήψεων 100. Σε κάθε επανάληψη υπολογίζονται οι έξοδοι και γίνεται σύγκριση αυτών με τις τιμές των επιθυμητών εξόδων (τιμές στόχοι). Η σύγκριση αυτή επιτυγχάνεται με την εισαγωγή της συνάρτησης υπολογισμού λάθους: *Loss Function*. Στην γραμμική παλινδρόμηση μια τέτοια συνάρτηση ονομάζεται **L2-norm** και εκφράζεται από τον εξής τύπο:

$$L2 - norm = \sum_i (y_i - t_i)^2$$

Η συνάρτηση λάθους **L2-norm** όπως φαίνεται αποτελεί το άθροισμα των διαφορών μεταξύ των τιμών εξόδου y_i και αυτών των τιμών στόχων t_i .

Με την χρήση της συνάρτησης λάθους λοιπόν βλέπουμε πόσο μακριά από το επιθυμητό αποτέλεσμα είμαστε και για αυτό τυπώνεται ώστε να παρακολουθούμε την εξέλιξη του αλγορίθμου. Φυσικά κάτι τέτοιο είναι χρήσιμο σε μεγάλο όγκο δεδομένων ή σε πολύ μικρό learning rate, όπου ο αλγόριθμος θα καθυστερήσει στην διεκπεραίωση του. Στο τέλος τις λούπας υπολογίζονται εκ νέου τα βάρη και το bias. Όταν ολοκληρωθεί η διαδικασία επανάληψης τυπώνουμε τα τελικά βάρη και το bias.

Στο τελευταίο βήμα του αλγορίθμου όπως βλέπουμε στο *Διάγραμμα 3.3.2.2* πως σχεδιάζεται η γραμμική παλινδρόμηση.

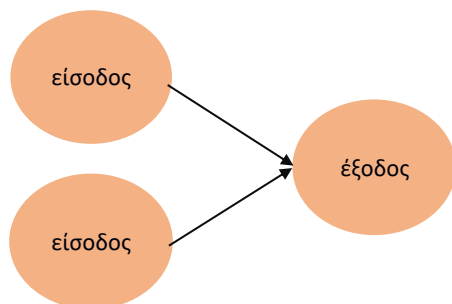
Γραμμική Παλινδρόμηση στην Βιοϊατρική

Η γραμμική παλινδρόμηση χρησιμοποιείται συχνά και στην επιστήμη της βιοπληροφορικής. Πολλές φορές το πρόβλημα έγκειται στο να προβλεφτεί η τιμή της εξόδου μια βιολογικής διαδικασίας, που λαμβάνει μέρος σε ένα συγκεκριμένο βιολογικό σύστημα, το οποίο μπορεί να προσδιοριστεί από γραμμικούς συνδυασμούς των επιπέδων έκφρασης των γονιδίων μέσα από το ίδιο το σύστημα. Το μείζον ζήτημα όπως είδαμε και στον παραπάνω αλγόριθμο άλλα όπως συμβαίνει και σε κάθε γραμμικό μοντέλο παλινδρόμησης είναι να καθοριστούν τα βάρη παλινδρόμησης. Στην περίπτωση μια βιολογικής διαδικασίας το ζήτημα είναι η εύρεση των τιμών των βαρών που αντιστοιχούν σε κάθε γονίδιο. Ο ρόλος των βαρών γονιδίων σε μια τέτοια διαδικασία δεν σταματά στον καθορισμό της κατάλληλης εξόδου, αλλά προσδιορίζει την συμαντικότητα των γονιδίων σε σχέση με την βιολογική διαδικασία.

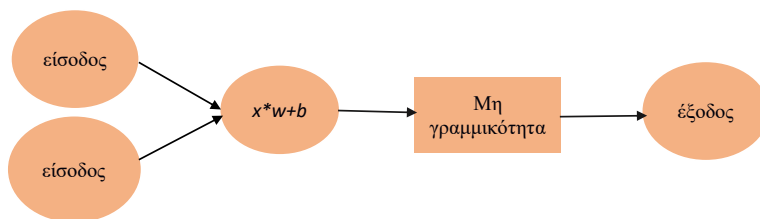
4.4 Νευρωνικά Δίκτυα

Όπως είδαμε παραπάνω η γραμμική παλινδρόμηση μπορεί να εφαρμοστεί για να αξιολογήσει την σχέση μεταξύ δύο μεταβλητών. Παρ' όλα αυτά στην πράξη τα προβλήματα στην πλειοψηφία τους δεν είναι γραμμικά. Σε πιο σύνθετα προβλήματα τα νευρωνικά δίκτυα, αποτελούν ένα ιδιαίτερα χρήσιμο εργαλείο. Τι είναι ωστόσο ένα νευρωνικό δίκτυο; Σύμφωνα με τους Mitchell, 1997 και Simon Haykin 1993: Νευρωνικό δίκτυο ονομάζεται ένα κύκλωμα διασυνδεδεμένων νευρώνων, είτε βιολογικών είτε τεχνητών. Στην περίπτωση των πρώτων, πρόκειται για ένα τμήμα νευρικού ιστού. Στην περίπτωση των τεχνητών νευρώνων, αναφερόμαστε σε ένα αφηρημένο αλγοριθμικό κατασκευάσμα το οποίο εμπίπτει στον τομέα της υπολογιστικής νοημοσύνης, όταν στόχος του νευρωνικού δικτύου είναι η επίλυση κάποιου υπολογιστικού προβλήματος, ή της υπολογιστικής νευροεπιστήμης, όταν στόχος είναι η υπολογιστική προσομοίωση της λειτουργίας των βιολογικών νευρωνικών δικτύων με βάση κάποιο μαθηματικό μοντέλο.

Σε ένα γραμμικό μοντέλο έχουμε μια αναπαράσταση όπως αυτή στο *Σχήμα 4.4.1*, όπου και μπορούμε να πούμε ότι είναι ένα νευρωνικό δίκτυο χωρίς **βάθος**.



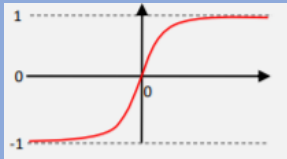
Σχήμα 4.4..1



Σχήμα 4.4..2

Στο Σχήμα 4.4.2, βρίσκεται μια περιληπτική αναπαράσταση του τι είναι ένα νευρωνικό δίκτυο. Το μοντέλο σε αυτό το σχήμα αλλάζει από εισόδους που είναι γραμμικά συνδυασμένες (από εδώ προκύπτει και η γραμμική συνάρτηση $x*w+b$), με αποτέλεσμα οι έξοδοι να αποτελούν νέες εισόδους. Οι νέες εισοδοι είναι γραμμικά ενωμένες και έπειτα περνούν από έναν μη-γραμμικό μετασχηματισμό που έχει ως αποτέλεσμα τις εξόδους.

Στην μηχανική μάθηση λοιπόν οι μη γραμμικοί μετασχηματισμοί ονομάζονται συναρτήσεις ενεργοποίησης. Οι συναρτήσεις ενεργοποίησης μετασχηματίζουν τις εισόδους σε εξόδους διαφορετικού τύπου. Υπάρχουν τέσσερις κύριες συναρτήσεις και όλες είναι μονότονες, συνεχείς και παραγωγίσιμες.

Όνομα	Τύπος	Παράγωγος	Γράφημα	Εύρος τιμών
Σιγμοειδής (logistic function)	$\sigma(\alpha) = \frac{1}{1 + e^{-\alpha}}$	$\frac{\theta \sigma(\alpha)}{\theta \alpha} = \sigma(\alpha)(1 - \sigma(\alpha))$		(0,1)
TanH (hyperbolic tangent)	$\tanh(\alpha) = \frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}}$	$\frac{\theta \tanh(\alpha)}{\theta \alpha} = \frac{4}{(e^{\alpha} + e^{-\alpha})^2}$		(-1,1)
ReLU (rectified linear unit)	$\text{relu}(\alpha) = \max(0, \alpha)$	$\frac{\theta \text{relu}(\alpha)}{\theta \alpha} = \begin{cases} 0, & \text{αν } \alpha \leq 0 \\ 1, & \text{αν } \alpha > 0 \end{cases}$		(0,∞)
SoftMax	$\sigma_i(\alpha) = \frac{e^{\alpha_i}}{\sum_j e^{\alpha_j}}$	$\frac{\theta \sigma_i(\alpha)}{\theta \alpha_i} = \sigma_i(\alpha)(\delta_{ij} - \sigma_i(\alpha))$ <i>Όπου δ_{ij} ισούται με 1 αν $i=j$, αλλιώς 0</i>		(0,1)

Πίνακας 4.4..2

Από τα online courses της ομάδας 365 Data Science

Οι συναρτήσεις ενεργοποίησης καλούνται και συναρτήσεις μετασχηματισμού και οι δύο όροι χρησιμοποιούνται στην μηχανική μάθηση ως ισοδύναμοι, παρ' όλο που σε άλλες επιστήμες διαφέρουν.

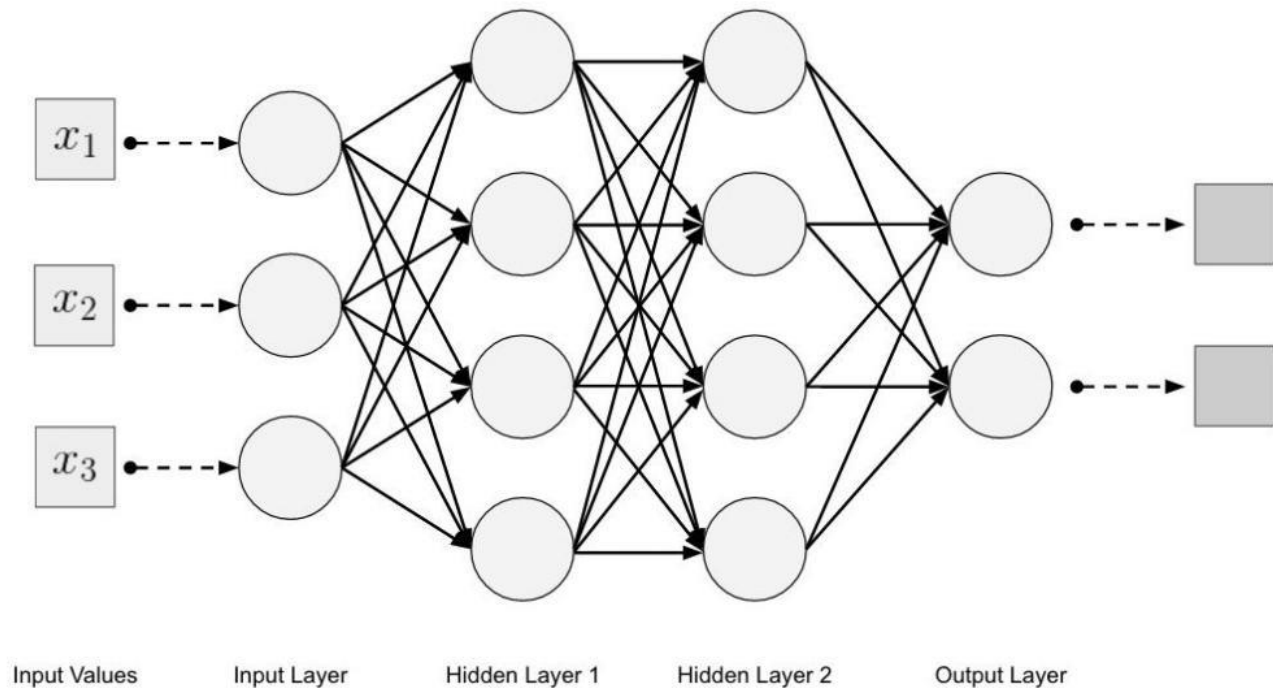
Συνάρτηση Softmax

Η συνάρτηση ενεργοποίησης SoftMax μετατρέπει ένα αριθμητικό σύνολο μεγάλων ή μικρών αριθμών σε μια έγκυρη κατανομή πιθανότητας. Αντιθέτως με τις άλλες συναρτήσεις ενεργοποίησης που παίρνουν μια τιμή εισόδου και την μετατρέπουν ανεξάρτητα των υπολοίπων στοιχείων, η SoftMax λαμβάνει υπόψιν της τα δεδομένα στο σύνολο τους. Για τον λόγο ότι αποτελεί μια κατανομή πιθανότητας οι τιμές εξόδου έχουν ένα εύρος τιμής από 0 έως 1 και το άθροισμα τους είναι πάντα ίσο με την μονάδα.

Η ιδιότητα της συνάρτησης SoftMax στο να παράγει ως έξοδο πιθανότητες είναι τόσο χρήσιμη ώστε να αποτελεί την πιο συχνά χρησιμοποιούμενη συνάρτηση ενεργοποίησης του τελικού σταδίου εξόδων ενός νευρωνικού δικτύου. Ωστόσο όταν χρησιμοποιείται πριν το τελικό στάδιο, το αποτέλεσμα δεν είναι τόσο αξιόπιστο και ικανοποιητικό. Ο λόγος που συμβαίνει κάτι τέτοιο είναι επειδή πολλές πληροφορίες που σχετίζονται με την μεταβλητότητα των δεδομένων χάνονται.

4.5 Στρώματα Νευρωνικών δικτύων

Τα μη γραμμικά νευρωνικά δίκτυα διαφέρουν από τα γραμμικά όπως ειπώθηκε και για τον λόγο ότι έχουν **βάθος**. Ένα νευρωνικό δίκτυο αποτελείται από στρώματα (layers) και έχει την αναπαράσταση του Σχήματος 4.5.3



Σχήμα 4.5.3

Πηγή: <https://www.oreilly.com/library/view/deep-learning/9781491924570/ch04.html>

Στην γραφική απεικόνιση ενός νευρωνικού δικτύου, παρατηρούμε ένα δίκτυο τεσσάρων στρωμάτων. Αρχικά πρώτο είναι το στρώμα εισόδου, εδώ παίρνανε οι τιμές εισόδου για την εκπαίδευση του δικτύου. Κάθε κύκλος συμβολίζει ένα κόμβο. Έπειτα μέσω των “βέλων” που αποτελούν τους μαθηματικούς μετασχηματισμούς, βρισκόμαστε στα “κρυφά στρώματα” (hidden layers) και τέλος οδηγούμαστε στο στρώμα εξόδου.

Ένα νευρωνικό δίκτυο χαρακτηρίζεται από το **βάθος (depth)** και το “**φάρδος (width)**”. Το βάθος ενός δικτύου ισούται με το πλήθος των επιπέδων ή με τον αριθμό των κρυμμένων στρωμάτων. Το width του δικτύου αναφέρεται στο πλήθος των κόμβων του μεγαλύτερου στρώματος του δικτύου. Τέλος τα δύο αυτά χαρακτηριστικά ονομάζονται και ως **hyperparameters** και είναι τιμές που επιλέγονται στην αρχή της κατασκευής ενός δικτύου.

Hyperparameters (εισαγόμενες στην αρχή)	Parameters (αλλάζουν μέσω του αλγορίθμου)
Width	Weights-βάρη (w)
Depth	Biases (b)
Learning rate	

Πίνακας 4.5.2

Διαφορές hyperparameters & απλών παραμέτρων

Πλεονεκτήματα Νευρωνικών δικτύων

Τα νευρωνικά δίκτυα χαρακτηρίζονται από τις ακόλουθες δυνατότητες:

1. **Μη γραμμικότητα.** Ένας νευρώνας έχει την δυνατότητα να χαρακτηρίζεται είτε ως γραμμικός είτε ως μη-γραμμικός. Ένα δίκτυο από μη γραμμικούς νευρώνες είναι επομένως και αυτό μη γραμμικό. Τα μη γραμμικά νευρωνικά δίκτυα μπορούν να εφαρμοστούν για την επίλυση μη γραμμικών προβλημάτων, όπως είναι η πρόβλεψη του καιρού, η ομιλία κ.ά.
2. **Αντιστοίχιση εισόδου-εξόδου:** Η αντιστοίχιση εισόδου εξόδου δημιουργείται μέσω της διαδικασίας μάθησης, προσπαθώντας να μειωθούν οι διαφορές της επιθυμητής τιμής εξόδου και της πραγματικής τιμής στην προσπάθεια μας να περιοριστούν οι αυθαίρετες εκτιμήσεις στο χώρο του σήματος εισόδου.
3. **Προσαρμοστικότητα:** Τα νευρωνικά δίκτυα έχουν την δυνατότητα να προσαρμόζουν τα συνοπτικά βάρη τους ανάλογα με τις αλλαγές που λαμβάνουν μέρος στο περιβάλλον τους. Ένα νευρωνικό δίκτυο που είναι εκπαιδευμένο να λειτουργεί σε ένα συγκεκριμένο περιβάλλον, μπορεί εύκολα να επανεκπαιδευθεί ώστε να λειτουργεί και σε κάποιο διαφορετικό περιβάλλον λειτουργίας. Η φυσική αρχιτεκτονική ενός νευρωνικού δικτύου για την ταξινόμηση προτύπων, την επεξεργασία σήματος και τις εφαρμογές ελέγχου, σε συνδυασμό με την προσαρμοστική του δυνατότητα, το καθιστά ιδιαίτερα χρήσιμο εργαλείο για την προσαρμοστική ταξινόμηση προτύπων, την προσαρμοστική επεξεργασία σήματος και τον προσαρμοστικό έλεγχο συστημάτων. Σαν γενικό κανόνα θα μπορούσαμε να πούμε ότι όσο πιο εύκολα προσαρμόζεται ένα σύστημα τόσο πιο χρήσιμο και πιο αποδοτικό θα είναι κυρίως όταν κληθεί να λειτουργήσει σε ένα ασταθές περιβάλλον.
4. **Ενδεικτική απόκριση:** Ένα νευρωνικό δίκτυο μπορεί να σχεδιαστεί ώστε να παρέχει πληροφορία όχι μόνο για το ποίο πρότυπο θα επιλεγεί, αλλά και σχετικά με τον βαθμό εμπιστοσύνης στην ληφθείσα απόφαση. Ο βαθμός εμπιστοσύνης μπορεί να χρησιμοποιηθεί για να απορρίψει την ακρίβεια της αναγνώρισης
5. **Ανοχές σε βλάβες:** Η κακή λειτουργία ή η καταστροφή για παράδειγμα ενός νευρώνα ή κάποιων συνδέσμων δεν είναι ικανή να διαταράξει σημαντικά τη λειτουργία τους καθώς η πληροφορία που εσωκλείουν δεν είναι εντοπισμένη σε συγκεκριμένο σημείο αλλά διάχυτη σε όλο το δίκτυο.

Overfitting

Τα νευρωνικά δίκτυα διαθέτουν τεράστια πληθώρα εφαρμογών και αποτελούν το κομβικό εργαλείο πολλών πεδίων. Ιδιαίτερη σημασία στα νευρωνικά δίκτυα έχει η φάση της εκπαίδευσης του μοντέλου. Το μοντέλο πάντα εκπαιδεύεται με βάση ένα σύνολο δεδομένων. Για παράδειγμα εάν χρησιμοποιηθεί το μοντέλο για την διεκπεραίωση διάγνωσης μια ασθένειας, θα πρέπει να προηγηθεί πρώτα η εκπαίδευση του μοντέλου και να ταξινομήσει μέσα από ένα σύνολο δεδομένων, τις διάφορες περιπτώσεις που μπορούν να προκύψουν, με βάση τα δεδομένα εκπαίδευσης, με απώτερο σκοπό όταν δοθούν στο μοντέλο τα δεδομένα του αντικειμένου που βρίσκεται υπό μελέτη εκείνη την στιγμή να τα ταξινομήσει στην αντίστοιχη ομάδα που ανήκει και να βγει ένα πόρισμα. Η επιτυχημένη διάγνωση μέσω ταξινόμησης ενός μοντέλου νευρώνων σε δίκτυο, απαιτεί την σωστή λειτουργία του ίδιου του

μοντέλου. Το πιο συχνό φαινόμενο στο να αποτρέψει ένα μοντέλο να λειτουργήσει ορθά είναι το φαινόμενο του **Overfitting**.

Το φαινόμενο του Overfitting παρατηρείται κατά την φάση της εκπαίδευσης όταν το μοντέλο έχει εστιάσει σε τέτοιο βαθμό στα δεδομένα της εκπαίδευσης που έχει χάσει το νόημα, σε γενικό πλαίσιο. Δηλαδή είναι τόσο καλό στο να ταξινομήσει το σύνολο δεδομένων προς εκπαίδευση, που όταν του δοθούν δεδομένα εκτός αυτών της εκπαίδευσης θα αποτύχει να φέρει εις πέρας την εργασία που του ανατέθηκε αποδοτικά, οδηγούμενο σε σφάλματα. Ο τυχαίος θόρυβος που περιέχουν τα δεδομένα συμπεριλαμβάνεται στο μοντέλο. Με άλλα λόγια κάποια δεδομένα που είναι τυχαία και δεν αντικατοπτρίζουν την πραγματικότητα, εμπεριέχονται και επηρεάζουν το μοντέλο, στο να παράγει εσφαλμένες προβλέψεις. Το αντίθετο φαινόμενο του overfitting παρατηρείται επίσης στην βαθιά μηχανική μάθηση και ονομάζεται **Underfitting**.

Το Underfitting παρατηρείται κυρίως στο μοντέλο της γραμμικής παλινδρόμησης. Ένα μοντέλο που αντιμετωπίζει αυτό το πρόβλημα δεν διαθέτει την υποκειμενική λογική ώστε να μπορέσει να ταξινομήσει τα δεδομένα αποδοτικά. Φυσικά ένα τέτοιο μοντέλο έχει την δυνατότητα να δώσει ένα αποτέλεσμα ωστόσο δεν διαθέτει την δύναμη να προβλέψει ορθά το αποτέλεσμα που απαιτείται. Οι τιμές εξόδου σε αυτήν την περίπτωση περιγράφονται από μεγάλη απώλεια καθώς είναι μακριά από τις τιμές στόχους και η ακρίβεια του μοντέλου είναι ιδιαίτερα χαμηλή.

Αντιμετώπιση Overfitting

Το overfitting μπορεί να καταστήσει το μοντέλο μας, “άχρηστο” και εμείς όχι μόνο να μην το καταλάβουμε, αλλά να έχουμε την εντύπωση ότι είναι πολύ αξιόπιστο και ακριβές καθώς έχει ταξινομήσει το σύνολο δεδομένων που του δώσαμε να εκπαιδεύσει άψογα. Υπάρχουν περισσότεροι από ένας τρόποι για να αποτρέψουμε κάτι τέτοιο να συμβεί στο μοντέλο μας, αλλά ο πιο αποδοτικός είναι να χωρίσουμε τα δεδομένα εκπαίδευσης σε τρία υποσύνολα. Το πρώτο από αυτά που θα πρέπει να είναι και το μεγαλύτερο σε μέγεθος θα είναι το υποσύνολο το οποίο θα χρησιμοποιηθεί για την εκπαίδευση, Training subset. Τα άλλα δύο υποσύνολα αποτελούνται από το Validation subset και το Test subset.

Ο ρόλος του πρώτου υποσυνόλου είναι ξεκάθαρος και αφορά την εκπαίδευση του μοντέλου. Καθώς εκπαιδεύουμε το μοντέλο σταματάμε σε κάθε Epoch, σε κάθε μια από αυτές τις παύσεις το μοντέλο μας είναι εκπαιδευμένο μέχρι ένα σημείο. Τότε δίνουμε στο μοντέλο το validation subset χωρίς να ανανεώσουμε τα βάρη και ουσιαστικά αυτό που συμβαίνει είναι μόνο να υπολογιστεί τη τιμή της συνάρτησης λάθους. Εδώ γίνεται σύγκριση της τωρινής τιμής σφάλματος – Validation loss και της τιμής που είχε η συνάρτηση λάθους όσο εκπαιδεύαμε το μοντέλο Training loss. Κατά μέσο όρο οι δύο τιμές θα πρέπει να είναι ίσες, επειδή τα υποσύνολα έχουν παρθεί από το ίδιο αρχικό σύνολο δεδομένων. Η διαδικασία αυτή γίνεται επαναληπτικά και κάθε νέα τιμή σφάλματος θα πρέπει να είναι ίση ή ελάχιστα μικρότερη από την προηγούμενη, μέχρι ένα σημείο. Αυτό το σημείο θα “έρθει” όταν η τιμή του Validation loss αυξηθεί. Τότε ο αλγόριθμος αρχίζει και προβλέπει πιο αποδοτικά το συγκεκριμένο σύνολο δεδομένων. Αυτό είναι και το σημείο που θα πρέπει να διακοπεί η εκπαίδευση του αλγορίθμου, καθώς έφτασε στο σημείο του Overfitting.

Το τελευταίο υποσύνολο είναι εξίσου μεγάλης σημασίας και “έρχεται” στο τέλος της παραπάνω διαδικασίας. Το Test subset υπολογίζει την τελική δύναμη πρόβλεψης που έχει το μοντέλο. Το αποτέλεσμα πρόβλεψης που θα έχουμε χάρη στο Test είναι και αυτό που ανταποκρίνεται στην πραγματικότητα. [66]

Για το πως χωρίζεται το σύνολο δεδομένων στα τρία υποσύνολα δεν υπάρχει ξεκάθαρος κανόνας. Ο πιο δημοφιλής διαχωρισμός είναι 80%-10%-10%. Το υποσύνολο εκπαίδευσης πρέπει πάντα να είναι αρκετά μεγάλο αλλά τόσο ώστε να υπάρχουν και δεδομένα για τα άλλα δύο υποσύνολα.

Οι εναλλακτικές μέθοδοι αντιμετώπισης του Overfitting δεν εγγυούνται την πλήρη αποτροπή του αλλά μπορεί να επιλεχθούν σε περιπτώσεις που το σύνολο δεδομένων δεν είναι αρκετά μεγάλο για να επιτρέψει τον διαχωρισμό του σε τρία υποσύνολα. Η μία μέθοδος είναι η ***N-fold Cross Validation*** όπου το Training subset και το Validation subset αποτελούν ένα και το Test subset παραμένει όπως και στην προηγούμενη μέθοδο. Εδώ χωρίζουμε το πρώτο υποσύνολο N φορές σε μια επαναληπτική διαδικασία N+1 Epoch, το ένα από τα N υπο-υποσύνολα αποτελεί το κομμάτι του Validation και τα υπόλοιπα N-1 αφορούν την εκπαίδευση. Στην πρώτη Epoch το υποσύνολο του Validation βρίσκεται στο πρώτο κομμάτι. Στην δεύτερη Epoch το υποσύνολο του Validation λαμβάνει χώρο στο δεύτερο από τα N κομμάτια. Η διαδικασία συνεχίζεται για N+1 Epoch και το Validation μετακινείται πάντα ένα κομμάτι μπροστά μέχρι να ξαναπάρει θέση στο πρώτο. Με αυτό το τρόπο δεν συγχέουμε τα δύο υπο-υποσύνολα και επίσης επιτυγχάνεται να χρησιμοποιηθεί όλο το σύνολο δεδομένων εκτός του υποσυνόλου Test. Ωστόσο η μέθοδος αυτή δεν είναι πλήρως αποτρεπτική και πιθανότατα να υπάρξει ποσοστό Overfitting στο μοντέλο μας.

Αρχικοποίηση – Initialization

Ένα πολύ σημαντικό σημείο ενός σωστά λειτουργικού νευρωνικού δικτύου είναι η αρχικοποίηση. Η αρχικοποίηση είναι η διαδικασία κατά την οποία θέτονται οι αρχικές τιμές που θα έχουν τα βάρη.

Όπως είδαμε στο νευρωνικό δίκτυο που κατασκευάστηκε για την απλή γραμμική παλινδρόμηση οι τιμές αρχικοποίησης που πήραν τα βάρη ήταν τυχαίες, μέσα σε ένα εύρος τιμών από -0.1 μέχρι +0.1. Η μέθοδος αυτή είναι αρκετά αποδοτική για νευρωνικά δίκτυα με μόνο ένα κρυφό επίπεδο, ωστόσο μια τόσο απλή μέθοδος δεν μπορεί να είναι εξίσου αξιόπιστη και στα βαθιά νευρωνικά δίκτυα. Ο δημοφιλής τρόπος αρχικοποίησης των τιμών αυτών ονομάζεται *Glorot Initialization* ή *Xavier Initialization* και προτάθηκε από τον καθηγητή Xavier Glorot το 2010 και υιοθετήθηκε καθολικά από την επιστημονική κοινότητα. Η κύρια ιδέα αναφέρεται στο ότι η μέθοδος που χρησιμοποιείται για την τυχαία επιλογή βαρών δεν είναι ιδιαίτερα σημαντική και ότι το πλήθος των εισόδων και των εξόδων είναι αυτό που διαδραματίζει το πιο κρίσιμο ρόλο. Με κάθε επίπεδο που περνάει η αρχικοποίηση του Xavier διατηρεί την διακύμανση μέσα σε ένα οριοθετημένο εύρος, με σκοπό να γίνει πλήρη εκμετάλλευση της συνάρτησης ενεργοποίησης.

Υπάρχουν δύο τύποι Αρχικοποίησης του Xavier.

Ο πρώτος είναι η ***Uniform Xavier Initialization***: Σχεδιάζουμε κάθε βάρος w από μια τυχαία ομοιόμορφη κατανομή εύρους,

$$[-x, x] \text{ για } x = \sqrt{\frac{6}{\text{είσοδοι} + \text{έξοδοι}}}$$

Ο δεύτερος τρόπος είναι η **Normal Xavier Initialization**: Σχεδιάζουμε κάθε βάρος w από μια κανονική κατανομή με μέση τιμή,

$$0 \text{ και } \sigma = \sqrt{\frac{2}{\text{είσοδοι} + \text{έξοδοι}}}$$

4.6 Πολυστρωματικά Τροφοδοτούμενα προς τα εμπρός Δίκτυα

Τα πολυστρωματικά τροφοδοτούμενα προς τα εμπρός δίκτυα-Multilayer Feedforward Networks ή αλλιώς MLFs ονομάζονται να νευρωνικά δίκτυα με αναπαράσταση όπως αυτή του σχήματος 4.5.3. Στα MLFs η εκπαίδευση γίνεται με επίβλεψη και όπως βλέπουμε στο σχήμα έχουν τριών ειδών στρώματα. Αυτό της εισόδου, τα κρυφά στρώματα εκεί όπου οι νευρώνες των κρυφών στρωμάτων πραγματοποιούν τους υπολογισμούς που συμβαίνουν στο δίκτυο και τέλος αυτό της εξόδου που αποτελείται όπως και τα δύο προηγούμενα από κόμβους-νευρώνες και εκτελούνται και εκεί υπολογισμοί. Όπως είναι φανερό και από το σχήμα, το στρώμα εισόδου βρίσκεται στα αριστερά οπότε στο δίκτυο τα σήματα εισόδου σε κάθε επαναληπτική διαδικασία εισέρχεται από τα αριστερά προς τα δεξιά και μεταδίδεται από στρώμα σε στρώμα. Ο πιο δημοφιλής αλγόριθμος που χρησιμοποιείται για την επίτευξη αυτής της μεθόδου είναι ο **Back Propagation** και αποτελεί μια μέθοδο του κανόνα μάθησης με διόρθωση σφάλματος. Όπως αναφέρθηκε και προηγουμένως στόχος της επαναληπτικής διαδικασίας στους αλγορίθμους των μοντέλων νευρωνικών δικτύων είναι να φτάσουμε στο ελάχιστον δυνατό ποσοστό λάθους. [67]

Back propagation

Τα πιο δημοφιλή νευρωνικά δίκτυα είναι τα πολυστρωματικά τα οποία εκπαιδεύονται με χρήση μάθησης με επίβλεψη και κυρίως των αλγόριθμο **Back propagation**. Στην πλειοψηφία τους χρησιμοποιούνταν ως μοντέλα για την επίλυση προβλημάτων του επιστημονικού πεδίου της χημείας. [65] Πλέον η ραγδαία εξέλιξη της τεχνολογίας έχει εξελιχθεί σε τέτοιο επίπεδο που η εφαρμογή του εργαλείου των πολυστρωματικών δικτύων με χρήση του συγκεκριμένου αλγορίθμου γίνεται σχεδόν καθολικά στο μεγαλύτερο μέρος των ερευνών της βιοϊατρικής και συγκεκριμένα στην διάγνωση, των ασθενειών.

Ο αλγόριθμος *Back propagation* είναι χρησιμοποιείται πολύ συχνά στην εκπαίδευση των δικτύων και περιλαμβάνει μια προς τα εμπρός φάση -από αριστερά προς τα δεξιά- και μία προς τα πίσω -από δεξιά προς τα αριστερά-. Ο λόγος που μπορεί να επιτευχθεί μια τέτοια εναλλαγή μεταβίβασης του σήματος και προς τα εμπρός και προς τα πίσω οφείλεται στο ότι οι νευρώνες εξόδου του ενός στρώματος αποτελούν τις εισόδους του επόμενου.

4.7 K-MEANS (γνώση πλήθους ομάδων)

Σε περίπτωση που υπάρχει γνώση του αριθμού των ομάδων στις οποίες πρέπει να διαχωριστούν τα δεδομένα, ο αλγόριθμος k-means αποτελεί ιδανική εναλλακτική πρόταση στις ιεραρχικές μεθόδους. Σύμφωνα με τον προαναφερθέν αλγόριθμο τα δεδομένα χωρίζονται σε προκαθορισμένο “k” αριθμό ομάδων-κλάσεων, έτσι ώστε εντός της κάθε ομάδας να υπάρχουν τα παρόμοια δεδομένα και αντίστοιχα τα δεδομένα διαφορετικών ομάδων να είναι ανόμοια. Ο αλγόριθμος k-means (k-μέσων) αποτελεί έναν αλγόριθμο (MacQueen 1967) που ομαδοποιεί αντικείμενα βάσει των χαρακτηριστικών μεριδίων k. Αποτελεί μεταβλητή του αλγορίθμου μεγιστοποίησης αναμονής (expectation-maximization algorithm-EM), όπου σκοπός είναι να οριστεί ο k-means δεδομένων που προήλθαν από Γκαουσιανές κατανομές.

Η ταξινόμηση σε ομάδες μέσω του αλγορίθμου K-means χρησιμοποιεί μη επιβλεπόμενη μάθηση, η οποία επιλέγεται όταν τα δεδομένα δεν διακρίνονται από χαρακτηριστικά που να τα κατηγοριοποιεί. Ο στόχος του είναι λοιπόν να βρει σε ποια από τις k σε πλήθος ομάδες ανήκει το κάθε δεδομένο. Ο αλγόριθμος πραγματοποιεί την ομαδοποίηση αυτή μέσα από μία επαναλαμβανόμενη εφαρμογή εντολών-βημάτων. Τα αποτελέσματα που δίνει ο αλγόριθμος είναι: αρχικά τα κέντρα των ομάδων που ονομάζονται κεντροειδή που χρησιμοποιούνται για να κατηγοριοποιηθούν τα δεδομένα, ουσιαστικά να προσαρτηθεί ένας τίτλος σε κάθε ομάδα, με βάση τα χαρακτηριστικά της. Δεύτερον και τελικά τα δεδομένα ταξινομούνται στις ομάδες. Κάθε κεντροειδές μιας ομάδας αποτελείται από μια συλλογή χαρακτηριστικών τιμών που ορίζουν την ομάδα που προκύπτει. Η εξέταση του βάρους του κεντρικού στοιχείου χρησιμοποιείται για την ποιοτική ερμηνεία του είδους της ομάδας που αντιπροσωπεύει. [44]

Όπως αναφέρθηκε ο αλγόριθμος K-means λειτουργεί επαναληπτικά. Αρχικά γίνεται εισαγωγή του πλήθους K των ομάδων και του σετ των δεδομένων. Το σετ δεδομένων αποτελεί μια συλλογή από χαρακτηριστικά τα οποία αντιπροσωπεύουν το κάθε σημείο δεδομένων. Με την εκτέλεση του αλγορίθμου υπολογίζεται η αρχική θέση των K κεντροειδών, η οποία παράγεται τυχαία ή επιλέγεται τυχαία από το σετ δεδομένων. Τότε ο αλγόριθμος ξεκινάει την επαναληπτική διαδικασία των δύο βημάτων. Στο πρώτο βήμα κάθε σημείο ανατίθεται στο κεντροειδές της ομάδας από το οποίο έχει την μικρότερη απόσταση. Η απόσταση είναι η Ευκλείδεια απόσταση. Το δεύτερο βήμα της επανάληψης αφορά τον επαναπροσδιορισμό των κεντροειδών της κάθε ομάδας. Αυτό επιτυγχάνεται υπολογίζοντας τον μέσο του συνόλου των σημείων της ομάδας. [44]

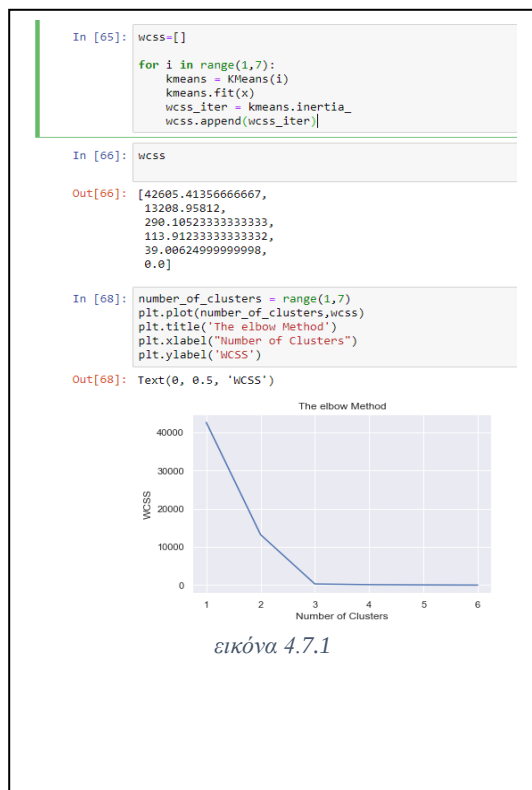
Ο αλγόριθμος για να φτάσει στην τελική του κατάσταση μετά την επανάληψη των δύο βημάτων αυτών, θα πρέπει να πληρούνται τα κριτήρια. Όταν για παράδειγμα τα δεδομένα παραμένουν στις ίδιες ομάδες (δηλαδή δεν αλλάζει ο μέσος του συνόλου επομένως οι κεντροειδή δεν μεταβάλλονται), όταν το άθροισμα των αποστάσεων βρίσκεται στο ελάχιστο δυνατό ή όταν κάποιο μέγιστο πλήθος επαναλήψεων έχει εκτελεστεί, τότε ο K-means βρίσκεται στην τελική του κατάσταση.

Τα βασικά βήματα του αλγορίθμου είναι τα εξής:

1. Επιλογή του αριθμού των ομάδων.
2. Τυχαία δημιουργία k ομάδων και ορισμός των κεντροειδών των ομάδων
3. Μετάβαση του κάθε σημείου στο κεντροειδές της κοντινότερης ομάδας.
4. Υπολογισμός των νέων κεντροειδών των ομάδων.
5. Επανάληψη μέχρι να συγκλίνει ο αλγόριθμος σε κάποιο κριτήριο.

Ο αλγόριθμος αυτός παραμένει δημοφιλής επειδή τείνει σε κάποιο όριο πολύ γρήγορα. Όσον αφορά την απόδοση ο αλγόριθμος δεν εγγυάται ότι θα αγγίξει το βέλτιστο. Η ποιότητα της τελικής λύσης εξαρτάται πολύ από το αρχικό σύνολο ομάδων και μπορεί να είναι πολύ χαμηλότερη από το συνολικό βέλτιστο. Επίσης ένα άλλο μειονέκτημα του αλγόριθμου είναι ότι ο αριθμός των ομάδων πρέπει να οριστεί εξαρχής. [7]

Πως ορίζει κανείς το σωστό πλήθος των ομάδων, ποιο είναι το βέλτιστο K για το κάθε σετ δεδομένων. Για να βρεθεί ο αριθμός K ο χρήστης θα πρέπει να τρέξει τον αλγόριθμο για διαφορετικές τιμές K και να συγκρίνει τα αποτελέσματα. Μια δημοφιλής μέθοδος που χρησιμοποιείται για να συγκρίνει κανείς τα αποτελέσματα εφαρμογών του αλγόριθμου στο ίδιο σετ δεδομένων που είχε ως είσοδο διαφορετικές τιμές K , είναι να πάρει τις μέσες τιμές απόστασης μεταξύ των σημείων δεδομένων που ανήκουν σε μία ομάδα, με τον κεντροειδή της. Όσο αυξάνεται το πλήθος των ομάδων ταξινόμησης τόσο θα μειώνεται η απόσταση μεταξύ των σημείων και του κεντροειδή. Για να επιλεγεί ο σωστός αριθμός K , δημιουργείται μια γραφική παράσταση, με τους άξονες αυτής να είναι η μέση απόσταση από τον κεντροειδή και το πλήθος των ομάδων, K . Το σημείο όπου η γραφική παράσταση μειώνεται απότομα ονομάζεται “elbow point” (σημείο αγκώνα) και η τιμή του K σε αυτό το σημείο είναι και η καταλληλότερη.



Στην εικόνα 4.7.1 παρουσιάζεται κώδικα σε Python 3.7 με χρήση των βιβλιοθηκών pandas, numpy, seaborn και matplotlib. Εδώ εμφανίζεται η τελευταία η οποία χρησιμοποιείται για την οπτικοποίηση των δεδομένων και στην προκειμένη περίπτωση φαίνεται ότι στο σετ δεδομένων που εφαρμόστηκε, το κατάλληλο πλήθος K των ομάδων ταξινόμησης είναι στο elbow point το οποίο είναι 3. Το σετ δεδομένων είναι από το παράδειγμα των διαδικτυακών μαθημάτων “The Data Science Course 2019: Complete Data Science Bootcamp” από την ομάδα “365 Careers Team” που βρίσκεται στο Udemy. Το σετ αυτό αποτελείται από έναν πίνακα 6 γραμμών και 4 στήλεων. Και αναφέρεται ενδεικτικά ως παραδειγμα.

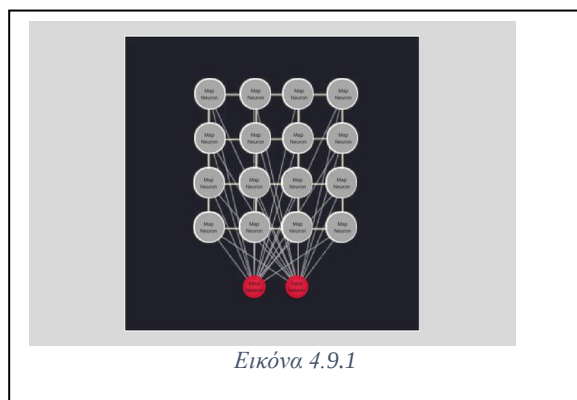
4.8 Τεχνικές Οπτικοποίησης

Κατά την τεχνική της οπτικοποίησης δημιουργούνται γραφήματα τα οποία μας επιτρέπουν να εξάγουμε συμπεράσματα για τα δεδομένα μας. Ο ανθρώπινος νους μπορεί να αντιλαμβάνεται εικόνες σε τρισδιάστατο χώρο μόνο. Όταν τα δεδομένα ξεπερνούν τις τρεις διαστάσεις δεν είναι εφικτό να πραγματοποιηθεί αποτελεσματική διάκριση χαρακτηριστικών και προτύπων και να γίνουν αντιληπτές οι τάσεις διαφοροποίησης μέσα στα δεδομένα. Αυτή είναι και η αιτία της εξέλιξης αλγορίθμων που μπορούν και επιτυγχάνουν μια πιο εύκολη αντίληψη των δεδομένων. Για να μπορέσουν να γίνουν ευκολότερα αντιληπτά τα δεδομένα και τα στοιχεία που έχουν μεγαλύτερο ενδιαφέρον, πραγματοποιείται ειδική μορφοποίηση και παρουσίαση τους. Οι τεχνικές οπτικοποίησης αποτελούν σημαντικό κομμάτι στην ανάλυση των δεδομένων στις επιστήμες ζωής και παρακάτω γίνεται η σύντομη περιγραφή συγκεκριμένων μεθόδων.

4.9 Self Organizing Maps

Οι SOM είναι αλγόριθμοι οι οποίοι βασίζονται στην χρήση Νευρωνικών Δικτύων ακολουθούν μάθηση χωρίς επίβλεψη και ανήκουν στις προσεγγίσεις διαχωρισμού. Τα δίκτυα αυτού του τύπου ονομάζονται και δίκτυα KOHONEN. Τα δίκτυα αυτά αποτελούνται από δύο στρώματα νευρώνων τα οποία είναι πλήρως συνδεδεμένα μεταξύ τους, λόγω του ότι κάθε νευρώνας εισόδου έχει μια προς τα εμπρός σύνδεση με κάθε νευρώνα εξόδου. Ο αλγόριθμος του μοντέλου ως εφαρμογή στην βιοπληροφορική έχει την δυνατότητα να διακρίνει τα γονίδια σε σειρά διαχωρισμών με κύριο κριτήριο την ομοιότητα των διανυσμάτων έκφρασης σε σύγκριση με διανύσματα αναφοράς τα οποία καθορίζονται για κάθε διαχωρισμό.

Αναλυτικότερα η εκπαίδευση του έχει στόχο την παραγωγή low-dimensional (συνήθως δισδιάστατες) αναπαραστάσεις των δειγμάτων εισόδων, τα οποία καλούνται “map” και αντίστοιχα η μέθοδος αυτή ονομάζεται “dimensionality reduction” (μείωση διαστάσεων). Ο αλγόριθμος SOM διαφέρει από τα άλλα τεχνητά νευρωνικά δίκτυα καθώς εφαρμόζει ανταγωνιστική μάθηση σε αντιπαράθεση με την μέθοδο “error-correction learning” (μάθηση με διόρθωση λαθών) που εφαρμόζεται σε αλγόριθμους όπως ο “backpropagation” δηλαδή χρησιμοποιούν μια λειτουργία γειτονικότητας για τη διατήρηση των τοπολογικών ιδιοτήτων του χώρου εισόδου. Αυτό θέτει τον αλγόριθμο ιδανικό για οπτικοποίηση δεδομένων πολλαπλών διαστάσεων.



Εικόνα 4.9.1

Όπως τα περισσότερα Τεχνητά Νευρωνικά Δίκτυα έτσι και ο αλγόριθμος SOM λειτουργεί σε δύο σκέλη: την εκπαίδευση και το “mapping” (χαρτογράφηση). Κατά την εκπαίδευση δημιουργείται ο “χάρτης” χρησιμοποιώντας παραδείγματα εισόδου και κατά το “mapping” γίνεται αυτόματη ταξινόμηση ενός νέου διανύσματος εισόδου. Ο λόγος που έχει πάρει την ονομασία “mapping” είναι προφανής αν δει κανείς μια εικόνα (Εικόνα 4.9.1) από ένα νευρωνικό δίκτυο. Ουσιαστικά αποτελούνται από φύλλα νευρωνικών δικτύων, των οποίων οι νευρώνες ενεργοποιούνται από σήματα εισόδου ποικίλων προτύπων και κλάσεων – κατηγοριών.

Το ορατό μέρος ενός αυτό-οργανωμένου χάρτη είναι ο χώρος του χάρτη, ο οποίος αποτελείται από τους νευρώνες που αποτελούν κόμβους και συνθέτουν το νευρωνικό δίκτυο. Ο χώρος του χάρτη

ορίζεται εκ των προτέρων, συνήθως ως μια πεπερασμένη δισδιάστατη περιοχή όπου οι κόμβοι είναι διατεταγμένοι σε κανονικό εξαγωνικό ή ορθογώνιο πλέγμα. Κάθε κόμβος συνδέεται με ένα διάνυσμα "βάρους", το οποίο είναι μια θέση στον χώρο εισόδου. δηλαδή, έχει την ίδια διάσταση με κάθε διάνυσμα εισόδου. Ενώ οι κόμβοι του χώρου χάρτη παραμένουν σταθεροί, η εκπαίδευση περιλαμβάνει τη μετακίνηση των βαρών προς τα δεδομένα εισόδου, χωρίς να αλλοιώνεται η τοπολογία που προκαλείται από το χώρο του χάρτη. Έτσι, ο αυτό-οργανωτικός χάρτης περιγράφει μια χαρτογράφηση από ένα χώρο υψηλότερων διαστάσεων εισόδου σε ένα χώρο με μικρότερες διαστάσεις. Μόλις εκπαιδευτεί, ο χάρτης μπορεί να ταξινομήσει ένα διάνυσμα από το χώρο δεδομένων, βρίσκοντας τον κόμβο με το διάνυσμα βάρους που βρίσκεται πλησιέστερα (με τη μικρότερη απόσταση μέτρησης) στον διανυσματικό χώρο δεδομένων.

Όπως φαίνεται και στην *Εικόνα 4.9.1* έχουμε μια απλή δομή ενός SOM (αυτό-οργανωτική δομή χάρτη) που αποτελείται από δύο νευρώνες εισόδου και απ' ό,τι φαίνεται τα δεδομένα εισόδου παρουσιάζονται από τρισδιάστατα διανύσματα. Πάνω από αυτά βρίσκονται οι κόμβοι του δικτύου, ο στόχος του συγκεκριμένου νευρωνικού δικτύου είναι να πάρει τα δεδομένα εισόδου και να τα παρουσιάσει ως δισδιάστατα δεδομένα. Με άλλα λόγια σε ένα παράδειγμα αλγορίθμου SOM σαν κι αυτό ο αλγόριθμος χρησιμοποιώντας μη-επιβλεπόμενη μάθηση κατηγοριοποιεί τα τρισδιάστατα δεδομένα και τα μετατρέπει σε δισδιάστατα.

Αυτό μπορεί να γίνει σε μεγαλύτερο εύρος, έχοντας οποιοδήποτε μέγεθος διαστάσεων στα δεδομένα εισόδου και αντίστοιχα και στα δεδομένα που παράγει ο αλγόριθμος κατά την ολοκλήρωση του ως έξοδο.

Όπως αναφέρθηκε νωρίτερα ο αλγόριθμος των SOM χρησιμοποιεί μη-επιβλεπόμενη μάθηση, η οποία ονομάζεται και ανταγωνιστική μάθηση και υπάρχει και αιτία αν προσέξει κανείς την δομή του αλγορίθμου. Στο πρώτο βήμα του αλγορίθμου γίνεται η αρχικοποίηση των βαρών των ενώσεων μεταξύ των κόμβων – νευρώνων. Στην συνέχεια επιλέγεται τυχαία από τα δεδομένα μια είσοδος για το δίκτυο. Έπειτα το δίκτυο υπολογίζει τα βάρη των νευρώνων, των οποίων η απόσταση είναι πιο κοντά στα δεδομένα εισόδου ή αλλιώς και διάνυσμα εισόδου.

$$\text{Απόσταση}^2 = \sum_{i=0}^n (\text{είσοδος}_i - \text{βάρος}_i)^2$$

Όπου n είναι το πλήθος των ενώσεων, δηλαδή των βαρών. Ο νευρωνικός χάρτης με το καλύτερο αποτέλεσμα ονομάζεται "Best Matching Unit" ή BMU (καλύτερα ταιριαστή μονάδα). Και αυτό το δίκτυο που αποτελεί η BMU το διάνυσμα εισόδου εκφράζεται καλύτερα από κάθε άλλο και για αυτό λέγεται και ανταγωνιστική μάθηση. Ο αλγόριθμος δεν σταματά εδώ, καθώς δεν υπολογίζει μόνο αυτό το σημείο αλλά κατά τη διάρκεια της μαθησιακής διαδικασίας προσπαθεί να φέρει το σημείο όσο το δυνατόν πιο κοντά είναι εφικτό. Δηλαδή ο υπολογισμός της απόστασης επαναλαμβάνεται ανανεώνοντας τα βάρη της ένωσης ώστε η απόσταση να ελαχιστοποιηθεί. Τα βάρη των ενώσεων των γειτόνων της BMU "παρασύρονται" υπο μία έννοια από την μεταβολή της BMU και πλησιάζουν και αυτά την τιμή που είχε το διάνυσμα της εισόδου. Λόγω αυτού όλος ο χάρτης μετακινείται προς αυτό το σημείο. Για αυτό η ακτίνα των γειτόνων που άλλαξαν θέση θα πρέπει να είναι γνωστή. Αρχικά η ακτίνα είναι μεγάλη αλλά σε κάθε επανάληψη της λούπας, μειώνεται. Έτσι το επόμενο βήμα του αλγορίθμου είναι ο υπολογισμός της τιμής της ακτίνας. Ο τύπος με τον οποίο υπολογίζεται είναι αυτός:

$$\sigma(t) = \sigma_0 e^{\frac{-t}{\lambda}}$$

Όπου το t είναι η επανάληψη εκείνη την στιγμή, το σ_0 αποτελεί την ακτίνα του χάρτη και το λ είναι η συνάρτηση που καθορίζεται από το εξής κλάσμα:

$$\lambda = \frac{k}{\sigma_0}$$

όπου k το συνολικό πλήθος των επαναλήψεων. Το λ ουσιαστικά εξυπηρετεί τον σκοπό της σμίκρυνσης της ακτίνας κατά την διάρκεια της εκπαίδευσης. Με λίγα λόγια σε κάθε επανάληψη του αλγορίθμου τα δεδομένα που εισέρχονται ως είσοδοι “παρασύρουν” τα σχετικά σημεία πιο κοντά στις τιμές των εισόδων.

Κατά τον επαναυπολογισμό της ακτίνας μέσα στην λούπα, τα βάρη των νευρώνων που βρίσκονται μέσα σ’ αυτήν την ακτίνα είναι αυτά που επαναπροσδιορίζονται. Το μέγεθος της αλλαγής του βάρους ενός νευρώνα, είναι αντιστρόφως ανάλογο της απόστασης από την BMU, δηλαδή όσο πιο κοντά βρίσκεται σε αυτήν τόσο μεγαλύτερη θα είναι η αλλαγή του βάρους του. Η εξίσωση που εκφράζει αυτήν την αλλαγή είναι η εξής:

$$w(t + 1) = w(t) + \theta(t)L(t)(\text{είσοδος}(t) - w(t))$$

Το w είναι το βάρος. Το L αποτελεί βαθμό μάθησης και ομοίως με την εξίσωση της ακτίνας εκφράζεται από το κλάσμα:

$$L(t) = L_0 e^{\frac{-t}{\lambda}}$$

Όπως αναφέρθηκε το βάρος των κόμβων επηρεάζεται αντίστροφα από την απόσταση που έχουν σε σχέση με την BMU, στην εξίσωση του βάρους αυτή η αλλαγή οφείλεται στην ύπαρξη του $\theta(t)$. Η τιμή του οποίου υπολογίζεται από:

$$\theta(t) = e^{-\text{απόστασηBMU}/2\sigma(t)^2}$$

Όπως είναι εμφανές στην εξίσωση όταν η απόστασηBMU μικραίνει η τιμή του $\theta(t)$ πλησιάζει στην μονάδα, με αποτέλεσμα η αλλαγή του βάρους του νευρώνα να είναι μεγαλύτερη. Η παραπάνω διαδικασία επαναλαμβάνεται.[42]

Ανακεφαλαιώνοντας την λειτουργία του αλγορίθμου SOM, τα κύρια βήματα που συνθέτουν τον αλγόριθμο είναι τα εξής:

1. Αρχικοποίηση βαρών
2. Επιλογή του διανύσματος εισόδου από τα δεδομένα ως είσοδος του δικτύου
3. Υπολογισμός της BMU

4. Υπολογισμός ακτίνας των γειτονικών βαρών που θα επαναυπολογιστούν
5. Κάθε βάρος των νευρώνων μέσω στην ακτίνα επαναπροσδιορίζεται και προσαρμόζεται στο διάνυσμα εισόδου
6. Επανάληψη των βημάτων 2-5 για κάθε διάνυσμα εισόδου που υπάρχει στα δεδομένα.

Οι Αυτό-Οργανωμένοι χάρτες διαφέρουν από τα υπόλοιπα νευρωνικά δίκτυα. Χρησιμοποιούν μη επιβλεπόμενη μάθηση για να δημιουργήσουν ένα “χάρτη” για τα δεδομένα εισόδου. Παρέχουν λύσεις σε μεγάλα και δύσκολα ερμηνεύσιμα σύνολα δεδομένων. Λόγω της εύκολης προσαρμογής τους χρησιμοποιούνται σε πολλά επιστημονικά πεδία, παρ’ όλο που ο αλγόριθμος δημιουργήθηκε για σκοπούς αναγνώρισης ομιλίας από τον Kohonen. Σήμερα οι εφαρμογές του αλγορίθμου βρίσκονται στην ταξινόμηση βιβλιογραφίας, στην αναζήτηση και ταξινόμηση εικόνων, στην κατανόηση δεδομένων, καθώς και στην Ιατρική Διάγνωση.

Ο αλγόριθμος SOM έχει αποδειχτεί ότι χρησιμεύει ως το ιδανικό εργαλείο οπτικοποίησης δεδομένων, ειδικά όταν αυτά είναι πολυδιάστατα. Σύμφωνα με τον J. Vesanto στο paper [43] ο αλγόριθμος παρέχει τρεις διαφορετικές κατηγορίες οπτικοποίησης των δεδομένων. Η πρώτη δίνει μια ιδέα για το συνολικό σχήμα και την δομή που πιθανόν να έχουν οι ομάδες στις οποίες έχουν ταξινομηθεί τα δεδομένα. Η δεύτερη απευθύνεται στην ανάλυση των βαρών, των χαρακτηριστικών των κλάσεων ταξινόμησης και στην συσχέτιση μεταξύ των διανυσμάτων. Τέλος αναφέρεται μέσα στο paper ότι ο αλγόριθμος χρησιμοποιείται για την οπτικοποίηση, εξέτασης νέων περιπτώσεων με σκοπό την διευκρίνιση και εύρεση καινοτομιών.

Η πιο δημοφιλής τεχνική οπτικοποίησης που χρησιμοποιεί ο αλγόριθμος είναι ο Unified Distance Matrix (ενοποιημένος πίνακας αποστάσεων).

4.10 3D Printing & U*Matrix

Οι αυτό-οργανωτικοί χάρτες όπως αναλύονται παραπάνω αποτελούν μέθοδο προβολής, η οποία επέτρεψε τη χρήση του U * -Matrix ως μια καθιερωμένη μέθοδος για την οπτικοποίηση των δεδομένων ως τοπία. Χάρη στη διαθεσιμότητα εκτύπωσης τρισδιάστατων μοντέλων, υπάρχει η δυνατότητα παρουσίασης της δομής των δεδομένων με διαισθητικό τρόπο. Για το σκοπό αυτό, πληροφορίες σχετικά με το ύψος των τοπίων χρησιμοποιείται για την παραγωγή ενός τρισδιάστατου τοπίου με έναν έγχρωμο εκτυπωτή 3D. Ομοιότητες μεταξύ μεγάλων διαστάσεων δεδομένων παρατηρούνται ως κοιλάδες και ανομοιότητες ως βουνά ή κορυφογραμμές. Μέσω των 3D εκτυπώσεων οι επίκαιροι εμπειρογνώμονες μπορούν αν έχουν μια αντιπροσωπευτική αντίληψη των δομών πολλαπλών διαστάσεων.

Με την χρήση λοιπόν του U*Matrix υπάρχει η ευκαιρία δημιουργίας ενός τοπογραφικού χάρτη με υψομετρικές αποχρώσεις, ο οποίος στον ανθρώπινο οφθαλμό θυμίζει ένα τρισδιάστατο τοπίο. Ωστόσο κάθε τρισδιάστατο μοντέλο οπτικοποίησης απαιτεί να μελετηθεί από πολλαπλές οπτικές γωνίες και συχνά υπόκειται σε σοβαρά προβλήματα απόφραξης, παραμόρφωσης και πλοήγησης [Jansen et al., 2013] & [Shneiderman, 2003] [8]

4.11 Principal Component Analysis

Κατά την εφαρμογή τεχνικών μέτρησης που χρησιμοποιούνται στις επιστήμες ζωής, συγκεντρώνονται δεδομένα για πολύ περισσότερες μεταβλητές ανά δείγμα, συγκριτικά με το πλήθος των δειγμάτων. Ένα τέτοιο παράδειγμα είναι οι μικροσυστοιχίες DNA, που μετρούν τα επίπεδα χιλιάδων mRNA από εκατοντάδες δείγματα. Το γεγονός ότι η συλλογή δεδομένων είναι πολυδιάστατη αποτελεί τροχοπέδη, στην εύκολη οπτικοποίηση των δειγμάτων και περιορίζει την εξερεύνηση των δεδομένων. Για να επιτευχθεί λοιπόν η οπτικοποίηση χρειάζεται η μείωση των διαστάσεων των δεδομένων μέσω ενός μαθηματικού αλγορίθμου που παρέχει την δυνατότητα ανάλυσης των κύριων ή πρωταρχικών συνιστωσών, ενώ ταυτόχρονα διατηρεί τα μεγαλύτερο μέρος της πληροφορίας, επομένως, διατηρεί τη δυνατότητα απόκτησης γνώσης για τις αποκλίσεις μεταξύ δεδομένων άρα και δειγμάτων.

Αυτό μπορεί να γίνει δυνατό με την αναγνώριση κατευθύνσεων οι οποίες ονομάζονται πρωταρχικές συνιστώσες, κατά μήκος της μέγιστης απόκλισης. Με βάση λίγες συνιστώσες, κάθε δείγμα αναπαρίσταται από σχετικά λίγα νούμερα σε σχέση με τις τιμές των χιλιάδων μεταβλητών που υπάρχουν στην πραγματικότητα. Τα δείγματα τώρα, μπορούν να αναπαρασταθούν σε διάγραμμα και είναι εφικτό να φανούν οι ομοιότητες και οι διαφορές μεταξύ τους, έτσι μπορούν και αν ομαδοποιηθούν.

Η PCA: Principal Component Analysis αποτελεί, δηλαδή μία μέθοδο οπτικοποίησης των δεδομένων από τον πολυδιάστατο στον γεωμετρικά τρισδιάστατο χώρο (στην πλειοψηφία των περιπτώσεων), έτσι ώστε τα αποτελέσματα και τα ευρήματα να μπορούν να γίνουν κατανοητά και ευπαρουσίαστα στο ανθρώπινο μάτι. Η PCA μετατρέπει τον πειραματικό χώρο σε λιγότερες διαστάσεις βάσει του πίνακα συσχέτισης (correlation matrix) των δεδομένων, στηριζόμενη σε γραμμικό μετασχηματισμό. Στόχος όπως έχει γίνει ξεκάθαρο είναι να μειωθούν οι διαστάσεις των δεδομένων με σκοπό την οπτικοποίηση τους, ενώ ταυτόχρονα να διατηρεί το μεγαλύτερο μέρος της διακύμανσης τους ώστε να έχει νόημα αυτό που βλέπει ο χρήστης, για να μπορέσουν να εντοπιστούν οι ομοιότητες και οι διαφορές των στοιχείων, όπως και οι ομαδοποιήσεις που προκύπτουν από το μεγάλο πλήθος των δεδομένων. Γίνεται προσπάθεια να χρησιμοποιηθούν όσο το δυνατόν λιγότερων μεταβλητών και συγκεκριμένα των πιο αντιπροσωπευτικών μεταβλητών του πειράματος, χωρίς να χαθεί σημαντική πληροφορία για το βαθμό συσχέτισης των αποτελεσμάτων, έτσι και μειώνεται η πολυπλοκότητα αλλά και μένει ανεπηρέαστο το τελικό αποτέλεσμα. Αυτό είναι περισσότερο εφικτό όταν τα πειραματικά δεδομένα σχετίζονται πολύ μεταξύ τους και μπορούν να παραλειφθούν κάποια ή να υπολογιστεί ένας μέσος όρος. Στην περίπτωση που η συσχέτιση των αρχικών μεταβλητών δεν είναι μεγάλη, τότε η ανάλυση με PCA καθίσταται ακατάλληλη μέθοδος.

Η PCA ξεκινάει με τη δημιουργία του πίνακα των δεδομένων στον οποίο οι σειρές αποτελούνται από τα δείγματα που εξετάστηκαν και αντίστοιχα οι στήλες του είναι οι εξεταζόμενες μεταβλητές (αναλυτικό σήμα, αναλυόμενες ενώσεις) Ο αλγόριθμος της PCA ανακατασκευάζει τον αρχικό πίνακα μεταβλητών. Οι τιμές των κύριων συνιστωσών είναι οι συντεταγμένες των παρατηρήσεων στους νέους κάθετους άξονες. Η πρώτη συνιστώσα είναι ένα διάνυσμα σε γωνία 90° μοιρών με την πρώτη συνιστώσα και εξηγεί τη δεύτερη μεγαλύτερη διακύμανση, δηλαδή τη μέγιστη διαφοροποίηση της διακύμανσης που δεν μπορεί να εξηγήσει η πρώτη συνιστώσα και ούτω καθεξής με τις υπόλοιπες συνιστώσες.

4.12 SNP Ανάλυση

Τα SNPs (single nucleotide polymorphisms) είναι μεταλλάξεις μίας μοναδικής νουκλεοτιδικής θέσης λόγω της εξέλιξης και διαβιβάζονται με κληρονομικότητα. Αντιπροσωπεύουν τα περισσότερα των γενετικών διαφορών μεταξύ των ατόμων και αποτελούν τη βάση πολλών μελετών σύνδεσης ασθενειών-γονιδίων. Δεδομένου ότι ο αριθμός των SNP στο γονιδίωμα είναι υψηλός και αριθμοί περίπου τα 7 εκατομμύρια (Kruglyak & Nickerson, 2001), είναι σημαντικό να εντοπιστεί το υποσύνολο των σχετικών SNPs που είναι επαρκώς πληροφορημένα για να διαφοροποιήσουν τους ασθενείς που επηρεάζονται από διαφορετικές παθολογίες. Αρκετοί υπολογιστικοί μέθοδοι επιλογής SNP έχουν προταθεί με βάση διαφορετικά κριτήρια.

4.13 Significance Analysis Of Microarrays

Η SAM είναι μια στατιστική μέθοδος η οποία χρησιμοποιείται, κυρίως, στη γονιδιωματική. Έχει την ιδιότητα να αναγνωρίζει γονίδια με στατιστικά σημαντικές αλλαγές στην έκφραση συνταιριάζοντας ένα σύνολο t tests τα οποία αφορούν συγκεκριμένα γονίδια. Κάθε ένα από αυτά παίρνει μια συγκεκριμένη τιμή που βασίζεται στην αλλαγή της έκφρασης του σχετικά με τυπική απόκλιση των επαναλαμβανόμενων μετρήσεων για το γονίδιο. Τα γονίδια με μεγαλύτερες τιμές από ένα δοσμένο όριο καθορίζονται ως, δυναμικώς σημαντικά. Το ποσοστό το γονιδίων που μπορούν να προκύψουν με τυχαιότητα, αποτελεί το ρυθμό λανθασμένης ανακάλυψης (FDR), ο οποίος προσδιορίζεται από ανάλυση πολλαπλών συνδυασμών των μετρήσεων.

4.14 SUPPORT VECTOR MACHINE

Ο αλγόριθμος Support Vector Machine μπορεί να εκπαιδευτεί από παραδείγματα και εκπαιδευτικά σύνολα δεδομένων, με σκοπό να αναγνωρίσει και να κατηγοριοποιήσει αντικείμενα και δείγματα. Βασική εφαρμογή του αλγορίθμου στο πεδίο της βιοϊατρικής είναι η κατηγοριοποίηση δεδομένων γονιδιακής έκφρασης που προκύπτουν μέσω πειραμάτων μικροσυστοιχιών. Ουσιαστικά είναι αλγόριθμος επιβλεπόμενης μάθησης, ο οποίος χρησιμοποιεί έναν αρχικό σύνολο δεδομένων για την εκπαίδευση της συνάρτησης του ώστε στην συνέχεια να είναι σε θέση να μπορεί να αναγνωρίσει και να διαχωρίσει με επιτυχία τα υπόλοιπα δεδομένα με βάση την έκφραση τους. Με λίγα λόγια ένα βασικό πλεονέκτημα είναι ότι κάνει χρήση βιολογικών πληροφοριών για να καθορίσει τα στοιχεία έκφρασης τα οποία είναι απαραίτητα για να χαρακτηριστούν οι ομάδες και έπειτα να είναι σε θέση να πραγματοποιήσει την ζητούμενη κατηγοριοποίηση οποιοδήποτε άλλων δειγμάτων ή αντικειμένων.

Ο αλγόριθμος Support Vector Machine δημιουργεί ένα υπερεπίπεδο (hyperplane) το οποίο έχει αυτήν την ονομασία λόγω του ότι βρίσκεται στον πολυδιάστατο χώρο και εξυπηρετεί στην παλινδρόμηση των δεδομένων καθώς και στην ταξινόμηση αυτών σε κλάσεις. Η διαδικασία παραγωγής του hyperplane πραγματοποιείται επαναληπτικά με σκοπό την ελαχιστοποίηση σφάλματος. Με την ταξινόμηση τα δεδομένα που αναπαρίστανται με σημεία στον χώρο, διαχωρίζονται από το hyperplane το οποίο αφήνει ένα περιθώριο σε σχέση με τα δεδομένα.

Support Vectors

Τα support vectors: διανύσματα υποστήριξης, αποτελούν τα σημεία δεδομένων σε ένα καρτεσιανό σύστημα αξόνων. Τα οποία είναι στην μικρότερη απόσταση από το hyperplane. Με τον υπολογισμό του περιθωρίου, καθορίζονται και τα σημεία αυτά που περιγράφουν τα δεδομένα.

Hyperplane

Ένα hyperplane μπορεί να χαρακτηριστεί και ως επίπεδο απόφασης καθώς διαχωρίζει και με αυτόν τον τρόπο ομαδοποιεί ένα σετ δεδομένων.

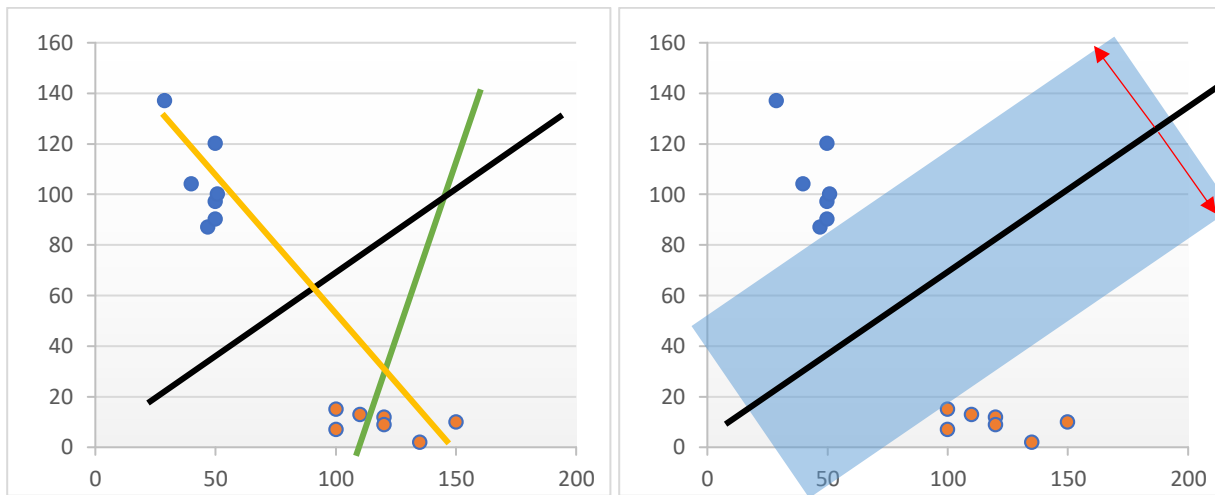
Margin

Το περιθώριο αποτελεί ένα κενό μεταξύ των δεδομένων που αναπαρίστανται ως σημεία. Το περιθώριο αυτό υπολογίζεται ως η κάθετη απόσταση από την γραμμή προς τα support vectors ή τα πλησιέστερα σημεία. Ένα το περιθώριο είναι μεγαλύτερο μεταξύ των κλάσεων θεωρείται καλό ενώ αντίστοιχα ένα μικρό περιθώριο θεωρείται ως κακό περιθώριο για το αποτέλεσμα της ταξινόμησης.

Σκοπός του αλγορίθμου όπως έχει γίνει ξεκάθαρο είναι να διαχωρίσει το δοσμένο σετ δεδομένων με τον βέλτιστο δυνατό τρόπο. Η απόσταση μεταξύ των δυο πλησιέστερων σημείων αποτελεί το λεγόμενο περιθώριο. Ο στόχος της μεθόδου αυτής είναι να επιλεγεί ένα hyperplane με το μέγιστο δυνατό περιθώριο μεταξύ των support vectors στο σετ δεδομένων που δόθηκε ως είσοδος στον αλγόριθμο. Ο SMV λοιπόν εκτελεί συγκεκριμένα βήματα για να βρει με επιτυχία ένα hyperplane όπως περιεγράφηκε.

Το πρώτο βήμα περιλαμβάνει την δημιουργία ενός hyperplane που διαχωρίζει τις κλάσεις με τον καλύτερο δυνατό τρόπο. Στο αριστερό διάγραμμα υπάρχουν τρία hyperplanes, ένα μαύρο, ένα πράσινο και ένα κίτρινο. Τα δύο τελευταία δηλαδή το πράσινο και το κίτρινο έχουν υψηλό ποσοστό λάθους, αλλά το μαύρο διαχωρίζει τα δεδομένα σωστά σε δύο κλάσεις.

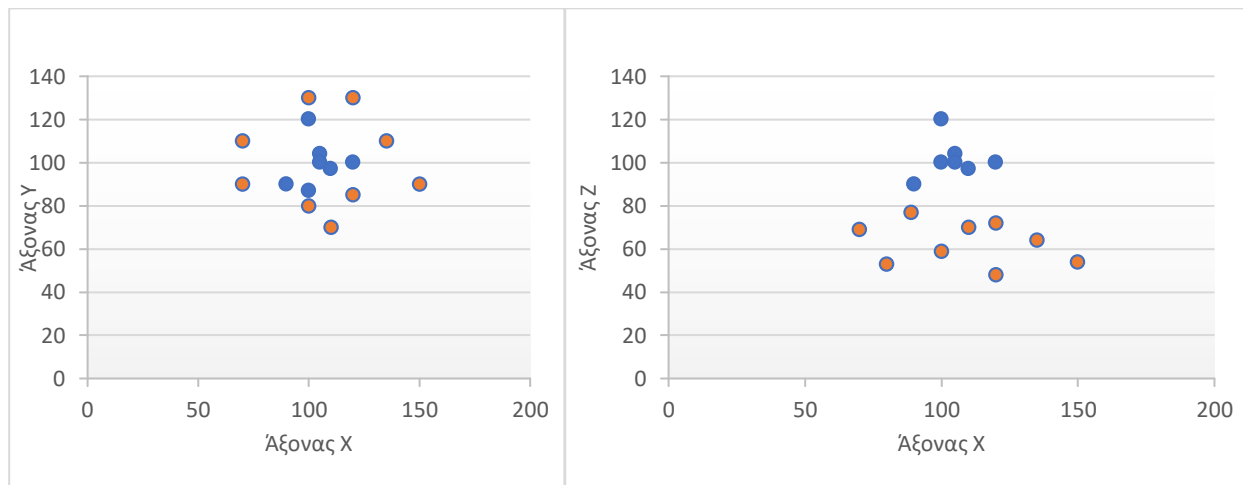
Στο δεύτερο βήμα όπως φαίνεται στο διάγραμμα δεξιά, επιλέγεται το σωστό hyperplane που είναι το μαύρο καθώς έχει την μέγιστη κάθετη απόσταση από τα πιο κοντινά σε αυτό δεδομένα. Η μέγιστη αυτή απόσταση αναπαρίσταται από το γαλάζιο διαφανές πλαίσιο με πλάτος ίσου μήκους του περιθωρίου



Μη γραμμικά και μη διαχωριζόμενα planes

Μερικά προβλήματα είναι αδύνατον να λυθούν εφαρμόζοντας την γραμμική προσέγγιση hyperplane, όπως έγινε παραπάνω. Ένα τέτοιο παράδειγμα είναι σαν αυτό που βρίσκεται στο διάγραμμα αριστερά.

Σε μια τέτοια περίπτωση, ο αλγόριθμος Support Vector Machine χρησιμοποιεί τέχνασμα πυρήνα για να μετασχηματίσει και να αύξηση τις διαστάσεις του χώρου.



Αυτό το τέχνασμα πυρήνα πραγματοποιείται, σχεδιάζοντας τα σημεία δεδομένων στους άξονες X και Z. Με το άξονα Z να ισούται με το τετράγωνο άθροισμα των σημείων X ή των Ψ. Έπειτα από την εφαρμογή του τεχνάσματος ο διαχωρισμός των στοιχείων σε κλάσεις γίνεται όπως προηγουμένως που το σετ δεδομένων μας ήταν γραμμικό.

$$Z = x^2 = \psi^2$$

Ο αλγόριθμος Support Vector Machine αποτελεί έναν ταξινομητή που προσφέρει υψηλό ποσοστό ακρίβειας και ταχύτητας πρόβλεψης, συγκριτικά με τον Naïve Bayes. Επιπρόσθετα γίνεται λιγότερο χρήση μνήμης λόγω του ότι χρησιμοποιεί υποσύνολα των σημείων εκπαίδευσης στη φάση λήψης απόφασης. Ο αλγόριθμος λειτουργεί καλά με ένα σαφές περιθώριο διαχωρισμού μεταξύ των δεδομένων και σε πολυδιάστατο χώρο. Παρ' όλα αυτά σε περιπτώσεις μεγάλων συνόλων δεδομένων ο αλγόριθμος δεν προτείνεται, καθώς σπαταλά πολύ χρόνο στην φάση της εκπαίδευσης. Επιπλέον η λειτουργία τους δεν είναι τόσο καλή όταν συναντά overlapping κλάσεις.

4.15 Naïve Bayes

Ένας ταξινομητής Naïve Bayes είναι ένα πικάντικο πρότυπο μηχανικής μάθησης. Η ουσία της ταξινόμησης που πραγματοποιεί ο Naïve Bayes βασίζεται στο θεώρημα Bayes

Οι τύποι ταξινόμησης με χρήση του Naïve Bayes είναι τρεις:

- **Multinomial Naïve Bayes**
- **Bernouli Naïve Bayes**
- **Gaussian Naïve Bayes**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Έστω δύο γεγονότα A και B του ίδιου δειγματικού χώρου. Η πιθανότητα να συμβεί το γεγονός A γνωρίζοντας ότι έχει συμβεί το γεγονός B ονομάζεται δεσμευμένη πιθανότητα και συμβολίζεται με: $P(A|B)$.

Θεώρημα Bayes

Multinomial Naïve Bayes (πολυονιμικός Naïve Bayes)

Αυτή η μέθοδος του Naïve Bayes είναι ιδιαίτερα δημοφιλής σε προβλήματα ταξινόμησης εγγράφων. Δηλαδή στην ανάθεση ενός εγγράφου με βάση το περιεχόμενο του σε κατηγορίες όπως η μουσική, ο αθλητισμός, η πολιτική κλπ. Τα χαρακτηριστικά που χρησιμοποιούνται σε αυτήν την περίπτωση αφορούν την συχνότητα των λέξεων που υπάρχουν στο έγγραφο.

Bernouli Naïve Bayes

Σε σχέση με τον Naïve Bayes, η μέθοδος Μπερνούλι Naïve Bayes, χρησιμοποιεί Boolean μεταβλητές. Οι παράμετροι που λαμβάνουν μέρος στην συγκεκριμένη μέθοδο με σκοπό την πρόβλεψη της μεταβλητής κλάσης, είναι μόνο “Ναι” και “Όχι”.

Gaussian Naïve Bayes

Όταν οι δείκτες πρόβλεψης λαμβάνουν μια συνεχή τιμή και δεν είναι διακριτοί, υποθέτουμε ότι αυτές οι τιμές λαμβάνονται από μια Γκαουσιανή κατανομή (καμπάνα)

Ο αλγόριθμος του Naïve Bayes εκτελείται με τρία βήματα.

1. Αρχικά γίνεται μετατροπή του συνόλου δεδομένων εισόδου σε πίνακα συχνοτήτων. Ένας πίνακας που περιγράφει πόσο συχνά εμφανίζεται ένα συγκεκριμένο δεδομένο (μια συγκεκριμένη τιμή).
2. Στο δεύτερο βήμα δημιουργείται ένας πίνακας πιθανοτήτων, αφού προηγηθεί η εύρεση των πιθανοτήτων μέσω του πίνακα συχνοτήτων.

3. Στο τελευταίο βήμα χρησιμοποιείται η εξίσωση Naïve Bayes για τον υπολογισμό της δεσμευμένης πιθανότητας της κάθε κλάσης. Η κλάση με την υψηλότερη δεσμευμένη πιθανότητα είναι και το αποτέλεσμα της πρόβλεψης.

Naïve Bayes vs Support Vector Machine

Στον παρακάτω πίνακα γίνεται η σύγκριση των δύο μεθόδων ταξινόμησης.

	Naïve Bayes	Support Vector Machine
Ακρίβεια	Μικρότερη	Μεγαλύτερη
Χρήση Μνήμης	Μεγαλύτερη	Μικρότερη
Πολυδιάστατος χώρος	Όχι	Ναι
Επικαλυπτόμενες κλάσεις	Ναι	Όχι
Διάρκεια εκπαίδευσης	Χαμηλή	Υψηλή
Αλληλεξαρτόμενα δεδομένα	Όχι	Ναι
Βέλτιστοι τύποι δεδομένων	Κατηγορηματικοί τύποι	Αριθμητικοί τύποι
Ταχύτητα	Μικρή	Μεγάλη
Μεγάλα σετ δεδομένων	Καλός	Κακός

5.1 Βιολογικές Βάσεις Δεδομένων

Οι αναλύσεις που πραγματοποιούνται στον επιστημονικό κλάδο της βιοπληροφορικής έχουν ως βάση ένα σύνολο βιολογικών δεδομένων. Τα δεδομένα αυτά αντλούνται από τις βιολογικές βάσεις δεδομένων έτσι ώστε οι ερευνητές να επιτύχουν την μελέτη τους. Επομένως οι βιολογικές βάσεις δεδομένων αποτελούν βασικό και αναγκαίο κομμάτι της σύγχρονης βιοπληροφορικής. Ακόμα και για όσους παράγουν τα πρωτογενή δεδομένα και δεν αντλούν πληροφορίες από τις βάσεις δεδομένων, χρειάζονται τις βάσεις δεδομένων και θεωρούνται μάλιστα και σημαντικές για την δουλειά τους καθώς θα πρέπει να καταθέσουν τα δεδομένα τους σε αυτές.

Όταν δημιουργήθηκαν οι πρώτες βάσεις δεδομένων ο όγκος της πληροφορίας ήταν μικρός, με αποτέλεσμα ο χρόνος και η προσπάθεια για συντήρηση και η ανανέωση των βάσεων ήταν πολύ μικρότερος σε σύγκριση με την τωρινή κατάσταση και ως αποτέλεσμα απαιτούσαν μικρό κόστος τόσο σε υποδομές όσο και σε ανθρώπινο δυναμικό. Η ραγδαία εξέλιξη της τεχνολογίας όμως άνοιξε τον δρόμο στις πειραματικές εργασίες και έκανε την διεκπεραίωση πολλών από αυτών πραγματικότητα με χρήση των νέων τεχνολογικών μέσων και μεθόδων που εμφανίστηκαν, που σε συνδυασμό με τον διαρκή προσδιορισμό των γωνιδιωμάτων διαφόρων οργανισμών, αύξησε σημαντικά τον όγκο της πληροφορίας σε όλα τα επίπεδα και ιδιαίτερα στο επίπεδο της αλληλουχίας. Αυτό είχε ως αποτέλεσμα οι βάσεις να περιέχουν τεράστιο όγκο δεδομένων με αποτέλεσμα η ανανέωση τους αποτελεί καθημερινή δραστηριότητα πλέον. Επίσης η συντήρηση των βάσεων απαιτεί μεγάλο πλήθος εξειδικευμένων επιστημόνων που θα ασχολούνται ενδελεχώς και καθημερινά με την επισήμανση ενδεχομένων σφαλμάτων καθώς και με το σχολιασμό των νεοεισερχομένων δεδομένων.

Δυο χαρακτηριστικά παραδείγματα βάσεων αποτελούν η UiprotKB/SWISS-PROT, η κύρια βάση πρωτεϊνικών αλληλουχιών που περιέχει 547.599 αλληλουχίες (Rel. 2015_02 – Φεβρουάριος 2015) και η EMBL Nucleotide Sequence Database που περιέχει νουκλεοτιδικές αλληλουχίες και έχει 510.014.239 εγγραφές (Rel. 122 - Νοέμβριος 2014). Κάθε ερευνητής μπορεί να έχει πρόσβαση στις βάσεις αυτές μέσω της χρήσης διαδικτύου. Αρκεί η επίσκεψη στην ιστοσελίδα της βάσης, η αναζήτηση των δεδομένων ενδιαφέροντος και στη συνέχεια η αποθήκευσή τους στον υπολογιστή. Παράλληλα έχουν δημιουργηθεί βάσεις στις οποίες η πληροφορία στο επίπεδο της αλληλουχίας και της δομής είναι ταξινομημένη με τέτοιο τρόπο ώστε η πληροφορία να είναι οργανωμένη για την εξαγωγή συμπερασμάτων ως προς την βιολογική τους σημασία.

Οι βιολογικές βάσεις δεδομένων, γενικά, μπορούν να διακριθούν σε 2 μεγάλες κατηγορίες, με επιμέρους κατηγοριοποιήσεις, όπως περιγράφονται παρακάτω.

Καταρχάς, υπάρχουν οι πρωτογενείς βάσεις δεδομένων, οι οποίες περιέχουν τα πρωτογενή πειραματικά δεδομένα και οι οποίες αναλύονται κυρίως σε:

- Βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών
- Βάσεις δεδομένων αμινοξικών αλληλουχιών πρωτεϊνών
- Βάσεις δεδομένων τρισδιάστατων βιολογιών δομών
- Βάσεις δεδομένων γονιδιακής έκφρασης
- Βάσεις δεδομένων γενετικής ποικιλομορφίας
- Βάσεις δεδομένων βιβλιογραφίας

Επιπλέον των πρωτογενών βάσεων δεδομένων υπάρχουν οι δευτερογενείς βάσεις δεδομένων, στις οποίες υπάρχουν κυρίως ταξινομήσεις των πρωτογενών δεδομένων και είναι χρήσιμες για αναλυτικούς σκοπούς οι οποίες διακρίνονται σε:

- Βάσεις δεδομένων οικογενειών (κυρίως πρωτεϊνών)
- Εξειδικευμένες Βάσεις δεδομένων (όλες οι άλλες κατηγορίες)

5.2 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΝΟΥΚΛΕΟΤΙΔΙΚΩΝ ΑΛΛΗΛΟΥΧΙΩΝ

Οι τρεις μεγαλύτερες Βάσεις δεδομένων νουκλεοτιδικών αλληλουχιών που είναι ελεύθερα διαθέσιμες στην ακαδημαϊκή κοινότητα είναι οι GENBANK (NCBI), DNA Data Bank of Japan (DDBJ) και EMBL Nucleotide Sequence Database (EBI). Οι τρεις αυτές βάσεις δεδομένων βρίσκονται σε συνεργασία, δηλαδή, έχουν θέσει κοινούς κανόνες ταξινόμησης και σχολιασμού δεδομένων με σκοπό την ανταλλαγή των καθημερινών εγγραφών τους.

1. **Genbank:** είναι μια βάση νουκλεοτιδικών αλληλουχιών (Benson et al., 2014), διατίθεται ελεύθερα στην επιστημονική κοινότητα και βρίσκεται και υπό την αιγίδα του Εθνικού Ινστιτούτου Υγείας των Η.Π.Α. (National Institutes of Health). Αποτελεί μια περιγραφική συλλογή όλων των δημόσια διαθέσιμων ακολουθιών DNA. Η GenkBank αποτελεί μέρος των της International Nucleotide Sequence Database Collaboration (Διεθνή Βάση Δεδομένων Νουκλεοτιδικών Αλληλουχιών), η οποία περιλαμβάνει την DNA DataBank of Japan (DDBJ) που θα περιγράψει παρακάτω, το European Nucleotide Archive (ENA) και την GenBank στο NCBI. Οι τρεις αναφερόμενες βάσεις δεδομένων ανταλλάσσουν δεδομένα καθημερινά. [60]

Για να επιτύχει κανείς την πρόσβαση στην βάση και να ανακτήσει δεδομένα υπάρχει μια ποικιλία τρόπων:

- Αναζήτηση στην βάση για αναγνωριστικά αλληλουχιών και σχολιασμούς με το Enterz Nucleotide.
- Αναζήτηση και ευθυγραμμίσει των ακολουθιών της βάσης με χρήση μια σειράς ερωτήσεων χρησιμοποιώντας το εργαλείο BLAST (Basic Local Alignment Search Tool)
- Αναζήτηση, συνδέσμου και λήψη αλληλουχιών προγραμματιστικά χρησιμοποιώντας τα ηλεκτρονικά βοηθήματα του NCBI
- Οι μορφές ASN.1 και flatfile διατίθενται στον ανώνυμο server του NCBI: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> και <ftp://ftp.ncbi.nlm.nih.gov/genbank>. [60]

2. **EMBL-Bank:** Η EMBL Nucleotide Sequence Database αποτελεί τη μεγαλύτερη βάση νουκλεοτιδικών αλληλουχιών στην Ευρώπη, βρίσκεται υπό την αιγίδα του Ευρωπαϊκού Εργαστηρίου Μοριακής Βιολογίας (EMBL) ενώ εδρεύει και συντηρείται από το Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής (EBI) στο Cambridge, UK. Όπως και η GenBank αντίστοιχα και η EMBL-Bank διαθέτει δεδομένα που ανταλλάσσονται μεταξύ των συνεργατικών βάσεων δεδομένων σε καθημερινή βάση. Οι κύριοι συντελεστές της βάσης δεδομένων EMBL-Bank είναι μεμονωμένοι συγγραφείς και ομάδες που έχουν αναλάβει πρότζεκτ γωνιδιωμάτων. Το WEBIN είναι το προτιμώμενο σύστημα υποβολής στο διαδίκτυο για μεμονωμένους υποψήφιους, ενώ οι αυτόματες διαδικασίες επιτρέπουν την ενσωμάτωση δεδομένων αλληλουχίας από μεγαλύτερης κλίμακας κέντρα προσδιορισμού αλληλουχίας γονιδιώματος και από το European Patent Office (EPO). Οι εκδόσεις της βάσης δεδομένων παράγονται κάθε νέο τρίμηνο. Οι υπηρεσίες δικτύου επιτρέπουν την ελεύθερη πρόσβαση στην

πλέον σύγχρονη συλλογή δεδομένων μέσω πρόσβασης στο διαδίκτυο και WWW interface. Το Sequence Retrieval System (SRS) της βάσης δεδομένων EMBL-Bank είναι το σύστημα ανάκτηση ακολουθιών. Είναι ένα πρόγραμμα περιήγησης του δικτύου για βάσεις δεδομένων στο πεδίο της μοριακής βιολογίας, που ενσωματώνει και συνδέει τις βασικές βάσεις δεδομένων νουκλεοτιδίων και πρωτεϊνών καθώς και πολλές εξειδικευμένες βάσεις δεδομένων. Τα εργαλεία αναζήτησης που χρησιμοποιούνται μέσα στην βάση για την εύρεση δεδομένων είναι τα BLITZ, FASTA, BLAST κ.ά. [61]

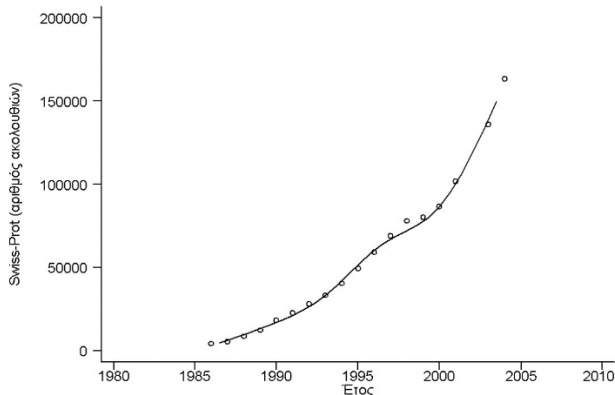
3. **DDBJ:** Η DNA Databank of Japan (DDBJ - <http://www.ddbj.nig.ac.jp/>) είναι η μοναδική διεθνώς αναγνωρισμένη βάση νουκλεοτιδικών αλληλουχιών στην Ιαπωνία. Ιδρύθηκε το 1986 στο Εθνικό Ινστιτούτο Γενετικής (NIG) και βρίσκεται υπό την αιγίδα του Υπουργείου Παιδείας, Επιστημών και Αθλητισμού της Ιαπωνίας. Το κέντρο παρέχει επίσης υπηρεσίες στο JGA (Japanese Genotype-phenotype Archive) με το Bioscience Database Center να αποτελεί την πηγή συλλογής δεδομένων παρμένες από μελέτες σε ανθρώπους, που υποβάλλονται από Ιάπωνες ερευνητές. [62]

Οι κυριότερες βάσεις δεδομένων με αλληλουχίες DNA στον διεθνή χώρο, η Genbank στις ΗΠΑ, η DDBJ στην Ιαπωνία και η EMBL Data Bank στην Ευρώπη, συνεργάζονται μέσω του International Nucleotide Sequence Collaboration, μιας οργάνωσης που οι ίδιοι δημιούργησαν, και έτσι μια αλληλουχία αφού καταχωρηθεί σε μια από αυτές μέσα από μια διαδικασία έγκρισης, καταχωρείται και στις άλλες. Πρακτική συνέπεια αυτού του γεγονότος, είναι ότι εκτός ελαχίστων εξαιρέσεων, οι 3 βάσεις περιέχουν τις ίδιες καταχωρήσεις, άρα δεν έχει και πολύ μεγάλη σημασία σε ποια από τις 3 βάσεις δεδομένων θα απευθυνθεί κανείς για μια έρευνα.

5.3 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΠΡΩΤΕΪΝΙΚΩΝ ΑΛΛΗΛΟΥΧΙΩΝ

Οι βάσεις δεδομένων πρωτεϊνικών αλληλουχιών, αποτελούν το δεύτερο μεγαλύτερο σε όγκο τμήμα του συνόλου των βιολογικών βάσεων δεδομένων.

Η **UniprotKB**, αποτελεί την κύρια, σε παγκόσμιο επίπεδο βάση δεδομένων πρωτεϊνικών αλληλουχιών (UniProt, 2014). Αποτελείται από δύο υποσύνολα, την Uniprot/SwissProt η οποία περιέχει τις καλά σχολιασμένες πρωτεϊνικές αλληλουχίες και την Uniprot/TrEMBL η οποία περιέχει τις πρωτεϊνικές αλληλουχίες που έχουν προκύψει από αυτόματη (ηλεκτρονική) μετάφραση γονιδιωμάτων αλληλουχιών



Η εκθετική αύξηση των αμινοξέων πρωτεϊνών οι οποίες είναι κατατεθειμένες στην Swiss-Prot, από το 1986 έως το τέλος του 2004

5.1.4 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΤΡΙΣΔΙΑΣΤΑΤΩΝ ΒΙΟΛΟΓΙΚΩΝ ΔΟΜΩΝ

Οι βάσεις δεδομένων τέτοιου τύπου περιέχουν όλα εκείνα τα δεδομένα που σχετίζονται με την τρισδιάστατη δομή βιολογικών μακρομορίων. Οι τρισδιάστατες δομές αποτελούν το τελικό στάδιο μιας επίπονης διαδικασίας η οποία μετά τη χρήση μοριακών τεχνικών, οδηγεί τελικά στην υπολογιστική επίλυση της δομής μέσω της διαδικασίας κρυσταλλογραφίας ακτίνων Χ, ή, σε πιο σπάνιες περιπτώσεις με φασματοσκοπία NMR. Το πιο σημαντικό κομμάτι, παρουσιάζεται στην δομές πρωτεϊνών, καθώς οι πρωτεΐνες είναι τα μοκρομόρια των οποίων η μεγάλη ποικιλομορφία της δομής συνδέεται άμεσα με την βιολογική δράση. Η μοναδική βάση παγκοσμίως είναι η PDB.

Η **Protein Data Bank** είναι η μοναδική βάση δεδομένων η οποία περιέχει τρισδιάστατες δομές βιολογικών μακρομορίων. Ιδρύθηκε το 1971 στα εργαστήρια Brookhaven National Laboratories (BNL) των ΗΠΑ.

5.4 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΓΕΝΕΤΙΚΗΣ ΠΟΙΚΙΛΟΜΟΡΦΙΑΣ

Ανάμεσα στις βάσεις γενετικής ποικιλομορφίας και σ' αυτές των αλληλουχιών DNA υπάρχει μια σύνδεση, παρ' όλα αυτά δεν αποτελούν παράγωγα τους, αλλά περισσότερο θα τις χαρακτηρίζε κανείς ως ανεξάρτητες οντότητες. Αυτό συμβαίνει λόγω του ότι σε μια δεδομένη θέση ενός γονιδιώματος ενός είδους, τα διαφορετικά άτομα είναι δυνατόν να έχουν διαφορετική γενετική πληροφορία (πχ. Οι άνθρωποι). Η βάση η οποία καταγράφει του πολυμορφισμούς και τις συχνότητες τους στους διάφορους πληθυσμούς είναι η dbSNP, ενώ η βάση στην οποία καταγράφονται πρωτογενώς τουλάχιστον οι αλληλοσυσχετίσεις των πολυμορφισμών αυτών είναι η HarMap.

1. Η **dbSNP** είναι η δημόσια βάση για τους νουκλεοτιδικούς πολυμορφισμούς (Sherry et al., 2001). Εκτός από νουκλεοτιδικούς πολυμορφισμούς (single nucleotide polymorphisms - SNPs), περιέχει και δεδομένα για πολυμορφικές θέσεις που αφορούν απαλοιφές ή εισαγωγές βάσεων (deletion insertion polymorphisms -DIPs), καθώς και για ένθετα μεταθετά στοιχεία και μικροδορυφορικές επαναλήψεις (short tandem repeats - STRs).
2. Ο International **HarMap** Project είναι το αποτέλεσμα μιας διεθνούς συνεργασίας σε μια προσπάθεια να εντοπισθούν και να καταγραφούν οι γενετικές διαφορές αλλά και οι ομοιότητες των ανθρώπινων πληθυσμών (HarMap, 2003). Ο σκοπός του προγράμματος είναι να συγκρίνει τις γενετικές αλληλουχίες διαφορετικών ατόμων (από διαφορετικούς πληθυσμούς) και να εντοπίσει με αυτόν τον τρόπο χρωμοσωμικές περιοχές στις οποίες οι γενετικές παραλλαγές (συνήθως, νουκλεοτιδικοί πολυμορφισμοί), κληρονομούνται μαζί.

5.5 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΒΙΒΛΙΟΓΡΑΦΙΑΣ

Αν και αυτές οι βάσεις δεν είναι ακριβώς “βιολογικές βάσεις δεδομένων”, συγκαταλέγονται σε αυτές. Ουσιαστικά περιλαμβάνουν καταχωρήσεις στοιχείων επιστημονικών δημοσιεύσεων. Η κυριότερη αυτών είναι η **PubMed** η οποία στεγάζεται στο NCBI και αριθμεί περισσότερες από 24 εκατομμύρια καταχωρήσεις επιστημονικών άρθρων από την βιοϊατρική βιβλιογραφία.

5.6 ΔΕΥΕΡΟΓΕΝΕΙΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Σε αυτήν την κατηγορία περιλαμβάνονται εκείνες οι βάσεις δεδομένων που περιέχουν διαφόρων ειδών ταξινομήσεις των πρωτογενών δεδομένων και χρησιμοποιούνται για σκοπούς ανάλυσης. Τα δύο “είδη” είναι οι βάσεις οικογενειών και οι εξειδικευμένες.

5.7 ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ ΟΙΚΟΓΕΝΕΙΩΝ

Οι πρωτεΐνες αποτελούνται από μια ή περισσότερες διακριτές λειτουργικές περιοχές, οι οποίες πολλές φορές είναι και δομικά αυτοτελείς. Οι βάσεις οικογενειών επιτελούν τον σημαντικό ρόλο ταξινόμησης αυτών αμινοξικών αλληλουχιών πρωτεϊνών σε οικογένειες. Οι κυριότερες είναι οι:

1. **PROSITE**
2. **PFAM**
3. **CATH**
4. **SCOP**

5.8 ΕΞΕΙΔΙΚΕΥΜΕΝΕΣ ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

Οι βάσεις αυτής της κατηγορίας συνήθως αφορούν τις αμινοξικές αλληλουχίες πρωτεϊνών και τις περισσότερες φορές, συντηρούνται από μικρές ή μεσαίου μεγέθους ερευνητικές ομάδες.

Οι κυριότερες είναι:

1. **TCDB**
2. **OMPdb**
3. **CAZy**
4. **MEROPS**
5. **neXtProt**
6. **KinG**
7. **EzCatDB**
8. **MACiE**
9. **ESTHER**
10. **ConoServer**
11. **CyBase**
12. **GPCRdb**
13. **IUPHAR/BPS Guide to PHARMACOLOGY**
14. **Kinase.com**
15. **Structure-Function Linkage Database**
16. **Histone Database**

1. **TCDB**

Transporter Classification Database: Η βάση δεδομένων περιέχει ένα πλήρες και λεπτομερή σύστημα ταξινόμησης πρωτεϊνών μεταφοράς μεμβράνης, γνωστό και ως Transporter Classification system -TC system (Σύστημα ταξινόμησης μεταφορέων), το σύστημα ταξινόμησης είναι εγκεκριμένο από το πανεπιστήμιο χημείας Queen Mary του Λονδίνου. Η βάση δεδομένων του TC system περιέχει 18500 πρωτεϊνικές αλληλουχίες, οι

οποίες ταξινομούνται σε 1384 κλάσεις με βάσει το σύστημα ταξινόμησης. Η επίσημη ιστοσελίδα της βάσης δεδομένων είναι η: <http://www.tcdb.org>

2. OMPdb

A data base of β -barrel outer membrane proteins: Οι πληροφορίες που περιέχει η OMPdb αποτελούνται από αλληλουχίες δεδομένων, όπως περιγραφή των δομικών χαρακτηριστικών, βιβλιογραφικές αναφορές, σύνδεση με άλλες δημόσιες βάσεις δεδομένων. Επίσης παρέχει πληροφορίες σχετικά με την ύπαρξη τρισδιάστατης δομής μιας δεδομένης πρωτεΐνης. [45]

3. CAZy

Carbohydrate-Active enZYmes Database: Η βάση δεδομένων CAZy είναι μια βάση δεδομένων η οποία αποτελείται από ενεργά ένζυμα και υδατάνθρακες. Η βάση περιγράφει οικογένειες συσχετιζόμενων δομικά καταλυτικών και συστοιχιών δέσμευσης υδατανθράκων, (ή λειτουργικών περιοχών) ενζύμων που αποικοδομούν, τροποποιούν ή δημιουργούν γλυκοσιδικούς δεσμούς. Τα δεδομένα της βάσης δεδομένων είναι προσβάσιμη είτε με περιήγηση σε οικογένειες που βασίζονται σε ακολουθίες είτε με περιήγηση στο περιεχόμενο των γωνιδιωμάτων σε ενεργά ένζυμα υδατάνθρακα. Τα νέα γονιδιώματα προστίθενται τακτικά λίγο μετά την εμφάνισή τους στις καθημερινές κυκλοφορίες της GenBank. Νέες οικογένειες δημοσιευμένων στοιχείων για τη δραστηριότητα τουλάχιστον ενός μέλους της οικογένειας και όλες οι οικογένειες ενημερώνονται τακτικά, τόσο στο περιεχόμενο όσο και στην περιγραφή. Η ιστοσελίδα της βάσης δεδομένων είναι : <http://www.cazy.org/> [46]

4. MEROPS

The Peptidase Database: Η βάση δεδομένων MEROPS είναι μια πηγή πληροφόρησης για τις πεπτιδάδες (που ονομάζονται επίσης πρωτεϊνάσες και πρωτεολυτικά ένζυμα) και οι πρωτεΐνες που τις περιορίζουν. Η περιγραφή μιας δεδομένης peptidase βρίσκεται σε μια σελίδα περίληψης και μπορεί να επιτευχθεί με τη χρήση ενός δείκτη κάτω από το όνομα της, με το αναγνωριστικό MEROPS ή τον οργανισμό προέλευσης. Η περίληψη περιγράφει την ταξινόμηση και την ορολογία της peptidase και προσφέρει links σε συμπληρωματικές σελίδες που δείχνουν αναγνωριστικά αλληλουχιών, τη δομή εάν αυτή είναι γνωστή, βιβλιογραφικές αναφορές και άλλα. Η βάση δεδομένων χρησιμοποιεί μια ιεραρχική δομή ταξινόμησης των peptidases. Κατά την ταξινόμηση κάθε peptidase αντιστοιχείται σε μια οικογένεια, με βάση σημαντικές ομοιότητες στην αλληλουχία των αμινοξέων και οικογένειες που πιστεύεται ότι είναι ομόλογες ομαδοποιούνται σε μία Φυλή. Για κάθε οικογένεια και κάθε φυλή υπάρχει και μια σελίδα περίληψης, [47]

5. neXtProt

Η neXtProt είναι μια βάση δεδομένων που μετρά 11 χρόνια ζωής και είναι σχεδιασμένη για να βοηθήσει τους ερευνητές στο να βγάλουν συμπεράσματα και να κατανοήσουν τι κάνουν οι πρωτεΐνες στο σώμα του ανθρώπου. Στο SIB, η neXtProt αναπτύχθηκε από την CALIPHO (Computer Analysis and Laboratory Investigation of Proteins of Human Origin), μια ομάδα ανθρώπων που δημιουργήθηκε από κοινού με το Πανεπιστήμιο της Γενεύης. Με επικεφαλής τον Amos Bairoch και την Lydie Lane η CALIPHO είναι μια διεπιστημονική ομάδα που στοχεύει σύμφωνα με τους ίδιους να χρησιμοποιήσει ένα εύρος μεθοδολογιών για να βοηθήσει στην αποκάλυψη της λειτουργίας των μη χαρακτηρισμένων ανθρώπινων πρωτεϊνών. Η ανάπτυξη της βάσης δεδομένων πραγματοποιείται σε συνεργασία με την

ομάδα Swiss-Prot, η οποία βρίσκεται υπό την καθοδήγηση του καθηγητή Bairoch για πάνω από δύο δεκαετίες. Η ομάδα της CALIPHO έχει θέσει και δουλεύει πάνω σε τρεις διαφορετικές προκλήσεις. Να προσθέσουν περισσότερες πληροφορίες πρωτεϊνών του ανθρώπινου σώματος στο σώμα των δεδομένων που ήδη υπάρχουν στην Swiss-Prot. Εκτός από τα ήδη διαθέσιμα δεδομένα της Swiss-Prot για τις ανθρώπινες πρωτεΐνες, προστίθενται δεδομένων που προέρχονται από μια ποικιλία προσεγγίσεων υψηλής απόδοσης, όπως η μικρό-συστοίχια, οι οθόνες αντισωμάτων, η πρωτεϊνωματική, η interatomics και η δομική γονιδιωματική, αυτές οι προσεγγίσεις αποτελούν τις αναφερόμενες προσθήκες. Όλα αυτά τα σετ δεδομένων επιλέγονται έτσι ώστε να παρέχουν μόνο δεδομένα υψηλής ποιότητας.

6. KinG

Η βάση δεδομένων KinG είναι μια περιεκτική συλλογή σερίνης/θρεονίνης/ tyrosine-specific kinases και των ομολόγων τους που ταυτοποιούνται σε ποικίλα ολοκληρωμένα γονιδιώματα χρησιμοποιώντας μεθόδους αναζήτησης αλληλουχίας και προφίλ. Η βάση που φιλοξενείται στο <http://hodgkin.mbu.iisc.ernet.in/> παρέχει τις αλληλουχίες αμινοξέων, τις ταξινομήσεις λειτουργικών περιοχών και την ταξινόμηση των γονιδιακών προϊόντων που περιέχουν περιοχές πρωτεϊνικής kinases. Ένα εργαλείο αναζήτησης που επιτρέπει την ανάκτηση πρωτεϊνικής kinases με καθορισμένες υποοικογένειες και συνδυασμούς τομέων και αποτελεί ένα από τα κύρια χαρακτηριστικά της. [48]

7. EzCatDB

Enzyme Catalytic-mechanism Database: Η βάση δεδομένων περιλαμβάνει καταλυτικούς

Μηχανισμούς ενζύμων υπο την έννοια των αλληλουχιών και τριτογενή δομές ενζύμων και πρωτεϊνόμενους καταλυτικούς μηχανισμούς, μαζί με την δομή των συνδέσεων. Η βάση δεδομένων EzCatDB ομαδοποιεί τα δεδομένα ενζύμων που υπάρχουν στην Protein Data Bank (PDB) και στην βάση δεδομένων SWISS-PROT με πανομοιότυπο τομέα συνδέσεων, Enzyme Commission – αριθμούς EC (επιτροπή ενζύμων) και καταλυτικούς μηχανισμούς. Η βάση δεδομένων EzCatDB μπορεί να αμφισβητηθεί από τον τύπο καλυτικού κατάλοιπου, το όνομα και ο τύπος του σύμπλοκου μορίου, το οποίο έρχεται σε διάδραση με ένα ένζυμο ως ένας συμπαράγοντας, ένα υπόστρωμα ή ένα προϊόν. Η βάση έχει την δυνατότητα παροχής πληροφοριών βιβλιογραφικού τύπου, άλλους κωδικούς βάσεων δεδομένων και EC αριθμούς. Επιπλέον η βάση δεδομένων παρέχει μια ιεραρχική ταξινόμηση των καταλυτικών μηχανισμών και θα πρέπει να σημειωθεί ότι αυτή η ομαδοποίηση της, ενσωματώνει τους καταλυτικούς μηχανισμούς και δομές ενεργού χώρου των ενζύμων καθώς και τις βασικές αντιδράσεις και αντιδραστικά τμήματα συνδετών μορίων. Η ιστοσελίδα της βάσης είναι η: <http://mbs.cbrc.jp/EzCatDB/> [49]

8. MACiE

Mechanism, Annotation and Classification in Enzymes: Η βάση δεδομένων MACiE είναι μια βάση δεδομένων που αποτελείται από μηχανισμούς αντίδρασης ενζύμων. Κάθε είσοδος στην βάση MACiE συνθέεται από μια συνολική αντίδραση που περιγράφει τις χημικές ενώσεις που εμπλέκονται, καθώς και το όνομα του είδους στο οποίο συμβαίνει η αντίδραση. Τα επιμέρους στάδια αντίδρασης για κάθε συνολική αντίδραση παρατίθενται με μηχανισμούς, εναλλακτικούς μηχανισμούς και εμπλεκόμενα αμινοξέα. [50]

9. ESTHER

Database of the Alpha/Beta-hydrolase fold superfamily of proteins: Η βάση δεδομένων ESTHER αποτελείται από Alpha/Beta-hydrolase fold υπεροικογένειες πρωτεϊνών. Είναι δωρεάν προς χρήση και για αυτό είναι ευρέως διαδεδομένη και χρησιμοποιούμενη, όπως αναφέρεται περιέχει πρωτεΐνες με α/β -hydrolase fold και περιέχει δεκάδες χιλιάδες επιλεγμένες χειρωνακτικά πρωτεΐνες. [51]

10. ConoServer

A database for conopeptides: Η βάση δεδομένων ConoServer εξειδικεύεται στην αλληλουχία και στην δομή κονοπεπτιδίων, τα οποία κονοπεπτίδια εκφράζονται υπό την μορφή σαρκοφάγων θαλάσσιων κωνικών σαλιγκαριών. Ένα μοναδικό χαρακτηριστικό αυτών των πεπτιδίων είναι η υψηλή ακρίβεια τους και η χημική συγγένεια τους σε σχέση με τα ιοντικά κανάλια των ανθρώπων, τους υποδοχείς και τους μεταφορείς του νευρικού συστήματος. Αυτά τα χαρακτηριστικά και κυρίως η ομοιότητα με τα ιοντικά κανάλια των ανθρώπων καθιστούν τα κονοπεπτίδια, μια ιδιαίτερα ενδιαφέρουσα πηγή για φυσιολογικές μελέτες των νευροϋποδοχέων και την πολλά υποσχόμενη φαρμακευτική αγωγή. Η ταξινόμηση των κονοπεπτιδίων τα οδηγεί στο να ομαδοποιούνται σε πλούσια σε δισουλφίδιο (κονοτοξίνες) και σε διάφορες άλλες κατηγορίες πεπτιδίων φτωχών σε δισουλφίδιο. Τα τρία συστήματα ταξινόμησης που χρησιμοποιούνται κατά την ταξινόμηση στην βάση δεδομένων ConoServer, είναι οι υπεροικογένειες γονιδίων, τα πλαίσια κυστεΐνης και οι φαρμακολογικές οικογένειες. Στην ιστοσελίδα της βάσης δεδομένων <http://www.conoserver.org/>, βρίσκονται αναλυτικά οι λεπτομέρειες για το πως λειτουργεί και τι περιλαμβάνει η βάση. [52]

11. CyBase

The Database of Cyclic Proteins: Η βάση δεδομένων CyBase αποτελείται από μια νέα κλάση πρωτεϊνών που περιέχουν κυκλική ραχοκοκαλιά στην οποία τα άκρα N και C συνδέονται με ένα συμβατικό αμιδικό δεσμό. Αυτά τα πρόσφατα χαρακτηρισμένα μόρια έχουν πλέον βρεθεί σε οργανισμούς από όλα τα βασίλεια ζωής και με δεδομένο τον τρέχοντα ρυθμό ανακάλυψης, ο αριθμός των αλληλουχιών θα μπορούσε σύντομα να αριθμεί εκατοντάδες. [53]

12. GPCRdb

Η βάση δεδομένων GPCRdb περιέχει δεδομένα, διαγράμματα και εργαλεία ιστού για συζευγμένους με την πρωτεΐνη G υποδοχείς (GPCRs). Οι χρήστες μπορούν να εξετάσουν όλες τις δομές που βρίσκονται στην βάση GPCR και όλες τις μεγάλες συλλογές υποδοχέων μεταλλάξεων. Τα διαγράμματα που εμπεριέχονται στην βάση μπορούν να σχεδιαστούν και να κατεβάσουν (downloaded) με σκοπό να απεικονίσουν υπολείμματα υποδοχέων και σχέσεις. Τα διαγράμματα αυτά έχουν τον σχηματισμού φιδιού και κουτιού έλικας, ενώ αντίστοιχα οι σχέσεις αποτελούνται από φυλογενετικά δένδρα. Οι ευθυγραμμίσεις αλληλουχιών που βασίζονται στη δομή αναφοράς λαμβάνουν υπόψη τα εξογκώματα και τις συστολές της έλικας και παρουσιάζουν στατιστικά στοιχεία για την διατήρηση των αμινοξέων και έχει αποδοθεί γενική αριθμηση καταλοίπων, για να επιτευχθεί η ισοδυναμία των υπολειμμάτων στους διάφορους υποδοχείς. [54]

13. IUPHAR/BPS Guide to PHARMACOLOGY

International Union of Basic and Clinical Pharmacology / British Pharmacological Society Guide to PHARMACOLOGY: Η βάση δεδομένων IUPHAR/BPS αποτελεί μια πηγή των σχέσεων συνδέσμου δραστηριότητας στόχου, η πλειοψηφία των σχέσεων

αυτών προέρχεται από φαρμακολογική και ιατρική βιβλιογραφία, υψηλής ποιότητας. Ο κύριος σκοπός της βάσης δεδομένων είναι η παροχή μιας φαρμακολογικής πληροφόρησης με βάση τη “one-stop shop” και ο κύριος στόχος της είναι να επιτευχθεί η παροχή μια βάσης δεδομένων που προσφέρει την δυνατότητα αναζήτησης με ποσοτικές πληροφορίες σχετικά με τα με τους στόχους για τα φάρμακα που συνταγογραφούνται και τα πειραματικά φάρμακα που δρουν επάνω τους. Οι πληροφορίες στην βάση δεδομένων παρουσιάζονται σε δύο επίπεδα: οι αρχικές σελίδες προβολής για κάθε οικογένεια στόχο, παρέχουν επισκόπηση με εξειδίκευση των βασικών ιδιοτήτων και των επιλεκτικών προσδεμάτων και των διαθέσιμων ενώσεων εργαλείων. Για τους επιλεγμένους στόχους στη βάση διατίθενται ενδελεχώς, εισαγωγικά κεφάλαια για κάθε οικογένεια μαζί με επιλεγμένες πληροφορίες σχετικά με τις φαρμακολογικές, φυσιολογικές, δομικές γενετικές και παθοφυσιολογικές ιδιότητες κάθε στόχου. Η βάση δεδομένων περιέχει συνδέσμους για πρόσθετες πληροφορίες σε άλλες βάσεις δεδομένων, συμπεριλαμβανομένων των Ensembl, UniProt, PubChem, ChEMBL και DrugBank, καθώς και επιμέρους χημικές πληροφορίες και βιβλιογραφικές παραπομπές στο PubMed. [55]

14. Kinase.com

Genomics, evolution and function of protein kinases: Η βάση δεδομένων της kinase.com διερευνά τις λειτουργίες, την εξέλιξη και την ποικιλομορφία των πρωτεϊνικών kinase, τους βασικούς ελεγκτές που είναι υπεύθυνοι για την συμπεριφορά των κύτταρων. Στην βάση δεδομένων επικεντρώνονται στο kinome, το πλήρες συμπλήρωμα των πρωτεϊνικών kinases σε οποιοδήποτε γονιδίωμα αλληλουχίας. Σε αυτήν περιλαμβάνεται η εκτεταμένη βάση δεδομένων KinBase και ποικίλα έγγραφα, υποστηρικτικό υλικό. Στην KinBase βρίσκει κανείς πληροφορίες για περισσότερα από 3.000 γονίδια πρωτεϊνικής kinases που βρίσκονται στα γονιδιώματα του ανθρώπου και σε πολλά άλλα γονιδιώματα αλληλουχίας. Μέσα στην βάση KinBase είναι εφικτό να έχει πρόσβαση κανείς στην αναζήτησή δεδομένων με πληθώρα διαφορετικών ονομάτων και προσθηκών γονιδίων ή σύμφωνα με την ταξινόμηση βάσει ακολουθίας. Η KinBase επιπρόσθετα μπορεί να αναζητηθεί από το Blast. (<http://kinase.com/blast/blast.html>) [56]

15. Structure-Function Linkage Database

Η Βάση δεδομένων SFLD είναι μια βάση που περιγράφεται από μια ιεραρχική ταξινόμηση των ενζύμων που σχετίζεται με συγκεκριμένα χαρακτηριστικά δομής αλληλουχίας με συγκεκριμένες χημικές δυνατότητες. Από τον Απρίλιο του 2019, η βάση δεδομένων είναι σε στατική μορφή και δεν ενημερώνεται. Οι τακτικοί χρήστες της βάσης SFLD ενδέχεται να παρατηρούν σημαντικές αλλαγές στη βάση δεδομένων, καθώς οι διαδραστικές λειτουργίες που υπήρχαν μέχρι τότε δεν είναι πλέον διαθέσιμες. Παρ' όλα αυτά, η ιεραρχία που χαρακτήριζε τις υπερ-οικογένειες, οι ευθυγραμμίσεις που βρίσκονται αρχειοθετημένες στην βάση, τα δίκτυα ομοιότητας αλληλουχιών και τα δίκτυα ομοιότητας αντίδρασης μπορούν να αναζητηθούν ακόμα και να ληφθούν. Αν και τα σετ αλληλουχιών της SFLD, οι ευθυγραμμίσεις και τα δίκτυα που αναφέρθηκαν δεν ενημερώνονται μπορούν να αποτελέσουν ένα χρήσιμο σημείο εκκίνησης για την εξερεύνηση και την παραγωγή υποθέσεων. [57]

16. Histone Database

Η βάση δεδομένων Histone αποτελεί μια βάση που μπορεί να χρησιμοποιηθεί για να εξερευνήσει ο χρήστης αυτής την ποικιλία των histone πρωτεϊνών και τις παραλλαγές των αλληλουχιών τους σε μια πληθώρα από οργανισμούς. Η πηγή δημιουργήθηκε για να

κατανοήσει καλύτερα πως η παραλλαγή της αλληλουχίας είναι εφικτή να επηρεάσει την λειτουργία αλλά και την δομή των χαρακτηριστικών των νουκλεοσωμάτων. Κυρίως ο χρήστης αναζητά έναν τύπο histone πρωτεΐνης και έπειτα του δίνεται η δυνατότητα να εξερευνήσει της διάφορες εκδοχές και παραλλαγές της. [58]

Παρακάτω παρουσιάζεται ο δεκάλογος της καλής βάσης δεδομένων έτσι όπως τον παρουσίασε ο καθηγητής Gert Vriend

1. Longevity - The one rule to rule them all. Gert asks that unless you can maintain your database for at least 10 years, then do not start.
2. Users - All databases need users and citations. To gain and keep users, you need to provide query and browsing interfaces as well as someone who answers emails.
3. Befriend Nucleic Acids Research and DATABASE journals - The descriptions of your database are essential to inform new users. But it is also essential to target publications to the readership.
4. Collaborate - Your collaborators may offer an exit strategy in the future.
- 4a. Be open - Nobody is going to steal your resource.
5. Give credit - There is more than 100% to go around.
6. Automate - Too much manual intervention makes for an unsustainable database leading to premature death. You need to automate roughly 90% of everything every year.
7. No new standards – Don't invent a new standard. Use what exists.
8. Keep it simple - Google is a model interface.
9. Visibility - Be at the right conferences and be recognizable. Use the same logo and present a poster.
10. Exit strategy - At some point you will retire. Start planning early to ensure your database continues.

6.1 Decision Trees

Σε πολλές περιπτώσεις, η έννοια της πληροφορίας συγχέεται λανθασμένα με αυτές της πληροφορίας και της γνώσης. Ωστόσο υπάρχει μια ακολουθία βημάτων πρωτού η πληροφορία μετατραπεί σε γνώση. Η πιο διαδεδομένη μέθοδος παραγωγής γνώσης μέσω των δεδομένων, είναι η ανάλυση αυτών και πρωταρχικά η ερμηνεία των αποτελεσμάτων. Αυτή είναι άλλωστε και η μέθοδος που χρησιμοποιείται καθημερινά από τους ανθρώπους στην λήψη αποφάσεων, επομένως βασιζόμενη σε αυτήν την μέθοδο στηρίζεται και η δημιουργία τεχνικών εργαλείων που προσομοιώνουν τον τρόπο που λαμβάνουν αποφάσεις οι άνθρωποι. Ένα τέτοιο εργαλείο είναι τα Δέντρα Απόφασης, με εφαρμογές σε μια μεγάλη πληθώρα επιστημονικών πεδίων. Ανάμεσα σε αυτά και η βιοπληροφορική, στην οποία πρόσφορο έδαφος δημιούργησαν οι τεχνολογίες μικροσυστοιχιών.

Οι τεχνολογίες μικροσυστοιχιών συνέβαλαν στην ανακάλυψη γονιδιακών εκφράσεων χιλιάδων γονιδίων ταυτοχρόνως. Τα πρότυπα που βρίσκονται “κρυμμένα” σε έναν τέτοιο όγκο δεδομένων είναι απολύτως απαραίτητα για την αποτελεσματική διάγνωση και την αντίστοιχα αποτελεσματική παρακολούθηση ασθενειών όπως ο καρκίνος, το αλτσχάιμερ και χρειάζονται μόνο ένα πολύ μικρό κομμάτι από το συνολικό γονιδίωμα για να λειτουργήσουν. Η αρχικοί τρόποι που χρησιμοποιούνταν στο επιστημονικό προσκήνιο της βιοπληροφορικής για την ανάλυση των δεδομένων ήταν βασισμένες κατά κύριο λόγο σε στατιστικές μεθόδους αλλά με την εμφάνιση των εργαλείων μηχανικής μάθησης στο επιστημονικό προσκήνιο της βιοπληροφορικής, πολλά προβλήματα του τομέα άρχισαν επιτέλους να αποδίδουν λύσεις πιο αποδοτικά και σε συντομότερο χρονικό διάστημα, κυρίως σε εργασίες ταξινόμησης των δεδομένων και στα πεδία που σχετίζονται με τη διαγνωστική φάση της ασθένειας, με τους πλήττοντες από την ασθένεια ανθρώπους, τις ομάδες ελέγχου ασθενών, τους τύπους των ασθενειών και την ανταπόκριση των φαρμάκων κατά την παρακολούθηση μέσω σύντομων χρονικών ακολουθιών γονιδιακών εκφράσεων. Η ιδιορρυθμία των παραπάνω εργασιών δεν έγκειται μονάχα στην πολυδιάστατη μορφή των δεδομένων τους αλλά και στην αξιολόγηση της ταξινόμησης των αποτελεσμάτων αυτών. Με άλλα λόγια σημαντικό ρόλο δεν παίζουν μόνο η ακρίβεια των ταξινομημένων μοντέλων, αλλά και η βιολογική τους σχετικότητα. [22] Τέτοια μοντέλα μπορούν να αποκαλύψουν βαθύτερες διαδικασίες, γονιδιακής αλληλεπίδρασης και γονιδιακών δεικτών. Για παράδειγμα τα **δέντρα απόφασης** μπορούν να παρέχουν πληροφορία σχετική με την γονιδιακή αλληλεπίδραση μέσω σταδιακής διάσπασης του συνόλου δεδομένων, κάθε μία από της διασπάσεις αποκαλύπτει ένα γονίδιο και η ιεραρχική δομή που ακολουθούν δείχνει την φύση της αλληλεπίδρασης τους.

Ο τρόπος με τον οποίο εκφράζονται οι μικροσυστοιχίες γωνιδιωμάτων προσφέρει την δυνατότητα ελέγχου και μελέτης των προφίλ των γονιδιακών εκφράσεων. Οι μικροσυστοιχίες αυτές αποτελούν ένα τεράστιο όγκο δεδομένων ο οποίος περιλαμβάνει εκατοντάδες ασθενείς και δεκάδες χιλιάδες γονίδια. Οι παραπάνω μικροσυστοιχίες μπορούν να φανούν χρήσιμες κατά την διάγνωση αλλά και τον έλεγχο – παρακολούθηση της ασθένειας που πλήττεται το αντικείμενο μελέτης. Ο όρος έλεγχος αναφέρεται στην παρακολούθηση του οργανισμού του ασθενή στην συνταγογραφούμενη φαρμακευτική αγωγή που του παρέχεται. Για να πραγματοποιηθούν όλες οι παραπάνω ενέργειες είναι καίριας σημασίας η αναγνώριση των προφίλ παρόμοιων γονιδιακών εκφράσεων, γεγονός που οδηγεί σε ομαδοποίηση των ασθενών-αντικειμένων μελέτης, σε αρρώστους/υγιείς, ή σε διαχωρισμό των ειδών καρκίνου κλπ. Τέτοιου είδους μεθόδων έχουν γίνει πρόσφατα δημοφιλής και χρειάζεται μεγάλη προσπάθεια για να επιτευχθεί η εύρεση της σωστής ανάλυσης των δεδομένων. Αυτός είναι και ο λόγος που η πλειοψηφία της επιστημονικής κοινότητας ψάχνει μεθόδους εφαρμογής μηχανικής μάθησης οι οποίες να μπορούν να αποδίδουν καλά πάνω σε συγκεκριμένα δεδομένα.

Η επιλογή των μεθόδων ταξινόμησης δεν είναι συγκεκριμένη και απόλυτη καθώς υπάρχει μια ποικιλία διαφορετικών αλγορίθμων ταξινόμησης που ταιριάζουν και αποδίδουν καλύτερα ο καθένας σε διαφορετικού τύπου προβλήματα. Όπως αναφέρθηκε και στην επεξεργασία και ανάλυση των βιολογικών συστημάτων ο αλγόριθμος Support Vector Machine (SVM) αποτελεί την πιο συνηθισμένη μέθοδο για να επιτευχθεί η μέγιστη δυνατή ταξινόμηση σε περιβάλλοντα δεδομένων πολλών διαστάσεων, με ή χωρίς δεδομένα που λείπουν. [19-21] Ωστόσο η ορθή προ-επεξεργασία δεδομένων μπορεί και παρέχει στον αναλυτή σημαντική βελτίωση σε άλλες μεθόδους. Τα **Δέντρα Απόφασης** έχουν πολλές φορές έχουν την ίδια αποτελεσματικότητα ταξινόμησης ή ακόμα και πιο αυξημένη σε σχέση με άλλους ταξινομητές όταν εφαρμόζονται σε συγκεκριμένα προβλήματα. Αρκετές φορές η χρήση είναι προτιμότερη λόγω της ικανότητας τους να παρέχουν σχετική επιλογή και επεκτασιμότητα γονιδίων, επιπλέον η λειτουργία τους προσφέρει ευκολότερη ερμηνεία λόγω του μοντέλου ακρίβειας.

6.2 Ταξινόμηση με Δέντρα Απόφασης

Τα δέντρα απόφασης με επαναλαμβανόμενο διαχωρισμό περίπτωσης χώρου κάνοντας χρήση των “hyperplanes” τα οποία είναι ορθογώνια σε άξονες. Η δημιουργία των μοντέλων αποφάσεων ξεκινά με την εισαγωγή ενός κόμβου που αποτελεί την ρίζα του δέντρου και αντιπροσωπεύει ένα χαρακτηριστικό, ο διαχωρισμός μια περίπτωσης χώρου που προαναφέρθηκε πραγματοποιείται με βάση μιας συνάρτησης των τιμών των χαρακτηριστικών, (οι διαχωριστικές τιμές επιλέγονται διαφορετικά από ξεχωριστούς αλγόριθμους). Τότε κάθε νέος υπο-χώρος που δημιουργείται αποτελείται από κάποια δεδομένα. Τώρα οι νέοι υπο-χώροι δεδομένων θα ξανά χωριστούν σε νέους μικρότερους υπο-χώρους, μια διαδικασία που επαναλαμβάνεται συνεχώς, μέχρι ένα από τα τελικά κριτήρια επιτευχθούν και οι τερματικοί κόμβοι του δέντρου (leaf-nodes: κόμβοι φύλλα) έχουν αναλάβει μια ταξινομημένη ομάδα οι οποία αντιπροσωπεύει το αποτέλεσμα της ταξινόμησης. Αυτή η ομάδα- κλάση (κατηγορία) αποτελεί την κατηγορία όλων ή της πλειοψηφίας των περιπτώσεων που περιέχονται στον υπο-χώρο. Το να αναθέσει κανείς τα σωστά τελικά κριτήρια αποτελεί μια από της σημαντικότερες διαδικασίες για την εφαρμογή των δέντρων απόφασης καθώς τα δέντρα που μπορεί να προκύψουν ίσως να είναι τεράστια και να υπερφορτώσουν, ή να είναι πολύ μικρά και να υποβαθμιστούν. Και στις δύο περιπτώσεις η ακρίβεια είναι πολύ δύσκολη να επιτευχθεί. Οι περισσότεροι αλγόριθμοι έχουν ένα μηχανισμό για να αντιμετωπίζουν το φαινόμενο ενός τεράστιου δέντρου απόφασης που ονομάζεται pruning, στα ελληνικά μεταφράζεται ως κουτσούρεμα αλλά έχει πιο πολύ την έννοια του κλαδέματος καθώς αυτή ακριβώς είναι η μέθοδος, να “κλαδεύει” μονοπάτια- κλαδιά του δέντρου που θεωρούνται άσκοπα να ακολουθήσουμε καθώς θα οδηγήσουν με βεβαιότητα σε περίπτωση πολύ διαφορετική από αυτήν που έχουμε ως επιθυμητή.

Κάθε νέα περίπτωση ταξινομείται μέσω ενός μονοπατιού που οδηγεί από τον αρχικό κόμβο-ρίζα του δέντρου μέχρι τον τελικό κόμβο-φύλλο του δέντρου, σύμφωνα με τα αποτελέσματα του τεστ κατά μήκος της διαδρομής.

Αν και τα δέντρα απόφασης μπορούν και παράγουν αποδοτικά μοντέλα, θεωρούνται ασταθή. Αυτό οφείλεται στην περίπτωση που τα δεδομένα εκπαίδευσης διαφέρουν έστω και λίγο. Τότε τα μοντέλα που θα προκύψουν ως αποτέλεσμα μπορεί να είναι τελείως διαφορετικά για αυτά τα δύο σετ. Λόγω αυτού, τα δέντρα αποφάσεων χρησιμοποιούνται συχνά σε ταξινομημένα σύνολα.

6.2.1 Single Decision Tree Classifiers

Οι πιο δημοφιλείς αλγόριθμοι που κατά την εκτέλεση τους, δημιουργούνται μοναδικά δέντρα απόφασης με σκοπό την ταξινόμηση είναι οι C4.5 και CART (Classification and Regression Trees). Τα δέντρα απόφασης προτάθηκαν αρχικά από την J. Ross Quinlan [23] ο οποίος περιέγραψε τον αλγόριθμο ID3 που χρησιμοποιήθηκε ως βάση για άλλα δέντρα απόφασης με στόχο την ταξινόμηση, όπου δημιουργήθηκαν μεταβαλλόμενες λειτουργίες αξιολόγησης και παράμετροι κατασκευής. Ο Αλγόριθμος C4.5 και ο CART χρησιμοποιούν παρόμοια μέθοδο για να διαιρέσουν τον χώρο χαρακτηριστικών, ωστόσο διαφέρουν στην δομή των δέντρων που παρέχουν, στα κριτήρια διάσπασης καθώς και στην μέθοδο κλαδέματος που εφαρμόζουν.

Ο αλγόριθμος C4.5 συνήθως κάνει χρήση του κέρδους πληροφοριών ή το λόγο του κέρδους ως κριτήριο επιλογής του χαρακτηριστικού, για κάθε διαχωρισμό χώρου που εφαρμόζει. Το κέρδος πληροφορίας αναφέρεται στην αλλαγή της εντροπίας των πληροφοριών εάν αλλάξει κάτι στην κατάσταση των τελευταίων. [23]

Με C να είναι το χαρακτηριστικό της κλάσης, με τιμές $\{c_1, c_2, \dots, c_n\}$ και A χαρακτηριστικό με τιμές $\{a_1, a_2, \dots, a_k\}$, η $H(C)$ θα είναι η εντροπία του χαρακτηριστικού κλάσης και η $H(A|C)$ θα είναι η υποθετική εντροπία που δείχνει την εντροπία του C εάν η κατάσταση του χαρακτηριστικού A είναι γνωστή. Απο αυτά προκύπτει ότι το κέρδος της πληροφορίας είναι:

$$I(C,A) = H(C) - H(C|A) \quad (1)$$

Οι CART και C4.5 έχουν και άλλες διαφορές μεταξύ τους όπως είναι η διαφορετικές μεθόδους κλαδέματος που εφαρμόζονται στους δυο αλγορίθμους, η έλλειψη χειρισμού τιμών και άλλα. Μέσω του κλαδέματος εξετάζεται και γίνεται αντικατάσταση των υπό-δέντρων που εμπεριέχονται μέσα σε ολόκληρο το δέντρο με ένα φύλλο ή ένα κλαδί του υποδέντρου όπου θεωρηθεί αναγκαίο. Στον C4.5 γίνεται χρήση κλαδέματος μειωμένων λαθών, καθώς προηγείται ανάλυση της πιθανής αντικατάστασης του υποδέντρου με ένα φύλλο ώστε να οδηγείται σε λιγότερα λάθη. Η παραπάνω τεχνική απαιτεί ξεχωριστό data set μόνο για την εφαρμογή του κλαδέματος, που από την μία πλευρά μπορεί να είναι μειονέκτημα, αλλά από την άλλη είναι πιο αποδοτική και γρήγορη τεχνική από άλλες καθώς εξετάζει κάθε υπό-δέντρο μια φορά. [24] Ο αλγόριθμος CART έχει το λιγότερο δυνατό κόστος ολυπλοκότητας κατά την εφαρμογή της τεχνικής κλαδέματος του. Η τεχνική αυτή αναθέτει το κόστος στα υπο-δέντρα με βάση το σφάλμα από το κλάδεμα και το μέγεθος του υπό-δέντρου. Και αντίθετα με τον C4.5 δεν κάνει χρήση ξεχωριστού data set για να υλοποιήσει το κλάδεμα. [25].

Χαρακτηριστικά	C4.5	CART
Τύπος δεδομένων	Συνεχής και Κατηγορηματικός	Συνεχείς και Ονομαστικά
Ταχύτητα	Μεσαία	Μεσαία
Boosting	Δεν διαθέτει το χαρακτηριστικό	Περιλαμβάνει το χαρακτηριστικό
Prunning / Κλάδεμα	Pre-Prunning	Post Prunning
Ελλείπουσες Τιμές	Δεν μπορεί να τις αποτρέψει	Μπορεί να τις αποτρέψει

Σύγκριση C4.5 με CART [59]

6.3 Decision Forests

Τα decision forests (απόφαση δάσους) είναι μια μεθοδολογία σύνοψης, η οποία δημιουργεί ένα μοντέλο το οποίο είναι δυνατόν να προβλεφθεί, με την ενσωμάτωση πολλαπλών μοντέλων (δέντρα αποφάσεων). Η παραπάνω μέθοδος μπορεί να χρησιμοποιηθεί για τη βελτίωση της απόδοσης πρόβλεψης καθώς και για τη σταθερότητα των ταξινομητών. [26] Οι πιο δημοφιλείς μέθοδοι είναι οι εξής: bagging, boosting και Random Forests.

Το bagging εισήχθη για πρώτη φορά από τον Breiman το 1996. Κατά το bagging κάθε δοκιμή $t = 1, 2, \dots, T$, ένα σύνολο εκπαίδευσης μεγέθους N δειγματίζεται με αντικατάσταση από τις αρχικές περιπτώσεις (το σετ κατάρτισης έχει το ίδιο μέγεθος με το αρχικό σετ, αλλά ορισμένες εμφανίσεις ενδέχεται να μην εμφανίζονται σε αυτό, ενώ ορισμένες περιπτώσεις εμφανίζονται περισσότερες από μία φορές). Στη συνέχεια, κατασκευάζεται ένας ταξινομητής για κάθε παραγόμενο σύνολο και ο τελικός ταξινομητής σχηματίζεται με τη συγκέντρωση των ταξινομητών T . Για να ταξινομηθεί μια νέα περίπτωση, μια ψήφος για την κλάση k καταγράφεται από κάθε ταξινομητή και η τελική ταξινόμηση είναι η τάξη με τις περισσότερες ψήφους [27].

Η μέθοδος boosting εισήχθηκε για πρώτη φορά από τους Freund και Schapire όταν πρότειναν τον αλγόριθμο AdaBoost. Το boosting διατηρεί ένα βάρος για κάθε περίπτωση - όσο υψηλότερο είναι το βάρος, τόσο περισσότερο η επίδραση επηρεάζει τον ταξινομητή. Σε κάθε δοκιμή, το διάλυμα των βαρών προσαρμόζεται ώστε να αντικατοπτρίζει την απόδοση του ταξινομητή, με αποτέλεσμα να αυξάνεται το βάρος των εσφαλμένων ταξινομημένων περιπτώσεων. Ο τελικός ταξινομητής συγκεντρώνει επίσης τους διδακτικούς ταξινομητές με ψηφοφορία, αλλά κάθε ψήφος είναι μία συνάρτηση της ακρίβειας τους.[27]

Τα Random forests χρησιμοποιούν ένα μεγάλο αριθμό μη επεξεργασμένων δέντρων αποφάσεων, τα οποία δημιουργούνται με τυχαίο διαχωρισμό σε κάθε κόμβο του δέντρου αποφάσεων. Ο αριθμός των χαρακτηριστικών που χρησιμοποιούνται για τον προσδιορισμό της απόφασης σε έναν κόμβο του δέντρου είναι προκαθορισμένος και είναι μικρότερος από τον αρχικό αριθμό των χαρακτηριστικών. Τα χαρακτηριστικά επιλέγονται τυχαία και ο καλύτερος διαχωρισμός μεταξύ αυτών των χαρακτηριστικών είναι εκλεκτός. Η ταξινόμηση ενός νέου δείγματος γίνεται με πλειοψηφία. [26]

6.4 Data Mining (Εξόρυξη Δεδομένων)

Η ταξινόμηση των δεδομένων είναι ιδιαίτερα χρήσιμη. Ακόμα και καίρια, ειδικά όταν αναφερόμαστε στην πολυτέλεια που δίνει κατά την διάγνωση των ασθενών. Εκεί τα ταξινομημένα δεδομένα μπορούν κυριολεκτικά να εντοπίσουν αν ένα αντικείμενο πλήττεται από μια ασθένεια ή όχι. Για να έχουν αυτά τα δεδομένα στην διάθεση τους ως είσοδο οι εφαρμογές και οι αλγόριθμοι ταξινόμησης θα πρέπει πρώτα να προηγηθεί ή ανάκτηση τους ή αλλιώς η εξόρυξη τους.

Η εξόρυξη δεδομένων αποτελεί χρήσιμο εργαλείο στην εξαγωγή απαραίτητων πληροφοριών από τον τεράστιο όγκο δεδομένων που συσσωρεύεται πλέον λόγω της ηλεκτρονικής-ψηφιακής μορφής της πληροφορίας και χρησιμοποιείται και για την εύκολη ερμηνεία τους μέσω της απεικόνισής τους.[39]

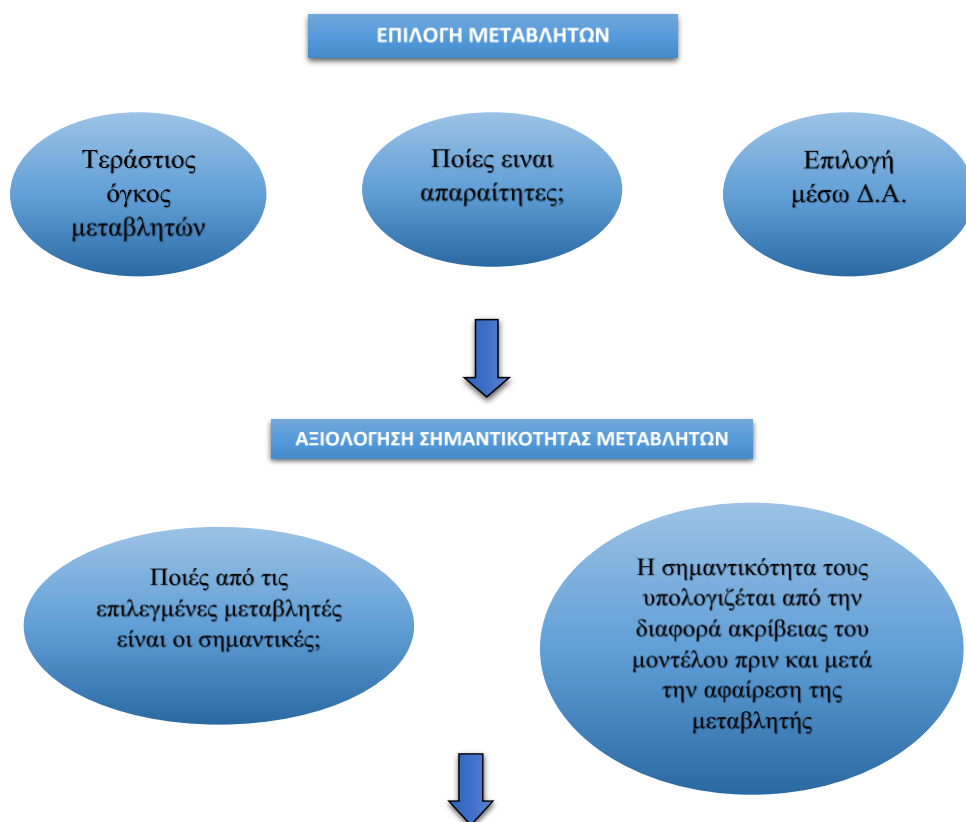
Συχνά γίνεται η σύγχυση μεταξύ της έννοιας της εξόρυξης δεδομένων και αυτής της ανακάλυψης γνώσης μέσα από τα δεδομένα (knowledge discovery from data – KDD). Ενώ η εξόρυξη δεδομένων

θα μπορούσε να θεωρηθεί ως ένα μέρος της ακολουθίας βημάτων που οδηγούν στην ανακάλυψη γνώσης. Τα βήματα της ανακάλυψης γνώσης είναι τα εξής.

1. **Data cleaning:** Εκκαθάριση δεδομένων, από μη σύμφωνα δεδομένα και θόρυβο
2. **Data integration:** Όρος που αναφέρεται στην ενσωμάτωση δεδομένων όπου πολλαπλές πηγές δεδομένων την καθιστούν δυνατή
3. **Data selection:** Επιλογή των δεδομένων που σχετίζονται και είναι χρήσιμα, ώστε να ανακτηθούν από την βάση δεδομένων
4. **Data transformation:** Μεταμόρφωση και ενοποίηση των δεδομένων, σε τέτοια μορφή ώστε να είναι εφικτή η εξόρυξη τους
5. **Data mining:** Από τα πιο σημαντικά βήματα για την ανακάλυψη γνώσης, εδώ γίνεται χρήση έξυπνων μεθόδων με σκοπό την εξόρυξη προτύπων δεδομένων.
6. **Pattern Evaluation:** Πρέπει να γίνει αξιολόγηση των προτύπων που ανακτήθηκαν για να βρεθούν τα πρότυπα εκείνα που θα είναι πιο χρήσιμα
7. **Knowledge presentation:** Η παρουσίαση της γνώσης γίνεται με χρήση τεχνικών οπτικοποίησης ώστε η γνώσεις που αποκτήθηκαν από την εξόρυξη να παρουσιαστεί στους χρήστες.

Από το βήμα 1 μέχρι το 4, υπάρχουν διαφορετικού τύπου επεξεργασία δεδομένων, όπου αυτά προετοιμάζονται για την διαδικασία της εξόρυξης. Στο βήμα της Εξόρυξης δεδομένων είναι πιθανόν να υπάρξει αλληλεπίδραση με τον χρήστη ή με μια βάση γνώσης. Επίσης σε μια βάση γνώσης θα αποθηκευτούν και τα πρότυπα εκείνα που θα κριθούν χρήσιμα.

Η ευχρηστία της εξόρυξης δεδομένων μέσω μοντέλων δένδρων απόφασης είναι μεγάλη και περιγράφεται παρακάτω.



ΔΙΑΧΕΙΡΙΣΗ ΕΛΛΕΙΠΟΥΝΤΩΝ ΤΙΜΩΝ

1. Ταξινόμηση τιμών που λείπουν σε μια δική τους κατηγορία η οποία θα αναλυθεί με το σύνολο των κατηγοριών.

2. Χρησιμοποίηση ενός δομημένου μοντέλου Δ.Α. το οποίο να ορίζει τη μεταβλητή με τις πολλές τιμές που λείπουν ως μεταβλητή στόχο ώστε να προβλέψει και να αντικαταστήσει τις τιμές που λείπουν με αυτές που προέκυψαν από την πρόβλεψη του μοντέλου



ΠΡΟΒΛΕΨΗ

Χρησιμοποιώντας το μοντέλο του δέντρου απόφασης το οποίο περιλαμβάνει δεδομένα αποτελεσμάτων από συγκεκριμένες καταστάσεις

Τότε το μοντέλο είναι εύκολο να προβλέψει τα αποτελέσματα σε περίπτωση αντίστοιχων καταστάσεων.



ΧΕΙΡΑΓΩΓΗΣΗ ΔΕΔΟΜΕΝΩΝ

Πολλές κατηγορίες κατηγοριολογικών μεταβλητών ή έντονα επικαλυμμένων συνεχών δεδομένων είναι κοινές στην ιατρική έρευνα.

Τα δέντρα αποφάσεων μπορούν να βοηθήσουν στην διάσπαση/κατάρευση των κατηγοριολογικών μεταβλητών έτσι ώστε να δημιουργηθεί ένα πιο εύχρηστο πλήθος κατηγοριών ή βοηθούν αντίστοιχα στον τρόπο υποδιαίρεσης των έντονων μεταβλητών σε εύρος.

Η εξόρυξη δεδομένων είναι μια τεχνολογία που μπορεί να εφαρμοστεί σε μια πληθώρα δεδομένων, ουσιαστικά μπορεί κανείς την χρησιμοποιήσει για σχεδόν όλους τους τύπους δεδομένων, υπό την προϋπόθεση ότι τα δεδομένα αυτά έχουν κάποια ουσία για την εφαρμογή που σκοπεύει να τα χρησιμοποιήσει. Κάποια παραδείγματα τύπων δεδομένων που μπορεί να εφαρμοστεί εξόρυξη είναι: οι αποθήκες δεδομένων (Data warehouses) τα data streams, οι ακολουθίες δεδομένων, τα network data, τα spatial data, τα δεδομένα κειμένου και πολυμέσο και φυσικά τα δεδομένα του διαδικτύου.

7 Omics and Translational Research Platforms

Η εξατομικευμένη ιατρική στοχεύει στη δημιουργία δεσμών μεταξύ των βιομοριακών χαρακτηρισμών, της κατάστασης που βρίσκεται ο ασθενής, την αποτελεσματικότητα της θεραπείας που ακολουθεί ο εκάστοτε ασθενής και τις δυσμενείς επιπτώσεις και μέσω αυτών των δεσμών ο ασθενής μπορεί να έχει την καλύτερη για αυτόν θεραπεία. [28] Κατά κύριο λόγο, οι εξελίξεις στην εξατομικευμένη ιατρική έχουν πραγματοποιηθεί χάρη στις πρωτοποριακές βελτιώσεις των βιομοριακών γνώσεων και των τεχνολογικών εξελίξεων στο ίδιο πεδίο κατά την διάρκεια της τελευταίας δεκαετίας. Σε αυτήν την χρονική περίοδο έχουν αναπτυχθεί πολλές τεχνολογίες υψηλής απόδοσης συγκριτικά με το τι προ-υπήρχε όσο αναφορά την διερεύνηση διαφορών πτυχών των κυτταρικών διεργασιών, όπως οι αλληλουχίες και οι δομικές παραλλαγές του γονιδιώματος, του μεταγραφικού, του επιγενόμενου, του πρωτεόμους και του interactome (όλα τα παραπάνω δεδομένα ονομάζονται omics ή ομικά δεδομένα). Υπάρχει μια πληθώρα από επιστημονικές μελέτες που παρέχουν σε βάθος πληροφορίες για αυτές τις τεχνολογίες. [29-35] Η ενσωματωμένη γονιδιωματική και η βιολογία συστημάτων, με γνώμονα αυτές τις νέες γνώσεις και τεχνολογίες που αναφέρονται στα παραπάνω άρθρα, έχουν βοηθήσει ιδιαίτερα στην κατανόηση των ανθρώπινων ασθενειών. [36] Για παράδειγμα, έχουν υπάρξει νέα στοιχεία όσο αναφορά τον μεταστατικό καρκίνο του παχέος εντέρου που οι μεταλλάξεις που προκύπτουν ενεργοποιούν το γονίδιο KRAS καταργώντας το θεραπευτικό αποτέλεσμα στην θεραπεία του υποδοχέα του αντί-επιδερμικού θεραπευτικού παράγοντα [37].

Ένας υπολογίσιμα μεγάλος όγκος από δεδομένα omics δημιουργούνται μέσω της χρήσης αυτών των τεχνολογιών, καλύπτοντας ένα ευρύ φάσμα τομέων. Αυτά τα δεδομένα omics για να μπορέσουν να φτάσουν στο μέγιστο του δυναμικού τους θα πρέπει να εξεταστούν στο πλαίσιο των κυκλοειδών διαδικασιών. Με μάλιστα παραπάνω από 1500 διαφορετικές βάσεις δεδομένων σχετιζόμενες με την βιομοριακή χρήση, οι ερευνητές χρειάζονται τα κατάλληλα εργαλεία για να μπορέσουν να βρουν την σχετική πληροφορία στον λαβύρινθο των διαθέσιμων βιολογικών δεδομένων. Ποικίλα συστήματα έχουν αναπτυχθεί με σκοπό την επίλυση του παραπάνω προβλήματος, δηλαδή υπάρχουν συστήματα που απευθύνονται στην ανάγκη των ερευνητών να μπορέσουν να δουλέψουν με τα δεδομένα omics., πχ. Gene Expression Omnibus, Array Express, PRIDE. Ωστόσο τα δεδομένα omics θα πρέπει να αναλύονται μαζί με τα κλινικά δεδομένα για να μπορούν να θεωρηθούν χρήσιμα στην translational έρευνα και κατ' επέκταση στην επωφέληση των ασθενών. Τέτοια συστήματα δεν παρέχουν λύση στο πρόβλημα της ολοκλήρωσης κλινικών και omics δεδομένων.

Τα τελευταία χρόνια εμφανίστηκαν νέες τάσεις στην διαχείριση και την ανάλυση κλινικών και omics δεδομένων. Έχουν ληφθεί αρκετές επιλογές για την εύρεση λύσεων σχετικά με μεθόδους πληροφορικής που συνδέουν μοριακές οντότητες με κλινικές οντότητες. Μεταξύ αυτών, αναπτύσσονται ενεργά translational ερευνητικές πλατφόρμες, ικανές να ενσωματώσουν μεγάλα σύνολα δεδομένων κλινικών πληροφοριών με δεδομένα omics.

Τέτοιες translational ερευνητικές πλατφόρμες θα πρέπει να είναι σε θέση να συνδυάσουν τη ροή εργασίας των ερευνητών για την επίτευξη της βέλτιστης χρήσης τους. Ως εκ τούτου, θα πρέπει να παρέχουν (i) την αποθήκευση και την ενσωμάτωση κλινικών και omics δεδομένων. (ii) ένα πλαίσιο ανάλυσης που επιτρέπει στους επιστήμονες να διερευνούν τα δεδομένα τους και να δημιουργούν υποθέσεις. και (iii) πρόσθετες πληροφορίες που διασταυρώνονται από εξωτερικές βάσεις δεδομένων (π.χ. σύνδεση με συγκεκριμένη περιγραφή γονιδίου σε δημοσιευμένη βιβλιογραφία ή δημόσιες βάσεις δεδομένων). [38]

Το 2014 στο πανεπιστήμιο της Οξφόρδης πραγματοποιήθηκε μια έρευνα πάνω στις Translational ερευνητικές πλατφόρμες, χρησιμοποιώντας το PubMed και ταυτοποιώντας 2359 επιστημονικά άρθρα που περιγράφουν τις ερευνητικές πλατφόρμες.[38]

Ουσιαστικά μέσα από την συγκεκριμένη μελέτη αναλύσανε 7 translational ερευνητικές πλατφόρμες.

BRISK

Biology-Related Information Storage Kit (2011) (βιολογικά συσχετισμένος αποθηκευτικό κιτ πληροφοριών)

Το BRISK είναι ένα πακέτο τριών web-based εφαρμογών ανοιχτού κώδικα που παρέχουν μια συνεκτική πλατφόρμα ολοκλήρωσης και διαχείρισης δεδομένων. Αρχικά αναπτύχθηκε για να προσφέρει μια λύση ανταλλαγής δεδομένων για τους ερευνητές στο consortium AllerGen (The Allergy, Genes and Environment Network) (<http://www.allergen-ncce.ca>). Το BRISK μπορεί να χειριστεί την περιγραφή κλινικών φαινοτύπων και τις πληροφορίες σωματικής μετάλλαξης (πολυμορφισμοί ενός νουκλεοτιδίου). Παρέχει στους ερευνητές δυνατότητες ανάλυσης μελετών συσχετισμού γονιδιώματος (GWAS). Αυτή η λύση περιλαμβάνει επίσης μια εργαστηριακή προσανατολισμένη εφαρμογή που διαχειρίζεται το φυσικό δείγμα, το θέμα και τα δεδομένα των δοχείων.[38]

CaTRIP

Η πλατφόρμα caTRIP αναπτύχθηκε ως συστατικό του έργου caBIG στις αρχές της δεκαετίας του 2000 για να επιτρέψει στους χρήστες να διερευνήσουν το δίκτυο caBIG. Το caBIG ήταν ένα πρόγραμμα του Εθνικού Ινστιτούτου Καρκίνου των ΗΠΑ. Ο στόχος της ήταν να αναπτυχθεί ένα δίκτυο ανοιχτού κώδικα σε όλες τις Ηνωμένες Πολιτείες για ασφαλείς ανταλλαγές σχετικά με την έρευνα για τον καρκίνο. Οι στόχοι του caTRIP περιλαμβάνουν να επιτρέπουν στους γιατρούς να βρίσκουν ασθενείς με παρόμοια προφίλ, να αναλύουν τα αποτελέσματά τους και να βρίσκουν πληροφορίες για επιτυχείς θεραπείες σε όλο το δίκτυο δεδομένων caBIG. Το σύστημα λειτουργεί με πολλές εφαρμογές caBIG, συμπεριλαμβανομένου του μητρώου όγκων, ενός κλινικού συστήματος που χρησιμοποιείται για τη συλλογή δεδομένων, το σύστημα εξαγωγής κειμένου (Text Information Extraction System) για τον καρκίνο, ένα εργαλείο φυσικής επεξεργασίας γλώσσας (Natural Language Processing - NLP) το οποίο σχεδιάστηκε με σκοπό να εξαγάγει τις κλινικές γνώσεις από την έκθεση ελεύθερου κειμένου της χειρουργικής παθολογίας χρησιμοποιώντας ελεγχόμενες ορολογίες, το caTissue CORE, ένα αποθετήριο τραπεζών ιστών, τη μηχανή σχολιασμού καρκίνου (Cancer Annotation Engine) και το caIntegrator, ένα εργαλείο αποθήκευσης, αναζήτησης και ανάλυσης δεδομένων.[38]

Cbio Cancer Genomics Portal (2012)

Το cBio Cancer Genomics Portal, που αναπτύχθηκε στο Memorial Sloan-Kettering Cancer Center (MSKCC), είναι μια πλατφόρμα ανοιχτού κώδικα που έχει σχεδιαστεί για να διευκολύνει την πρόσβαση των translational ερευνητών σε σύνολα δεδομένων που παράγονται από μεγάλης κλίμακας έργα γονιδιωματικής του καρκίνου, όπως το Atlas του Καρκίνου του Γονιδιώματος : (<http://cancergenome.nih.gov/>) και τη Διεθνή Κοινοπραξία για τον Καρκίνο του Καρκίνου (<http://icgc.org/>). Ενσωματώνει κλινικά δεδομένα από-ταυτοποίησης, όπως η φαινοτυπική περιγραφή, η επιβίωση ή τα διαστήματα επιβίωσης χωρίς ασθένεια, με σημαντικά ομοιώματα υψηλής απόδοσης (DNA, messenger RNA - mRNA και πρωτεΐνες). Επιπρόσθετα, οι εικόνες παθολογίας μπορούν να προσεγγιστούν μέσω ενσωματωμένης απεικόνισης αρχειοθέτησης ψηφιακού διαύλου καρκίνου TGCA (<http://cancer.digitalslidearchive.net/>). Οι εικόνες μπορούν να αποκτήσουν πρόσβαση μέσω ενσωματωμένης απεικόνισης αρχειοθέτησης ψηφιακών αρχείων καρκίνου TGCA (<http://cancer.digitalslidearchive.net/>). Παρέχονται προηγμένες λειτουργίες οπτικοποίησης, ανάλυσης

και εξαγωγής. Η δημόσια διαδικτυακή έκδοση αποθηκεύει κυρίως δημοσιευμένα σύνολα δεδομένων μεγάλης κλίμακας για τη γονιδιωματική του καρκίνου, ενώ μια ιδιωτική παρουσία της πύλης μπορεί να δημιουργηθεί σε τοπικό επίπεδο από ερευνητικές ομάδες που επιθυμούν να εισαγάγουν τα δικά τους σύνολα δεδομένων έρευνας.[38]

G-DOC Georgetown Database of Cancer (2012)

Το Georgetown Database of Cancer (G-DOC), που αναπτύχθηκε στο πλήρες κέντρο καρκίνου του Lombardi στο Πανεπιστήμιο του Georgetown, είναι μια translational πληροφορική υποδομή που στοχεύει στη διευκόλυνση του translate και της βασισμένης σε συστήματα ιατρικής. Σχεδιάστηκε ειδικά για την αντιμετώπιση του φραγμού ενεργοποίησης για τη χρήση εργαλείων βιοϊατρικής πληροφορικής από βασικούς, κλινικούς και translational ερευνητές. Το G-DOC ενσωματώνει τα χαρακτηριστικά του ασθενούς (π.χ. δημογραφικά στοιχεία, δομημένα δεδομένα κλινικής έρευνας) και δεδομένα κλινικών αποτελεσμάτων με τέσσερα σημαντικά ομόκεντρα υψηλής απόδοσης (DNA, mRNA, microRNA και μεταβολίτες) σε ένα ενοποιημένο περιβάλλον. Το σχετικό πλαίσιο, ο Μηχανισμός Ανάπτυξης Κλινικών και Omics Georgetown (G-CODE) περιέχει ένα ευρύ φάσμα εργαλείων βιοπληροφορικής και βιολογίας συστημάτων που είναι αφιερωμένα στην ανάλυση και την απεικόνιση δεδομένων.[38]

Icod Intergratd Clinical Omics Database(2010)

Η Βάση Δεδομένων (iCOD) αναπτύχθηκε για να συνδυάζει ολοκληρωμένες κλινικοπαθολογικές και μοριακές πληροφορίες ασθενών για να παρέχει μια ολιστική κατανόηση των ασθενειών που πλήττουν τα αντικείμενα μελέτης. Το iCOD μπορεί να χειριστεί δεδομένα omics όπως προφίλ γονιδιακής έκφρασης και ετερογενείς κλινικές πληροφορίες, όπως λεπτομερείς φαινότυποι, εικόνες ακτινολογίας ή αποτελέσματα εργαστηριακών εξετάσεων. Οι τοπικοί χάρτες ολοκληρωμένης απεικόνισης των ασθενειών παρέχονται για να συνοψίσουν την αλληλεπίδραση των κλινικών και omics δεδομένων και αντιπροσωπεύουν εύλογες οδούς ασθένειας. [38]

DASH: Integrating data for analysis, anonymization and sharing (2011)

Το iDASH είναι Εθνικό Κέντρο Βιοϊατρικής Πληροφορικής των Η.Π.Α.. Το iDASH παρέχει στους ερευνητές σε όλες τις Ηνωμένες Πολιτείες μια ισχυρή υπολογιστική υποδομή που απαιτείται για την ενοποίηση δεδομένων και την ανάλυση δεδομένων. Το iDASH διανέμει επίσης εργαλεία και αλγόριθμους που επικεντρώνονται στην ανταλλαγή δεδομένων με τρόπο που προστατεύει την προστασία της ιδιωτικής ζωής. Το iDASH παρέχει στους βιοϊατρικούς και συμπεριφορικούς ερευνητές πρόσβαση σε δεδομένα, λογισμικό και περιβάλλον πληροφορικής υψηλής απόδοσης, δίνοντάς τους τη δυνατότητα να παράγουν και να δοκιμάζουν νέες υποθέσεις. [38]

TranSMART (2010)

Αυτή η πλατφόρμα αναπτύχθηκε αρχικά ως μια προανταγωνιστική πλατφόρμα συνεργασίας για φαρμακευτικές εταιρείες από μια ιδιωτική κοινοπραξία προτού κυκλοφορήσει στην κοινότητα ανοιχτού κώδικα (το Ίδρυμα tranSMART είναι τώρα υπεύθυνο για τη βιωσιμότητα και την ανάπτυξη αυτού του κώδικα). Η πλατφόρμα βασίζεται στο i2b2 CDW ανοιχτού κώδικα. Κατασκευάζεται για να βοηθήσει τους επιστήμονες να αναπτύξουν και να βελτιώσουν τις ερευνητικές υποθέσεις με τη διερεύνηση συσχετισμών μεταξύ φαινοτυπικών και omics δεδομένων. Το TranSMART μπορεί να χειριστεί δομημένα δεδομένα από κλινικές δοκιμές (δημογραφικά στοιχεία, αποτελέσματα, εργαστηριακά αποτελέσματα και κλινικούς φαινότυπους) και ευθυγραμμισμένα δεδομένα βιοδεικτών υψηλής περιεκτικότητας όπως προφίλ γονιδιακής έκφρασης, γονότυπους, μεταβολικοποιητικά και πρωτεϊνωματικά δεδομένα. Παρέχει στους ερευνητές εργαλεία ανάλυσης ικανά να παράγουν προηγμένα στατιστικά στοιχεία περιγραφής και ανάλυσης.[38]

8.1 Ο ανθρώπινος εγκέφαλος (δομή)

Όπως θα γίνει παρακάτω ξεκάθαρο η νόσος του Αλτσχάιμερ όπως και η άνοια αποτελεί αρρώστια που πλήττει τον ανθρώπινο εγκέφαλο και την λειτουργία αυτού.

Ο εγκέφαλος θεωρείται από τα πιο περίπλοκα όργανα του ανθρώπινου σώματος.

Αποτελεί ένα άθροισμα συσχετιζόμενων νευρωνικών συστημάτων, που κάθε ένα ρυθμίζει την δική του δραστηριότητα αλλά και την δραστηριότητα των άλλων, με έναν δυναμικό και πολυσύνθετο τρόπο. Τα μεγάλα ανατομικά τμήματα του εγκεφάλου μας, παρέχουν έναν αδρό τρόπο ταξινόμησης των λειτουργιών αυτού.

8.1.1 Εγκεφαλικός φλοιός

Ο φλοιός του εγκεφάλου όπως αναφέρθηκε συντίθεται από νευρώνες οι οποίοι αριθμούν περίπου τους 50 δισεκατομμύρια. Οι περιοχές του φλοιού ομαδοποιούνται και ταξινομούνται με ποικίλους τρόπους, ανάλογα με το είδος της πληροφορίας που επεξεργάζονται και την θέση που καταλαμβάνουν στον εγκέφαλο. Ο οποίος χωρίζεται σε δύο ημισφαίρια τα οποία καταλαμβάνουν το 85% του εγκεφάλου. Κάθε ημισφαίριο έχει 4 λοβούς: τον μετωπιαίο λοβό, η περιοχή που σχετίζεται με την κίνηση και την προσωπικότητα του ανθρώπου, ο πλευρικός ή και βρεγματικός λοβός ο οποίος είναι υπεύθυνος για την αίσθηση την αντίληψη και την επεξεργασία και βρίσκεται πίσω από τον μετωπιαίο, ο ιανικός λοβός που βρίσκεται στο πίσω μέρος του εγκεφάλου και ευθύνεται για την όραση καθώς σε αυτόν εμπεριέχεται ο οπτικός φλοιός και τέλος ο κροταφικός λοβός που παίρνει θέση στο πλαϊνό μέρος του εγκεφάλου κάτω από τον μετωπιαίο και πλευρικό και από αυτόν εξαρτώνται οι λειτουργίες την ομιλίας της ακοής και της μνήμης.

Στο υπόλοιπο 15% του εγκεφάλου περιλαμβάνεται η παρεγκεφαλίδα που βρίσκεται πάνω από το εγκεφαλικό στέλεχος και κάτω από τον ιανικό λοβό. Οι λειτουργίες που εκτελούνται εκεί αφορούν τη στάση του σώματος και την ισορροπία. Επιπλέον, το εγκεφαλικό στέλεχος το οποίο συνδέεται με την σπονδυλική στήλη και καθορίζει τις διαδικασίες της αρτηριακής πίεσης, τον καρδιακό παλμό, την αναπνοή, τον ύπνο και τα όνειρα. Επίσης το λιμβικό σύστημα, το οποίο αναπτύσσεται ραγδαία στα πρώτα έτη της ζωής του ανθρώπου και είναι υπεύθυνο για τις λειτουργίες της μάθησης, της μνήμης της αντίληψης της γεύσης και του άλγους.

Το μέρος του εγκεφάλου που καθορίζει τα ισχυρά συναισθήματα όπως ο φόβος βρίσκεται στον κροταφικό λοβό μαζί με τον ιππόκαμπο. Το κομμάτι του εγκεφάλου που είναι υπεύθυνο για τις συναισθηματικές λειτουργίες ονομάζεται αμυγδαλή, ενώ ο ιππόκαμπος διαχειρίζεται την μάθηση και την πρόσφατη μνήμη. Τέλος μια ακόμη λειτουργία του εγκεφάλου είναι η θερμοκρασία του σώματος και ο έλεγχος της ποσότητας τροφής, ενέργειες που ελέγχονται από τον προθάλαμο.

8.2 Κυτταρική Οργάνωση Του Εγκεφάλου

Η κυτταρική οργάνωση του ανθρώπινου εγκεφάλου διακρίνεται σε 3 μοντέλα συνδέσεων μεταξύ των νευρώνων.

Το πρώτο μοντέλο αναφέρεται στις μακρές-ιεραρχικές νευρωνικές οργανώσεις όπως χαρακτηριστικά βρίσκονται στις μακρές αισθητικές κι κινητικές οδούς. Σε αυτό το μοντέλο η ροή της πληροφορίας ακολουθεί μια αυστηρά καθορισμένη ιεραρχία, σύμφωνα με την οποία πραγματοποιούνται οι συνδέσεις των νευρώνων. Ο ιεραρχικός τρόπος ξεκινά από τους περιφερικούς

υποδοχείς, μετά περνά στον δεύτερο αισθητικό νευρώνα και τέλος καταλήγει στον αισθητικό φλοιό. Αντίστροφα για την κινητική οδό η ροή της πληροφορίας ξεκινά από τον κινητικό φλοιό στους κινητικούς νευρώνες του νωτιαίου μυελού.

Το δεύτερο μοντέλο οργάνωσης περιλαμβάνει νευρώνες, με τις συνδέσεις αυτών να γίνονται στο άμεσο περιβάλλον τους και ως εκ τούτου να δημιουργούνται μικρά τοπικά νευρωνικά κυκλώματα.

Το τρίτο μοντέλο οργάνωσης χρησιμοποιείται από ορισμένα νευρωνικά συστήματα του υποθαλάμου και του μυελεγκεφάλου. Τα συστήματα αυτά περιέχουν είτε μια μονοαμίνη ή κάποιο πεπτίδιο. Οι συνδέσεις των νευρώνων στο εξής μοντέλο δεν ακολουθούν κάποιου τύπου ιεραρχική δομή και προβάλλουν με πολλαπλές διακλαδιζόμενες αποκλίνουσες ίνες σε πολλά κύτταρα στόχους.

8.3 Νευρωνικές Συνάψεις

Όπως αναφέρεται παραπάνω οι νευρώνες αποτελούν τις βασικές δοκιμές μονάδες του ανθρώπινου εγκεφάλου καθώς ο αριθμός αυτών είναι τεράστιος. Οι νευρώνες έχουν πολλά κοινά με τα υπόλοιπα κύτταρα καθώς γενετικά οργανώνονται με τον ίδιο τρόπο έχουν τις ίδιες βιοχημικές συσκευές και αποτελούνται από τα ίδια γονίδια. Παρ' όλα αυτά είναι ιδιαίτερα επειδή διαθέτουν μοναδικά χαρακτηριστικά που κάνουν την λειτουργία του εγκεφάλου την πιο ζωτική και περίπλοκη σε σύγκριση με τις λειτουργίες των υπόλοιπων οργάνων του ανθρώπινου σώματος. Τα μοναδικά αυτά χαρακτηριστικά αναφέρονται στο ιδιόμορφο κυτταρικό τους σχήμα, στην εξωτερική τους μεμβράνη που είναι ικανή να δημιουργεί και να μεταδίδει, στην μοναδική τους κατασκευή, η σύναψη, μέσω της οποίας μεταδίδονται οι πληροφορίες μεταξύ των νευρώνων.

Η σύναψη αποτελεί το σημείο όπου μέσω χημικών μεταβιβαστών η πληροφορία μεταφέρεται από τον ένα νευρώνα στον άλλο. Συνήθως οι συνάψεις αυτές πραγματοποιούνται μεταξύ του νευράξονα ενός κυττάρου και του δενδρίτη ενός άλλου, ωστόσο υπάρχουν κι άλλα είδη συναπτικών συδέσεων.

9.1 Εισαγωγή στο Αλτσχάιμερ

Το Alzheimer αποτελεί την πιο συχνή μορφή άνοιας και για την ακρίβεια μια προοδευτική νευροεκφυλιστική μορφή της. Ανακαλύφθηκε από τον Γερμανό νευρολόγο Alois Alzheimer το 1907. Ο A. Alzheimer παρατήρησε αλλοιώσεις στον εγκέφαλο μιας ασθενούς που πέθανε από την νόσο αυτή. Στη σημερινή εποχή, 26.000.000 άτομα πάσχουν από Alzheimer, ενώ στην Ελλάδα ο αριθμός ανέρχεται στα 160.000 άτομα. Μετά την ηλικία των 65, ένας στους είκοσι πάσχει από την νόσο, ενώ μετά τα 80 έτη η αναλογία ανέρχεται σε τέσσερις στους δέκα. Εμφανίζεται πιο συχνά στον γυναικείο πληθυσμό, αφενός γιατί το προσδόκιμο ζωής είναι μεγαλύτερο και αφετέρου διότι μετά την εμμηνόπαυση από τα οιστρογόνα, το νευρικό σύστημα των γυναικών είναι απροστάτευτο.[10]

Η νόσος του Αλτσχάιμερ θεωρείται ως μια ασθένεια νευρωγενετικής φύσεως η οποία έχει την δυνατότητα να προκαλέσει προβλήματα στην συμπεριφορά του ασθενή, στην μνήμη και στην σκέψη του και όπως είναι γνωστόν καταλήγει στον θάνατο του. Αποτελεί την πιο συχνή μορφή άνοιας και μάλιστα το ποσοστό των ασθενών με αλτσχάιμερ σε σχέση με το συνολικό πλήθος των ασθενών που πλήττονται από άνοια κυμαίνεται περίπου στο 60 με 80 %. [11] Ως ασθένεια θεωρείται η αυτή που στοιχίζει περισσότερο στις χώρες του δυτικού πολιτισμού καθώς ο ασθενής χάνοντας την μνήμη του και την λογική του οδηγείται σε κοινωνικά κατακριτέες συμπεριφορές χωρίς τον έλεγχο του, με αποτέλεσμα η αρρώστια να αποτελεί φόβο για τα μέλη πολλών κοινωνικών ομάδων των ανεπτυγμένων χωρών. Οι πληγέντες από την νόσο έφτασαν τους 26.6 εκατομμύρια το 2006 σε όλον το κόσμο και υπολογίζεται ότι 1 στους 85 ανθρώπους παγκοσμίως θα πλήττεται από την νόσο του Αλτσχάιμερ μέχρι το 2050[12]. Στο World Alzheimer Report που πραγματοποιήθηκε το 2010, [13] παρουσιάστηκε μία μελέτη που προέβλεπε ότι μέχρι το 2030 θα θύματα με άνοια θα έχουν φτάσει τα 65.7 εκατομμύρια ενώ 20 χρόνια αργότερα ο αριθμός αυτός θα έχει αυξηθεί στους 115.4 εκατομμύρια πάσχοντες. Στην ίδια αναφορά παρουσιάζεται το ότι τα δύο τρίτα των θυμάτων άνοιας θα προέρχονται από χαμηλές ή μεσαία εισοδηματικές χώρες, οι οποίες όπως αναφέρεται θα πρέπει να περιμένουν μεγάλη αύξηση τα επόμενα χρόνια στους πάσχοντες της νευροεκφυλιστικής αρρώστιας. Το γεγονός αυτό αποτελεί μεγάλη πρόκληση για ποικίλους λόγους, ένας εκ των οποίων είναι ότι σε αυτές τις χώρες οι ασθενείς ζητούν μη επίσημη θεραπεία, δηλαδή συμβουλευονται δωρεάν ιατρική βοήθεια από τους κοντινούς τους ανθρώπους, μία πρακτική που μπορεί να αποδειχθεί εξαιρετικά δύσκολη στη διατήρηση καθώς το γηραιότερο μέρος αυτών των πληθυσμών αυξάνεται σε αριθμό και η ασθένεια ακολουθεί την ανοδική πορεία του αριθμού αυτού. [13] Η άνοια προκαλεί τεράστιο κόστος στην κοινωνία έτσι όπως αυτή υφίσταται σήμερα, το οποίο αντιστοιχεί για το 1.01% του συνολικού αθροίσματος των παγκόσμιων Ακαθάριστων Εθνικών Προϊόντων. [13] Πιστεύεται ότι το παραπάνω πρόβλημα επιδεινωθεί στα χρόνια που πρόκειται να ακολουθήσουν, μάλιστα προβλέπεται ότι σε ένα ποσοστό 85% παγκοσμίως θα αυξηθεί το κοινωνικό κόστος μέχρι το 2030, υποθέτοντας φυσικά ότι δεν υπάρχουν πιθανοί παράγοντες στο υπόβαθρο όπως οι μακροοικονομικοί παράγοντες ή η αλλαγή στην διαθεσιμότητα ή την αποδοτικότητα της θεραπείας.

Στην νόσο Αλτσχάιμερ κάποια συμπτώματα που εμφανίζονται θυμίζουν κάποιες τυπικές και φυσιολογικές συμπεριφορές που προκαλούνται στους ανθρώπους λόγω προχωρημένης ηλικίας, αλλά πρέπει να τονιστεί ότι το Αλτσχάιμερ δεν είναι ένα φυσιολογικό κομμάτι της γήρανσης αλλά μια μορφή άνοιας. Μάλιστα όσο η αρρώστια προχωράει με τον καιρό τα συμπτώματα της άνοιας χειροτερεύουν και γίνονται πιο εμφανή. Σήμερα δεν υπάρχει συγκεκριμένη θεραπεία για την πλήρη απαλλαγή του ασθενή από το Αλτσχάιμερ. Ο στόχος των θεραπειών είναι να μειωθεί η εξάπλωση της νόσου, να βελτιωθεί ο ασθενής ως προς τα συμπτώματα, η αντιμετώπιση των προβλημάτων συμπεριφοράς και φυσικά να επιτευχθεί η βελτίωση του βιοτικού επιπέδου του ασθενή. Ωστόσο, οι

θεραπείες που εφαρμόζονται σήμερα μπορούν να επιβραδύνουν προσωρινά την ανάπτυξη των συμπτωμάτων άνοιας εάν η νόσος διαγνωστεί σε πρώιμο στάδιο. Ενώ πιο αποτελεσματικές θεραπείες και εν τέλει η πρόληψη ή ακόμα και η πλήρη θεραπεία είναι πολύ σημαντικοί στόχοι, πρώιμες διαγνώσεις μπορούν να προσφέρουν στον ασθενή πολύ καλύτερα αποτελέσματα θεραπείας, σε σύγκριση με αυτά που θα έχει η θεραπεία μιας πιο αργής διάγνωσης.

Οι ακριβείς λόγοι που προκαλούν το Αλτσχάιμερ είναι ακόμα και σήμερα άγνωστοι με ελάχιστες εξαιρέσεις που προκαλούνται λόγω αναγνωρίσιμων γενετικών διαφορών. Σύγχρονες έρευνες δείχνουν, ωστόσο ότι σχετίζεται με τις νευριτικές πλάκες και τα νευροϊνιδιακά πλέγματα του εγκεφάλου. [14] Ενώ το αμυλωιδές βήτα, μία πρωτεΐνη που αποτελεί το “υλικό” με το οποίο είναι κατασκευασμένες οι νευριτικές πλάκες, είναι γνωστό ότι εμπλέκεται με την ανάπτυξη της ασθένειας και γύρω από αυτό γίνεται διάλογος κατά πόσο είναι η αιτία της αρρώστιας όπως πιστεύουν μερικοί. Γενικότερα η πρωτεΐνη αναγνωρίζεται ως βιοδείκτης της νόσου.

Η τρέχουσα τάση αναφέρεται στο ότι υπάρχει συνεχώς αυξανόμενη ικανότητα να εντοπίσει κανείς την εμφάνιση της νόσου και να παρακολουθήσει την πρόοδο της πριν καν προκύψουν τα συμπτώματα. Αυτό αποτελεί μεγάλη εξέλιξη της στην ερεύνα και φυσικά με τον όρο αυτό εννοείται η ταυτοποίηση των βιοδεικτών, που επιτρέπουν την ταυτοποίηση και την οπτικοποίηση των διαδικασιών που σχετίζονται με το Αλτσχάιμερ μήνες, χρόνια ακόμη και δεκαετίες πριν από την εμφάνιση των όποιων κλινικών συμπτωμάτων προκύψουν.

Οι βιοδείκτες για το Αλτσχάιμερ μπορούν να διαχωριστούν σε πρόωρους βιοδείκτες, οι οποίοι τυπικά μετρούν την αποβολή αμυλωειδούς στον εγκέφαλο (πχ. PET imaging, CSF αμυλωειδές) και μεταγενέστερους βιοδείκτες, οι οποίοι με την σειρά τους τυπικά μετρούν το νευροεκφυλισμό. (πχ. Δομική μαγνητική τομογραφία -MRI-, FDG, PET CSF).

Οι σαρώσεις εικόνων του εγκεφάλου συχνά χρησιμοποιούνται για να αποκλειστούν άλλες αιτίες από τα συμπτώματα, αλλά επίσης μπορούν να θεωρηθούν ενδείξεις ύπαρξης της νόσου ή όχι. Σε ένα επιστημονικό άρθρο που εκδόθηκε το 1996 από τον Laakso αναφέρεται ότι: “ Στις κλινικές πρακτικές, η διάγνωση της νόσου του Αλτσχάιμερ βασίζεται σε τυπικά χαρακτηριστικά και σε αποκλεισμό άλλων συνθηκών που προκαλούν άνοια”[17]

Ο μόνος γνωστός τρόπος για να μπορεί κανείς να είναι απόλυτα βέβαιος ότι ένα άτομο πάσχει από την νόσο είναι δυστυχώς η μεταθανάτια εξέταση του εγκεφαλικού ιστού του. Παρ’ όλα αυτά όπως προαναφέρθηκε τόσο τα νευροϊνιδιακά πλέγματα όσο και οι νευριτικές πλάκες φαίνεται να παίζουν ρόλο στην ανάπτυξη και την εξέλιξη της νόσου του Αλτσχάιμερ.

Αν και έχουν πραγματοποιηθεί ποικίλες προσπάθειες έρευνας σχετικά με το Αλτσχάιμερ, εξακολουθεί να υπάρχει ανάγκη για ένα πρώιμο μη επεμβατικό διαγνωστικό εργαλείο για την νόσο.

9.2 Ήπια Γνωστική Εξασθένηση (Mild Cognitive Impairment)

Η ήπια γνωστική εξασθένηση (MCI) θεωρείται ως ένα πρώιμο στάδιο της νόσου Αλτσχάιμερ, αν και είναι κάπως αμφισβητήσιμο εάν το MCI αντιστοιχεί σε κάποιο διαφορετικό στάδιο της διαγνωστικής ή σε προδρομικό στάδιο του Αλτσχάιμερ.

Σε αυτή την φάση των ασθενών, οι αλλαγές στον εγκέφαλο έχουν ήδη αρχίσει για κάποιο χρονικό διάστημα και τα συμπτώματα αρχίζουν σιγά σιγά να εμφανίζονται. Επίσης δεν οδηγεί ακόμα σε μεγάλα

προβλήματα που να είναι αρκετά σοβαρά ώστε να έχουν σημαντικό αντίκτυπο στην καθημερινή λειτουργία των ασθενών, δηλαδή να εμφανίζει συμπτώματα τέτοια ώστε να μπορεί ο ασθενής να θεωρηθεί ότι πάσχει από άνοια.

9.3 Παράγοντες Κινδύνου

Ενώ οι ξεκάθαροι λόγοι που πλήττονται οι ασθενείς από την νόσο δεν είναι εντελώς γνωστοί, υπάρχουν συγκεκριμένοι παράγοντες που σχετίζονται άμεσα με τις πιθανότητες ενός ανθρώπου να αποκτήσει Αλτσχάιμερ. Αυτοί οι παράγοντες είναι οι εξής :

Ηλικία

Η ηλικία είναι ξεκάθαρα σχετιζόμενη με την νόσο Αλτσχάιμερ. Μετά από μία συγκεκριμένη ηλικία συνήθως τα 65, οι πιθανότητες να αποκτήσει κάποιος Αλτσχάιμερ διπλασιάζονται κάθε πενταετία και μετά τα 85 έτη ένας άνθρωπος έχει σχεδόν 50% πιθανότητες να αποκτήσει την νόσο.

Οικογενειακό Ιστορικό

Εάν ένα οικογενειακό μέλος πλήττεται από την νόσο, τότε όσοι είναι οικογενειακά κοντινά μέλη έχουν περισσότερες πιθανότητες να αναπτύξουν Αλτσχάιμερ. Εάν πάνω από ένα μέλος έχει ή είχε την αρρώστια, τότε οι πιθανότητες αυξάνονται.

Γνωσιολογία

Η γνωσιολογία διαδραματίζει σημαντικό ρόλο στην ανάπτυξη Αλτσχάιμερ. Υπάρχουν μάλιστα δύο κατηγορίες γονιδίων που επηρεάζουν την ανάπτυξη Αλτσχάιμερ.

- Γονίδια ρίσκου (αναπτύσσονται στην παιδική ηλικία)
- Ντετερμινιστικά Γονίδια (άμεση αιτία της αρρώστιας)

Παθοφυσιολογία

Το Αλτσχάιμερ θεωρείται μια ιδιαίτερα περίπλοκη ιατρική περίπτωση καιόπως έχει ειπωθεί πολλές φορές τα αίτια εμφάνισης του παραμένουν άγνωστα.

9.4 Βιοχημεία

Η νόσος αναγνωρίζεται ως πρωτεϊνικά λανθασμένη αναδίπλωση (Protein misfolding) λόγω της λανθασμένα αναδιπλωμένης πρωτεΐνης αμυλοειδές βήτα (β -amyloid). Η πρωτεΐνη αυτή είναι ένα μικρό πεπτίδιο που αποτελεί ένα μη φυσιολογικό παραπροϊόν της πρόδρομης πρωτεΐνης (APP), της οποίας οι λειτουργίες δεν είναι ξεκάθαρες, αλλά πιστεύεται ότι εμπλέκεται στην ανάπτυξη των νευρώνων. Τα θραύσματα αμυλοειδούς τα οποία είναι κολλώδη, συσσωρεύονται μαζί και σχηματίζουν πλάκες (νευριτικές πλάκες). Αυτές οι πλάκες εμποδίζουν την επικοινωνία μεταξύ των κυττάρων, η οποία παρεμπόδιση φαίνεται να προκαλεί ανοσολογικές αντιδράσεις που οδηγούν σε προγραμματισμένο κυτταρικό θάνατο των μη λειτουργικών νευρώνων.

Οι αμυλοειδικές πλάκες είναι “ένα σήμα κατατεθέν χαρακτηριστικού μιας παθοφυσιολογικής διάγνωσης του Αλτσχάιμερ.” [16] και αντικατοπτρίζουν σε βιοδείκτες οι οποίοι είναι εφικτό να εντοπίσουν και να ποσοτικοποιήσουν την συσσώματωση του αμυλοειδούς στον ανθρώπινο εγκέφαλο. Η πρωτεΐνη μπορεί να μετρηθεί άμεσα στο εγκεφαλονωτιαίο υγρό (CSF) και στο πλάσμα. Η

τομογραφία εκπομπής ποζιτρονίων (PET) είναι επίσης δυνατόν να χρησιμοποιηθεί. Τα κλινικά κριτήρια της ήπιας γνωστικής εξασθένησης που οφείλεται στην κατάσταση του Αλτσχάιμερ είναι ακόλουθη της συσχέτισης του αμυλοειδούς:

“Τα τρέχοντα στοιχεία υποδεικνύουν ότι οι δείκτες αμυλοειδούς παθολογίας (PET και CSF) προηγούνται των ενδείξεων νευρολογικής βλάβης. Αυτό δεν αποδεικνύει ότι η Αβ είναι αρχικός παράγοντας για την ασθένεια. Ωστόσο, προτείνει ότι αυτές οι διαφορετικές κατηγορίες βιοδεικτών φαίνεται να παρέχουν διαφορετικά είδη πληροφοριών σχετικά με την εξέλιξη της νόσου στον εγκέφαλο” [16]

Το Αλτσχάιμερ επίσης θεωρείται μια ταυτοπάθεια εξαιτίας της μη φυσιολογικής συσσωμάτευση πρωτεΐνης T (tau-ταφ), ένας μικρό σωλήνας συσχετισμένος με την πρωτεϊνική έκφραση σε νευρώνες η οποία σε φυσιολογική κατάσταση αντιδρά με τρόπο τέτοιο ώστε οι μικρό σωλήνες στον κυτταροσκελετό κυττάρων να βρίσκονται σε μια σταθερή κατάσταση. Συνήθως οι T πρωτεΐνη κάνουν τους μικροσωληνίσκους ίσιους, γεγονός που επιτρέπει στα μόρια να περάσουν ελεύθερα. Στην περίπτωση του Αλτσχάιμερ ωστόσο η πρωτεΐνη καταρρέει σε στριμμένα σκέλη, όπου και οι σωλήνες αποσυντίθενται εμποδίζοντας την τις θρεπτικές ουσίες να φτάσουν τα νευρικά κύτταρα, γεγονός που οδηγεί στον θάνατο των κυττάρων.

Τα κριτήρια MCI παρουσιάζουν διαφωνία στο ότι η Ταφ και η φωσφορυλιωμένη-ταφ έχουν την δυνατότητα να αντικατοπτρίσουν την γενική βλάβη στους νευρώνες και στις συνάψεις.

Η συσσωμάτωση των β-αμυλωειδών πλακών και των νευροϊνιδικών πλεγμάτων καταστρέφει την δομή του εγκεφάλου πιο συγκεκριμένα οδηγούν στην απώλεια των νευρώνων και των συνάψεων, η οποία έλλειψη οδηγεί στην απώλεια μνήμης, προκαλώντας επίσης κι άλλες βλάβες γνωστικού τύπου.

Παρόλα αυτά οι πλάκες και τα πλέγματα έχουν επίσης βρεθεί στους εγκεφάλους μεμονωμένων ανθρώπων που κατά την διάρκεια ζωής τους δεν είχαν δείξει σημάδια ότι πάσχουν από Αλτσχάιμερ.

9.5 Νευροπαθολογία

Τα συμπτώματα που προκύπτουν κατά την διάρκεια της ανάπτυξης του Αλτσχάιμερ προκαλούνται από αλλαγές που λαμβάνουν χώρα στον εγκέφαλο του ασθενή. Ο ιππόκαμπος του ασθενή που πάσχει από την νόσο, παρουσιάζει ατροφία σε πρώιμο στάδιο του Αλτσχάιμερ και η ανίχνευση αυτού του φαινομένου μπορεί να πραγματοποιηθεί αξιόπιστα με χρήση ογκομετρικής μαγνητικής τομογραφίας, που επιλέγεται για διαγνωστικούς σκοπούς [15]. Η απώλεια των νευρώνων και των συνάψεων λόγω της παρουσίας της νόσου οδηγεί σε εμφανή διαφορές στον ιστό του εγκεφάλου.

Το γεγονός ότι εκφυλίζεται ο εγκεφαλικός φλοιός και ο ιππόκαμπος οδηγεί σε φλοιώδη ατροφία σε κροταφικές, μετωπιαίες και βρεγματικές περιοχές.

Οι μικροσκοπικές αλλαγές στον εγκέφαλο ξεκινούν πολύ πριν τα πρώτα σημάδια της απώλειας μνήμης, με μια πρόσφατη έρευνα να διαπιστώνει την παρουσία μεσοποριακών αλλοιώσεων έως και 5,6 έτη πριν την κλινική διάγνωση του Αλτσχάιμερ. [17] Στην ίδια μελέτη αναφέρεται ότι υπήρχε απουσία ατροφίας στους μετωπικούς λοβούς, γεγονός που ήταν σύμφωνο με μελέτες που έδειχναν ότι είναι ευκολότερα επηρεαζόμενοι από τον χρόνο κατά τον οποίο γίνεται η διάγνωση.

9.6 Βιοδείκτες

Όσο η νόσος του Αλτσχάιμερ προχωράει με τον χρόνο, τα μεγέθη των βιοδεικτών φτάσουν σε μη λογικά επίπεδα σε μια προβλέψιμη σειρά.

Στην παρακάτω εικόνα εμφανίζονται οι βιοδείκτες ως ενδείξεις άνοιας, οι καμπύλες υποδεικνύουν τις αλλαγές που προκλήθηκαν από 5 βιοδείκτες που μελετήθηκαν.

1. Απεικόνιση βήτα αμυλοειδούς ανιχνεύεται σε απεικόνιση αμυλοειδούς σε CSF και PET αμυλοειδούς απεικόνισης.
2. Νευροεκφυλισμός ανιχνεύθηκε με άνοδο των ειδών tau-CSF και της συναπτικής δυσλειτουργίας μετρούμενη μέσω FDG-PET.
3. Η ατροφία του εγκεφάλου και η απώλεια νευρώνων μετρούνται με μαγνητική τομογραφία (κυρίως ανιχνεύονται στον ιππόκαμπο, στον πυρήνα του αυχένα και στον κροταφικό λοβό)
4. Απώλεια μνήμης μετρούμενη από την γνωστική εκτίμηση
5. Γενική γνωστική άρνηση μετρούμενη από την γνωστική εκτίμηση

Οι πρώτοι τρεις βιοδείκτες στην παραπάνω λίστα μπορούν να παρατηρηθούν πριν την διάγνωση της άνοιας, ενώ οι δύο τελευταίοι είναι οι κλασσικές ενδείξεις της διάγνωσης άνοιας όπως αναφέρεται στο [18]

Για να μπορεί να επιτευχθεί η παροχή σωστής θεραπείας όπου είναι δυνατόν, οι βιοδείκτες περιλαμβάνονται και σε εκείνο το πλαίσιο διάγνωσης, αν και ουσιαστικά είναι για να χρησιμοποιηθούν στην έρευνα. Εδώ περιλαμβάνονται και βιοδείκτες που αντικατοπτρίζουν άμεσα την παθολογία της νόσου του Αλτσχάιμερ, όπως η πρωτεΐνη βήτα αμυλοειδούς Αβ και ταφ, καθώς και βιοδείκτες που παρέχουν λιγότερο άμεσες ή μη ειδικές ενδείξεις της νόσου, παρακολουθώντας δείκτες νευρωνικής βλάβης, οι οποίοι είναι κάπως συγκεκριμένοι για την νόσο λόγω περιφερειακού προτύπου ανωμαλιών. Ενώ πρωταρχικά συνδυάζονται με το Αλτσχάιμερ, αυτοί οι βιοδείκτες μπορούν να παρατηρηθούν και για άλλες ασθένειες. Όταν οι δύο παρατηρηθούν στο ίδιο αντικείμενο μελέτης υπάρχουν πολλές πιθανότητες το αντικείμενο να πάσχει από Αλτσχάιμερ.

9.7 Συμπτώματα

Τα συμπτώματα του Αλτσχάιμερ τυπικά μπορεί να πει κανείς ότι ξεκινούν όταν ο ασθενής αδυνατεί να θυμηθεί πληροφορίες που μαθαίνει στην καθημερινότητα του και συνεχίζουν σε ολοένα και πιο έντονα συμπτώματα όπως ο αποπροσανατολισμός, οι απότομες αλλαγές διάθεσης και συμπεριφοράς, η έντονη σύγχυση σχετικά με τα γεγονότα, τον χρόνο και τον τόπο, αβάσιμες υποψίες για την οικογένεια, τους φίλους και τους επαγγελματίες φροντιστές. Επίσης καταλήγει σε πιο σοβαρές απώλειες μνήμης και αλλαγές συμπεριφοράς, δυσκολία ακόμη και στην ομιλία, την κατάποση και το περπάτημα, όσο η νόσος προχωράει. Τα συμπτώματα συνήθως δεν είναι εμφανή τόσο στους πάσχοντες όσο στους κοντινούς ανθρώπους του.

Ο καταστροφικός συνδυασμός των πλακών και των πλεγμάτων αρχίζει από τον ιππόκαμπο ο οποίος έχει την λειτουργία να δημιουργία αναμνήσεων. Η βραχυπρόθεσμη απώλεια μνήμης είναι συνήθως το πρώτο σύμπτωμα του Αλτσχάιμερ. Οι πρωτεΐνες τότε επεκτείνονται και εισβάλλουν σταδιακά σε άλλα μέρη του εγκεφάλου, δημιουργώντας μοναδικές αλλαγές, σηματοδοτώντας ποικίλες φάσεις την νόσου.

Στο μπροστινό κομμάτι του εγκεφάλου, καταστρέφεται η ικανότητα επεξεργασίας των λογικών σκέψεων. Έπειτα μειώνεται ο έλεγχος των συναισθημάτων όπου και από 'κει προκύπτουν οι απότομες και χωρίς λόγο αλλαγές διάθεσης που παρατηρούνται στον ασθενή. Στην συνέχεια οι πρωτεΐνες φτάνουν στο πάνω κομμάτι του εγκεφάλου και ο ασθενής οδηγείται σε καταστάσεις παράνοιας και σε παραισθήσεις. Μετά στο πίσω μέρος του εγκεφάλου θα φτάσουν οι πρωτεΐνες και ο συνδυασμός τους θα οδηγήσει στο να διαγραφτούν οι βαθύτερες αναμνήσεις του νου. Τέλος τα κέντρα ελέγχου του εγκεφάλου που είναι υπεύθυνα για τον καρδιακό παλμό και την αναπνοή εξουδετερώνονται με αποτέλεσμα των θάνατο.

9.8 Διάγνωση Αλτσχάιμερ

Η πιο ασφαλής μέθοδος που εφαρμόζεται για την διάγνωση της νόσου είναι η αυτοψία του εγκεφαλικού ιστού και των χαρακτηριστικών παθολογοανατομικών ευρημάτων της νόσου. Το πρόβλημα με την συγκεκριμένη διαδικασία διάγνωσης είναι ότι αυτή μπορεί να πραγματοποιηθεί μόνο μετά το θάνατο του ασθενή και την καθιστά μη πρακτική για το ίδιο τον ασθενή.

Ωστόσο είναι εφικτή η διάγνωση μέσω κλινικής εξέτασης από νευρολόγο αξιοποιώντας το ιστορικό του ασθενή, με νευροψυχολογικό έλεγχο, εξέταση μνήμης και άλλων λειτουργιών όπως είναι η μάθηση, η προσοχή και η παρακολούθηση των καθημερινών δραστηριοτήτων του αντικειμένου. Η κυριότερη εξέταση που χρησιμοποιείται είναι η MMSE (Mini Mental State Examination), που μελετά με απλό και γρήγορο τρόπο χαρακτηριστικά όπως η προσοχή, ο προσανατολισμός, η ικανότητα παρακολούθησης και η γλώσσα του ασθενούς. Άλλες εξετάσεις που γίνονται είναι το CDS (Clinical Dementia Rate) και GDS (Global Deterioration Scale). Επίσης, γίνεται εξέταση αίματος, έλεγχος της βιταμίνης 12, της θυρεοειδούς λειτουργίας και της ομοκυστεΐνης, παράγοντες που έχουν σχέση με την λειτουργία της μνήμης. Ωστόσο πρέπει να εξεταστεί και η ατροφία του ιππόκαμπου που είναι υπεύθυνη για την μνήμη, διάγνωση που γίνεται με Μαγνητική τομογραφία (MRI) και Αξονική τομογραφία (CT).

Τα τελευταία χρόνια, έχει επιτευχθεί η μετάβαση από έναν μοναδικό βιοδείκτη σε μια προσέγγιση πολλαπλών βιοδεικτών. Σύμφωνα με τον Di Luca, η ταυτόχρονη χρήση διάφορων βιοδεικτών, δίνει την δυνατότητα ώστε κάθε βιοδείκτης να έχει ως στόχο την μελέτη και διερεύνηση ενός συστατικού, πιο συγκεκριμένα ενός μοναδικού σύνθετου παθο-φυσιολογικού μονοπατιού. Αυτή η μέθοδος έχει το εγγενές πλεονέκτημα για να συλλάβει περισσότερες πληροφορίες που συνδέονται με την ασθένεια υπό μελέτη, σε σύγκριση με την προσέγγιση ενός μεμονωμένου βιοδείκτη. Η ταυτόχρονη χρήση διαφορετικών παραγόντων μπορεί να μειώσει τον κίνδυνο τυχαίας μεταβολής του μεμονωμένου παράγοντα καλύπτοντας το αληθινό σήμα. Τα νευρωνικά δίκτυα μπορούν να εισάγουν πολλαπλούς παράγοντες ταυτόχρονα, συνδυάζοντας και ανασυνδυάζοντάς τα μέσα με διαφορετικούς τρόπους σύμφωνα με συγκεκριμένες (γενικά μη γραμμικές) εξισώσεις. Έτσι, οι υψηλότερες προγνωστικές τιμές που λαμβάνονται από τα ΤΝΔ μπορούν να εξηγηθούν λόγω της εξέτασης των παραμέτρων που μπορεί να μην φτάσουν σε σημαντικό βαθμό για ολόκληρο τον πληθυσμό, αλλά να είναι πολύ σημαντικές εντός των υποομάδων. Από την άλλη πλευρά, το ότι οι συμβατικές στατιστικές αποκαλύπτουν μόνο τις παραμέτρους που είναι σημαντικές για ολόκληρο τον πληθυσμό. Δεδομένα από διάφορες μελέτες έχουν επισημάνει την ύπαρξη μιας ισχυρής συσχέτισης μεταξύ της νευροπαθολογίας της νόσου του Alzheimer (AD) και της νοητικής κατάστασης. Ωστόσο, λόγω της εξαιρετικά πολύπλοκης και μη γραμμικής τους σχέσης, είναι δύσκολο να αναπτυχθεί ένα πρότυπο πρόβλεψης για την ταξινόμηση μεμονωμένων ασθενών μέσω των παραδοσιακών προσεγγίσεων που πραγματοποιούνται με χρήση στατιστικών στοιχείων. Όταν εκτίθενται σε σύνθετα σύνολα δεδομένων, τα τεχνητά νευρικά δίκτυα

(ΤΝΔ) μπορούν να αναγνωρίσουν πρότυπα, να μάθουν τη σχέση των διαφόρων μεταβλητών και την επίλυση καθηκόντων ταξινόμησης.

Επιπρόσθετα κάποια νέα διαγνωστικά κριτήρια εισάχθηκαν για την Ήπια Γνωστική Βλάβη (MCI), όπως για την άνοια χάρη στην μελέτη της νόσου του Αλτσχάιμερ το 2011, από ειδικές διεθνή ομάδες εργασίας που συγκάλεσε η Alzheimer 's Association και το National Institute on Aging (μια υπηρεσία της U.S. National Institute of Health – NIH). Ο λόγος της σύγκλισης της υπηρεσίας οφείλεται στην αυξανόμενη επικρατούσα άποψη στο ιατρικό γίνεσθαι ότι η νόσος περιλαμβάνει μια φάση κατά την οποία τα άτομα που πλήττονται από την νόσο βιώνουν μια βαθμιαία επιδείνωση των γνωστικών ικανοτήτων τους που προκύπτει από τη συσσωμάτωση της παθολογίας της νόσου του Αλτσχάιμερ στον εγκέφαλο τους. Ο ασθενής διαγιγνώσκεται με άνοια-Αλτσχάιμερ όταν τα συμπτώματα είναι επαρκώς έντονα. Οι ειδικοί γράψαν ότι : “ καθώς το Αλτσχάιμερ αποτελεί μια αργή, προοδευτική διαταραχή, χωρίς σταθερά γεγονότα που καθορίζουν την εκδήλωση της, είναι ιδιαίτερα δύσκολο για τους ιατρούς να εντοπίσουν τα σημεία μετάβασης μεμονωμένων ασθενών” [24] και επομένως είναι δύσκολο να εντοπιστεί το σημείο στον χρόνο κατά το οποίο η ασθένεια σε έναν συγκεκριμένο ασθενή διαμορφώνει τις μεταβάσεις, για παράδειγμα μεταβάσεις τύπου από την ασυμπτωματική φάση στη φάση της συμπτωματικής προ-αιτιολογίας ή από τη φάση της συμπτωματικής προ-αιτιολογίας στην έναρξη της άνοιας.

Κατά το Αλτσχάιμερ προκαλούνται σε μεγάλο βαθμό αλλαγές στην δομή αλλά και στην λειτουργία του εγκεφάλου. Αυτές οι αλλαγές έχουν διαγνωστική και προγνωστική αξία τόσο στην άνοια όσο και στην MCI και από ότι φαίνεται αντικατοπτρίζουν την ζημιά που έχουν υποστεί οι νευρώνες και οι συνάψεις. Τα διαγνωστικά κριτήρια “δηλώνουν” ότι : ”Πολλές από αυτές τις αλλαγές έχουν τοπογραφική ακρίβεια για την νευρολογική ζημιά ή δυσλειτουργία που έχει προκαλέσει η ύπαρξη του Αλτσχάιμερ” [16]. Αυτοί οι νευρωνικοί τραυματισμοί μπορούν να μετρηθούν με δομικά και/ή λειτουργικά μέτρα, συμπεριλαμβανομένων των απεικονίσεων MRI, PET και SPECT.

Μερικά δομικά και λειτουργικά πρότυπα είναι χαρακτηριστικά και του Αλτσχάιμερ. Η έλλειψη όγκου από τον ιππόκαμπο, για παράδειγμα που μπορεί να εντοπιστεί μέσω MRI και η μείωση της γλυκόζης του μεταβολισμού ή η διάχυση στον μετεγχειρητικό φλοιό που μπορούν να παρατηρηθούν μέσω σάρωσης PET ή SPECT αποτελούν τέτοια χαρακτηριστικά κριτήρια. Όπως υπογράμμισαν και οι ειδικοί: “Αν και αυτοί οι βιοδείκτες έχουν συσχετιστεί με την νευροπαθολογία του Αλτσχάιμερ, η περιφερειακή ατροφία, η ολική ατροφία, ο περιφερειακός υπομεταβολισμός και η υποδιήθηση δεν είναι συγκεκριμένα για το Αλτσχάιμερ. Τα μέτρα αυτά φαίνεται να παρέχουν στοιχεία σχετικά με το στάδιο ή τη σοβαρότητα της ασθένειας που μπορεί να μην παρέχεται από βιοδείκτες Αβ” [16]. Υπάρχει επίσης το ζήτημα των βιοχημικών γεγονότων που χαρακτηρίζουν την νόσο του Αλτσχάιμερ, συμπεριλαμβανομένου του οξειδωτικού στρες και της φλεγμονής.

10 Συμπέρασμα

Στην παρούσα πτυχιακή εργασία μελετήθηκε η σύμπραξη των επιστημονικών πεδίων της Ιατρικής και της Βιοπληροφορικής. Έγινε εκτενή αναφορά σε συγκεκριμένα εργαλεία προγραμματισμού και πληροφορικής που εφαρμόζονται στην Ιατρική. Κύρια μελέτη έγινε πάνω στα δεδομένα που προκύπτουν από Ιατρικές έρευνες ασθενειών και όχι μόνο και πως μπορούν αυτά να χρησιμοποιηθούν από προγραμματιστικά εργαλεία, ώστε να βοηθήσουν σε αυτές τις έρευνες, ή και ακόμα στην καθημερινή εφαρμογή της Ιατρικής, όπως και αναφέρθηκε πολλές φορές, για παράδειγμα στην διάγνωση μιας ασθένειας.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- [1] Ιατρική Πληροφορική. (2018). Ιατρική Πληροφορική. [online] Available at: <https://iwannakop.wordpress.com/> [Accessed 27 Sep. 2018]
- [2] Strimbu, K. and Tavel, J. (2018). What are biomarkers?.
- [3] Medeiros, F. (2018). Biomarkers and Surrogate Endpoints: Lessons Learned From Glaucoma.
- [4] Cancer Kaiser J. First pass at cancer genome reveals complex landscape. Science. 2016;313:1370.
- [5] Chan T.A., Wolchok J.D., Snyder A. Genetic basis for clinical response to CTLA-4 blockade in melanoma. N Engl J Med. 2015;373:1984
- [6] Gpb.big.ac.cn. (2018). [online] Available at: <http://gpb.big.ac.cn/articles/download/596> [
- [7] Nemertes.lis.upatras.gr. (2018). [online] Available at: http://nemertes.lis.upatras.gr/jspui/bitstream/10889/1208/1/Nimertis_Dimitrakopoulou.pdf
- [8] Wscg.zcu.cz. (2018). [online] Available at: <http://wscg.zcu.cz/wscg2016/short/A43-full.pdf>
- [9] Wires.wiley.com. (2018). Machine learning for bioinformatics and neuroimaging - WIREs Data Mining and Knowledge Discovery. [online] Available at: <http://wires.wiley.com/WileyCDA/WiresArticle/wisId-WIDM1248.html>
- [10] Artificial Neural Networks in the Discrimination of Alzheimer's disease Using Biomarkers Data
- [11] Alzheimer's Disease & Dementia | Alzheimer's Association. url: http://www.alz.org/alzheimers_disease_what_is_alzheimers.asp (visited on 04/02/2014).
- [12] Ron Brookmeyer et al. "Forecasting the global burden of Alzheimer's disease". In: Alzheimer's & dementia: the journal of the Alzheimer's Association 3.3 (July 2007), pp. 186 – 191. issn: 1552-5279. doi: 10.1016/j.jalz.2007.04.381.
- [13] Anders Wimo and Martin Prince. World Alzheimer Report 2010: The Global Economic Impact of Dementia. 2. Available at <http://www.alz.co.uk/research/files/WorldAlzheimerRepo> Alzheimer's Disease International (ADI), Sept. 21, 2010. url: <http://www.alz.co.uk/research/world-report-2010>.
- [14] P Tiraboschi et al. "The importance of neuritic plaques and tangles to the development and evolution of AD". In: Neurology 62.11 (June 8, 2004), pp. 1984 – 1989. issn: 1526 - 632X.
- [15] M P Laakso et al. "Hippocampal volumes in Alzheimer's disease, Parkinson's disease with and without dementia, and in vascular dementia: An MRI study". In: Neurology 46.3 (Mar. 1996), pp. 678 – 681. issn: 0028 - 3878.
- [16] Marilyn S Albert et al. "The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease". In: Alzheimer's & dementia: the journal of the Alzheimer's Association 7.3 (May 2011), pp. 270 – 279. issn: 1552-5279. doi: 10.1016/j.jalz.2011.03.008.
- [17] Charlotte Bernard et al. "Time course of brain volume changes in the preclinical phase of Alzheimer's disease". In: Alzheimer's & Dementia 10.2 (Mar. 2014), 143–151.e1. issn: 15525260. doi: 10.1016/j.jalz.2013.08.279. url: <http://ezproxy.hiof.no:2162/pubmed/?term=Time+course+of+brain+volume+changes+in+the+preclinical+phase+of+Alzheimer%27s+disease> (visited on 04/23/2014).

- [18] ADNI | Background & Rationale. url: <http://adni.loni.usc.edu/study-design/background-rationale/>.
- [19] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy. "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis." *Bioinformatics*, Vol. 21, Issue 5, pp. 631 - 643, Mar. 2005.
- [20] B. Twala, M. Phorah. "Predicting incomplete gene microarray data with the use of supervised learning algorithms." *Pattern Recognition Letters*, In Press, Corrected Proof, Available online 11 May 2010.
- [21] M. Z. Man, G. Dyson, K. Johnson, B. Liao. "Evaluating Methods for Classifying Expression Data." *Journal of Biopharmaceutical Statistics*, Vol. 14, Issue 4, pp. 1065-1086, Nov. 2004.
- [22] Y. Lu, J. Han. "Cancer classification using gene expression data." *Information Systems* Vol. 28, Issue 4, pp. 243 - 268, Jun. 2003.
- [23] J. R. Quinlan. "Induction of Decision Trees." *Machine learning*, Vol. 1, Issue 1, pp. 81-106, 1986.
- [24] J. R. Quinlan. "Simplifying decision trees." *International Journal of Man-Machine Studies*, Vol. 27, Issue 3, pp. 221 - 248, Sep. 1987.
- [25] R. Kohavi, J. R. Quinlan. "Decision-tree discovery," in *Handbook of Data Mining and Knowledge Discovery*, W. Klossen and J. M. Zytkow, Eds. Oxford: Oxford University Press, 2002, pp. 267 - 276.
- [26] L. Rokach, O. Z. Maimon. *Data mining with decision trees: theort and applications*. Singapore: World Scientific, 2008
- [27] American Association for Artificial Intelligence. *The Thirteenth National Conference on Artificial Intelligence*, August 4-8, 1996, Portland, Oregon, USA. Menlo Park, CA: AAAI Press; Cambridge, MA: MIT Press, 1996.
- [28] Hood L, Flores M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *N Biotechnol*. 2012;29(6):613–24. [PubMed]
- [29] Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet*. 2010;11(1):75 – 87. [PMC free article] [PubMed]
- [30] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57 – 63. [PMC free article] [PubMed]
- [31] Farnham PJ. Insights from genomic profiling of transcription factors. *Nat Rev Genet*. 2009;10(9):605 – 16. [PMC free article] [PubMed]
- [32]. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet*. 2009;10(10):669 – 80. [PMC free article] [PubMed]
- [33] Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet*. 2010;11(1):31 – 46. [PubMed]
- [34]Beyer A, Bandyopadhyay S, Ideker T. Integrating physical and genetic maps: from genomes to interaction networks. *Nat Rev Genet*. 2007;8(9):699 – 710. [PMC free article] [PubMed]
- [35] Laird PW. Principles and challenges of genomewide DNA methylation analysis. *Nat Rev Genet*. 2010;11(3):191 – 203. [PubMed]
- [36]Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. *Nat Rev Genet*. 2010;11(7):476 – 86. [PMC free article] [PubMed]
- [37]De Roock W, Claes B, Bernasconi D, et al. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. *Lancet Oncol*. 2010;11(8):753 – 62. [PubMed]

- [38] Canuel, V, et al. "Translational Research Platforms Integrating Clinical and Omics Data: a Review of Publicly Available Solutions." Briefings in Bioinformatics., U.S. National Library of Medicine, Mar. 2015, www.ncbi.nlm.nih.gov/pubmed/24608524.
- [39] Yan-yan SONG, Y. (2018). Decision tree methods: applications for classification and prediction. [online] PubMed Central (PMC). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/>
- [40] Παντελής Ι. Μπάγκος, (2015) Βιοπληροφορική, Πανεπιστήμιο Θεσσαλίας Available at: eclass.uth.gr/eclass/modules/document/file.../Lecture8-pHMMs.pdf
- [41] Κεφάλαιο 2: Βιολογία Συστημάτων, Τεχνολογίες Omics
Γεώργιος Θεοδωρίδης Available at: <https://repository.kallipos.gr/handle/11419/3668>
- [42] RAMLI, M. R., MALEK, S. AND MILOW, P.
Application of Self Organizing Map, SAHN clustering and spatial analysis on the diversity of plant resource in homegardens: Case study in Perak, Malaysia
In-text: (Ramli, Malek and Milow, 2018)
Your Bibliography: Ramli, M., Malek, S. and Milow, P. (2018). Application of Self Organizing Map, SAHN clustering and spatial analysis on the diversity of plant resource in homegardens: Case study in Perak, Malaysia.
- [43] J. Vesanto. SOM-based data visualization methods. *Intelligent Data Analysis*, 3(2):111 – 126, 1999
- [44] Pooja, B., Kulvinder, S.: Big data mining: analysis of genetic K-Means algorithm for big data clustering. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* 6(7), 223 – 228 (2016)
- [45] Tsirigos, K., Bagos, P. and Hamodrakas, S. (2010). OMPdb: a database of α -barrel outer membrane proteins from Gram-negative bacteria. *Nucleic Acids Research*, 39(Database), pp.D324 - D331.
- [46] Cazy.org. (2019). CAZY - Home. [online] Available at: <http://www.cazy.org/> [Accessed 4 Sep. 2019].
- [47] Ebi.ac.uk. (2019). MEROPS - the Peptidase Database. [online] Available at: <https://www.ebi.ac.uk/merops/cgi-bin/pepsum?id=S09.003> [Accessed 4 Sep. 2019].
- [48] Krupa, A. (2004). KinG: a database of protein kinases in genomes. *Nucleic Acids Research*, 32 (90001), pp.153D - 155.
- [49] Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2 - 41 - 6 Aomi, Koto-ku, Tokyo 135-0064, Japan. n.nagano@aist.go.jp
- [50] Ebi.ac.uk. (2019). Identifiers.org < EMBL-EBI. [online] Available at: <https://www.ebi.ac.uk/miriam/main/datatypes/MIR:00000077> [Accessed 5 Sep. 2019]
- [51] Lenfant, N., Hotelier, T., Velluet, E., Bourne, Y., Marchot, P. and Chatonnet, A. (2012). ESTHER, the database of the α/β -hydrolase fold superfamily of proteins: tools to explore diversity of functions. *Nucleic Acids Research*, 41(D1), pp.D423 - D429.
- [52] Kaas, Q., Westermann, J., Halai, R., Wang, C. and Craik, D. (2007). ConoServer, a database for conopeptide sequences and structures. *Bioinformatics*, 24(3), pp.445 - 446.
- [53] Cybase.org.au. (2019). Cybase. [online] Available at: <http://www.cybase.org.au/> [Accessed 6 Sep. 2019].
- [54] Gpcrdb.org. (2019). GPCRdb. [online] Available at: <https://gpcrdb.org/> [Accessed 6 Sep. 2019].
- [55] Alexander SPH, Kelly E, Marrion NV, Peters JA, Faccenda E, Harding SD, Pawson AJ, Sharman JL, Southan C, Buneman OP, Cidlowski JA, Christopoulos A, Davenport AP, Fabbro D, Spedding M, Striessnig J, Davies JA; CGTP Collaborators. (2017) The Concise Guide to PHARMACOLOGY 2017/18. *Br J Pharmacol.* 174 (Suppl 1): S1 - S446
- [56] Kinase.com. (2019). Protein Kinases, Kinomes and Evolution, at Kinase.Com. [online] Available at: <http://kinase.com/> [Accessed 6 Sep. 2019].
- [57] Sfld.rbvi.ucsf.edu. (2019). SFLD - About. [online] Available at: <http://sfld.rbvi.ucsf.edu/archive/django/web/about/index.html> [Accessed 6 Sep. 2019].
- [58] Sullivan, S. (2000). The Histone Database: a comprehensive WWW resource for histones and histone fold-containing proteins. *Nucleic Acids Research*, 28(1), pp.320 - 322.
- [59] Robinson, A., S. C. Punitha, S. and Thangaiyah, D. (2011). Hybrid Attribute Selection Process for Decision Tree Based Classification Algorithms. *Indian Journal of Applied Research*, 1(9), pp.34 - 36.
- [60] Ncbi.nlm.nih.gov. (2019). GenBank Overview. [online] Available at: <https://www.ncbi.nlm.nih.gov/genbank/> [Accessed 10 Sep. 2019].
- [61] Stoesser, G. (2001). The EMBL nucleotide sequence database. *Nucleic Acids Research*, 29(1), pp.17 - 21.
- [62] Tateno, Y. (1998). DNA Data Bank of Japan at work on genome sequence data. *Nucleic Acids Research*, 26(1), pp.16 - 20.
- [62] Repository.kallipos.gr. (2019). [online] Available at: https://repository.kallipos.gr/bitstream/11419/3382/1/02_chapter_04.pdf [Accessed 19 Sep. 2019].

- [63] Chang, M. and Chang, M. (2017). iWordNet: A New Approach to Cognitive Science and Artificial Intelligence. *Advances in Artificial Intelligence*, 2017, pp.1-10.
- [64] Yu, D. (2015). Softmax function based intuitionistic fuzzy multi-criteria decision making and applications. *Operational Research*, 16(2), pp.327-348.
- [65] J. Zupan, J. Gasteiger, *Neural Networks for Chemists*, VCH, New York, 1993.
- [66] Svozil, D., Kvasnicka, V. and Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems*, 39(1), pp.43-62.
- [67] Qeethara Kadhim Al-Shayea MIS Department, Al-Zaytoonah University of Jordan Amman, Jordan (2011). *Artificial Neural Networks in Medical Diagnosis*.