ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΕΡΓΑΣΤΗΡΙΟ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ

Π.Μ.Σ. «ΒΙΟΛΟΓΙΑ ΣΥΣΤΗΜΑΤΩΝ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Εξόρυξη Βιοτεχνολογικών Δεδομένων με Μεθόδους Βιοπληροφορικής

ΓΙΩΡΓΟΣ Α. ΠΑΠΑΗΛΙΟΥ

ΑΘΗΝΑ 2018

Επιβλέπων:

Καθηγητής Χατζόπουλος Πολυδεύκης

ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΕΡΓΑΣΤΗΡΙΟ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ

Π.Μ.Σ. «ΒΙΟΛΟΓΙΑ ΣΥΣΤΗΜΑΤΩΝ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Εξόρυξη Βιοτεχνολογικών Δεδομένων με Μεθόδους Βιοπληροφορικής

ΓΙΩΡΓΟΣ Α. ΠΑΠΑΗΛΙΟΥ

ΑΘΗΝΑ 2018

Επιβλέπων:

Καθηγητής Χατζόπουλος Πολυδεύκης

Η ανάθεση της παρούσας μεταπτυχιακής μελέτης έγινε με απόφαση της Γενικής Συνέλευσης του Τμήματος Βιοτεχνολογίας κατά την οποία εγκρίθηκαν το θέμα και η τριμελής συμβουλευτική και εξεταστική επιτροπή της μελέτης.

Τύπος εργασίας: Μεταπτυχιακή Εργασία

Τίτλος εργασίας: Εξόρυξη Βιοτεχνολογικών Δεδομένων με Μεθόδους Βιοπληροφορικής

Ον/μο: Γιώργος Α. Παπαηλιού

Εξεταστική επιτροπή:

1. Καθηγητής Χατζόπουλος Πολυδεύκης (Επιβλέπων)

2. Καθηγητής Κωνσταντίνος Γιαλούρης (μέλος)

3. Αναπληρωτής Καθηγητής Εμμανουήλ Φλεμετάκης (μέλος)

Ευχαριστίες

Ευχαριστώ την εξεταστική επιτροπή, τους κυρίους Εμμανουήλ Φλεμετάκη, Πολυδεύκη Χατζόπουλο και Κωνσταντίνο Γιαλούρη για την υπομονή που επέδειξαν κατά την διάρκεια της μελέτης και συγγραφής της παρούσας εργασίας.

Ιδιαίτερα θα ήθελα να ευχαριστήσω τον κύριο Γιαλούρη για την επιμέλεια και την επιμονή του στη διόρθωση της εργασίας.

Τέλος, ευχαριστώ θερμά τον κύριο Χατζόπουλο και την κυρία Κούρτη για την ηθική υποστήριξή τους.

Περιεχόμενα

Περίληψη

Η Βιοτεχνολογία αποτελεί τον κατ' εξοχήν κλάδο εφαρμογών της βιολογικής γνώσης. Οι εφαρμογές εμπλέκουν γνώση από άλλα πεδία ή επιστήμες, δημιουργώντας νέους κλάδους, με νέους στόχους. Παράδειγμα αποτελεί η Βιοπληροφορική, που αναλαμβάνει να οργανώσει και να διαχειριστεί το πληροφοριακό πλούτο των βιολογικών πειραμάτων, καθώς και να εντάξει στο βιολογικό πλαίσιο εργαλεία, μεθόδους, αλλά και τη νοοτροπία της Πληροφορικής. Ένα τέτοιο σύνολο εργαλείων και μεθόδων αποτελεί η Εξόρυξη από Δεδομένα, η οποία αναπτύσσει τη δυναμική της κυρίως στο πλαίσιο της Βιολογίας Συστημάτων.

Η παρούσα εργασία αποτελεί μια προσπάθεια εισαγωγής στις έννοιες, στις διαδικασίες και στις δυνατότητες της Εξόρυξης από Δεδομένα, τόσο ως προς το θεωρητικό υπόβαθρο όσο και ως προς τη τρέχουσα ερευνητική εφαρμογή της. Ως εκ τούτου αποτελεί μια εμπλουτισμένη θεωρητικά βιβλιογραφική ανασκόπηση.

Αρχικά, γίνεται μια σκιαγράφηση του εύρους εφαρμογών τις οποίες αναπτύσσουν οι κλάδοι της Βιοτεχνολογίας και Βιοπληροφορικής, καθώς και οι προεκτάσεις τους μέσω της Βιολογίας Συστημάτων. Παράλληλα, παρουσιάζονται και τα πρώτα προβλήματα ή σκοτεινά πεδία που φέρνουν στο φως οι νέες τεχνολογίες. Προβλήματα διαχείρισης, απεικόνισης ή διαλογής μεγάλου όγκου δεδομένων σχετίζονται σε πληροφοριακό επίπεδο με μεθοδολογίες Ταξινόμησης, Αναγνώρισης Συστάδων ή μοτίβων γενικότερα, καθώς και Εύρεσης Κανόνων Συσχέτισης.

Αφού γίνει μια περιεκτική παρουσίαση και ορισμός των βασικών εννοιών του Δεδομένου και της Δομής Δεδομένων, που βρίσκονται στον πυρήνα της Εξόρυξης από Δεδομένα, επιχειρείται μια αναλυτική παρουσίαση των βασικών κατηγοριών Εξόρυξης. Έτσι αναλύονται οι επιμέρους τεχνικές για κάθε κατηγορία Ταξινόμησης, Αναγνώρισης Συστάδων και Εύρεσης Κανόνων Συσχέτισης, σε λειτουργικό επίπεδο χωρίς δημοσιευμένα παραδείγματα. Επιπροσθέτως, και αφού παρουσιαστούν οι τεχνικές, γίνεται και η σύνδεσή τους με το πεδίο της

Εξόρυξης από Κείμενο, που αποτελεί κλάδο αιχμής κυρίως για την αυτοματοποιημένη διαχείρισης της επιστημονικής βιβλιογραφίας.

Τέλος, παρουσιάζονται παραδείγματα τεχνικών και αποτελέσματα από τη σύγχρονη βιβλιογραφία (2009-2017). Στην πλειονότητα των δημοσιεύσεων αυτών η διαδικασία Εξόρυξης έχει μερικό ή επικουρικό ρόλο. Η παρουσίαση των δημοσιεύσεων χωρίζεται βάσει του είδους των δεδομένων στα οποία έγινε η Εξόρυξη, τα οποία διακρίνονται σε πρωτεομικά, χημικο-γονιδιακά ή δεδομένα μικρο-συστοιχιών, ενώ γίνεται και μια ξεχωριστή κατηγοριοποίηση για της δημοσιεύσεις που επιστράτευσαν τεχνικές Ενσωματωμένης Εξόρυξης σε περιεχόμενο κειμένου. Από την κάθε δημοσίευση παρουσιάζονται τα σημαντικότερα αποτελέσματα που αφορούν στην Εξόρυξη από Δεδομένα, και συγκεκριμένα σχηματικές απεικονίσεις των δεδομένων ή των διαδικασιών τις οποίες υπέστησαν κατά την επεξεργασία.

Λέξεις-κλειδιά: Εξόρυξη από Δεδομένα, Εξόρυξη από Κείμενο, Ενσωματωμένη Εξόρυξη, Βιολογία Συστημάτων, Βιοπληροφορική, Βιοτεχνολογία, Ταξινόμηση, Συσταδοποίηση, Κανόνες Συσχέτισης, Πρωτεομικά Δεδομένα, Γονιδιακά Δεδομένα, Δεδομένα Μικρο-Συστοιχιών

Abstract

Biotechnology constitutes the primary scientific field through which biological knowledge is applied. Its application employs knowledge from existing fields to generate new ones with novel aims. Bioinformatics is an example of such a novel interdisciplinary field that is responsible for handling and organizing the complex output of biological experiments by introducing the tools, methodology and mentality of informatics in the biological framework. One such set of tools is Data Mining which is predominantly applied in the context of Systems Biology.

The present thesis attempts to introduce the concepts, processes and capabilities of Data Mining referring both to the theoretical background of the field and to its current research applications. The result is a theoretically enriched literature review.

Initially, the breadth of applications of biotechnology and bioinformatics and their extensions in the field of System Biology in particular are illustrated. At the same time, some emerging problems of these technologies are presented that include difficulties with handling, presentation or selection of great volume data (Big Data). The core of any Big Data problems relates to issues of classification, clustering and the discovery of correlation rules, which are methodologies adopted from informatics.

After defining core elements and concepts such as “data” and “data set”, an extensive presentation of the basic categories of mining is attempted. Each category is examined with respect to the techniques of classification, clustering and discovery of correlation rules that are examined on a functional level and without initially taking published examples into considerations. These particular categories and techniques are subsequently considered in the context of text mining processes, which are crucial for automated filtering of published scientific work.

Finally, examples of published techniques and outcomes of recent literature (2009-2017) that utilize Data Mining processes are presented. In their majority of these published examples the use of data mining has only a partial or subsidiary role. The published articles that are presented are sorted according to the nature

of the data that were mined, namely: proteomic data, chemical-genomic data or micro-array data. A separate section is dedicated to publications that used embedded data mining. From each publication only the most important data mining results are described and figures of these data or the procedures though which they were processed are presented.

Keywords: Data Mining, Text Mining, Embedded Data Mining, Systems Biology, Bio-Informatics, Biotechnology, Classification, Clustering, Correlation Rules, Proteomic Data, Genomic Data, Micro-Array Data

Εισαγωγή

Κατά γενική ομολογία ζούμε τις ημέρες της Πληροφορικής Επανάστασης, σε αντιστοιχία πάντα με τον όρο Βιομηχανική Επανάσταση. Η έννοια της Πληροφορίας ως οντότητα, είναι βασικό αντικείμενο με το οποίο ασχολείται ο άνθρωπος και αναμένεται η ενασχόλησή του με αυτήν να είναι αυξητική. Η καλύτερη ένδειξη τις αξίας της Πληροφορίας για την ανθρώπινη παραγωγικότητα είναι οι συνέπειες και η διαφορά που θα επιφέρει σε αυτήν, η υποτιθέμενη απουσία του Διαδικτύου (Word Wide Web) έστω και για λίγες ώρες από την παραγωγική διαδικασία (Nuwer, 2017). Το Διαδίκτυο είναι ένα σύστημα καναλιών που συνδέει όλους όσους έχουν μία συσκευή (υπολογιστή, κινητό ή tablet) και έναν τρόπο σύνδεσης (ενσύρματο ή όχι) και μπορεί να γίνει, θεωρητικά τουλάχιστον, απείρως εκτενές και αρκετά συγχρονισμένο. Παρ' όλ' αυτά, το μοναδικό πράγμα που δύναται να διακινείται μέσω αυτού του συστήματος είναι η Πληροφορία.

Υπ' αυτήν την έννοια, και σε σχέση με το πόσο επηρεάζει την παραγωγική διαδικασία, οφείλουμε να θέσουμε την πληροφορία, ως αντικείμενο μελέτης πλέον, στο μικροσκόπιο την επιστημονικής ανάλυσης. Να συσχετίσουμε δηλαδή την φύση της με αντίστοιχες απτές λειτουργίες, όπως η αποθήκευση, η μεταφορά, η αποκωδικοποίηση, παραγωγή, καταστροφή κ.α. Ο συσχετισμός των παραπάνω λειτουργιών αφορά σε ποικίλες πτυχές της επιστήμης, με αποτέλεσμα ένα μεγάλο μέρος του ανθρωπίνου δυναμικού της επιστημονικής κοινότητας και του χρόνου που διαθέτει, να αφιερώνεται στην πληροφοριακή διαχείριση. Μεγάλο ρόλο στην διαμόρφωση της κατάστασης φέρνουν οι Νέες Τεχνολογίες Πληροφορικής και Επικοινωνιών, που γιγαντώνουν τις δυνατότητες μας στην παραγωγή, επεξεργασία, αποθήκευση και μεταφορά των πληροφοριών που σχετίζονται με την ανθρώπινη δραστηριότητα. Έτσι, δίνεται η ευκαιρία να τις χειριστούμε όπως και τα υλικά προϊόντα, χωρίς όμως να έχουν υλική υπόσταση.

Βρισκόμαστε μόλις στην αυγή μιας εποχής διαφορετικών προσέγγισεων των βιολογικών δεδομένων, η οποία αναδεικνύει νέες προκλήσεις, προβληματισμούς και κατευθύνσεις έρευνας. Στα πλαίσια αυτής της εργασίας θα σημειωθούν μερικά απ' τα ζητήματα, που καλείται και ήδη αντιμετωπίζει ο τομέας της

Εξόρυξης από Δεδομένα. Η μεγάλη πρόκληση είναι η δόμηση ενός πλαισίου που να επιτρέπει στους ερευνητές να αλληλεπιδρούν με τα δεδομένα που τους παρέχονται, προκειμένου να «διατυπώνουν ερωτήσεις» που σχετίζονται με αυτά και να είναι σε θέση να τις απαντούν άμεσα και εύστοχα. Ερωτήσεις του τύπου: «Δείξε ομοιότητες, διαφορές, ανωμαλίες μιας συγκεκριμένης συλλογής δεδομένων», μπορούν να οδηγήσουν στην ανακάλυψη νέων προτύπων. *Ποιο μαθηματικό πλαίσιο απαιτείται;* Μια πρόκληση είναι να μπορεί μέσα σε αυτό το πλαίσιο να αλληλεπιδρούν ερευνητές, χωρίς πρότερη μαθηματική ή υπολογιστική παιδεία (Holzinger et al., 2014). Υπάρχει ανάγκη η Τεχνητή Νοημοσύνη να αντιμετωπίσει την πληθώρα δεδομένων και έχει πολλά να διδαχθεί απ' την λειτουργία του ανθρώπινου εγκεφάλου την οποία καλείται προσομοιάσει ή να ξεπεράσει.

Η παρούσα εργασία είναι μια προσπάθεια απεικόνισης της Εξόρυξης από Δεδομένα, τόσο σε επίπεδο μεθοδολογίας και δυνατοτήτων, όσο και σε εφαρμογής στην Βιοτεχνολογία και Βιοπληροφορική. Στο Πρώτο Κεφάλαιο εκτίθενται τα βασικά πλαίσια των κλάδων της Βιοτεχνολογίας και της Βιοπληροφορικής. Στο Δεύτερο Κεφάλαιο γίνεται εκτενής περιγραφή της Εξόρυξης από Δεδομένα ως διαδικασία ανακάλυψης γνώσης και αναλύεται το τεχνικό και μαθηματικό υπόβαθρο. Επίσης, γίνεται ειδική αναφορά στην Εξόρυξη Κειμένου (Text Mining), στα νέα εργαλεία που προσφέρει και στην συμβολή τους στην διαχείριση του μεγάλου όγκου δεδομένων που βρίσκονται στις ηλεκτρονικές βιβλιοθήκες. Το Τρίτο Κεφάλαιο αφιερώνεται στην έκθεση εργασιών που έχουν δημοσιευθεί στον τομέα της Εξόρυξης, στον ευρέως φάσματος συσχετισμό γονιδίων ή γενικότερα βιοδεικτών με ασθένειες ή φαινοτύπους. Τέλος, το Τέταρτο Κεφάλαιο καταλήγει σε συμπεράσματα και σχόλια για τις αδυναμίες των μεθόδων Εξόρυξης. Επιπλέον, παρουσιάζονται τα πεδία των οποίων οι ανάγκες εξωθούν την τεχνολογία, τη μεθοδολογία και την επιστημονική πρακτική σε νέα βήματα.

Η έννοια της επιστημονικής προσέγγισης της πραγματικότητας δια μέσου του υπολογιστή, είναι μεν καινούρια αλλά έχει υπάρξει αντικείμενο αρκετά παλαιότερων οραματισμών. Σε καθαρά θεωρητικό επίπεδο η προσέγγιση αυτή έχει οριστεί ως το Τέταρτο Παράδειγμα (Fourth Paradigm) βάσει του οποίου ο άνθρωπος μπορεί να αναζητήσει γνώση σχετικά με την Φύση. Αναφορικά, το Πρώτο Παράδειγμα είναι ο εμπειρικός τρόπος, που σχετίζεται με την μνήμη και

την γνώση ως ανθρώπινη εμπειρία, το Δεύτερο Παράδειγμα είναι το πειραματικό και σχετίζεται με την ανθρώπινη πρωτοβουλία και την επί τούτου κατάστρωση πειράματος για την απόκτηση γνώσης. Το Τρίτο Παράδειγμα είναι το θεωρητικό, το οποίο αντλεί γνώση μέσω συλλογιστικής και συνδυασμού αξιωμάτων. Το Τέταρτο Παράδειγμα στο οποίο αναφερόμαστε είναι η παρατήρηση μέσω του νέου εργαλείου του υπολογιστή, που για πρώτη φορά ιστορικά έχει δυνατότητες παρόμοιες με του ανθρώπινου εγκεφάλου, όπως Αναγνώριση Προτύπων, Ταξινόμηση πολλών μεταβλητών, μάθηση και μπορεί να αντικαταστήσει μερικώς τον ερευνητή (Tansley & Tolle, 2009).

1. Βιοτεχνολογία και Βιοπληροφορική

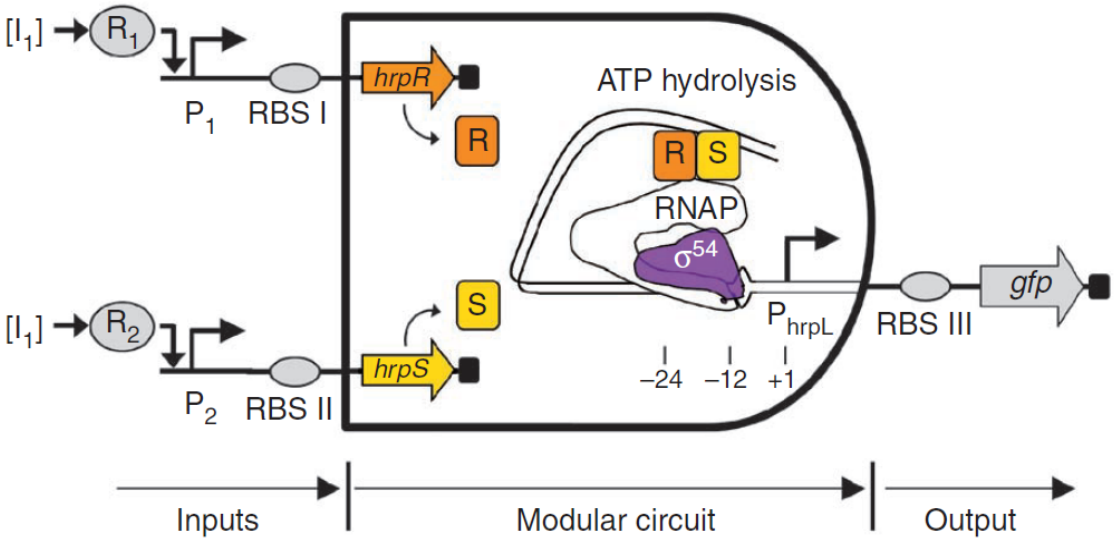
1.1. Βιοτεχνολογία

Στην πιο απλή της μορφή, η Βιοτεχνολογία είναι ένα σύνολο γνώσεων, διαδικασιών και υλικών αποτελεσμάτων που βασίζονται στην Βιολογία. Στόχος της είναι η τιθάσευση ενδοκυτταρικών και βιομοριακών διαδικασιών για την ανάπτυξη τεχνολογιών και προϊόντων, που να βελτιώνουν τη ζωή του ανθρώπου και του περιβάλλοντος στο οποίο ζει. Όσο και σύγχρονη να φαίνεται αυτή η ιδέα αριθμεί πολλά χρόνια παρουσίας στην ανθρώπινη εξέλιξη. Εδώ και 6.000 χρόνια ο άνθρωπος χρησιμοποιεί βιολογικές διεργασίες, που εκτελούν μικροοργανισμοί, για την παραγωγή διαφόρων προϊόντων, όπως ψωμί, κρασί, τυρί κ.α. Η Βιοτεχνολογία εκφράζει μια νοοτροπία εκμετάλλευσης των κατανενοημένων διαδικασιών που επιτελεί ένα σύστημα, ένας οργανισμός, ένα όργανο ή ένας μηχανισμός. Κάθε τεχνολογική εφαρμογή που εκμεταλλεύεται ένα βιολογικό σύστημα, με σκοπό να δημιουργήσει ή να τροποποιήσει παραγόμενα αγαθά ή διαδικασίες παραγωγής, για ερευνητική, σε πρώτη φάση, χρήση.

Το ευρύ πλαίσιο της Βιοτεχνολογίας εγκολπώνει φάσμα λειτουργιών και διαδικασιών, για τροποποίηση της έμβιας φύσης προς όφελος του ανθρώπου. Πρόκειται για μια διαδικασία βελτιστοποίησης που συνεχώς εστιάζει σε πιο δυσπρόσιτους βιολογικούς πρωταγωνιστές, ξεκινώντας απ’ τα οικόσιτα ζώα και τα καλλιεργούμενα φυτά, μέχρι τα ένζυμα ταχείας κομποστοποίησης (Miyatake & Kazunori, 2005). Σχετίζεται άμεσα με την Παραγωγική Διαδικασία καθώς και αποτελεί πεδίο κοινής δράσης πολλών επιστημών όπως η Βιοπληροφορική, η Βιορομποτική και η Χημική Μηχανική.

Η σύγχρονη Βιοτεχνολογία παρέχει ρηξικέλευθα προϊόντα και τεχνογνωσία για την αντιμετώπιση νέων και πρωτοφανών προβλημάτων που αντιμετωπίζει πλέον η ανθρωπότητα. Χαρακτηριστικό παράδειγμα είναι οι λογικές πύλες που αποτελούνται από βιολογικά δομικά στοιχεία (**Εικόνα 1**). Μια τέτοια πύλη παράγει αποτέλεσμα (στην προκειμένη περίπτωση φθορισμό [έξοδος RBSIII]), μόνο όταν συνυπάρχουν «και» οι δύο είσοδοι (είσοδος_1 το γονίδιο hrpR [RBSI]

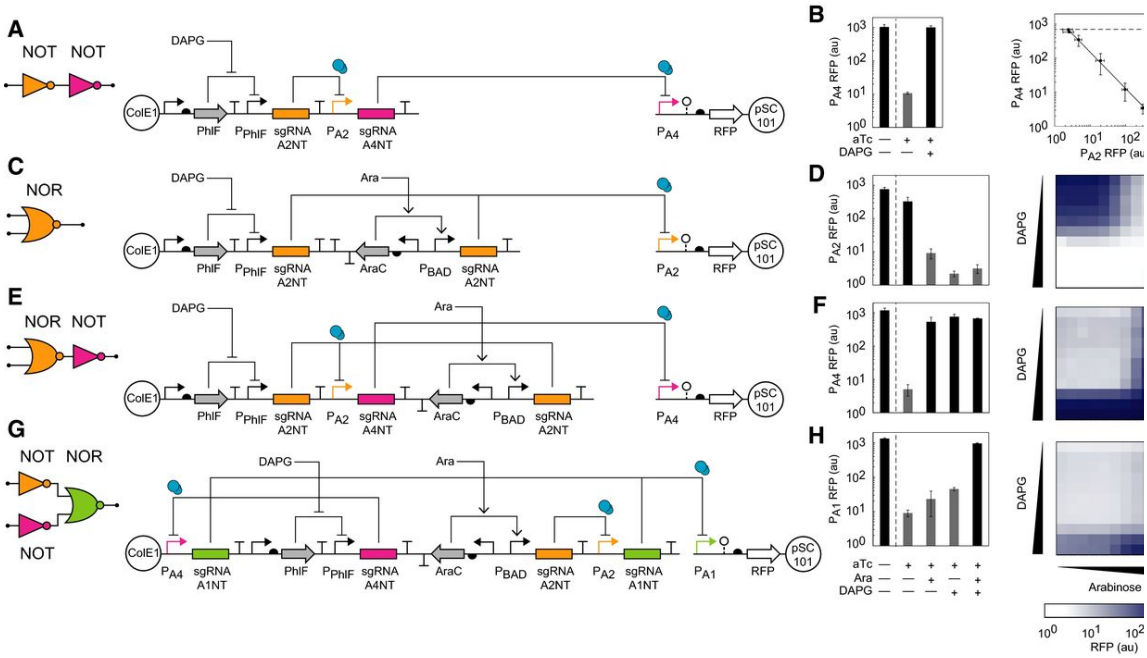
και είσοδος_2 το γονίδιο hrpS [RBSII]). Εκτός απ' την πύλη AND μπορούν να κατασκευαστούν όλες οι λογικές πύλες και δίνεται η δυνατότητα να προκύψει ένα πλήρες βιολογικό κύκλωμα (**Εικόνα 2**), σε αναλογία με το ηλεκτρονικό κύκλωμα. Το βασικότερο εμπόδιο στην διαδικασία είναι ο τρόπος με τον οποίον θα συνδέονται οι πύλες, καθώς κάθε μια μπορεί να λειτουργεί άριστα, αλλά σε διαφορετικό βιολογικό υπόβαθρο, εν γένει, απ' τις υπόλοιπες πύλες.



Εικόνα 1. *Λογική πύλη AND που αποτελείται από βιολογικά στοιχεία.*
(Η εικόνα ελήφθη από το διαδίκτυο: https://www.tomshw.it/computer-biologici-batteri-dal-dna-modificato-36378)

Η Βιοτεχνολογία στοχεύει σε προβλήματα όπως η αντιμετώπιση σπάνιων ασθενειών, η μείωση του αποτυπώματος που έχουν οι καταστροφικές ανθρώπινες ενέργειες στο περιβάλλον, η κάλυψη των βασικών αναγκών σε ύλες, τρόφιμα και ενέργεια, καθώς και η βελτιστοποίηση της βιομηχανικής παραγωγικής διαδικασίας. Προς το παρόν, έχουμε στην διάθεσή μας πάνω από 250 βιοτεχνολογικά προϊόντα, που αφορούν στην Υγεία και πολλά απ' αυτά σχετίζονται με ασθένειες που στο παρελθόν θεωρούνται μη ιάσιμες. Πάνω από 13 εκατομμύρια αγροτών ανά τον κόσμο χρησιμοποιούν αγροτεχνολογικά προϊόντα για την βελτίωση των σοδειών τους, την αποφυγή ζημιών και την μείωση της οικολογικής καταστροφής που επιφέρει η γεωργία. Ήδη στην Βόρειο

Αμερική δεκάδες καινοτόμες βιομηχανίες έχουν αρχίσει την διΰλιση βιοκαυσίμων, προκειμένου να μπορέσει να ελεγχθεί η βιωσιμότητα ενός νέου καυσίμου που δεν θα επιβαρύνει τον Κύκλο του Άνθρακα, η αποσταθεροποίηση του οποίου είναι η αιτία του Φαινομένου του Θερμοκηπίου.

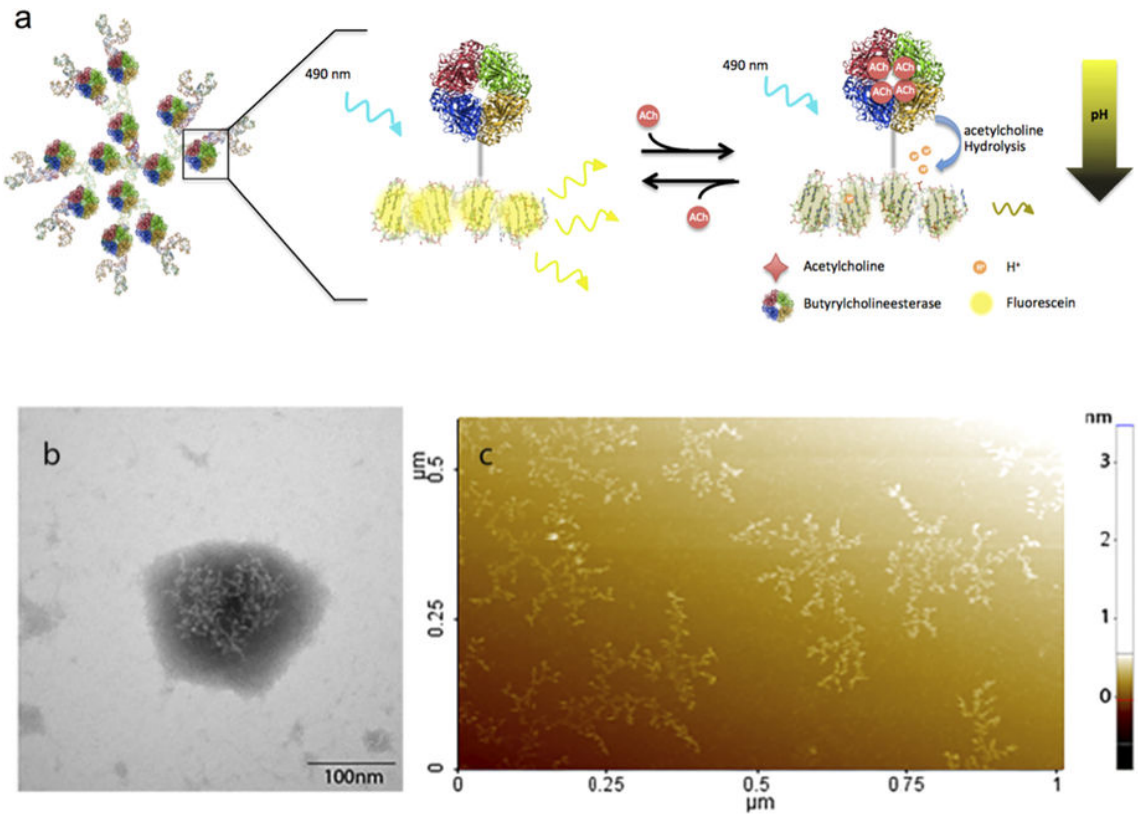


Εικόνα 2. *Παραδείγματα βιολογικών κυκλωμάτων και οι αποκρίσεις τους σε διάφορα βιολογικά σήματα.*

(Η εικόνα ελήφθη απ' το διαδίκτυο: http://2014.igem.org/wiki/index.php?title=Team:Fudan/pro&oldid=311336)

Όσον αφορά στην Υγεία είναι δυνατό να μειωθούν τα ποσοστά θνησιμότητας από μολυσματικές ασθένειες και κυρίως η μείωση της παιδικής θνησιμότητας. Επίσης μπορεί να βελτιωθούν οι συνθήκες διαβίωσης οι οποίες επιβαρύνουν την υγεία μεγάλων πληθυσμών του πλανήτη, μόλυνση υδάτων-αέρα και ερημοποίηση γης. Νέες δυνατότητες διαφαίνονται στην τροποποίηση των θεραπειών, ώστε να εξειδικεύονται ανά ασθενή με σκοπό την αποφυγή κινδύνων και παρενεργειών, που προκύπτουν απ' την περίθαλψή. Τέλος, εφευρίσκοντας πιο ακριβή, φθηνά και εύχρηστα σε μη ειδικούς εργαλεία και μεθόδους ανίχνευσης ασθενειών, αυξάνεται η ταχύτητα διάγνωσης, που εν γένει στον αναπτυσσόμενο κόσμο είναι μακριά από ιατρικά κέντρα.

Στον τομέα του επισιτισμού βελτιώσεις μπορούν να γίνουν στην Φυτική Παραγωγή, με την χρήση πιο βιώσιμων λύσεων στον τομέα της Φυτοπαθολογίας, Φυτοπροστασίας και Φυτανθεκτικότητας. Στόχος είναι η μείωση του χημικού φορτίου λίπανσης και παρασιτοκτονίας που απαιτεί η σύγχρονη καλλιέργεια. Σημαντικά βήματα γίνονται και στην βελτίωση της σοδειάς ανά φυτό. Νέες τεχνολογίες παράγουν φυτά και άλλους οργανισμούς που είναι εμβολιασμένα με νάνο-αισθητήρες, οι οποίοι καταγράφουν διάφορες παραμέτρους, που δίνουν τη δυνατότητα στους ερευνητές να παρακολουθούν την ανάπτυξη του οργανισμού, αποδίδοντας ένα τεράστιο όγκο δεδομένων (Charaniya et al., 2008). Χαρακτηριστικό παράδειγμα αποτελεί η τεχνολογία των βιο-αισθητήρων και των βιο-ανιχνευτών, επιτρέπει την μη εργαστηριακή και επιτόπου ανάλυση βιολογικών δειγμάτων (**Εικόνα 3**).



Εικόνα 3. *Βίο-ανίχνευση. (a) Υπολογιστική αναπαράσταση του νάνο-αισθητήρα και σχηματική απεικόνιση του προτεινόμενου μηχανισμού. (b) Εικόνα της δομής του βίο-αισθητήρα, διαμέτρου 130 ± 24 nm. (c) Μακροσκοπική επίδειξη της λειτουργίας του με ανάλυση AFM του ασυναρμολόγητου ακόμα DNA.*

(Η εικόνα αποτελεί μέρος της δημοσίευσης: Walsh et al., 2015)

Επιπλέον, αναπτύσσονται πρώτες τροφικές ύλες με εξειδικευμένα θρεπτικά χαρακτηριστικά, ανάλογα με τις ανάγκες και τις ελλείψεις των καταναλωτών, που σχετίζονται είτε με το περιβάλλον στο οποίο ζουν είτε με τις ασθένειες ή αλλεργίες απ' τις οποίες πάσχουν.

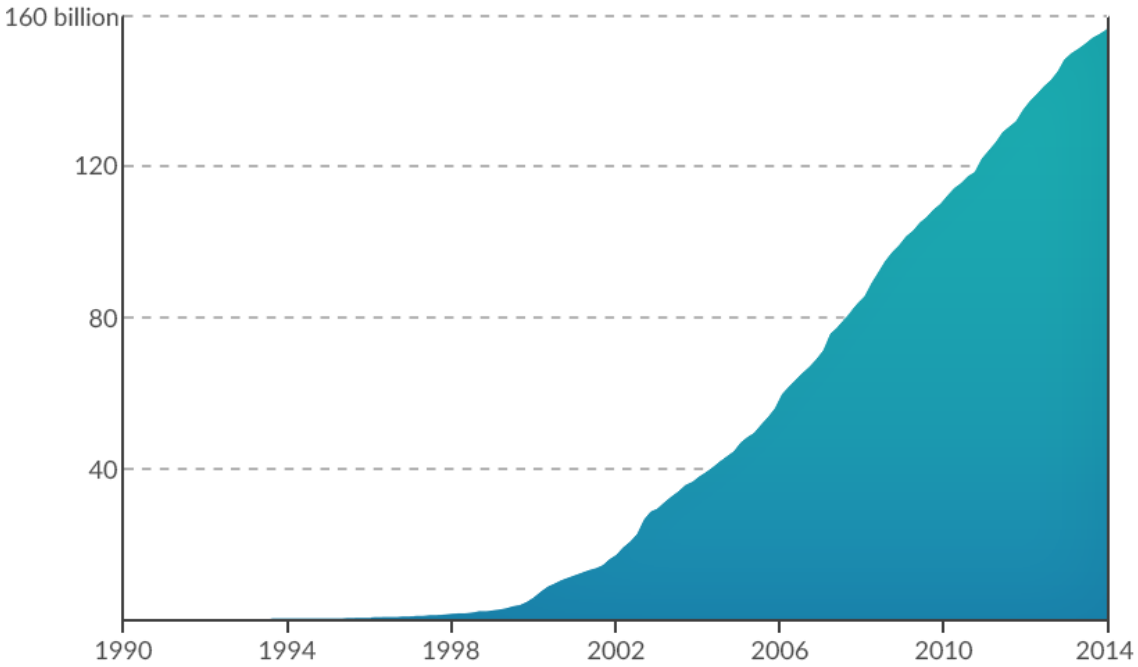
Επιπροσθέτως, ένζυμα, ζύμες και άλλοι μικροοργανισμοί μπορούν να επιστρατευθούν ως βίο-καταλύτες και με την εκμετάλλευση των διεργασιών ζύμωσης, που από την φύση τους επιτελούν, να αποτελέσουν για τον άνθρωπο νέα μικροσκοπικά «εργοστάσια» παραγωγής ή εξοικονόμησης ενέργειας. Παρακάμπτοντας πολλά στάδια της χημικής επεξεργασίας καυσίμων, βελτιστοποιείται ιλιγγιωδώς η γραμμή παραγωγής. Επίσης, μειώνοντας την θερμοκρασία πλύσης ρούχων, λόγω της βιολογικής μηχανικής μεταχείρισης που έχει υποστεί το απορρυπαντικό, εξοικονομούνται τεράστιες ποσότητες ενέργειας, που μειώνουν τις εκπομπές αερίων του θερμοκηπίου στο μισό και αντιμετωπίζουν την υπερκατανάλωση νερού. Ενώ μελλοντικός στόχος είναι η παραγωγή απόβλητης βιομάζας, που πληροί τις παραδοσιακές προδιαγραφές και είναι φιλική προς το περιβάλλον, στρατηγική που απαιτεί υψηλά ποσοστά πέψης των αποβλήτων.

1.2. Βιοπληροφορική

Η Βιοπληροφορική είναι κοινό πεδίο της Πληροφορικής και της Βιολογίας, όπου κυρίως βασίζει την μεθοδολογία και τις εφαρμογές στην Πληροφορική, ενώ εμπνέεται και βρίσκει λύσεις στην Βιολογία, την οικολογία και τις επιστήμες υγείας. Κατασκευάζει ή επεκτείνει υπάρχοντα εργαλεία λογισμικού τύπου για τον χειρισμό και την κατανόηση βιολογικών δεδομένων. Χρησιμοποιεί και συνδυάζει εργαλεία από την Επιστήμη Υπολογιστών, την Στατιστική, τα Μαθηματικά και την Φυσική.

Η τεχνολογική εξέλιξη που επήλθε στην βιολογική πειραματική διαδικασία, έδωσε την δυνατότητα για περισσότερη πληροφορία, αλλάζοντας σημαντικά την συσχέτιση μεταξύ ερμηνευμένης και παραγόμενης πληροφορίας. Αρχικά η συσσώρευση χρόνων έρευνας, αλλά ακόμα περισσότερο οι νέοι τρόποι παραγωγής πρωτογενούς πληροφορίας και δεδομένων, επιβαρύνουν την επιστημονική δράση με θηριώδες πλήθος ήδη κεκτημένων δεδομένων (**Εικόνα**

4), που όμως δεν έχουν ακόμα εκπληρώσει τον σκοπό για τον οποίον παρήχθησαν. Τα δεδομένα αυτά προορίζονται να περάσουν απ' τις διαδικασίες της ερμηνείας και της αποκωδικοποίησης προκειμένου να εξαχθεί γνώση, υπό την δυσκολία όμως ότι οι διαδικασίες ερμηνείας και αποκωδικοποίησης είναι χρονικά αργές και διεξοδικές. Υπάρχουν, παράγονται και προβλέπεται να παραχθούν τόσα δεδομένα, που οι λογικές ερμηνείας και κωδικοποίησης κρίνονται εμφανώς εξελικτέες και χρείζουν καινοτόμων και εφευρετικών προσθηκών. Μέριμνα της Βιοπληροφορικής είναι να χειριστεί τα δεδομένα και την παρουσίασή τους, έτσι ώστε να βελτιστοποιηθεί η παραγωγή γνώσης από αυτά.



Εικόνα 4. *Χαρακτηριστική εικόνα που δείχνει την ιλιγγιώδη αύξηση της αποθηκευμένης πληροφορίας (σε μονάδες αζωτούχων βάσεων) που καλείται να αντιμετωπίσει η Βιοπληροφορική. Χαρακτηριστική είναι η αύξηση στα έτη 2002 και 2003 που άρχισαν να αποδίδονται τα πρώτα αποτελέσματα απ' την αποκωδικοποίηση του ανθρωπίνου γονιδιώματος.*
(Η εικόνα ελήφθη απ' το διαδίκτυο: http://www.micronautomata.com/bioinformatics)

Βασικός στόχος της Βιοπληροφορικής είναι να αυξήσει την κατανόηση των βιολογικών διεργασιών. Αυτό που την ξεχωρίζει από τις άλλες προσεγγίσεις είναι η επιμονή της στην ανάπτυξη και εφαρμογή εντατικών υπολογιστικών τεχνικών στην διαδικασία ανάλυσης (Lesk, 2009). Αναγνώριση μοτίβων, εξόρυξη από

δεδομένα, αλγόριθμοι εκπαίδευσης μηχανών και νέες οπτικοποιήσεις αποτελούν τα κύρια πεδία δράσης. Με μεγάλη προσφορά στην συναρμολόγηση αλληλουχίας, τον εντοπισμό γονιδίου, τη γονιδιωματική συναρμολόγηση, τον σχεδιασμό και ανακάλυψη φαρμάκων, στην αλληλούχηση πρωτεϊνικής δομής, στην πρόβλεψη γονιδιακής έκφρασης και πρωτεϊνικής αλληλεπίδρασης. Επίσης, οι δια-γονιδιωματικές συγηριτικές μελέτες, η μοντελοποίηση της εξέλιξης και της κυτταρικής διαίρεσης-μίτωσης είναι πιο μακρόπνοα σχέδια μελέτης και απαιτούν συνεργασία πολλών επιστημόνων με διαφορετική πορεία, αλλά κάτω από τη στέγη της συστηματικής προσέγγισης.

Προκειμένου να μελετηθεί το πως συνήθεις κυτταρικές λειτουργίες μεταλλάσσονται σε διάφορες καταστάσεις παθολογίας, τα βιολογικά δεδομένα πρέπει να συνδυαστούν για να προκύψει μια πιο συνολική εικόνα. Έτσι, το πεδίο της Βιοπληροφορικής έχει εξελιχθεί ώστε να επιτελεί μια απαιτητική εργασία που να χρήζει ανάλυσης και ερμηνείας από πλειάδα τύπων δεδομένων. Η πραγματική διαδικασία ανάλυσης και διερμηνείας των δεδομένων ορίζεται ως Υπολογιστική Βιολογία. Τα βασικά καθήκοντα της Βιοπληροφορικής και της Υπολογιστικής Βιολογίας είναι δύο. Πρώτο, η αναβάθμιση και εγκατάσταση λογισμικών προγραμμάτων που επιτρέπουν επαρκή πρόσβαση σε χρήση και διαχείριση διαφόρων τύπων πληροφορίας. Δευτερευόντως, είναι αναγκαίο να αναπτυχθούν νέοι αλγόριθμοι και στατιστικές μετρήσιμες διαδικασίες που να αξιολογούν τους συσχετισμούς μεταξύ τμημάτων δεδομένων ενός μεγάλου όγκου πακέτου ή πακέτων δεδομένων (Lesk, 2009).

Στην προσπάθεια να βρεθούν νέοι τρόποι ερμηνείας, εγείρεται η ανάγκη μιας πιο καθολικής αντιμετώπισης της. Η αντίληψη ότι κάθε μέρος κληρονομεί κάποιες απ’ τις ιδιότητες του συστήματος στο οποίο υπόκειται, δίνει νέα κατεύθυνση που πρέπει, αλλά πιο σημαντικά μπορεί σήμερα να έχει η βιολογία. Είμαστε πλέον στην εποχή της Βιολογίας Συστημάτων ή Systems Biology, όπου οφείλουμε να συμπεριλάβουμε την επίδραση οποιουδήποτε περιβάλλοντος στην μελέτη του ατόμου και διεργασιών που συντελούνται στο εσωτερικό του. Και νέοι τρόποι πρόσβασης σε αποτελέσματα προκύπτουν από την νέα ολοκληρωμένη εικόνα. Τα γονιδιακά μονοπάτια ενός οργάνου αλληλεπιδρούν με τα μονοπάτια του διπλανού οργάνου και πιθανόν τα δύο μαζί να εξαρτώνται από κάποια εξωτερική παράμετρο, συνθήκη ή διαδικασία. Συνεπώς, στη συστηματική προσέγγιση

επαναπροσδιορίζονται το αντικείμενο μελέτης, η διαδικασία, το ερώτημα και ο τρόπος που θα τεθεί η απάντηση.

1.3. Η Εξόρυξη από Δεδομένα στα πλαίσια του Systems Biology

Η Βιολογία Συστημάτων είναι είναι ο τομέας όπου η πληροφορία προκύπτει από επεξεργασία ανομοιογενών δεδομένων. Καίριο ζήτημα είναι η αναζήτηση και αναγνώριση μοτίβων, συμμετριών που διέπουν τα δεδομένα. Η ανάδυση τέτοιων ευνοείται από την αύξηση του πλήθους των δεδομένων. Πολλές φορές αυτό δεν είναι αυτονόητο, καθώς χρειάζεται να συνδυαστούν στοιχεία από διαφορετικές μετρήσεις ή διαφορετικά πειράματα. Ο τρόπος συνεισφοράς του καθενός πρέπει να καθοριστεί από βιολογικά κριτήρια και νοοτροπία, γεγονός που ανάγει την ερευνητικότητα της βιολογικής πρακτικής σε καθολικό επίπεδο.

Η μεταγενομική επανάσταση αντιμετωπίζει την νέα γενιά βιολογικών δεδομένων της τάξης petabytes ετησίως (1 petabyte=1000terrabytes), με ευρύ φάσμα ενδιαφέροντος από την εξελικτική θεωρία, την αναπτυξιακή βιολογία, την αγροκαλλιέργεια και την αντιμετώπιση παθογενειών. Η επιστήμη και τεχνολογία της μετατροπής μιας πρωτόγνωρης συλλογής δεδομένων σε νέα γνώση. Ο τεράστιος όγκος οργανώνεται γύρω από δύο αλληλοκαλυπτόμενα θέματα, δικτυακές και λειτουργικές παρεμβολές.

Η Εξόρυξη από Δεδομένα είναι κλάδος της Πληροφορικής που ασχολείται με την εξερεύνηση Βάσεων Δεδομένων και Ψηφιακών Βιβλιοθηκών. Η Εξόρυξη από Δεδομένα μπορεί να εφαρμοστεί σε βιολογικά δεδομένα με διάφορους τρόπους.

- Πρώτον, ως εκμετάλλευση των πειραματικών δεδομένων από αναλύσεις υψηλής απόδοσης (high-throughput analysis) και από μεθόδους παρεμβολής, με σκοπό την ανακατασκευή δικτύων.
- Δεύτερον, τον χειρισμό της βιβλιογραφίας, καθώς κανένας πλέον ερευνητής δεν είναι ικανός να συλλάβει το σύνολο των σύγχρονων ερευνών και χιλιάδες ήδη υπαρχόντων δημοσιεύσεων, που περιέχουν βιολογικά φαινόμενα. Μια μεγάλη ποικιλία λογισμικών είναι διαθέσιμη για δευτερογενή ανάλυση των άρθρων που υπάρχουν διαθέσιμα.
- Τρίτον, πολλές συγκριτικές διαδικτυακές Βάσεις Δεδομένων που είναι διαθέσιμες, συνεπώς η πρόσβαση και η πλοήγηση, στοχευμένη ή

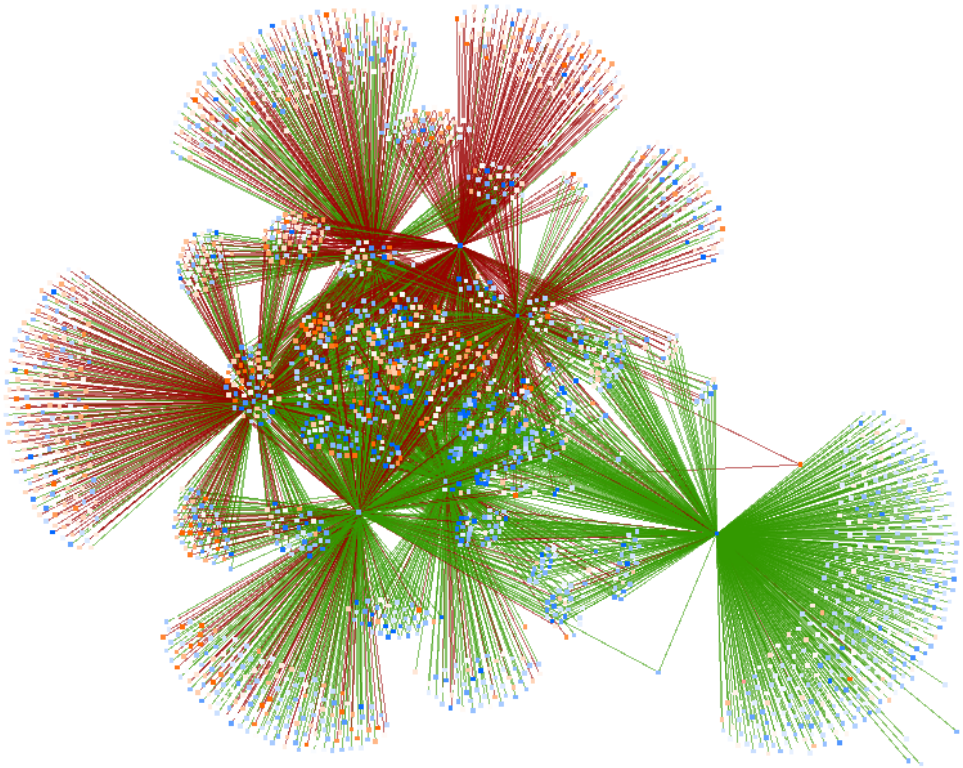
περιπλανώμενη αναζήτηση, επ' αυτών αποτελούν καθημερινότητα για τους χρήστες τέτοιων συστημάτων (Mamitsuka, 2013). Σημειωτέον, η διαδικασία αυτή της απεικόνισης τέτοιων δεδομένων, αποτελεί μέθοδο κατ' εξοχήν φιλική για το χρήστη.

Ένα πολύ χειροπιαστό παράδειγμα εξόρυξης από δεδομένα, που έχει ως σκοπό την ανάλυση της δομής ενός δικτύου και των σχέσεων μεταξύ των κόμβων που το αποτελούν, είναι αυτό που το 2003 μια ομάδα απ' το πανεπιστήμιο του Κολοράντο, ανέπτυξε έναν αλγόριθμο (PathMiner, 2003) για την εξαγωγή ή εύρεση κανόνων στην Γονιδιακή Εγκυκλοπαίδεια του Κυότο (KEGG), η οποία είναι μία απ' τις πιο διάσημες διαδικτυακά διαθέσιμες βιβλιοθήκες γονιδιακών μονοπατιών, γονιδίων, καθώς και γονιδιακών εκφράσεων. Κατά την πλοήγηση στην βιβλιοθήκη KEGG, η ομάδα εξήγαγε μια λίστα από 110 κανόνες αναμόρφωσης που διέπουν την αλληλεπίδραση συγκεκριμένων συνθετικών μονοπατιών, σε μεταβολικό επίπεδο. Στη συνέχεια, οι κανόνες χρησιμοποιήθηκαν, από κοινού μαζί με τους μαθηματικούς αλγορίθμους, για να προβλέψουν πως τα μονοπάτια αποτοξίνωσης θα μεταβολίσουν την αιθυλική και φουρφουρυλική αλκοόλη (ethyl and furfuryl alcohol). Η πρόγνωση του μοντέλου που προκύπτει βασίζεται σε γνωστά μοτίβα του αλκοολικού μεταβολισμού (McShan et al., 2003).

Διαφορετικές τεχνικές της εξόρυξης δεδομένων που εφαρμόζονται σε δίκτυα, ακόμα κι αν δεν ανήκουν στις παραδοσιακές τεχνικές εξόρυξης στο πεδίο εφαρμογής, μπορούν να έχουν πλεονεκτήματα λόγω της συστηματικής προσέγγισης, γεγονός που φαίνεται στην παραγωγικότητα γνώσης. Από το 2004, παράλληλα πολλές τεχνολογίες μαζικής αλληλούχησης DNA έκαναν την εμφάνισή τους, προσφέροντας εξαιρετικά μεγάλης κλίμακας ανάλυση και χαμηλού ανά βάση κόστους σε σχέση με το παρελθόν. Η συνεργασία των αποτελεσμάτων από διαφορετικές διαδικασίες είναι ζήτημα που αντιμετωπίζεται από στοχαστικές μεθόδους, γνωστές απ' τη θεωρία, που πηγάζουν απ' τη μοντελοποίηση της Βιολογίας Συστημάτων.

Στο καθαρά κατασκευαστικό σκέλος, αυτοματοποιημένα εργαλεία εξόρυξης έχουν ήδη αρχίσει να κάνουν την εμφάνισή τους. Η αύξηση του πλήθους τους είναι χαρακτηριστική κυρίως μετά το 2002, όπου οι συστηματικές, δικτυακές και ομικές προσεγγίσεις άρχισαν να γίνονται επίκαιρες. Εργαλεία που διευκολύνουν

την ανάλυση δικτύων (π.χ: CytoScape) και χαρακτηρίζονται από πλήθος συναρμολογικών δυνατοτήτων και χειρισμό επί των δεδομένων. Τα γραφήματα που κατασκευάζονται με τέτοια λογισμικά μπορεί να είναι ιδιαιτέρως περίπλοκα στην όψη, αλλά η αξία τους έγκειται στο ότι μπορούν να αναδείξουν κρυμμένη γνώση σε μεγάλο σύνολο πληροφοριών (**Εικόνα 5**). Επιπροσθέτως, τα εργαλεία υποστηρίζουν Διαφορική Μοντελοποίηση, Μοντελοποίηση Ασαφούς Λογικής (Fuzzy Logic Modeling), ή πλεονεκτήματα αρχειοθέτησης (Su et al., 2014). Δίνεται επίσης έμφαση και στην κατασκευή εργαλείων για τον διαμοιρασμό δεδομένων και τον συντονισμό των βιβλιοθηκών, με αποτέλεσμα η διεπεμβατικότητα, δηλαδή η ικανότητα πολλών συγγραφέων να εργάζονται πάνω στο ίδιο αρχείο ταυτόχρονα.



Εικόνα 5. *Ένα τυχαίο δίκτυο απεικονισμένο απ' το λογισμικό CytoScape. Στο σχέδιο φαίνονται οι κόμβοι καθώς και οι ακμές που τους ενώνουν. Όπως φαίνεται οι περισσότεροι κόμβοι επικοινωνούν μόνο έναν άλλον κόμβο, ενώ υπάρχουν και λίγοι οι οποίοι είναι υπέρ-συνδεδεμένοι (bottleneck Knots). Η διαφορές στον χρωματισμό των ακμών υποδηλώνουν μία επιπλέον ταξινόμηση ως προς τις ακμές.*

(Η εικόνα ελήφθη απ' το διαδίκτυο: https://www.biostars.org/p/240268/)

Πρόσφατα, εφαρμογές της Μαζικής Παράλληλης Αλληλούχησης (Massive Paralel Sequencing) έχουν επιστρατευθεί για αρχειοθέτηση μελετών από αλληλεπιδράσεις μεταξύ πρωτεϊνών-DNA, μεθυλιώσεις σε κυττοσίνες, γενετικές ποικιλομορφίες, γονιδιωματικές ανακατατάξεις. Τέτοιες πλατφόρμες λογισμικού (όπως τα Roche (454) GS FLX sequencer, Illumina Genome Analyzer και τον Applied Biosystems SOLiD sequencer) είναι ικανές να παράγουν εκατομμύρια διαβασμάτων βραχέων αλληλουχικών (Liu et al,. 2012). Ο πρώτος τύπος προϊόντος είναι κατάλληλος για γονιδιακή αλληλούχηση, πλήρη ανάκτηση γονιδιώματος, εντοπισμός microRNA, παρεμβολή από μεθυλίωση, πειραμάτων τύπου αλληλούχισης μετά από χρωματινική ανοσοκαθίζηση (ChIPSeq) και ανακάλυψη πολυμορφισμών μοναδιαίων βάσεων (SNP). Ο δεύτερος τύπος μπορεί να φανεί χρήσιμος για την αλληλούχηση ολόκληρου γονιδιώματος, αξιολογώντας ποσοτικά τις δομικές ανακατατάξεις και εναλλαγές του αριθμού αντιτύπων του DNA σε οργανισμούς με πολυπλοειδία, καθώς επίσης και για την ανακάλυψη των πολυμορφισμών SNP. Τέτοια δεδομένα αναμένεται να ωθήσουν τις μεθόδους του Data Mining έξω απ’ τα σημερινά τους όρια, στην χρήση μεγαλύτερου πλήθους διαφορετικών ειδών, θεμελιώνοντας διπλά την συγκριτική προσέγγιση. Ένα τέτοιο παράδειγμα είναι η από κοινού μελέτη του βακτηριώματος του ανθρώπινου σώματος, καθώς είναι γνωστό ότι αποτελείται από περίπου 20 χιλιάδες διαφορετικά βακτήρια και πρέπει να βρεθεί τεχνικά τρόπος να εξεταστούν ταυτόχρονα ως σύστημα βακτηρίων (Liu et al,. 2012). Η πρόκληση είναι να κατασκευαστούν ευσταθείς, βέλτιστες και γρήγορες μέθοδοι, με χαμηλές ανάγκες σε εκπαίδευση και λειτουργική πολυπλοκότητα. Οι προοπτικές μπορούν επίσης να συγκρίνονται σε δίκτυα από πειράματα αλληλούχησης με διάφορες καταστάσεις.

Όπως αναφέρθηκε και νωρίτερα, ο στόχος της Βιολογίας Συστημάτων βασίζεται στην καλύτερη ικανότητα πρόγνωσης. Πολλές εξέχουσες χρησιμοποιούμενες μέθοδοι έχουν προκύψει από τον μαθηματικό τομέα των Δυναμικών Συστημάτων, όπως Στοχαστική Προσομοίωση του Monte Carlo. Ο Poincaré, στην αυγή του 19ου αιώνα απέδειξε ότι τέτοια συστήματα εξομοίωσης, δεν είναι ικανά να συγκλίνουν και αυτό γίνεται χειρότερο όταν αυξάνονται οι φορείς που αλληλεπιδρούν σε ένα σύστημα (Mamitsuka, 2013). Ο Von Bertallanffy λίγο αργότερα (Bertalanffy,1968), ισχυρίζεται ότι οι ζώντες

οργανισμοί μπορούν να απεικονισθούν ως ανοιχτά συστήματα, μη-χαοτικά και ανεξάρτητα από τις αρχικές συνθήκες, τις συνθήκες που επικρατούν στο σύστημα την στιγμή που ξεκινάμε να το μελετάμε. Ο βασικός υπεύθυνος για αυτήν την μη-χαοτικότητα είναι οι ρυθμιστικοί μηχανισμοί. Παρόλα αυτά, παραμένουν σημαντικά ερωτήματα, όπως για παράδειγμα: Ποια γονίδια και αλληλεπιδράσεις απαιτούνται από ένα τέτοιο μοντέλο; Πώς προσεγγίζονται οι παράμετροι του μοντέλου; Ένα συνδυασμός υποθέσεων που αναδεικνύονται από την εξόρυξη δεδομένων, πειραματικών προσομοιώσεων και αν είναι δυνατόν, επιβεβαίωση στο Εργαστήριο συνθέτουν την συστηματική αντιμετώπιση και υπόσχονται, αν όχι πολλά, τουλάχιστον ότι θα εξαντλήσουμε τις υπάρχουσες δυνατότητες.

1.3.1 Το Πρόγραμμα Γονιδιακής Οντολογίας (Gene Ontology Project)

Το πόνημα για την χαρτογράφηση της γονιδιακής οντολογίας είναι μια συλλογική προσπάθεια να καλυφθούν οι ανάγκες την συστηματικής περιγραφής των προϊόντων και συνεπειών της έκφρασης γονιδίων, που υπάρχουν αλληλουχημένα σε Βάσεις Δεδομένων. Το 1998 ξεκίνησε σαν συνεργασία μεταξύ τριών βάσεων δεδομένων για οργανισμούς μοντέλα, την FlyBase για τη Drosophila (http://flybase.org/), την Saccharomyces Genome Database (SGD, http://www.uniprot.org/database/DB-0095) για το ζαχαρομύκητα και την Mouse Genome Database (MGD, http://www.uniprot.org/database/DB-0060) για το ποντίκι. Η Κοινοπραξία Γονιδιακής Οντολογίας (Gene OntologyConsortium, http://www.geneontology.org/) έχει έκτοτε συμπεριλάβει πολλές νέες βάσεις, συμπεριλαμβανομένων μερικών απ’ των κυριότερων γονιδιακών αποθεμάτων για φυτά, ζώα και μικρόβια. Δημιουργήθηκε έτσι μια νέα βάση δεδομένων που προσπαθεί να συμπεριλάβει όλα τα δεδομένα, συσχετισμούς, μοτίβα, κατηγοριοποιήσεις, που εκλείπουν για να γίνει κατανοητό πως παράγεται ο φαινότυπος απ’ το γονιδίωμα. Βασική είναι η προϋπόθεση ότι καμία απ’ αυτές τις λειτουργίες δεν θα ταυτοποιεί κάποιον οργανισμό, αλλά θα είναι βασικού βιολογικού ενδιαφέροντος.

Αναγνωρίζονται τρεις διαφορετικές προοπτικές σχετικά με την προσπαθεία, δεδομένου ότι δεν ξεκίνησε με κάποιο σαφή στόχο, αλλά ακόμα και σήμερα οι

στοχεύσεις παραμένουν ανοιχτές. Η πρώτη είναι η ανάπτυξη και η διατήρηση των οντολογιών ως δεδομένα, ενώ η δεύτερη είναι ο σχολιασμός και περιγραφή των γονιδιακών παραγώγων, που προϋποθέτει την καταγραφή συσχετισμών μεταξύ οντολογιών-γονιδίων-γονιδιακών παραγώγων. Τρίτη προοπτική είναι ο σχεδιασμός εργαλείων που να διευκολύνουν την δημιουργία, συντήρηση και χρήση των οντολογιών.

Έτσι, κατασκευάζονται οντολογικά λεξικά που μπορούν να κάνουν σύνθετες αναζητήσεις. Τα λεξικά αυτά είναι ανοιχτά και ικανά να αναδιαμορφωθούν από οποιονδήποτε το χρησιμοποιεί. Η διαδραστικότητα αυτή δίνει ένα δυνατό προβάδισμα προς την ενοποίηση των βιολογικών βιβλιοθηκών σε μία βιβλιοθήκη πολλών επιπέδων με εγκυκλοπαιδικό χαρακτήρα. Πλέον θα μπορούν να ταξινομούνται μαζί γονίδια, πρωτεΐνες, μεταβολίτες και άλλοι κυτταρικοί παράγοντες και η αναζήτηση να γίνεται στο οντολογικό επίπεδο της λειτουργίας.

Από την σκοπιά της αλληλούχησης φαίνεται σαφώς ότι ένα μεγάλο κομμάτι των γονιδίων που επιτελούν τις βασικές βιολογικές λειτουργίες είναι εν πολλοίς κοινό σε όλους τους ευκαρυώτες. Η γνώση για τον ρόλο μιας πρωτεΐνης σε έναν οργανισμό, μπορεί να δώσει απαντήσεις στην λειτουργία της σε έναν άλλο οργανισμό. Ο στόχος του λεξικού GO είναι να παράξει δυναμική ελεγχόμενη ετυμολογική βάση, η οποία μπορεί να εφαρμοστεί στο σύνολο των ευκαρυωτών, ακόμα και αν η γνώση για τον ρόλο γονιδίων και πρωτεϊνών συνεχώς αλλάζει και συσσωρεύεται. Μέχρι τώρα τρεις ανεξάρτητες κατηγορίες με βάση την λειτουργία, βιολογικές διεργασίες (ένζυμα), μοριακή λειτουργία και κυτταρικό δομικό στοιχείο (δομικές πρωτεΐνες), έχουν χαρακτηριστεί και τα στοιχεία τους είναι προσβάσιμα στον ισότοπο (http://www.geneontology.org).

1.3.2 Θεωρία Ασαφών Συνόλων (Fuzzy Set Theory)

Στα Μαθηματικά, τα Ασαφή Σύνολα είναι σύνολα των οποίων τα στοιχεία έχουν βαθμούς συμμετοχής-συσχέτισης. Τα Ασαφή Σύνολα ορίστηκαν για πρώτη φορά το 1965 (Zadeh, 1965), ως μια προέκταση στην κλασική έννοια του συνόλου. Επιχειρήθηκε ο ορισμός ενός πιο γενικού είδους δομής, που αρχικά ονόμασε L-συσχέτιση. Οι ασαφείς σχέσεις έχουν εφαρμογή σε διάφορα πεδία, όπως φιλολογία-γλωσσολογία, Λήψη Αποφάσεων (Decision Making),

Συσταδοποίηση (Clustering). Στα πεδία αυτά L-συσχετίσεων το L συμβολίζει την μονάδα συσχέτισης που ανήκει στο σύνολο [0,1].

Στην κλασική μελέτη Συνόλων η συμμετοχή των στοιχείων εκτιμάται με δυαδικό τρόπο, δηλαδή ένα στοιχείο είτε ανήκει είτε δεν ανήκει σε ένα σύνολο, ή 0 ή 1. Ενώ, στην Ασαφή Λογική επιτρέπεται βαθμωτή ένταξη σε ένα σύνολο, πράγμα το οποίο περιγράφεται μορφοποιημένα με την καμπύλη συμμετοχής, που παράγεται από μια πραγματική συνάρτηση με σύνολο τιμών το [0,1]. Τα Ασαφή Σύνολα αποτελούν γενίκευση των κλασικών συνόλων μιας και το πεδίο τιμών τους είναι υπερσύνολο των ακραίων λύσεων 0 ή 1. Η Θεωρία Ασαφών Συνόλων μπορεί να χρησιμεύσει σε μεγάλο εύρος προβλημάτων, όπου η πληροφορία είναι ατελής και αφηρημένη με χαρακτηριστικότερο παράδειγμα τη Βιοπληροφορική.

2. Εξόρυξη από Δεδομένα (Data Mining)

2.1. Εισαγωγικοί Ορισμοί

Από την εφεύρεση της γραφής μέχρι την κωδικοποίηση σε mp3 και την κωδικοποιημένη αποθήκευση των πρωτεϊνών στον πυρήνα του κυττάρου, η έμβια Φύση, συνειδητά ή ασυνείδητα, εκφράζει παντού την ανάγκη της για μνήμη. Η ανάγκη αυτή έχει συσχετιστεί σε φιλοσοφικό επίπεδο με την έννοια του έμβιου, της συνειδητότητας και της αναγέννησης. Η παρούσα εργασία θα αναφερθεί σε μία άμεση τεχνική συνέπεια, που έρχεται να καλύψει την ανάγκη αυτή, την έννοια της Αποθήκευσης. Αποθήκευση είναι μνήμη σε ποικίλες μορφές, όπως DNA, αντισώματα, μία λίστα για ψώνια, μία φωτογραφία, ένα καρδιογράφημα και οτιδήποτε άλλο μπορεί να διατηρείται στο χρόνο και να αναπαράγει σε διαφορετικές χρονικές στιγμές την πληροφορία που περιέχει.

Κομβικές έννοιες που χαρακτηρίζουν τη φύση του δεδομένου είναι αποθήκευση, διατήρηση, αναπαραγωγή, κωδικοποίηση και σήμανση επί κάποιου φυσικού αντικειμένου. Η γενικότητα των εννοιών μπορεί να καταδείξει την ευρύτητα του νέου αυτού κλάδου της Εξόρυξης από Δεδομένα. Εκτός απ’ την τεράστια εφαρμογή στον κλάδο της Βιολογίας μπορεί να έχει εφαρμογές στις υπόλοιπες θετικές επιστήμες, στις ανθρωπιστικές επιστήμες, στην οικονομία, το μάρκετινγκ και την διοίκηση επιχειρήσεων, ομάδων, κρατών. Οπότε εξόρυξη είναι η εξαγωγή κανόνων, ιδιοτήτων, συμμετριών από ένα σύνολο αποθηκευμένων δεδομένων. Μια τέτοια εξόρυξη εκπόνησαν οι μελετητές της ελληνιστικής περιόδου, μελετώντας την αρχαία ελληνική γραμματεία στην βιβλιοθήκη της Αλεξάνδρειας, με αποτέλεσμα την σύνταξη της *Αρχαίας Ελληνικής Γραμματικής*. Βεβαίως η μεθοδολογία έχει διευρυνθεί από τότε, με την πιο μεγάλη αλλαγή να είναι η χρήση υπολογιστών, που οι δυνατότητές άλλαξαν ριζικά το σκηνικό.

Ο ορισμός του *δεδομένου* απ’ το λεξικό της Οξφόρδης (https://en.oxforddictionaries.com/definition/data), είναι:

«Δεδομένο είναι η ποσότητα, χαρακτήρας ή σύμβολο το οποίο επεξεργάζεται ένας υπολογιστής και μπορεί να είναι ικανό για αποθήκευση και μεταφορά υπό

μηχανική, οπτική ή ηλεκτρομαγνητική μορφή». Για την χρήση του ορισμού στα πλαίσια της παρούσας εργασίας, απλώς συμπληρώνεται ότι το Δεδομένο έχει προκύψει από ένα βιολογικό πείραμα και συμβολίζει μία βιολογική φυσική πραγματικότητα.

Η γενίκευση είναι εύκολη, κάθε συλλογή δεδομένων ιδίου τύπου που αφορούν σε ίδιου τύπου οντότητα μπορεί να αποτελέσει μία Δομή Δεδομένων. Η Δομή Δεδομένων είναι το πεδίο πάνω στο οποίο διενεργείται η ανάλυση και το πλήρες σύστημα το οποίο μελετά η Εξόρυξη Δεδομένων. Η πληροφορία που κρύβει μέσα του ένα βιολογικό σύστημα θεωρείται ότι μεταβαίνει, μέσω της παραμετροποίησης στην δομή που διαθέτει προς την επεξεργασία. Από εκεί και πέρα αναλαμβάνει η Πληροφορική, τα Μαθηματικά, η Στατιστική και πλέον η Βιολογία έχει συμβουλευτικό χαρακτήρα. Έτσι λοιπόν, το πεδίο εξόρυξης είναι ένα Σύνολο Δεδομένων. Κάθε τι που μπορεί να αποθηκευτεί σε υπολογιστή είναι εν δυνάμει μια συλλογή δεδομένων και στην συνέχεια με κατάλληλη επεξεργασία μια Δομή Δεδομένων. Λογιστικά φύλλα πειραματικών αποτελεσμάτων (Excel), εικόνες από gel ή φωτογραφίες μικροσκοπίου, βίντεο με κυτταρικές μιτώσεις, αρχεία αλληλούχησης τύπου FASTA, γραφήματα ήχου, καθώς και οποιοδήποτε νέο ψηφιακό δεδομένο μπορεί να προκύψει απ' την επεξεργασία των προηγουμένων.

Το κλειδί στην ευρύτητα των πιθανών εισροών αυτής της διαδικασίας της Εξόρυξης είναι η κωδικοποίηση, ο τρόπος με τον οποίον είναι γραμμένα ή αποθηκευμένα τα δεδομένα. Ο τρόπος συνδέεται άμεσα με την χρήση του υπολογιστή. Κάθε αποθηκευμένο δεδομένο, ανεξαρτήτως αν τελικά θα αναπαραχθεί ως εικόνα, ήχος, η λίστα, για τον υπολογιστή δεν είναι παρά μια επεξεργάσιμη δομή π.χ. ένας πίνακας στοιχείων, και βασιζόμενοι στις παλαιότερες γνώσεις της Πληροφορικής και των Μαθηματικών, μπορεί να έχει εξαιρετικά αναλυτική, ενδελεχή και λεπτή επεξεργασία επί της δομής αυτής.

2.2. Η Δομή των Δεδομένων

Δομή Δεδομένων δεν είναι απλώς μια συλλογή δεδομένων, αλλά φέρει επιπλέον και τους κανόνες διάταξης των Δεδομένων που εμπεριέχονται σε αυτήν. Η μορφή που πρέπει να έχει μία δομή δεδομένων (data set). Σε πρώτη

φάση η δομή είναι ένας πίνακας δύο διαστάσεων $n \times m$, με n γραμμές και m στήλες. Κάθε γραμμή αποτελεί ένα στιγμιότυπο (instance) και κάθε στήλη μία παράμετρο (attribute) (Witten et al., 2016). Κάθε κελί του πίνακα έχει μία τιμή η οποία μπορεί να είναι αριθμός, χαρακτήρας, λογική μεταβλητή ή σύμβολο, έτσι ώστε κελιά που βρίσκονται στην ίδια στήλη, να έχουν ιδίου τύπου τιμές. Το σύνολο των κελιών που βρίσκονται στην ίδια γραμμή ονομάζονται στιγμιότυπο, υπό την έννοια ότι αποτελούν ένα σύνολο μετρήσεων που αφορούν στο ίδιο πράγμα. Η ονομασία των γραμμών ως στιγμιότυπα δεν είναι αφελής. Η δομή των δεδομένων αναπαριστά μια πραγματικότητα, οπότε κάθε γραμμή είναι μια διαφορετική εκδοχή ή στιγμιότυπο αυτής. Τέλος, κάποιες από τις στήλες, που είναι, όπως αναφέρθηκε, παράμετροι, ονομάζεται κλάση. Η κλάση είναι η μοναδική στήλη που τα κελιά της δεν έχουν απαραιτήτως τιμές που έχουν προκύψει από μέτρηση. Για τυπικούς λόγους η κλάση είναι συνήθως η τελευταία στήλη. Την κλάση την ορίζει συνήθως είτε αυτός που κατέγραψε τα δεδομένα είτε ο ερευνητής που εργάζεται στην εξόρυξη και έχει άμεση σύνδεση με το τι πρέπει και τι μπορεί να παραχθεί ως αποτέλεσμα απ’ την εργασία.

Πίνακας 1. *Παράδειγμα στιγμιότυπου με 5 παραμέτρους, 4 μεταβλητές και μία κλάση.*

Περιγραφή καιρού	Θερμοκρασία (°C)	Υγρασία (%)	Άνεμος	Ματαίωση παιχνιδιού
ηλιοφάνεια	26	75	όχι	όχι
ψιχάλα	18	97	όχι	όχι
βροχή	16	99	ναι	ναι

Ένα στιγμιότυπο (**Πίνακας 1**) από ένα σύνολο n ημερών, οπότε η μεταβλητή περιγραφή θα έπαιρνε πιθανόν τιμές “ηλιοφάνεια”, “ψιχάλα”, “βροχή” κτλ., ενώ οι παράμετροι είναι κάποιες μεταβλητές, $m=5$ σε πλήθος, που αξιολογούν την κάθε ημέρα. Εδώ για παράδειγμα αυτό που παριστά ο πίνακας είναι η κατάσταση μιας ημέρας (4: Περιγραφή, Θερμοκρασία, Υγρασία, Άνεμος) συν την επιπλέον πληροφορία για το αν τελικά μια ενέργεια ματαιώθηκε ή όχι (+1: κλάση, Ματαίωση), που στην προκειμένη περίπτωση είναι αρνητική, “όχι”. Από κάτω μπορούν να προστέθουν όσα στιγμιότυπα ημερών επιθυμούνται.

Αυτή είναι η μορφή με την οποία μπορούν να χειριστούν τα δεδομένα και αν τα δεδομένα δεν την έχουν πρέπει να την κατασκευάσει αυτός που κάνει την

μελέτη. Ας αναφερθεί ένα άλλο παράδειγμα απ’ την Βιολογία που έχει ενδιαφέρον, διότι τα δεδομένα δεν έχουν την μορφή αυτή. Έστω ότι είναι γνωστή η αρχιτεκτονική n πρωτεϊνών, από μία μέθοδο κρυσταλλογραφίας και παράλληλα είναι γνωστή η εργασία που εκτελεί. Οπότε υπάρχουν n εικόνες και αρμοδιότητες για κάθε μία από αυτές. Το σύνολο των παραμέτρων που γνωρίζουμε για κάθε πρωτεΐνη είναι 2, η εικόνα της και η αρμοδιότητά της. Δεν έχει νόημα να δομηθεί ένας πίνακας $n \times 2$ διότι δεν μπορούν να εισάγχθουν σε μία στήλη οι φωτογραφίες ως έχουν, ο υπολογιστής δεν θα μπορεί να τις χειριστεί. Επίσης οι αρμοδιότητες μπορεί να ποικίλουν τόσο που να μην μπορεί να γίνει ταξινόμηση. Όπως είναι προφανές δεν έχει νόημα να ταξινομηθεί ένας πλήθος αντικειμένων σε περίπου ίδιο αριθμό κλάσεων, έτσι ώστε οι κλαεις να είναι ίσες με το πλήθος. Θα πρέπει το πλήθος των κλάσεων να είναι κατά πολύ μικρότερο απ’ το πλήθος των αντικειμένων. Αυτό δεν είναι κάποιος θεωρητικός κανόνας, αλλά τίθεται απ’ την κοινή λογική (Witten et al., 2016).

Άρα τα προβλήματα που ενυπάρχουν, πριν ακόμα ξεκινήσει η εξόρυξη είναι δύο. Πρώτον, να επαναπροσδιοριστούν τις κλάσεις, ομαδοποιώντας τις ήδη υπάρχουσες σύμφωνα με την γνώμη ενός βιολόγου, έτσι ώστε πρωτεΐνες με συγγενικές λειτουργίες να αποτελέσουν μια κλάση. Για παράδειγμα, μπορεί να κατασκευαστεί μια κατηγορία με τις δομικές πρωτεΐνες, που θα περιλαμβάνει πρωτεΐνες που δομούν κυτταρικό τοίχωμα, την μεμβράνη του πυρήνα ή έχουν απλώς συγκολλητικό ρόλο σε κάποιο σύμπλοκο. Άλλη κατηγορία μπορεί να σχηματίζεται απ’ τις πρωτεΐνες που κόβουν DNA, RNA. Άλλη κατηγορία μπορεί να είναι οι πρωτεΐνες που καταβολίζουν υπολείμματα στο εσωτερικό του κυτταροπλάσματος, έτσι ώστε τελικά να απομειωθεί ο αριθμός των κλάσεων σε ένα εύλογο πλήθος. Όπως φαίνεται ο καθορισμός της κλάσης μπορεί να επαφίεται σε αυτόν που μελετά τα δεδομένα και όχι απαραίτητα σε αυτόν που τα καταγράφει. Αυτό έχει να κάνει με το γεγονός ότι οι κλάσεις είτε είναι άγνωστες την στιγμή που λαμβάνονται τα δεδομένα και καθορίζονται στην πορεία της διαδικασίας Εξόρυξης, είτε είναι σχετικές με την διαδικασία Εξόρυξης, άρα και υποκειμενικές του τρόπου που γίνεται η Εξόρυξη. Έτσι, η στήλη των κλάσεων δεν είναι κατ’ ανάγκη μέρος τις αρχικής δομής προς επεξεργασία.

Η δεύτερη και πιο απαιτητική εργασία που μένει είναι η μετατροπή όλων των εικόνων σε κάτι που να είναι διαχειρίσιμο από μία αναλυτική διαδικασία, να

μετατραπούν οι εικόνες σε Δομή Δεδομένων. Πρακτικά αυτό οδηγεί στην κατάστρωση ενός πίνακα με γραμμές όσες είναι οι εικόνες και με στήλες όσες είναι οι παράμετροι που θα χαρακτηρίσουν την ίδια εικόνα. Να παραμετροποιηθούν όλες εικόνες ξεχωριστά ώστε να δημιουργηθεί ένα στιγμιότυπο, μία γραμμή στον πίνακα με τις τιμές από ποσότητες.

Από μία εικόνα μπορεί να υπολογιστεί αδρά το μέγεθος μιας πρωτεΐνης, το πλήθος των α-ελίκων και το πλήθος των β-πτυχώσεων, σύμφωνα με την δευτεροταγή τους δομή, καθώς επίσης αν έχουν ινώδη ή σφαιρική μορφή. Η εργασία αυτή μπορεί να γίνει με παρατήρηση ή με την βοήθεια λογισμικού που κάνει οπτική ανάλυση. Με αυτήν την παραγοντοποίηση μετατρέπεται μία εικόνα σε μία σειρά από μεταβλητές. Οι τιμές των μεταβλητών αυτών για την εκάστοτε πρωτεΐνη θα γεμίσουν τα κελιά της κάθε γραμμής, που αποτελεί όπως αναφέρθηκε ένα στιγμιότυπο.

Πίνακας 2. *Δομή δεδομένων που αποθηκεύει βιολογικά στιγμιότυπα. Το κάθε ένα από αυτά αποτελείται από τις παραμέτρους που χαρακτηρίζουν την εικόνα μιας πρωτεΐνης.*

Μέγεθος	Πλήθος α-ελίκων	Πλήθος β-ελίκων	Σφαρική δομή	Λειτουργία
14.2	8	19	όχι	δομική
6.7	3	12	όχι	ανήκει_σε_σύμπλοκο
...	...	...	...	...

Αυτό μπορεί να γίνει για όλες της πρωτεΐνες, έτσι ώστε τελικά μία δομή δεδομένων έτοιμη για επεξεργασία (Πίνακας 2). Μπορεί για παράδειγμα να παραμετροποιηθεί το πλήθος των πρωτεϊνών που συνεργάζονται σε μία βιολογική διαδικασία, δημιουργώντας μια δομή στην οποία θα αναζητήσουμε τυχούσες συμμετρίες, ομοιότητες ή μοτίβα, εφαρμόζοντας τεχνικές Εξόρυξης.

2.3. Κατηγοριοποίηση των Τεχνικών Εξόρυξης

Η Εξόρυξη από Δεδομένα είναι ζωτικής σημασίας για την Βιοπληροφορική προσφέροντας στον ερευνητή πολλά περισσότερα από μια γονιδιακή αναζήτηση, όπως το Ensembl ή το UCSC Genome Browser. Μέσω της εξόρυξης μπορούν

να απαντηθούν πιο σύνθετα ερωτήματα, όπως το ποια είναι η βιολογική σημασία των αποτελεσμάτων που εξάγει μια πλατφόρμα μικρο-συστοιχιών ή ακόμα πώς ταυτοποιείται, μεταξύ άλλων, ένα βραχύ μοτίβο κοντά στην περιοχή και ανωφορικά (upstream) ενός γονιδίου (Birney, 2008).

Έχοντας μια δομή ή διάταξη δεδομένων στην μορφή που αναφέρθηκε παραπάνω, είναι πλέον δυνατή η εργασία πάνω σε αυτά. Η Εξόρυξη από Δεδομένα είναι η διαδικασία αναζήτησης κανόνων, μοτίβων ή γενικότερα συμμετριών, που περιγράφουν φαινομενολογικά μία διάταξη δεδομένων. Σε πρώτη ανάλυση, οι βασικές λειτουργίες είναι τρεις, η Ταξινόμηση (Classification), η Αναγνώριση Συστάδων (Clustering Recognition) και η Εύρεση Κανόνων Συσχέτισης (Association Rules)(Witten et al., 2016). Όλες οι τεχνικές που εμπεριέχονται στην ομπρέλα του Data Mining μπορούν να ταξινομηθούν σε μία απ' τις τρεις παραπάνω κατηγορίες. Η κάθε μια κατηγορία διαχωρίζεται με την σειρά της σε επιμέρους τεχνικές, που πηγάζουν από διαφορετικές προσεγγίσεις και εκτελούνται με διαφορετικούς αλγόριθμους.

Οι προσεγγίσεις εξόρυξης μπορούν γενικά να κατηγοριοποιηθούν δευτερευόντως και με ένα άλλο τρόπο σε δύο κατηγορίες, περιγραφικές προσεγγίσεις και προγνωστικές προσεγγίσεις. Οι περιγραφικές προσεγγίσεις στοχεύουν στην αναγνώριση μοτίβων που χαρακτηρίζουν δεδομένα, ενώ οι προγνωστικές προσεγγίσεις στην κατασκευή μοντέλων που προσομοιώνουν την λειτουργία ενός συστήματος βάσει της εκπαίδευσής τους σε γνωστά δεδομένα (Charaniya et al., 2008).

2.3.1 Ταξινόμηση

Σχεδόν σε όλα τα προβλήματα, που έχουν να κάνουν με πληθώρα αντικειμένων ή δεδομένων, η μεθοδολογία της λύσης έχει ένα προπαρασκευαστικό στάδιο ταξινόμησης. Εξυπηρετεί γιατί οργανώνει τα πολυπληθή αντικείμενα και μειώνει την πολυπλοκότητα του προβλήματος. Αυτό συμβαίνει διότι διαφορετικά αντικείμενα έχουν διαφορετικό χειρισμό. Γενικά, ως ταξινόμηση ορίζεται η συστηματική διευθέτηση αντικειμένων σε ομάδες βάσει κριτηρίων που αφορούν τα χαρακτηριστικά τους.

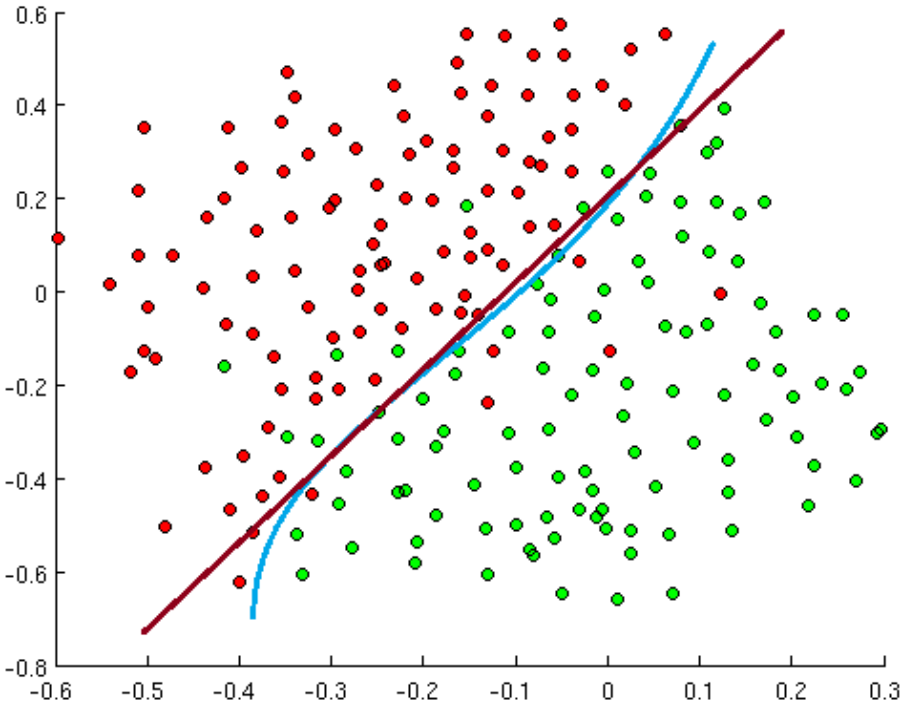
Οι τεχνικές ταξινόμησης βασίζονται στην ιδέα της Στατιστικής, ότι ένα αντιπροσωπευτικό δείγμα ενός συνόλου, περιέχει το ίδιο ποσό πληροφορίας με το σύνολο στο οποίο ανήκει. Έτσι απλά και μόνο μελετώντας το δείγμα μπορεί να εξαχθεί ένα ικανοποιητικό κλάσμα της πλήρους γνώσης. Προφανώς, και δεν είναι πάντα στο μικροσκόπιο το πλήρες σύστημα, οπότε ένα δείγμα αυτού αρκεί. Άρα η ταξινόμηση γίνεται με τον τρόπο που υποδεικνύεται από το δείγμα. Το δείγμα σε αυτή την περίπτωση είναι μια δομή όπως περιγράφηκε παραπάνω, ένας πίνακας δηλαδή $n \times m$ με καθορισμένες κλάσεις. Από αυτόν μπορεί να δομηθεί μια μεθοδολογία ή ένας αλγόριθμος που να ταξινομεί τα στιγμιότυπα με βέλτιστο τρόπο ως προς τις κλάσεις του. Αυτή η μεθοδολογία είναι μια διαδικασία μάθησης, που μπορεί να περιλαμβάνει είτε την εκπαίδευση κάποιου μοντέλου, την κατασκευή Νευρωνικών Δικτύων ή Δέντρων Αποφάσεων, είτε την στατιστική παρατήρηση της υπό μελέτη δομής.

Υπάρχει μια πληθώρα από τεχνικές ταξινόμησης απ’ τις οποίες είναι δυνατό να επιλεγούν, επιτελώντας πολλές διαφορετικές εργασίες και έχοντας συμπεριφορά διαφορετική ανάλογα με τα δεδομένα και τη διαδικασία μάθησης. Το σημαντικό είναι να γίνει μια διάκριση των χαρακτηριστικών κάθε αλγορίθμου, χωρίς ιδιαίτερη εμβάθυνση, ώστε να γίνει κατανοητή η ποικιλία των αποτελεσμάτων.

Η γενικότερη κατηγορία ονομάζεται επιβλεπόμενη μάθηση η οποία υποστηρίζει από κοινού την γραμμική και μη-γραμμική ταξινόμηση. Το ποιο είδος ταξινόμησης θα επιλεγεί εξαρτάται απ’ την δομή των δεδομένων και την μορφή που έχουν αυτά, ως προς τη λήψη απόφασης. Βασικό χαρακτηριστικό της ταξινόμησης είναι ότι εργάζεται βάσει κάποιου προτύπου, το οποίο ακολουθεί. Θεωρείται ότι επιβλέπεται από κάποιο δείγμα και ως εκ τούτου καλείται επιβλεπόμενη διαδικασία.

Στη γραμμική ταξινόμηση η κλάση κάθε τμήματος απ’ το δείγμα, προκειμένου να χαρακτηριστεί η τιμή συγκρίνεται με ένα σταθερό κατώφλι. Για παράδειγμα, έστω ένα σύνολο σημείων που είναι γνωστό εκ των προτέρων ότι ανήκει σε δύο κλάσεις. Η αναπαράσταση γίνεται στις δύο διαστάσεις. Η γραμμική ταξινόμηση προσπαθεί να διαχωρίσει τα σημεία με μια ευθεία γραμμή, ενώ η μη-γραμμική με μια καμπύλη (**Εικόνα 6**). Ο τρόπος ταξινόμησης ενδέχεται να μην είναι πλήρης, δηλαδή ένα στιγμιότυπο ή σημείο μπορεί να ταξινομηθεί σε κλάση που δεν

ανήκει, χωρίς αυτό να επηρεάζει καθοριστικά το αποτέλεσμα. Αυτό που ενδιαφέρει είναι το ποσοστό των ταξινομημένων σωστά και η σχέση τους με το ποσοστό της λάθος ταξινομημένων.

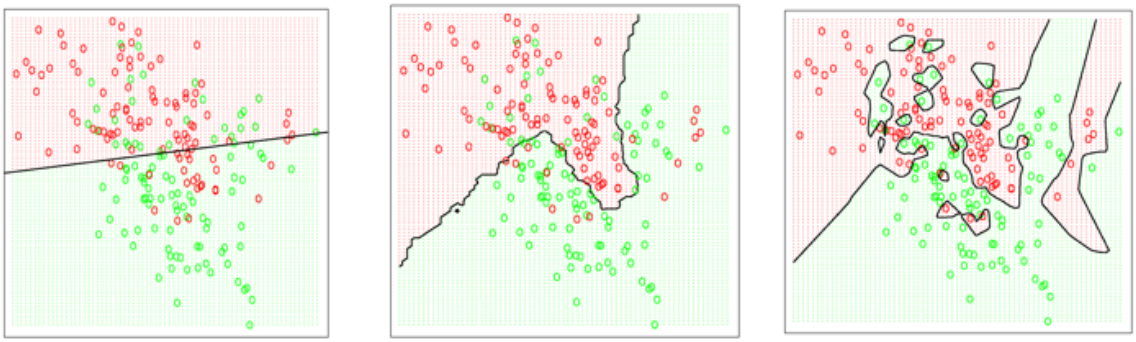


Εικόνα 6. *Γραμμική και μη-γραμμική ταξινόμηση. Στην εικόνα φαίνεται ένα σύνολο στιγμιοτύπων ταξινόμησης σε δύο τάξεις. Τα σημεία με διαφορετικό χρώμα είναι γνωστό εκ των προτέρων ότι ανήκουν δύο κλάσεις. Οι δύο ειδών ταξινομήσεις αναφέρονται: γραμμική (μπορντό γραμμή) και μη-γραμμική (κυανή γραμμή).*

(Η εικόνα ελήφθη απ' το διαδίκτυο: https://cs.stackexchange.com/questions/16545/what-is-the-difference-between-a-neural-network-a-deep-learning-system-and-a-de/44485)

Η εφαρμογή εποπτευόμενης μάθησης στην Βιοπληροφορική μπορεί να έχει διπλό όφελος. Πρώτον, για να περιορίσει τον ασαφή χαρακτήρα των δεδομένων. Μια διάταξη δεδομένων τείνει να είναι πιο ασταθής, όταν ο διαμοιρασμός των στοιχείων δεν είναι ίσος ή υπάρχουν κλάσεις που συγκεντρώνουν μεγάλο πλήθος. Δεύτερο και μεγαλύτερο είναι το πρόβλημα της υπέρ-προσαρμογής (overfitting) (**Εικόνα 7**), κατά την οποία το μοντέλο ταξινόμησης προσκολλάται στα δεδομένα εκπαίδευσης. Αριστερά στην εικόνα (**Εικόνα 7**) ένα παράδειγμα γραμμικής ταξινόμησης, όπου τα στοιχεία κάθε συνόλου διαχωρίζονται από μία βέλτιστη ευθεία και ο διαχωρισμός δεν είναι ακριβής. Κεντρικά στην εικόνα (**Εικόνα 7**) η ταξινόμηση γίνεται βάση μιας ιδιαίτερης συνάρτησης, όπου τα

διαφορετικά στοιχεία είναι καλύτερα διαχωρισμένα, αλλά όχι πλήρως. Δεξιά στην εικόνα (**Εικόνα 7**) ένα παράδειγμα overfitting και ενώ τα στοιχεία είναι ταξινομημένα με βέλτιστο τρόπο (πλήρης ταξινόμηση), η συνάρτηση διαχωρισμού δεν εκφράζει κάποια φυσική σημασία, αντιθέτως μοιάζει να έχει γίνει επιτηδευμένα στα πλαίσια του συγκεκριμένου παραδείγματος. Η προσκόλληση αυτή δεν επιτρέπει να παράξει ταξινομήσεις σε άλλα δεδομένα και αυτό οφείλεται στο ότι δεν εκπαιδεύτηκε απ' τα παραδείγματα ώστε να φτιάξει τις κλάσεις, αλλά ταύτισε τις κλάσεις με τα παραδείγματα αυτά. Έτσι, το αποτέλεσμα που δίνει το μοντέλο είναι κατά ισχυρά εξαρτώμενο απ' το σύνολο δεδομένων εκπαίδευσης (training set), με αποτέλεσμα να κάνει πολλές λάθος ταξινομήσεις σε τυχαία δείγματα. Το πρόβλημα της υπέρ-προσαρμογής πηγάζει απ' την κακή εκπαίδευση του μοντέλου να ακολουθεί με προσήλωση την στοχαστικότητα των δεδομένων και τον θόρυβο του δείγματος (Fayyad et al., 1996). Το πρόβλημα αυτό είναι συχνά εγγενές του συνδυασμού μεθόδων και δεδομένων και χρίζει ριζικής αντιμετώπισης.



Εικόνα 7. *Υπέρ-προσαρμογή (Overfitting). Τρία παραδείγματα εκπαίδευσης σε ταξινόμηση.*
(Η εικόνα ελήφθη απ' το διαδίκτυο: https://mathbabe.org/2012/11/20/columbia-data-science-course-week-12-predictive-modeling-data-leakage-model-evaluation/)

Το κατά πόσο είναι λειτουργικό ένα μοντέλο και μπορεί να προσομοιάσει την πραγματικότητα, μπορεί να αποτυπωθεί ως προς τρεις ιδιότητες: Την ικανότητα να καταστέλλει τα μεροληπτικά συστηματικά σφάλματα (bias) στην ταξινόμηση, την ικανότητα να προσαρμόζεται σε πολλά προβλήματα και την οικονομία της μεθόδου ως προς την πολυπλοκότητα του μοντέλου. Αυτά οι τρεις ιδιότητες στο βαθμό που υπάρχουν, αποτελούν σημεία κλειδιά στην λειτουργία ενός μοντέλου.

Η εποπτευόμενη εκπαίδευση ή αλλιώς Ταξινόμηση είναι μια τεχνική εξόρυξης αρχικά εφαρμόστηκε στην αναζήτηση σε βιβλιοθήκες, όπου η δομή ήταν ήδη γνωστή. Τέτοιες ενέργειες απαιτούν δύο φάσεις, εκπαίδευση και επαλήθευση-αξιολόγηση εκπαίδευσης. Κατά την εκπαίδευση, κατασκευάζεται μοντέλο βασισμένο σε αντιπροσωπευτικά δείγματα, που απαιτούν την εισροή παραμέτρων προκειμένου να επέλθει μάθηση. Με αυτόν τον τρόπο τα μοντέλα μπορούν αμέσως και αποδοτικά να προγνώσουν με βάση την διαδικασία εκπαίδευσής τους. Στη συνέχεια εξάγονται χαρακτηριστικά ενδιαφέροντος απ' το μοντέλο. Εδώ επιβεβαιώνεται ότι η λίστα με τα χαρακτηριστικά-παραμέτρους δεν θα είναι πολύ μεγάλη, με σκοπό να αποφευχθεί η πολυπλοκότητα των διαστάσεων. Όταν ολοκληρωθεί η φάση της εκπαίδευσης, είναι δυνατή η αξιολόγηση του μοντέλου. Κατά την επαλήθευση το εκπαιδευμένο μοντέλο καλείται να ταξινομήσει τυχαίο δείγμα δεδομένων. Επίσης, εκτελούνται ρουτίνες αξιολόγησης όπως η holdout και η k-fold, cross-validation (Liew et al., 2005). Οι ρουτίνες αυτές είναι τυποποιημένοι αλγόριθμοι ικανοί να ελέγξουν την ικανότητα ταξινόμησης του μοντέλου.

Η κατασκευή του μοντέλου αυτού μπορεί να γίνει με διάφορους τρόπους ανάλογα με τις ανάγκες. Ο περισσότερο εύληπτος ως διαδικασία τρόπος μοντελοποίησης είναι τα δέντρα αποφάσεων. Είναι μια επαλληλία ερωτήσεων διαλογής, των οποίων οι απαντήσεις διακλαδίζονται, στην οποία υπάγονται όλα τα στοιχεία των δεδομένων προκειμένου να ταξινομηθούν. Άλλος τρόπος μοντελοποίησης είναι ο στατιστικός. Όπου ο νόμος των δεσμευμένων πιθανοτήτων μετατρέπεται σε αλγόριθμο, με το όνομα Naive Bayes, για να ταξινομήσει με πιθανοκρατικό τρόπο τα στοιχεία, χρησιμοποιώντας το δείγμα εκπαίδευσης σαν στατιστικό δείγμα (Liew et al., 2005). Τέλος, όταν καμία μεθοδολογία δεν δουλεύει και το σύνολο στιγμιοτύπων εκπαίδευσης είναι ανεπαρκές, είναι προτιμητέα η τροφοδότηση των δεδομένων σε ένα Τεχνητό Νευρωνικό Δίκτυο. Τα Νευρωνικά Δίκτυα είναι ικανά δίνουν αποτελέσματα, εκεί που απουσιάζει μεθοδολογία τόσο επεξεργασίας όσο και μοντελοποίησης.

Πλέον είναι γνωστή η δομή των δεδομένων. Αποτελούνται από γραμμές με τιμές γνωστών μεταβλητών που ονομάζονται στιγμιότυπα και είναι στοιβαγμένα κατακόρυφα σχηματίζοντας ένα δισδιάστατο πίνακα. Αν υποτεθεί ότι κάθε στιγμιότυπο αντιπροσωπεύει μία εκδοχή, τότε για την πληρέστερη περιγραφή της

εκδοχής θα προτιμούνται όλο και περισσότερα στοιχεία, γεγονός που θα έκανε την γραμμή να έχει μεγαλύτερο μήκος, κάτι που θα επηρέαζε και τον πίνακα δεδομένων. Η σκέψη αυτή εκ πρώτης φαίνεται καλή αφού αυξάνει την υπό επεξεργασία πληροφορία (Papalexakis & Faloutsos, 2016). Υπάρχει ένας περιορισμός στον αριθμό των παραμέτρων που σχετίζεται μια κλάση. Αν η απόφαση για την κλάση απαιτεί πολύ μεγάλο αριθμό παραμέτρων αυξάνεται η πολυπλοκότητα και καθυστερεί να συγκλίνει το μοντέλο.

Η ερευνητικές εργασίες που έχουν γίνει παρουσιάζουν τις συνέπειες της αύξησης των διαστάσεων μια δομής δεδομένων. Επιχειρήθηκαν σχεδιασμοί αλγορίθμων με εξειδίκευση στα μεγάλα δεδομένα. Η ανάλυσή τους στόχευε στο να αποδείξει, πως η αύξηση των δειγμάτων σε μια διάταξη δεδομένων επιφέρει μείωση στη διακύμανση, για την ανάπτυξη αλγορίθμων που παρέχουν πιο ενδιαφέροντα παράγωγα. Η διακύμανση είναι το μέγεθος που μετράει τον βαθμό της εκάστοτε πρόγνωσης, που εκτελείται κατά την ταξινόμηση από το εκπαιδευμένο μοντέλο και διαφέρει αναλόγως τα δεδομένα. Σε περίπτωση που ο αριθμός των δεδομένων εκπαίδευσης είναι μικρός, τότε εφαρμόζεται ο σχετικιστικός παράγοντας αυτών, ωστόσο δεν είναι αρκετός για να αναπαραστήσει όλο τον πληθυσμό του δείγματος. Έτσι, μπορεί να αναμένεται μεγάλη διακύμανση στα επίπεδα εκπαίδευσης (Liew et al., 2005).

Το στατιστικό συμπέρασμα που εξάγεται από τα πειράματα ισχυροποιεί την υπόθεση και έτσι η ταξινόμηση διεκπεραιώνεται καλύτερα με την αύξηση των δειγμάτων που χρησιμοποιήθηκαν για την εκπαίδευσή τους. Αυτό αποτελεί απόδειξη ότι η μείωση της διακύμανσης στην συμπεριφορά των μοντέλων συνδέεται άμεσα με την ευρεία διακύμανση τιμών κατά την εκπαίδευσή τους. Επομένως η ευστάθεια ενός μοντέλου χτίζεται κατά την διαδικασία εκπαίδευσης (Fayyad et al., 1996).

Στατιστικά Μοντέλα, τεχνική Naive Bayes

Ο νόμος του Bayes για τις δεσμευμένες πιθανότητες ενέχει μια δυναμική, αφού προκειμένου να υπολογιστεί απαιτεί την γνώση των πιθανοτήτων από δύο χρονικές στιγμές του συστήματος. Η μέθοδος αυτή ονομάζεται Naive Bayes, καθώς βασίζεται στον κανόνα του Bayes, αλλά με τον τρόπο της μεθόδου του Naive υποτίθεται η ανεξαρτησία των πιθανοτήτων. Η υπόθεση ότι οι παράμετροι

είναι ανεξάρτητες στον πραγματικό κόσμο φαντάζει πολύ απλουστευτική, αλλά δίνει ικανοποιητικά αποτελέσματα σε πραγματικά προβλήματα. Σαν μελανό σημείο η μέθοδος αυτή έχει μια αριθμητική ιδιαιτερότητα. Στον συνδυασμό πιθανοτήτων μπορεί πολύ εύκολα να προκύψει πολλαπλασιασμός με το μηδέν, που «σκοτώνει» μεταφορικά την πληροφορία. Όπως παραστατικά λέγεται οι μηδενικές πιθανότητες ασκούν βέτο. Όταν όμως κάτι τέτοιο συμβαίνει τα πράγματα δεν βαίνουν καλώς. Το πρόβλημα αυτό μπορεί εύκολα να λυθεί με τις λιγότερες προσθήκες στη μέθοδο, υπολογίζοντας τις πιθανότητες των συχνοτήτων (Wittenet al., 2016).

Έστω ότι σε ένα σύστημα μπορούν να συμβούν δύο τύποι γεγονότων Α, Β που είναι ασυσχέτιστα μεταξύ τους. Η εξίσωση μας δίνει την πιθανότητα να γίνει το Α, υπό την συνθήκη ότι έχει προηγηθεί το Β και αυτό που επιτυγχάνει είναι ο συσχετισμός των δύο υπό συνθήκη συμβάντων. Η εξίσωση του Bayes για τις δεσμευμένες πιθανότητες είναι:

$$P[A \vee B] = \frac{P[B \vee A] \cdot P[A]}{P[B]}$$

Η εξίσωση λειτουργεί μόνο για δύο ενδεχόμενα Α και Β. Μπορεί να γίνει γενίκευση σε πολλά γεγονότα. Η γενίκευση στηρίζεται στην ιδέα πως το δεύτερο ενδεχόμενο μπορεί να περιγράφει όλα τα ενδεχόμενα εκτός απ’ το Α, περιγράφοντας πλήρως το τι θα συμβεί αν όχι Α. Έτσι κάθε πρόβλημα αποκτά μορφή στον δυαδικό χώρο.

Δέντρα Λήψης Αποφάσεων

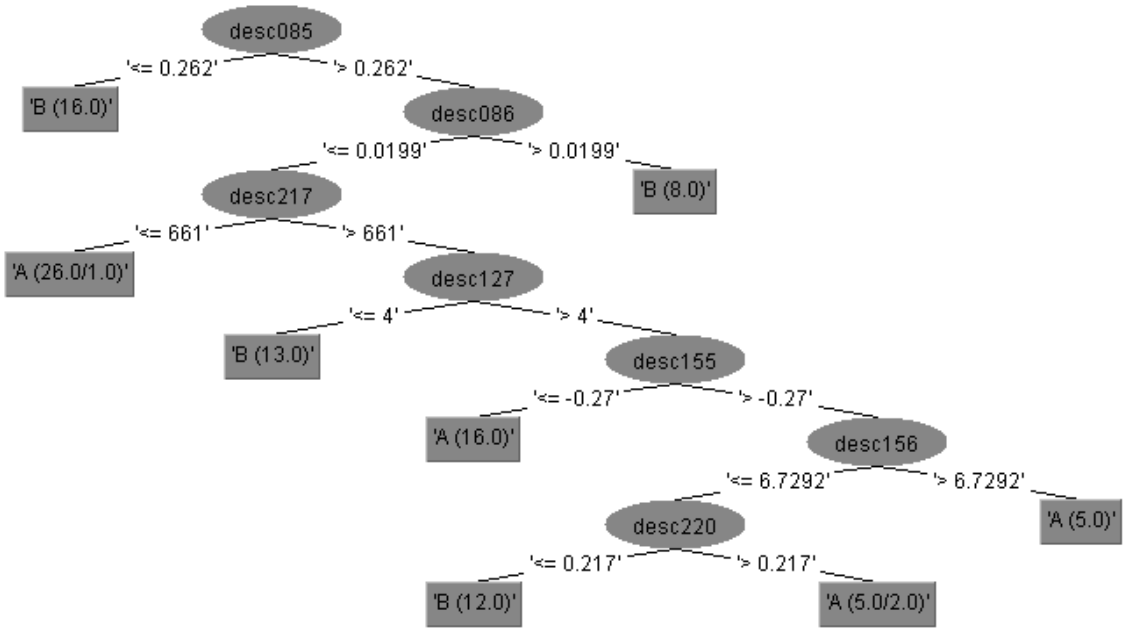
Τα Δέντρα Λήψης Απόφασης αποτελούν μια σειριακή διαδικασία με αλλεπάλληλες ερωτήσεις ως προς τις παραμέτρους της διάταξης, που διαχωρίζουν τα στοιχεία ανάλογα με την τιμή που έχουν σε κάθε παράμετρο. Η κάθε ερώτηση συμβολίζεται με διακλάδωση, αφού κάθε στοιχείο ακολουθεί διαφορετικό δρόμο ανάλογα με την τιμή που λαμβάνει. Όταν δεν μπορεί να γίνει επιπλέον διαχωρισμός, θεωρείται ότι έχει επέλθει ταξινόμηση για το σύνολο των στοιχείων, λέμε ότι το σύνολο αυτό αποτελεί ένα φύλλο. Αυτή είναι η προσομοίωση μιας διεργασίας, που αρχικά παίρνει σε δέσμη όλη τη διάταξη και την υποχρεώνει σε διακλαδώσεις, μέχρι που τις ταξινομεί σε φύλλα. Εκ των

υστέρων η μέθοδος ελέγχου και αξιολόγησης του μοντέλου είναι καταγραφή του ποσοστού των κακώς ταξινομημένων, στα φύλλα στοιχείων.

Προτεραιότητα αποτελεί η επιλογή της παραμέτρου που θα τεθεί στην ρίζα του δέντρου, πρακτικά αυτό σημαίνει ως προς ποια μεταβλητή θα παραμετρικοποιηθούν και θα χωριστούν τα στιγμιότυπα. Στην συνέχεια πρέπει να οριστούν τα επιμέρους υποσύνολα που προκύπτουν. Στην προσπάθεια αυτή αξιολογείται η γνώση που αποταμιεύτηκε από κάθε ερώτηση, σύμφωνα με κανόνες από την Θεωρία Πληροφορίας. Το πρόβλημα της κατασκευής δένδρων απόφασης μπορεί να εκφραστεί σε διάφορα στάδια. Πρώτον, επιλέγεται η παράμετρος που θα τεθεί στην ρίζα ή αρχή του δέντρου και δημιουργείται ένα κλαδί για κάθε πιθανή τιμή που δίνουν διαφορετικά στιγμιότυπα. Έτσι, με κανόνα την τιμή που φέρει το κάθε στιγμιότυπο μπορούμε να τα χωρίσουμε σε επιμέρους κατηγορίες. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να εξαντληθούν οι παράμετροι των δεδομένων, άρα πρακτικά να γίνουν οι ερωτήσεις για κάθε μία παράμετρο. Κατά την διαδικασία, όλοι κλάδοι έχουν δεδομένα με την ίδια τιμή με την παράμετρο που μόλις ερωτήθηκε στην διακλάδωση, και θα διαφέρουν μόνο ως προς τις παραμέτρους που έπονται να ερωτηθούν. Αν οποιαδήποτε στιγμή όλα τα στιγμιότυπα που φτάνουν σε έναν κόμβο ανήκουν στην ίδια κλάση, τότε το συγκεκριμένο κλαδί δεν έχει άλλη ανάπτυξη, δηλαδή δεν διακλαδίζεται επιπλέον. Αποτελείται ιδανικά από στιγμιότυπα μιας κλάσης και ονομάζεται φύλλο. Σε περίπτωση που περιέχει στιγμιότυπα από άλλες κλάσεις αυτό είναι ανεκτό, υπό την συνθήκη ότι αυτά αποτελούν μειονότητα.

Η διαδικασία είναι απλή και απαιτείται μόνο ένας τρόπος με τον οποίο θα ιεραρχήθουν τις ερωτήσεις. Ως προς ποια παράμετρο θα διαχωριστούν σε πρώτη φάση τα στιγμιότυπα και πως θα επιλεγεί η επόμενη παράμετρος. Για αυτή την εργασία απευθυνόμαστε στη Θεωρία Πληροφορίας, θέτοντας το ερώτημα: Ποιες απαντήσεις θα μας δώσουν την μεγαλύτερη ποσότητα πληροφορίας; Με άλλα λόγια βάσει ποιας μεταβλητής το σύνολο των στοιχείων θα χωριστεί με τρόπο ώστε τα υποσύνολα που θα προκύψουν να είναι όσο το δυνατόν πιο ομοιογενή; Και έτσι στην ιεραρχία των ερωτήσεων προηγούνται αυτές που διαχωρίζουν καλύτερα τα στιγμιότυπα. Το μέτρο ταξινόμησης που χρησιμοποιείται ονομάζεται πληροφορία και μετριέται σε μονάδες που ονομάζονται bits. Σχετιζόμενο με τον κάθε κόμβο το μέγεθος αυτό αναπαριστά το

αναμενόμενο ποσό πληροφορίας, που θα χρειαστεί για να καθοριστεί εάν ένα νέο στιγμιότυπο ανήκει στη μία κλάση ή στην άλλη. Σε αντίθεση με τα bits του υπολογιστή, η τιμή τους είναι κλάσματα της μονάδας, στο σύνολο [0,1], διότι προκύπτουν από πιθανότητες (Witten, Frank, & Hall, 2016). Πάντα προτιμούνται αυτές οι διακλαδώσεις που δίνουν μεγαλύτερο ποσό πληροφορίας, άρα τιμή πιο κοντά στο 1, μέχρι να ταξινομηθεί το σύνολο των υπό επεξεργασία δεδομένων.



Εικόνα 8. Ταξινόμηση αντικειμένων με την χρήση Δενδρικού διαγράμματος Λήψης Απόφασης. Τα αντικείμενα ταξινομούνται βάσει της τίμης που λαμβάνουν για διάφορες παραμέτρους (ελλειπτικά πλαίσια).

(Η εικόνα ελήφθη απ' το διαδίκτυο: http://weka.8497.n7.nabble.com/SOLVED-j48-tree-display-number-of-instances-and-errors-in-leaves-for-test-set-instead-of-training-set-td37617.html)

Αναλύοντας ένα παράδειγμα ενός τέτοιου Δέντρου Ταξινόμησης είναι ευκολότερα κατανοητή η λειτουργία του. Το παρόν δενδρικό διάγραμμα έχει δημιουργηθεί με το λογισμικό Weka **(Εικόνα 8)** και ταξινομεί τα στιγμιότυπα που βρίσκονται στα ορθογώνια πλαίσια. Πριν καταλήξει ένα στοιχείο σε ορθογώνιο πλαίσιο περνάει υποχρεωτικά από ένα ελλειπτικό πλαίσιο. Κάθε ελλειπτικό πλαίσιο αποτελεί μια μεταβλητή βάση της οποίας γίνεται ο διαχωρισμός των στιγμιοτύπων. Ο διαχωρισμός των στιγμιοτύπων συμβολίζεται με τις διακλαδώσεις που ακολουθούν τις ελλείψεις. Στιγμιότυπα που καταφθάνουν σε

ένα ορθογώνιο δεν θα υποστούν περαιτέρω διαχωρισμό, έτσι τα ορθογώνια ως απολήξεις του δενδρικού διαγράμματος καλούνται και φύλλα, σε πλήρη αναλογία με τα δέντρα του φυσικού κόσμου. Στο συγκεκριμένο παράδειγμα διαχωρίζονται στιγμιότυπα τύπου Α και Β με βάση 7 απ' τις μεταβλητές ή ιδιότητες που τα χαρακτηρίζουν. Στην προκειμένη περίπτωση οι μεταβλητές μας είναι: desc085, desc086, desc217, desc127, desc155, desc156, desc220 (**Εικόνα 8**).

Τεχνητά Νευρωνικά Δίκτυα

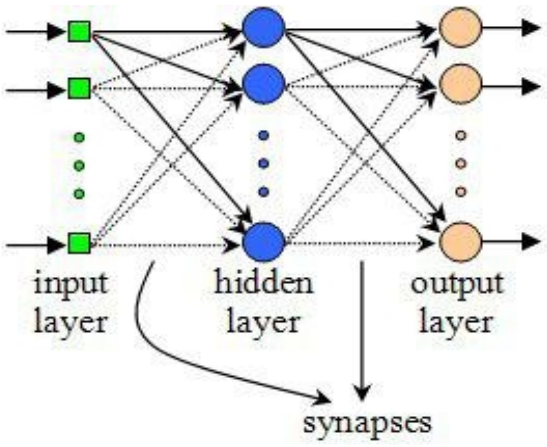
Τα Τεχνητά Νευρωνικά Δίκτυα αποτελούν μια νέα προσέγγιση στην μοντελοποίηση και προσομοίωση λειτουργιών που επιτελούνται από περίπλοκα συστήματα. Τέτοιες δύσκολες λειτουργίες είναι η αναγνώριση προτύπων, ο υπολογισμός συναρτήσεων, η βελτιστοποίηση, η πρόβλεψη, ο αυτόματος έλεγχος, οι οποίες αποτελούν απαιτήσεις όλο και περισσότερων επιστημών αλλά και εμπορικών εφαρμογών (Αργυράκης, 2001). Η βασική ιδέα για την δομή τους πηγάζει από τον τρόπο λειτουργίας των νευρικών συστημάτων στους έμβιους οργανισμούς.

Το 1943 οι McCulloch και Pitts δημιουργούν ένα πρωταρχικό υπολογιστικό μοντέλο Νευρικών Δικτύων που βασίζεται στην λογική της ενεργοποίησης του κάθε νευρώνα υπό τη συνθήκη ότι η ηλεκτρική τάση που δέχεται υπερβαίνει ένα κατώφλι. Όταν υπερβαίνει το κατώφλι αυτό τότε ο νευρώνας ενεργοποιείται και μεταφέρει την τάση στον επόμενο νευρώνα (Carpenter, 1989). Έτσι το πλήθος των νευρώνων που βρίσκονται σε σύνδεση μεταξύ τους, μπορεί να παρομοιαστεί με ένα δίκτυο ηλεκτρικών διακοπτών που συνδέονται με ηλεκτρικούς αγωγούς (καλώδια). Σε πλήρη αναλογία με τους έμβιους οργανισμούς τα δίκτυα αυτά έχουν μια κατεύθυνση, στην ροή της οποίας κάποιοι νευρώνες προηγούνται και άλλοι έπονται, δημιουργώντας μια πολύπλοκη αλυσίδα που ενώνει τον τελικό αποδέκτη (εγκέφαλο) με το περιβάλλον (εκτός του έμβιου σώματος). Στην περίπτωση των Τεχνητών Νευρωνικών η αλυσίδα αυτή ενώνει την είσοδο των δεδομένων (input) με την απόφαση (output) (**Εικόνα 9**).

Η αρχιτεκτονική τέτοιων δικτύων χαρακτηρίζεται από τρία επίπεδα, το επίπεδο εισόδου (input layer) αποτελούμενο απ’ τους νευρώνες που δέχονται την εισροή, το επίπεδο εξόδου (output layer) που εξέρχεται η απόφαση απ’ την επεξεργασία που πραγματοποιήθηκε και το κρυφό επίπεδο (hidden layer) που

αποτελείται απ’ τους νευρώνες που παρεμβάλλονται μεταξύ εισόδου και εξόδου. Σημειωτέον, το κρυφό επίπεδο μπορεί και να απουσιάζει (παράδειγμα δικτύων Kohonen).

Βασικό χαρακτηριστικό για την λειτουργία της δομής αυτής είναι το κατώφλι ενεργοποίησης των νευρώνων το οποίο δεν είναι για όλους τους νευρώνες το ίδιο. Άλλοι νευρώνες είναι πιο «ευαίσθητοι» και άλλοι πιο «αδρανείς» και στην γλώσσα των τεχνικών το μέγεθος αυτό αποκαλείται βάρος του νευρώνα και συμβολίζεται με w. Επίσης πολύ βασικό χαρακτηριστικό της λειτουργίας είναι και ο τρόπος που κάθε νευρώνας ενεργοποιεί τον επόμενό του. Στη γενική του μορφή ο κανόνας είναι μια συνάρτηση η οποία επιλέγεται καταλλήλως απ’ τον δημιουργό του δικτύου.

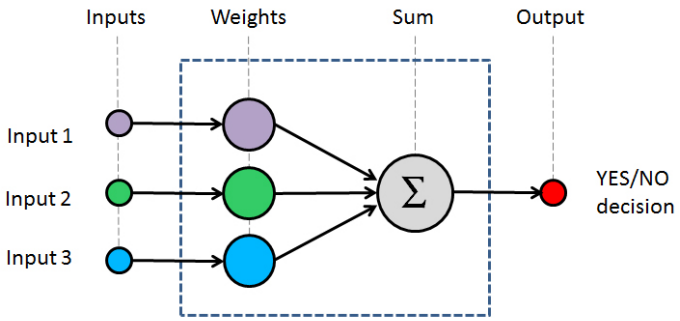


Εικόνα 9. *Σχηματική αναπαράσταση ενός Τεχνητού Νευρωνικού Δικτύου.*
(Η εικόνα ελήφθη απ' το διαδίκτυο: https://www.codeproject.com/Articles/9447/Neural-Network-Classifier)

Όπως και στους έμβιους οργανισμούς, η πιο εξεζητημένη και κομβική διαδικασία είναι η εκπαίδευση των νευρώνων και όχι η δημιουργία τους. Πρακτικά ως εκπαίδευση εννοείται ο καθορισμός των βαρών (w) που χαρακτηρίζουν το κατώφλι ενεργοποίησης του κάθε νευρώνα (**Εικόνα 10**). Για την εκπαίδευση είναι απαραίτητα κάποια δεδομένα ιδίου τύπου και δομής με εκείνα τα οποία θα μπορεί να επεξεργάζεται, όταν θα είναι πλέον εκπαιδευμένο. Η διαδικασία εκπαίδευσης αξιολογείται από το πόσο αποκλίνει απ’ το επιθυμητό αποτέλεσμα.

Για παράδειγμα, σε μία διαδικασία αναγνώρισης χειρόγραφων ψηφίων, από το 0 μέχρι το 9, τα οποία βρίσκονται αποθηκευμένα σε εικόνες, η διαδικασία έχει ως εξής (Nielsen, 2015):

1. Εισάγονται οι n παράμετροι που χαρακτηρίζουν την εικόνα (έστω το ψηφίο 5) στο επίπεδο εισόδου, που μοιραία θα έχει n νευρώνες. Το σύνολο των παραμέτρων αποτελεί ένα στιγμιότυπο.
2. Το δίκτυο επεξεργάζεται την είσοδο στους νευρώνες του κρυφού επιπέδου, βάσει των βαρών w_i που του έχουν ανατεθεί, η τιμή των οποίων δεν έχει ιδιαίτερη σημασία.
3. Τελικά ενεργοποιούνται οι νευρώνες εξόδου, οι οποίοι είναι 10 σε πλήθος όσες και οι κατηγορίες ψηφίων (από 0 μέχρι 9). Ακριβέστερα, δεν ενεργοποιούνται όλοι οι νευρώνες αλλά μόνο ένας και η ενεργοποίησή του ισοδυναμεί με την «απόφαση» για το ποιο ψηφίο εισήχθη στην είσοδο. Η απόφαση αυτή θα είναι κατά πάσα πιθανότητα λάθος. Δεν υπάρχει πρόβλημα γιατί η εκπαίδευση μόλις ξεκίνησε.
4. Τώρα ο αλγόριθμος μπορεί να αλλάξει τις τιμές των βαρών έτσι ώστε η έξοδος να είναι ο νευρώνας 5 και να συμπίπτει με την έξοδο και αυτή η διαδικασία ονομάζεται Οπισθία Διάδοσης Λάθους.
5. Με τα βάρη των νευρώνων w_i του δικτύου να έχουν τις τιμές που δόθηκαν απ' την προηγούμενη διαδικασία, στην είσοδο τίθενται οι παράμετροι μίας νέας εικόνας για τον αριθμό 5 και επαναλαμβάνονται όλα τα προηγούμενα βήματα (Kumar & Abhishek, 2012).



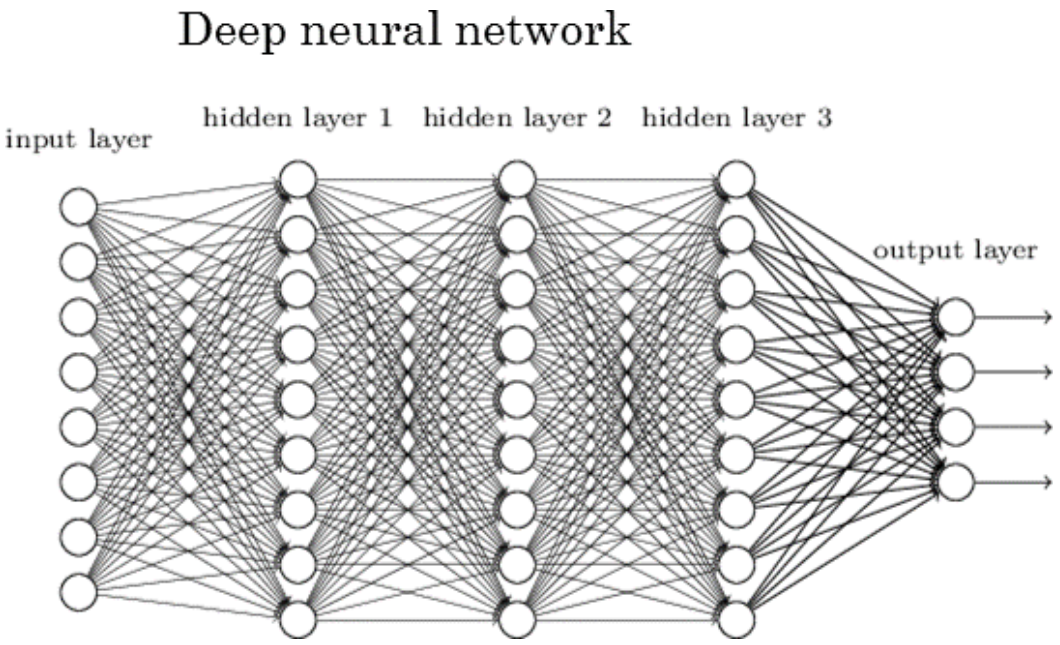
Εικόνα 10. *Σχηματική απεικόνιση τεχνητού νευρώνα (εντός των διακεκομμένων γραμμών).*
(Η εικόνα ελήφθη απ' το διαδίκτυο: http://projectgokturk.blogspot.gr/2015/05/yapay-yapay-zekalar-vol-2.html)

Ύστερα από αρκετές επαναλήψεις, το πλήθος των οποίων καθορίζεται εμπειρικά απ' τον ερευνητή, υπάρχει βάσιμος ισχυρισμός ότι το Νευρωνικό Δίκτυο έχει μάθει να αναγνωρίζει τον χειρόγραφο αριθμό 5. Η διαδικασία επαναλαμβάνεται για τα υπόλοιπα ψηφία. Κατά την εκπαίδευση, φορτώνουμε στους νευρώνες ένα προς ένα όλα τα στιγμιότυπα και επιβάλλεται η απαίτηση να ενεργοποιηθεί μόνο ο νευρώνας με την ορθή κλάση. Το δίκτυο σε περίπτωση που δεν έχει προβλέψει σωστά, διορθώνει κάποιους αριθμούς που επηρεάζουν τη δομή του και προσπαθεί εκ νέου, μέχρι να προσεγγίσει τη λύση με την απόκλιση που του επιτρέπεται. Λόγω του ότι η εκπαίδευση αξιολογείται από το κατά πόσο το αποτέλεσμα του δικτύου ταυτίζεται με συγκεκριμένο στόχο, η εκπαίδευση καλείται εποπτευόμενη. Ύστερα απ’ το πέρας μιας τέτοιας εποπτευόμενης εκπαίδευσης, όπου το δίκτυο θα έχει μάθει να ενεργοποιεί τον σωστό νευρώνα κλάσης για κάθε στιγμιότυπο εκπαίδευσης, είναι έτοιμο να δοκιμαστεί σε άγνωστα δεδομένα. Τα άγνωστα δεδομένα διατηρούν τη μορφή των στιγμιοτύπων που χρησιμοποιήθηκε για την εκπαίδευση (Nielsen, 2015).

Σύμφωνα με την σχηματική περιγραφή που δόθηκε απαιτείται ένα μεγάλο πλήθος δεδομένων, εικόνων στην προκειμένη περίπτωση, που λειτουργούν ως παραδείγματα στην εκμάθηση και προφανώς, όσο περισσότερα είναι τα παραδείγματα τόσο καλύτερη θα είναι η εκπαίδευση. Παρόλα αυτά, το μεγάλο πλήθος παραδειγμάτων είτε δεν είναι πάντα διαθέσιμο, είτε καθυστερεί την εκπαίδευση, χωρίς να προσφέρει μεγάλη βελτίωση στον λειτουργία που καλείται να προσομοιάσει το Νευρωνικό Δίκτυο. Άρα είναι αναγκαίο να οριστεί ένα μέγεθος που να μπορεί να αξιολογεί την εκπαίδευση όταν αυτή ολοκληρωθεί. Έτσι ορίζεται το ποσοστό επιτυχίας, που είναι ο λόγος των σωστά ταξινομημένων παραδειγμάτων προς το πλήθος αυτών. Η παραπάνω διαδικασία εκπαίδευσης είναι η πιο κύρια και επικρατούσα στις εφαρμογές, αλλά δεν είναι η μόνη. Υπάρχουν κι άλλες μέθοδοι όπως η μέθοδος Boltzmann, η μέθοδος Ειδικής Θερμότητας ή μη-Γραμμική Βελτιστοποίηση, που αντλούν την έμπνευσή τους απ’ τις έννοιες και τον τρόπο προσέγγισης της Στατιστικής Φυσικής (Αργυράκης, 2001).

Εξέλιξη της δομής που ήδη περιγράφηκε αποτελούν τα Βαθιά Νευρωνικά Δίκτυα ή Νευρωνικά Δίκτυα Πολλαπλών Επιπέδων (Deep Neural Networks). Η μοναδική διαφορά είναι ότι διαθέτουν περισσότερα από ένα κρυφά επίπεδα

μεταξύ εισόδου και εξόδου (**Εικόνα 11**). Στα Βαθιά Νευρωνικά Δίκτυα κάθε επίπεδο νευρώνων εκπαιδεύεται σε ένα ξεχωριστό σύνολο χαρακτηριστικών, που έχουν προκύψει από την έξοδο του προηγούμενου επιπέδου. Όσο πιο κοντά είμαστε στην έξοδο του δικτύου, τόσο τα χαρακτηριστικά αυτά μπορούν να γίνονται πιο σύνθετα, καθώς ενέχουν μέσα τους αυτά απ' τα προηγούμενα στάδια.



Εικόνα 11. *Βαθιά Νευρωνικά Δίκτυα με πολλά επίπεδα αλληλεπίδρασης (Deep Neural Network).* (Η εικόνα ελήφθη από το διαδίκτυο: https://www.codeproject.com/Articles/1206388/Build-Simple-AI-NET-Library-Part-Artificial-Neural)

Αυτή η Ιεραρχία Χαρακτηριστικών (Feature Hierarchy) και η αύξηση την πολυπλοκότητας, δίνουν στα Βαθειά Νευρωνικά Δίκτυα την ικανότητα να χειρίζονται δεδομένα από πολύ μεγάλες και πολυδιάστατες δομές. Επιδεικνύουν μεγάλη ικανότητα στην εύρεση λανθανόντων ιδιομορφιών και μοτίβων σε αδόμητα και αχαρακτήριστα δεδομένα. Έτσι αποτελούν από τους πιο δυνατούς αλγόριθμους στην ταξινόμηση τεράστιων δομών καθώς και στην ανακάλυψη ανωμαλιών μέσα σε αυτές. Εν πολλοίς, τα δίκτυα αυτά αποτελούν τον μοναδικό ωφέλιμο τρόπο χειρισμού των δεδομένων του Διαδικτύου σχετικά πρόσφατα και αρχίζουν να χρησιμοποιούνται στον χειρισμό και την Εξόρυξη από βιολογικά δεδομένα.

Η λειτουργία των Νευρωνικών Δικτύων δεν έχει αναλυτική περιγραφή, γεγονός που σημαίνει ότι δεν είναι γνωστό σε βάθος πως επιτυγχάνουν τα θαυμάσια και ταχύτατα επιτεύγματά τους. Αυτό είναι ένα θέμα για το οποίο το πλήθος της επιστημονικής κοινότητας που τα χρησιμοποιεί δεν δίνει ιδιαίτερη βάση, μιας και συχνά δίνουν ένα αξιόλογο αποτέλεσμα και εντός πολύ καλού χρονικού ορίου. Έτσι το ζήτημα της πλήρους θεμελίωσης του μαθηματικού μηχανισμού που τα διέπει, απασχολεί μόνο του μαθηματικούς (Nielsen, 2015).

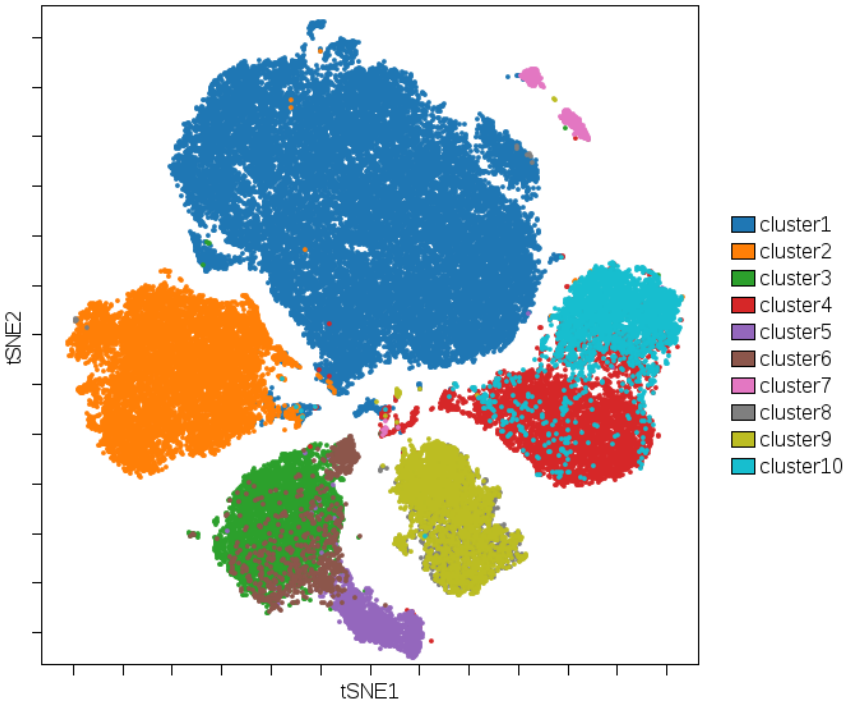
2.3.2 Συσταδοποίηση

Συσταδοποίηση καλείται η ταξινόμηση χωρίς επίβλεψη, όπου τα διάφορα στιγμιότυπα (instances) τείνουν να ομαδοποιηθούν βάσει των χαρακτηριστικών τους. Οι συστάδες (clusters) αποτελούν ομάδες που τα στοιχεία ή στιγμιότυπά τους μοιράζονται κάποια κοινά χαρακτηριστικά. Το πλήθος των ομάδων μπορεί να είναι γνωστό από την αρχή ή να προκύπτει από την εξέλιξη της διαδικασίας.

Πολλές φορές η διαθέσιμη δομή στερείται της στήλης των κλάσεων. Υπάρχει δηλαδή, μια δομή από στιγμιότυπα είτε δεν έχουν χαρακτηριστεί σε κλάσεις, είτε οι κλάσεις που έχουν δεν εμφανίζουν ενδιαφέρον για την μελέτη, οπότε απαιτείται να βρεθούν καινούριες. Η κλάση, αν υπάρχει, δεν παίζει κανένα ρόλο και είτε αφαιρείται, είτε η στήλη των κλάσεων λαμβάνεται υπ’ όψιν ως μια απλή παράμετρος. Στόχος είναι να βρεθεί η κλάση, με την έννοια να βρεθεί ένας τρόπος διακριτοποίησης. Οι διάφορες διαδικασίες προσπαθούν να καταστήσουν διακριτή την διαφοροποίηση των στιγμιοτύπων ως προς μια παράμετρο (**Εικόνα 12**). Αν αυτό είναι δυνατό τότε αυτή η παράμετρος μπορεί να ονομαστεί κλάση.

Η Συσταδοποίηση χρησιμοποιείται για τον διαχωρισμό αντικειμένων σε κατηγορίες βάσει ομοιοτήτων και διαφορών. Η διακριτική ικανότητα διαχωρισμού των κλάσεων εξαρτάται τόσο απ’ την ομοιομορφία των μετρήσεων, όσο και απ’ την ικανότητα να ανακαλυφθούν το σύνολο ή μέρος των κρυφών μοτίβων. Αντικείμενα τα οποία τείνουν να συσσωματώνονται στην ίδια ομάδα, θα πρέπει να έχουν κοινά μεταξύ τους. Μια καλή μέθοδος, που φέρνει αποτέλεσμα αποτελείται από τέσσερα στάδια, την συλλογή χαρακτηριστικών, τον σχεδιασμό του αλγόριθμου που θα επιχειρήσει την αναγνώριση, την επαλήθευση-

αξιολόγηση, αλλά και την οπτικοποίηση και τελική αξιολόγηση του αποτελέσματος (Wunsch and Xu, 2005).



Εικόνα 12. *Συσταδοποίηση ενός συνόλου σημείων όπου το καθένα από αυτά αποτελεί ένα στιγμιότυπο. Τα σημεία έχουν ταξινομηθεί σε 10 συστάδες. Η Συσταδοποίηση έχει γίνει βάσει της θέσης του κάθε σημείου στο παρόν επίπεδο, αλλά και ως προς άλλους χώρους ή διαστάσεις ή ιδιότητες που δεν απεικονίζονται στο σχήμα. Έτσι κοντινά σημεία δεν ανήκουν απαραίτητα στην ίδια συστάδα.*

(Η εικόνα ελήφθη απ' το διαδίκτυο: https://support.cytobank.org/hc/en-us/articles/223232388-Cluster-Channel-and-Cluster-Gates)

Οι τεχνικές Συσταδοποίησης έχουν εφαρμογή στην ανάλυση μικρο-συστοιχιών για να υπερβούν τους περιορισμούς από εξερεύνηση των άγνωστων κλάσεων, ιδίως όταν οι κλάσεις δεν είναι γνωστές απ' την αρχή. Για παράδειγμα, ο καθορισμός του κατά πόσο μια ασθένεια σε έναν συγκεκριμένο ιστό ή κάτω από συγκεκριμένη κατάσταση, μπορεί να επιφέρει την έκφραση ενός γονιδίου, δεν είναι μόνο ζήτημα ύπαρξης ή όχι, αλλά μπορεί να έχει και επιμέρους ιδιαιτερότητες, που να φανούν στη πορεία των πειραμάτων. Επίσης, μπορεί σε κάθε κλάση να υπάρχουν και επιμέρους υπό-κλάσεις.

Η ανάλυση συστάδων ανακαλύπτει πιθανές κλάσεις στις οποίες μπορεί να αναλύεται ένα μεγάλο σύνολο δεδομένων. Οι πιθανές κλάσεις υπολογίζονται είτε

χρησιμοποιώντας μια ιεραρχική δομή, είτε μερίζοντας τα δεδομένα σύμφωνα με έναν ήδη γνωστό κανόνα. Ο κανόνας αυτός περιλαμβάνει κυμαινόμενα σκαλοπάτια από προεργασία στην ανακάλυψη συστάδων και η εύρεσή του αποτελεί μεγάλο κίνητρο για τους ερευνητές. Παρ' όλο που οι διαδικασίες αυτές επιλύουν μερικά προβλήματα, υπάρχουν σοβαροί περιορισμοί. Συνήθως επιβάλλονται προσεγγίσεις που ενέχουν την δημιουργία συστηματικών σφαλμάτων (bias), με αποτέλεσμα καμία απ' τις διαδικασίες να μην αποτελεί την βέλτιστη λύση για όλους τους τύπους προβλημάτων. Έτσι, η επιλογή της τεχνικής που θα ακολουθηθεί αποτελεί ζήτημα του ερευνητή (Uchiyama & Arbib, 1994).

Πλήθος αλγορίθμων συσταδοποίησης έχουν αναπτυχθεί, καθένας απ' αυτούς βασίζεται σε διαφορετική προσέγγιση του προβλήματος. Όλοι τους μοιράζονται ως κοινό χαρακτηριστικό την ίδια μορφή εισροών (inputs), ένα σύνολο παραμέτρων που μπορεί να είναι το πλήθος των ομάδων, διανύσματα αρχικοποίησης που απαιτούνται από τον αλγόριθμο κάποιες υποθέσεις για την πυκνότητα των διανυσμάτων στο χώρο (Estivill-Castro, 2002). Ανάλογα με το ποιες είναι οι παράμετροι αυτοί και πάντα σε συμφωνία με τις ανάγκες του προβλήματος, ο ερευνητής έχει να επιλέξει ανάμεσα σε τρεις κατηγορίες αλγορίθμων:

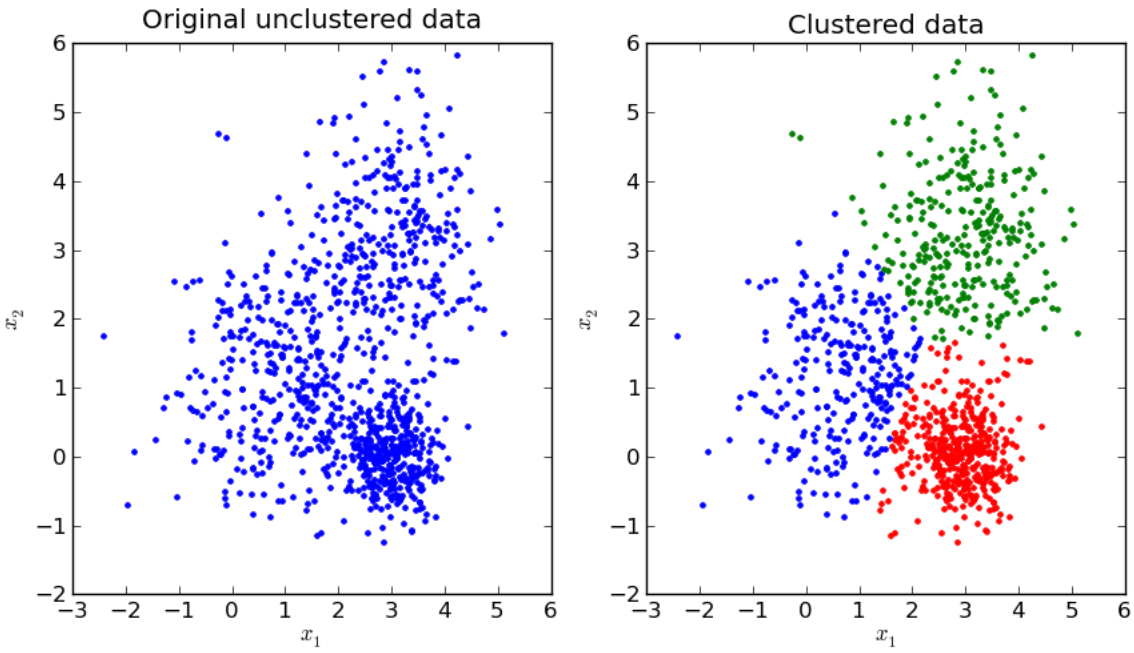
- Αλγόριθμοι Μέσου Όρου (τύπου k-mean),
- Ιεραρχικοί Αλγόριθμοι και
- Αλγόριθμοι Ανταγωνιστικής Μάθησης.

Πρακτικά, για να επιτευχθεί η Συσταδοποίηση θα πρέπει όλες οι τιμές των στιγμιοτύπων να αναπαρασταθούν σε έναν χώρο, όπου η θέση του κάθε στιγμιότυπου στο χώρο καθορίζεται σαφώς απ' τις παραμέτρους που το χαρακτηρίζουν. Ο χώρος αυτός δεν έχει κανένα περιορισμό σε διαστάσεις, αλλά συνήθως είναι δύο ή τριών διαστάσεων. Έτσι, προκύπτει ένας χώρος στον οποίο υφίστανται διατεταγμένα σημεία, τα στιγμιότυπα, και επί του χώρου αυτού δυνητικά χωρίζονται τα σημεία σε ομάδες (Manning et al., 2009).

Αλγόριθμοι Μέσου Όρου

Η λειτουργία των Αλγορίθμων Μέσου Όρου στηρίζεται στην ιδέα ότι τα σημεία-μέλη μιας συστάδας είναι εγγύτερα στο κέντρο ή κεντροειδές της συστάδας στην οποία ανήκουν, από τα κέντρα των υπόλοιπων συστάδων. Απαραίτητα για την εργασία είναι:

- Η γνώση του πλήθους k των συστάδων, η οποία γίνεται απ' τον ερευνητή και σχετίζεται με την φύση του προβλήματος και
- η αρχικοποίηση των θέσεων για κάθε κεντροειδές καθεμίας από τις k συστάδες.



Εικόνα 13. *Ένα σύνολο σημείων που αναπαριστούν στιγμιότυπα (αριστερά) και το ίδιο σύνολο με χρωματισμένες τις τρεις διαφορετικές συστάδες στις οποίες ομαδοποιούνται. Η Συσταδοποίηση σε αυτή την περίπτωση έχει γίνει με βάση την χωρική απόσταση και συγκεκριμένα με τον αλγόριθμο k-means.*

(Η εικόνα ελήφθη απ' το διαδίκτυο: https://mubaris.com/2017-10-01/kmeans-clustering-in-python)

Στη συνέχεια θα πρέπει να οριστεί ο τρόπος με τον οποίο μετρούνται οι αποστάσεις του εκάστοτε σημείου με όλα τα κεντροειδή. Αν ο χώρος είναι ομοιογενής τότε η απόσταση είναι η Ευκλείδεια Απόσταση, αν αντιθέτως ο χώρος είναι ανομοιογενείς τότε υπάρχει μια συνάρτηση που περιγράφει την ανομοιογένειά του η οποία παραλλάσσει την Ευκλείδεια Απόσταση. Τελικά, για

κάθε σημείο προκύπτει ένα διάνυσμα k διαστάσεων. Η διαδικασία του αλγορίθμου μετακινεί τελικά τα κεντροειδή, έτσι ώστε κάθε σημείο-στιγμιότυπο να έχει εμφανώς μία διάσταση μικρότερη από τις άλλες. Η μικρότερη διάσταση είναι δείκτης της συστάδας που ανήκει το σημείο. Η διαδικασία μετακίνησης των κεντροειδών επικυρώνεται από μια Συνάρτηση Κόστους που αξιολογεί την κάθε μετακίνηση (Jain, 2010).

Ο πιο διαδεδομένος τέτοιος αλγόριθμος ονομάζεται k-mean (**Εικόνα 13**) και αποτελεί πηγή έμπνευσης και βάση εξέλιξης για πολλούς άλλους. Η αρχική ιδέα των Lloyd στα 1957 και Forgy το 1965 είχε προταθεί για την επεξεργασία σήματος ως πνευματική ιδιοκτησία των Bell Laboratories και δημοσιοποιήθηκε το 1982. Οι δυνατότητές του είναι αυτές της επίπεδης συσταδοποίησης, μιας και παράγει ένα σύνολο συστάδων οι οποίες δεν έχουν κάποια ιδιαίτερη δομή-σχέση μεταξύ τους.

Ιεραρχικοί Αλγόριθμοι

Οι Ιεραρχικοί Αλγόριθμοι δημιουργούν μια ιεραρχία συσταδοποιήσεων. Η ιεραρχία προκύπτει απ’ το γεγονός ότι υπάρχουν συστάδες που εμπεριέχουν άλλες συστάδες. Όταν η Α συστάδα εμπεριέχει την Β, τότε η Β ονομάζεται εμφωλευμένη. Το πλήθος των συστάδων που περιέχονται σε μία συστάδα είναι ανάλογο με την θέση της στην ιεραρχία. Όσο περισσότερες συστάδες περιέχονται σε μία συστάδα τόσο πιο ψηλά τοποθετείται αυτή στην ιεραρχία. Έτσι προκύπτουν συστάδες που περιέχουν μεμονωμένα στοιχεία και άλλες συστάδες, οι οποίες με τη σειρά τους μπορεί να περιέχουν και αυτές άλλες, μικρότερες συστάδες, δημιουργώντας έτσι τα επίπεδα της ιεραρχίας.

Οι Ιεραρχικοί Αλγόριθμοι διακρίνονται σε δύο υποκατηγορίες:

- Τους *Συσσωρευτικούς* και
- Τους *Διαιρετικούς*.

Η βασική διαφοροποίησή τους είναι ότι οι μεν Συσσωρευτικοί εκκινούν την διαδικασία με έναν αρχικό αριθμό συστάδων και σε κάθε βήμα του αλγορίθμου το πλήθος τους μειώνεται κατά ένα, συγχωνεύοντας δύο συστάδες σε μία, έως καταλήξουν σε μία μοναδική που να εμπεριέχει όλα τα σημεία. Οι δε διαιρετικοί κάνουν ακριβώς το αντίστροφο (Murtagh, 1983).

Τα αποτελέσματα τέτοιων αλγόριθμων μπορούν να αναπαρασταθούν σε δενδρικά διαγράμματα, τα οποία παρουσιάζουν τη διάταξη των συστάδων που δημιουργήθηκαν από την ιεραρχική συσταδοποίηση, έτσι ώστε κάθε επίπεδο του διαγράμματος να ορίζει κι ένα βήμα του αλγορίθμου. Το βασικό πλεονέκτημα των Ιεραρχικών Αλγορίθμων είναι ότι δεν χρειάζεται να υποθέσουμε ένα συγκεκριμένο αριθμό συστάδων, αφού οποιοσδήποτε αριθμός μπορεί να επιτευχθεί, διακλαδίζοντας το δενδρικό διάγραμμα στο κατάλληλο επίπεδο.

Το μέτρο της απόστασης δύο συστάδων μπορεί να οριστεί με αρκετούς τρόπους. Οι πιο βασικοί είναι και οι πιο εύκολα εφαρμόσιμοι, με κυρίαρχους την Απόσταση Κεντροειδών και την Ελάχιστης Απόστασής ή Απλού Δεσμού. Ενώ η μέτρηση της απόστασης των κεντροειδών είναι προφανείς από το όνομα, στην ελάχιστη απόσταση μετράμε τα δύο κοντινότερα σημεία που να ανήκουν σε διαφορετικές συστάδες. Το μέτρο της απόστασης συστάδων δεν είναι εμφανώς χρήσιμο στους Αλγορίθμους Μέσου Όρου, αλλά εκτός ότι αποτελεί μια παραπάνω πληροφορία για τα δεδομένα, είναι και δομικό κομμάτι της ιεραρχικοποίησης (Manning et al., 2009).

Εκτός απ’ τα δύο κοινά μέτρα υπάρχουν και άλλα, όπως η Μέγιστης απόστασης, η Μέσου Όρου Συστάδας και η Μέθοδος Ward. Τονίζεται ότι ο Μ.Ο. Θέσης Συστάδας χαρακτηρίζεται εν γένει από διαφορετική τιμή από την Θέση Κεντροειδούς και ενώ είναι ορθότερο από μαθηματικής σκοπιάς, απαιτεί μια παραπάνω υπολογιστική πράξη, καθώς η Θέση Κεντροειδούς αποτελεί ήδη υπολογισμένη μεταβλητή του αλγορίθμου, είναι αποθηκευμένη και προσβάσιμη απ’ τον υπολογιστή.

Οι Συσσωρευτικοί Αλγόριθμοι ξεκινάνε με n συστάδες. Σε κάθε βήμα, συγχωνεύονται οι δύο πιο κοντινές συστάδες. Ο τρόπος αξιολόγησης της απόστασης είναι ένας απ’ τους προαναφερθέντες και η επιλογή του δεν έχει να κάνει τόσο με τη φύση του προβλήματος όσο με το σχήμα των συστάδων. Τέλος η επαναληπτική διαδικασία τερματίζει όταν ο αλγόριθμος καταλήξει σε μια μοναδική συστάδα, η οποία θα τις εμπεριέχει όλες. Το δενδρικό διάγραμμα αποτελείται από n-1 επίπεδα, όπου το κάθε ένα αντιστοιχεί σε ένα βήμα του αλγορίθμου.

Αντιθέτως, οι Διαιρετικοί Αλγόριθμοι ξεκινάνε με μια συστάδα που εμπεριέχει όλα τα σημεία. Με την πάροδο των επαναλήψεων η συστάδα διασπάται σε

επιμέρους και ο αριθμός συστάδων σε κάθε επανάληψη αυξάνεται γεωμετρικά, αυτό γίνεται έως ότου καταλήξουμε σε n συστάδες. Η πολυπλοκότητα των διαιρετικών αλγορίθμων είναι μεγαλύτερη από αυτή των συσσωρευτικών, αφού η διάσπαση μιας συστάδας σε δύο μπορεί να γίνει κατά $2^{n-1}-1$ τρόπους. Η επιλογή της βέλτιστης διάσπασης πρακτικά είναι αδύνατη ακόμα και για μικρό n. Στην πράξη η διάσπαση γίνεται, αλλά όχι κατά τον βέλτιστο τρόπο. Η όλη διαδικασία του αλγορίθμου μπορεί να αναπαρασταθεί, όπως και στους συσσωρευτικούς, με δενδρικό διάγραμμα.

Ένας από τους πιο βασικούς ιεραρχικούς αλγορίθμους είναι ο DBSCAN. Χαρακτηριστικό του είναι ότι διαδικασία του βασίζεται στο μέγεθος της Χωρικής Πυκνότητας Σημείων, δημιουργώντας ένα διαχωριστική επιφάνεια μεταξύ δύο πυκνωμάτων και ξεχωρίζοντας τις δύο Συστάδες. Η απλότητα καθώς και η ικανότητα προσαρμογής του σε πλήθος προβλημάτων τον κάνουν έναν απ' τους πιο διάσημους σε αναφορές. Το μόνο ουσιαστικό του μειονέκτημα είναι ότι δουλεύει σε δύο διαστάσεις.

Αλγόριθμοι Ανταγωνιστικής Μάθησης

Η Ανταγωνιστική Μάθηση είναι μια μορφή μη επιβλεπόμενης μάθησης στα Νευρωνικά Δίκτυα, των οποίων οι νευρώνες ή κόμβοι ανταγωνίζονται για το δικαίωμα βέλτιστης απόκρισης σε ένα υποσύνολο των δεδομένων εισόδου (inputs). Στα προηγούμενα χρόνια ο κλάδος των Νευρωνικών Δικτύων, ιδιαίτερα οι Χάρτες με Ικανότητα αυτό-Οργάνωσης (Self Organizing Maps, Kohonen 1990), τα Δίκτυα Ανταγωνιστικής Μάθησης και τα Δίκτυα από θεωρία Προσαρμοστικής Αντήχησης (Adaptive Resonance Theory, ART), αντιμετωπίζει νέες προκλήσεις στην επίλυση προβλημάτων Συσταδοποίησης.

Στα πλαίσια της παρούσας εργασίας θα παρουσιαστεί η Ανταγωνιστική Μάθηση με το παράδειγμα της μη-εποπτευόμενης εκπαίδευσης των Δικτύων Kohonen, που αποτελεί μια απ' τις πιο διάσημες μεθόδους. Χαρακτηριστικό ενός τέτοιου δικτύου είναι ότι μπορεί να ταξινομεί τα σημεία που απεικονίζουν τα δεδομένα, με την βοήθεια ενός αλγορίθμου αυτόνομης μάθησης. Η δυνατότητα οργάνωσης έγκειται στο γεγονός στην ικανότητα του αλγορίθμου να ρυθμίζει έτσι τα βάρη των κόμβων του δικτύου, του μοναδικού κρυφού επιπέδου, ώστε να

είναι σε θέση να αναγνωρίζει τα διανύσματα εισόδου. Είναι δηλαδή ένα φίλτρο πολλαπλής κατακράτησης που διαχωρίζει τα διανύσματα (σημεία του χώρου) ανάλογα με τη Συστάδα που αυτά ανήκουν (Uchiyama & Arbib, 1994).

2.3.3 Εύρεση Κανόνων Συσχέτισης

Η εύρεση ή εξαγωγή Κανόνων Συσχέτισης (Assocciation Rules) είναι μια διαδικασία ανακάλυψης «κρυφών» σχέσεων μεταξύ γεγονότων, που αρχικά φαίνονται ασυσχέτιστα. Η γενική ιδέα είναι ότι ύστερα απ' την διαδικασία Εξόρυξης, θα είναι σε θέση να εκφραστεί μια υπόθεση για την συσχέτιση δύο γεγονότων. Ιστορικά το πρώτο πεδίο που εφαρμόστηκε ήταν η κατανόηση της αγοράς ή η Ανάλυση Καλαθιού (Market Basket Analysis) (Giudici & Figini, 2009), όπου απ' το πλήθος των γεγονότων αγοράς έγινε μια προσπάθεια να βρεθούν τα μοτίβα που ακολουθούνται με σκοπό την πρόβλεψη της κατανάλωσης. Με την είσοδο του Η/Υ στην διαδικασία λιανικής πώλησης μπορούν να αποθηκευτούν όλα τα δεδομένα απ' τις αγορές που έχουν γίνει σε ένα κατάστημα. Ύστερα αυτό που ενδιαφέρει είναι να βρεθούν υποθέσεις γεγονότων που συσχετίζονται, π.χ. το Α προϊόν:

- Συχνά αγοράζεται σε συνδυασμό με το Β ή ότι σπάνια αγοράζεται σε συνδυασμό με το Β.
- Αγοράζεται πιο συχνά τους χειμερινούς μήνες.
- Το προτιμούν γυναίκες 28-35 ετών.
- Ανεξάρτητα, το αποτέλεσμα μιας τέτοιας διαδικασίας είναι μία ρητή υπόθεση.

Η Υπόθεση ως λογική οντότητα είναι μαθηματικώς θεμελιωμένη στο πεδίο της Στατιστικής, απ' όπου δανείζεται τις μετρικές, τρόπους ποσοτικής αξιολόγησης της υπόθεσης. Επίσης, προκειμένου να γίνουν διαχειρίσιμα τα δεδομένα απ' τα μαθηματικά της Στατιστικής θα πρέπει να εισαχθεί η έννοια του στοιχειο-συνόλου. Στοιχειο-σύνολο είναι κάθε υποσύνολο του αρχικού συνόλου γεγονότων. Για παράδειγμα σε μία αγορά οικιακών αγαθών που αποτελεί ένα σύνολο γεγονότων {τρόφιμο_α, τρόφιμο_β, τρόφιμο_γ, τρόφιμο_δ, αναλώσιμο_α, αναλώσιμο_β}, όπου τα πιθανά στοιχειο-σύνολα είναι το Α={τρόφιμο_α,

τρόφιμο_β} ή Β={τρόφιμο_α, τρόφιμο_γ} ή Γ={τρόφιμο_δ, τρόφιμο_β, αναλώσιμο_β} ή απλώς {τρόφιμο_γ}. Ο ορισμός του στοιχειο-συνόλου (Α) είναι σημαντικός διότι για κάθε τέτοιο προκύπτει η πιθανότητα της Υποστήριξης (Support, sup(A)). Η Υποστήριξη είναι μέτρο της συχνότητας που συναντάται ένα στοιχειο-σύνολο στην αρχική δομή δεδομένων. Ο τρόπος να υπολογίσουμε την Υποστήριξη μιας υπόθεσης είναι να βρούμε την πιθανότητα να συμπίπτουν δύο στοιχεία σε ένα πλήθος στοιχειο-συνόλων. Ή αλλιώς, η πιθανότητα σε ένα καλάθι να βρίσκονται δύο προϊόντα. Με δεδομένο το στοιχειο-συνόλου (Α) είναι πλέον δυνατό να δομηθεί μια υπόθεση που το σχετίζει με ένα άλλο γεγονός (Z). Την υπόθεση αξιολογεί ποσοτικά ένα δεύτερο μέγεθος αυτό της Εμπιστοσύνης (Confidence, comf(A => Z)), που μετρά το πόσο ισχύει η υπόθεση στα πλαίσια του στοιχειο-συνόλου. Ο τρόπος να υπολογίζεται η Εμπιστοσύνη, είναι να βρεθεί η δεσμευμένη πιθανότητα της αγοράς ενός προϊόντος, με την προϋπόθεση ότι κάποιο υποψήφιο συσχετιζόμενο έχει αγοραστεί προηγουμένως.

Πρακτικά, η διαδικασία αυτή έχει δύο στάδια:

- Βρίσκονται τα συχνά στοιχειο-σύνολα, όσα έχουν υψηλή Υποστήριξη (sup),
- Ελέγχονται όλες οι πιθανές υποθέσεις που δημιουργούν τα συχνά στοιχειο-σύνολα με όλα τα πιθανά λοιπά στοιχεία ή στοιχειο-σύνολα και διατηρείται μόνο αυτά που είναι πάνω απ’ το ελάχιστο όριο Εμπιστοσύνης (minconf).

Η επιτέλεση του δευτέρου σταδίου είναι τετριμμένη διαδικασία χωρίς ιδιαίτερες δυσκολίες. Στο πρώτο στάδιο συναντάται η κύρια δυσκολία της μεθόδου, μιας και εδώ γίνεται η πρώτη διαλογή των στοιχειο-συνόλων που θα θεωρηθούν σημαντικά. Τα σημαντικά αυτά στοιχειο-σύνολα εκτός από συχνά καλούνται και υποψήφια. Για την αντιμετώπιση του προβλήματος της διαλογής των σημαντικών στοιχειο-συνόλων ο πιο βασικός αλγόριθμος που αναπτύχθηκε είναι ο Apriori (Agrawal & Srikant, 1994), ενώ στην συνέχεια διαμορφώθηκαν και διάφορες παραλλαγές του όπως ο Partition, ο FP-Growth και ο Elcat. Αξίζει να σημειωθεί ότι τόσο ο FP-Growth όσο και ο Elcat κάνουν χρήση δενδρικών διαγραμμάτων προκειμένου να δομήσουν την διαδικασία επίλυσής τους. Από βιολογικής άποψης υπάρχει μια πλειάδα από εφαρμογές, όπως το BioMart Web

Interface, που μπορούν να επεξεργάζονται ταυτόχρονα δεδομένα από δύο ή και περισσότερες βάσεις με διαφορετικά δεδομένα, καθώς ο συνδυασμός των πληροφοριών από διαφορετικά πεδία, μπορεί να αποβεί πολύ εύφορος και να παράξει αποτελέσματα σε μερικές ώρες, πολύ πιο γρήγορα από κάθε προηγούμενη μέθοδο. Τέτοιο παράδειγμα είναι ο συνδυασμός μιας πρωτεϊνικής βάσης δεδομένων και μίας βάσης που έχει αποθηκευμένα μεταβολικά μονοπάτια. (Durinck et al., 2008).

2.4. Εξόρυξη από Κείμενο (Text Mining)

Ως Εξόρυξη από Κείμενο (Text Mining) ορίζεται η υπολογιστική ανακάλυψη νέων, άγνωστων πληροφοριών, με αυτοματοποιημένη την λειτουργία εξαγωγής σε πολλές διαφορετικές γραπτές πηγές (Jensen et al., 2006). Γενικά, η Κειμενική Εξόρυξη περιλαμβάνει δύο μεγάλα στάδια, την ανάκτηση δεδομένων και την εξαγωγή από αυτά (Rebholz-Schuhmann et al., 2005). Το στάδιο της ανάκτησης, αναζητά και βρίσκει βιβλιογραφία ή περιλήψεις σχετικές με κάποιες Λέξεις Κλειδιά, με παράλληλο στόχο οι μηχανές αναζήτησης ή ειδικότερα σχεδιασμένα εργαλεία αναζήτησης να είναι όλο και πιο φιλικά προς τον χρήστη (Ananiadou et al., 2006). Στην ανάκτηση υφίστανται δύο προσεγγίσεις με αρκετά κοινά γνωρίσματα. Η μία βασίζεται σε μορφοποιημένη από κανόνες, γνώση και η άλλη στη στατιστική, με το όνομα Μηχανική Μάθηση (Machine-Learning) (Hirschman et al., 2002).

Καθώς η βιβλιογραφία στο χώρο της Βιολογίας συνεχώς αυξάνεται και η κωδικοποίηση της γνώσης στο πεδίο αυτό είναι κατ’ ανάγκη μη φορμαλισμένη, γεννάται η ανάγκη τέτοια κείμενα να μπορούν να γίνουν αναγνώσιμα και επεξεργάσιμα από Η/Υ. Ενδεικτικά, απ’ το 1997 έως το 2006 ο αριθμός των δημοσιεύσεων που αφορούν στη Βιοϊατρική αυξήθηκε κατά 500 φορές, αύξηση που σίγουρα δεν μπορεί να ακολουθήσει το ανθρώπινο ερευνητικό δυναμικό, το οποίο παραδοσιακά ήταν επιφορτισμένο με την επεξεργασία τέτοιων δεδομένων (Ananiadou et al., 2006). Έτσι ανοίγεται ένα νέο πεδίο στην Εξόρυξη από Δεδομένα, όπου τα δεδομένα εισροής είναι πλέον, όχι απλά λέξεις ή αριθμοί αλλά ολόκληρα κείμενα, που μέχρι πρότινος ήταν αποκλειστικότητα των ερευνητών. Πλέον, μέρος του φόρτου αυτού επαφίεται σε καινοτόμα λογισμικά. Η

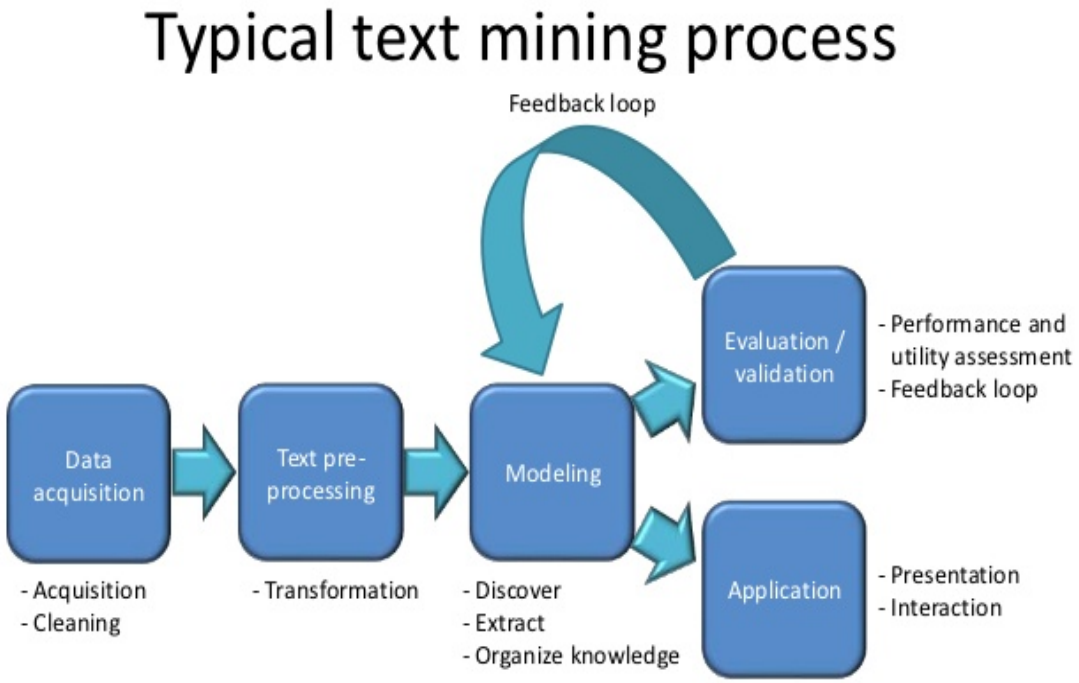
Εξόρυξη από Δεδομένα Κειμένου έχει ως στόχο την αυτοματοποίηση της εξαγωγής νοήματος μιμούμενη τον άνθρωπο.

Αρχικά με την χρήση προτύπων που βασίζονται μόνο σε γνώσεις και όρους βιολογίας, όπως για παράδειγμα ορολογίες: «προστάτης» ή «μεμβράνη», χρησιμοποιούνται οι Λέξεις Κλειδιά για την ανεύρεση της σχετικής βιβλιογραφίας. Η δεύτερη προσέγγιση χρησιμοποιεί δέντρα συντακτικής σύναψης, τα οποία μπορεί να βασίζονται σε κανόνες ή ταξινομητές, για κατηγοριοποιούν τα ευρήματα της βιοϊατρικής βιβλιογραφίας. Η Αναγνώριση Ονόματος Οντολογικής Καταχώρησης (Name Entity Recognition, NER), αποτελεί προϋπόθεση για την επόμενη λειτουργία, την εξαγωγή γνώσης και στηρίζεται σε εργαλεία ή μεθόδους, αυτόματης αναγνώρισης όρου, έτσι θα μπορέσει να εξάγει επιπλέον σχετικά γονίδια, πρωτεΐνες, φάρμακα ή άλλα μόρια. Το εγχείρημα iHOP (information Hyperlink Over Proteins, http://ihop-net.org/), που επιχειρεί να ενοποιήσει της πληροφορίες του διαδικτύου σχετικά με πρωτεΐνες, είναι ένα εξαιρετικό παράδειγμα, περιγράφοντας προτάσεις από περιλήψεις δημοσιεύσεων της Medline με βάση τις οντολογικές καταχωρήσεις που εμφανίζονται. Η εξαγωγή πληροφορίας χρησιμοποιείται στην συνέχεια για να ταυτοποιήσει ή να συνοψίσει τις σχετικές οντολογικές καταχωρήσεις ή γεγονότα, που έχουν διασωθεί από αρχειακό υλικό (Jensen et al., 2006).

Η διαδικασία της Εξόρυξης από Κείμενο μπορεί να αναλυθεί σε 4 στάδια (**Εικόνα 14**). Το πρώτο στάδιο είναι ο καθορισμός του ποια χωρία από το κείμενο είναι αυτά που αποτελούν δεδομένα. Το δεύτερο στάδιο είναι η μετατροπή των δεδομένων ώστε να μπορούν να είναι ομογενή και ικανά να μοντελοποιηθούν. Το τρίτο και κυρίως στάδιο η χρήση τους για την εκπαίδευση και την κατασκευή του μοντέλου. Το τέταρτο στάδιο η αξιολόγηση της μοντελοποίησης και ανατροφοδότηση του προηγούμενου σταδίου μέχρι το μοντέλο να ικανοποιεί τις ανάγκες της έρευνας.

Η εξαγωγή πληροφορίας ως διαδικασία μπορεί με τη σειρά της να αναλυθεί περαιτέρω με δύο προσεγγίσεις. Η πρώτη και απλούστερη αντιλαμβάνεται ως Λέξεις Κλειδιά βάσει σύμπτωσης (co-occurence), αναγνωρίζοντας οντότητες συμπίπτουν στο ίδιο κείμενο. Επιπλέον, η σύμπτωση (ή συνύπαρξη) μπορεί να χρησιμοποιηθεί για την εξαγωγή σχέσεων κάθε τύπου, όπως φυσιολογική πρωτεϊνική αλληλεπίδραση (Jensen et al., 2006). Η δεύτερη προσέγγιση εξάγει

συσχετίσεις όπως γονίδιο με γονίδιο, πρωτεΐνη με πρωτεΐνη ή ολόκληρα βιολογικά μονοπάτια, που εκτείνονται πέρα απ' τους απλούς όρους αναζήτησης/αναγνώρισης. Η Επεξεργασία της Φυσικής Γλώσσας (Natural Lagnuage Prossesing, NLP), μια τεχνολογία που συνδυάζει συντακτικά και σημασιολογικά, έχει ευρέως εφαρμοστεί για το δεύτερο στάδιο επεξεργασίας (Cohen & Hunter, 2008).



Εικόνα 14. *Διαδικασία της Εξόρυξης από Κείμενο.*
(Η εικόνα ελήφθη απ' το διαδίκτυο:
https://www.picquery.com/v/url-check_zA0PEcx*va0z2PkvgQ%7CsFpNzVJK*oM05XVEH1EapVW0/)

Η εφαρμογή αυτού του εργαλείου είναι τεράστια κυρίως στην αναδυόμενη συνδυαστική επιστήμη της Βιολογίας Συστημάτων. Η βασική ανάγκη είναι η εξαγωγή υποθέσεων. Τέτοιες υποθέσεις μπορεί να σχετίζουν πρωτεΐνες μεταξύ τους, πρωτεΐνες με γονίδια, γονίδια με λειτουργίες και τελικά ασθένειες. Πιο συγκεκριμένα, η διαδικασία αναλύεται επιμέρους σε τρία στάδια (Ananiadou et al., 2006):

- Πρώτο, σε ανάκτηση πληροφορίας, την εύρεση δηλαδή των κειμένων που αφορούν και την επιλογή τους απ' το πλήθος της ψηφιοποιημένης βιβλιογραφίας στην οποία μπορεί να έχει πρόσβαση ο Η/Υ.
- Δεύτερο, σε εξαγωγή πληροφορίας, που αποτελεί την εύρεση της ωφέλιμης πληροφορίας και την αποθήκευσή της, με σκοπό να δημιουργηθεί μια δομή, απ' τα επιμέρους κείμενα που επιλέχθηκαν στην πρώτη φάση, η οποία να είναι επεξεργάσιμη.
- Και τρίτο, σε επεξεργασία αυτής της δομής μέσω τεχνικών εξόρυξης. Έτσι, τα δύο πρώτα στάδια μπορούν να θεωρηθούν ως προπαρασκευαστικά για το τρίτο, το οποίο είναι η κατ' εξοχήν εξόρυξη και αποτελούν ιδιαιτερότητα της κειμενικής εξόρυξης, που οφείλεται στην εγγενή αδυναμία των γλωσσικών κειμένων να είναι αναγνώσιμα από τον υπολογιστή.

Μία παράπλευρη, αλλά σημαντική, αρμοδιότητα της εξόρυξης από κείμενο βασίζεται στις λέξεις-όρους που χρησιμοποιούνται σε αναφορά με βιολογικές έννοιες. Πολλές φορές οι όροι που χρησιμοποιούνται πρέπει να θεωρηθούν ή να επικυρωθούν (validate) καθώς η ασάφεια που μπορεί να ενέχουν δυσκολεύει την αυτοματοποιημένη αναζήτηση. Οι άμεσοι στόχοι είναι η δυνατότητα επεξεργασίας ολόκληρου κειμένου, η επιπλέον εμβάθυνση των τεχνικών της Εξόρυξης σε κείμενα, η βελτίωση της απεικόνισης των αποτελεσμάτων και η βελτίωση του σε ορόσημο, άρα και πιθανή αξιόλογη πληροφορία σε ένα κείμενο (Ananiadou et al., 2006).

2.5. Η διαδικασία Ανακάλυψης Γνώσης ως επέκταση της Εξόρυξης από Δεδομένα

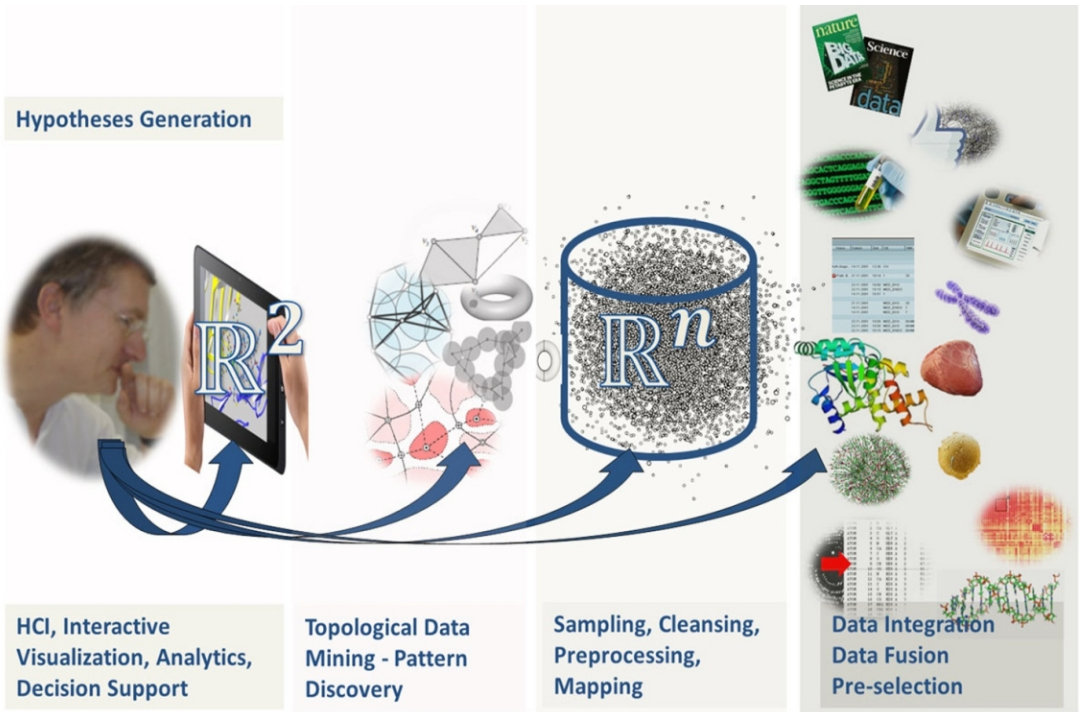
Η παραδοσιακή μέθοδος για την μετατροπή των δεδομένων σε γνώση βασίζεται σε χειροκίνητη ανάλυση και ερμηνεία από κάποιον ειδικό στο αντικείμενο μελέτης, προκειμένου να αναδειχθούν νέες ιδιότητες του υπό μελέτη συστήματος, με σκοπό να υποστηρίξουν και να υποβοηθήσουν την λήψη μιας απόφασης.

Σήμερα, πέρα απ' την αναγνώριση προτύπων, η διαδικασία αυτή έχει ονομαστεί με διάφορους τρόπους: Εξόρυξη από Δεδομένα (Data Mining), Εξαγωγή Γνώσεως (Knowledge Extraction), Πληροφοριακή Ανακάλυψη (Information Discovery), Συγκομιδή Πληροφοριών (Information Harvesting), Αρχαιολογία Δεδομένων (Data Archeology) και Επεξεργασία Μοτίβων στα δεδομένα (Fayyad et al., 1996b). Στο κλασσικό έργο του Fayyad και των συνεργατών του (Fayyad et al., 1996a), η διαδικασία αυτή περιγράφεται σε διακριτά βήματα, ξεκινώντας από την συλλογή δεδομένων, την προεργασία, αναμόρφωση τους και ερμηνεία της κατάστασής που περιγράφουν. Εκεί ορίζεται η Εξόρυξη από Δεδομένα ως υποσύνολο της Ανακάλυψη Γνώσης, παρόλο που η αυθεντική ονομασία είναι Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων (Knowledge Discovery Databases, KDD). Πλέον, προκειμένου να δώσουμε έμφαση στο ότι η Εξόρυξη είναι το χαρακτηριστικότερο υποσύνολο στην Επεξεργασία της Γνώσης, χρησιμοποιούνται επίσης οι ορολογίες Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων και Εξόρυξη από Δεδομένα (Data Mining). Αξίζει να σημειωθεί, ότι η Ανακάλυψη Γνώσης σε Βάσεις Δεδομένων μπορεί να θεωρηθεί ως μια διαδικασία που ενσωματώνει την πλήρη αλυσίδα προστιθέμενης αξίας από τον φυσικό κόσμο των δεδομένων στο χώρο της ανθρώπινης διανόησης, όπου υφίσταται η γνώση, η οποία ορίζεται από γνωσιακή άποψη: Η γνώση ως σύνολο προσδοκιών.

Περαιτέρω, επεκτείνεται και ο αρχικός ορισμός που δόθηκε από τον Fayyad και τους συνεργάτες του (Fayyad et al., 1996b), με την συμμετοχή του ανθρώπου στην διαδικασία Ανακάλυψης. Η διάδραση, η επικοινωνία και η νοηματοδότηση είναι από τις θεμελιώδεις επικεφαλίδες στην Διάδραση Ανθρώπου-Υπολογιστή (Human-Computer Interaction, HCI), και έτσι, μια καινούρια προσέγγιση είναι ο συνδυασμός HCI και KDD (Holzinger, 2013). Η κεντρική υπόσχεση της ζεύξης των δύο χώρων HCI-KDD είναι να επιτρέψει στους χρήστες να βρίσκουν και να χαρακτηρίζουν, άγνωστα μέχρι στιγμής και δυνητικά αξιοποιήσιμες πληροφορίες. Με την κλασική έννοια, μπορεί να οριστεί ως διαδικασία ταυτοποίησης νέων προτύπων, με στόχο την κατανόησή τους. Ο ειδικός σε κάποιον τομέα, κατέχει εξεζητημένη γνώση και όταν του επιτραπεί από την τεχνολογία να εξερευνά με διαδραστικό τρόπο ένα σύνολο δεδομένων, καθίσταται ικανότατος στο να

ταυτοποιήσει και να κατανοήσει νέα πρότυπα και συμμετρίες (Holzinger et al., 2012).

Η παρουσία της πλήρους διαδικασίας ανεύρεσης γνώσης σχηματοποιείται για να περιγραφούν κάποιες απ' τις προκλήσεις και τα προβλήματα, σε όλο το φάσμα από απ' τον υπολογιστή προς τον άνθρωπο, δηλαδή από δεξιά προς τ' αριστερά **(Εικόνα 15**). Ο ερευνητής έχει μια πολύ δεδομένη αντιληπτική ικανότητα που περιορίζεται από τα μέσα και τις δυνατότητές του, που δεν είναι εν γένει ικανές για την αντιμετώπιση ή κατανόηση του πλήρους φάσματος των δεδομένων. Στο ενδιάμεσο παρεμβάλλονται έννοιες που μπορούν να γίνουν αντιληπτές και διαχειρίσιμες από τον υπολογιστή, έννοιες όπως τα μοτίβα ή πολυδιάστατοι χώροι και λειτουργίες όπως η διαχείριση τεραστίων δειγμάτων, πολύπλοκων σχημάτων και η προσπέλαση αλγορίθμων. Τα πεδία χωρίζονται όχι μόνο λόγω της θεωρητικής ανάγκης για διάκριση και ταξινόμηση, αλλά και γιατί αναμένεται να αποτελέσουν νέους χώρους έρευνας με καινούριους στόχους και μεθοδολογίες.



Εικόνα 15. *Η απόσταση μεταξύ δεδομένων και ερευνητή δεν είναι τόσο μικρή όσο μοιάζει. Στο σχεδιάγραμμα αυτό τοποθετούνται στα δύο άκρα ο ερευνητής (αριστερά) και τα δεδομένα (δεξιά) σε απλή ή συγχωνευμένη μορφή.*

(Η εικόνα αποτελεί μέρος της δημοσίευσης: Holzinger et al., 2014.)

Το πρώτο ζήτημα είναι πώς θα επιτευχθεί από την αρχή μια αυστηρή διαλογή των δεδομένων, η οποία θα είναι συγχωνευμένη και ταυτόχρονη με την διαδικασία καταγραφής, σκοπό να περιορίζεται η πολυπλοκότητα σε μεγάλους όγκους δεδομένων. Η έμβια φύση συχνά κάνει την υπολογιστική μας ισχύ να αδυνατεί να επαρκέσει, προκαλώντας διάφορα προβλήματα που συνοψίζονται σε πέντε κατηγορίες:

- Ετερογένεια πηγών δεδομένων, που απαιτεί ομογενοποίηση δεδομένων,
- Πολυπλοκότητα δεδομένων,
- Θόρυβος,
- Αβεβαιότητα,
- Ασυμβατότητα μεταξύ δεδομένων, πληροφορίας και γνώσης (Jurisica et al., 1998).

Σε σύγκριση με τα ερευνητικά συστήματα, τα εμπορικώς διαθέσιμα Συστήματα Διαχείρισης Πληροφοριών έχουν μόνο μερικές δυνατότητες συγχώνευσης δεδομένων (Bleiholder & Naumann, 2008). Αποτελεί μεγάλη πρόκληση να ενσωματωθούν και να συγχωνευθούν τα βιολογικά δεδομένα με άλλου τύπου δεδομένα, όπως αρχεία ασθενών, δεδομένα φυσιολογίας ή οπτικό υλικό (Wiltgen et al., 2006)(Viceconti et al., 2007). Τα ζητήματα αυτά είναι τόσο κρίσιμα, ώστε έχουν δημιουργήσει την δική τους αλληλουχία Συνεδρίων, με τίτλο «Data integration in life-science» (Baker CJ, Butler G, Jurisica, 2013).

Το δεύτερο ζήτημα είναι αυτό της συγχώνευσης πολλαπλών δομών δεδομένων, οι οποίες περιλαμβάνουν κοινές τιμές (όπως μεγέθη, ονόματα, λογικές μεταβλητές) και είναι απ’ τα πιο συχνά αναδυόμενα στον χώρο της Εξόρυξης. Το πρόβλημα αυτό ονομάζεται Συγχώνευσης/Απομόνωσης και είναι δύσκολο να αντιμετωπιστεί ταυτόχρονα σε όλο το εύρος των δομών και με την απαιτούμενη ακρίβεια (Hernández & Stolfo, 1998). Η διαλογή δεδομένων από την πρωτογενή συγκεχυμένη μορφή, με την οποία παράγονται ή είναι αποθηκευμένα, είναι έργο καίριας σημασίας στην καθημερινότητα της επεξεργασίας δεδομένων και έχει οδηγήσει στην ανάπτυξη ενός φάσματος μεθόδων βελτίωσης της ακρίβειας και δευτερογενώς την χρηστικότητα των υπαρχόντων δεδομένων (Müller & Freytag, 2005). Πολλοί αλγόριθμοι δεν τα καταφέρνουν καλά με πολυδιάστατα δεδομένα, όπως για παράδειγμα κάποιοι

απ’ τους αλγορίθμους της Μηχανικής Μάθησης. Πολλοί την χαρακτηρίζουν ως την κατάρα της πολυπλοκότητας των διαστάσεων (Catchpoole et al., 2010).

Ένα τρίτο ζήτημα είναι ότι τα περισσότερα κλινικά δεδομένα είναι ατελή, με κενά στην πληροφορία και ασυνέπειες στην ορολογία. Ακόμα ενδέχεται να απαιτούν εντοπισμό και εξαίρεση διπλών καταχωρήσεων (Lee et al., 1999). Έτσι, ο βασικός στόχος της ποιότητας των δεδομένων θέτει τα δικά του ζητήματα και προκλήσεις (Elloumi & Zomaya, 2013)(Jarke et al., 2013). Η ποιότητα των δεδομένων τελικώς επηρεάζει άμεσα την ποιότητα της πληροφορίας.

Ένα τελευταίο ζήτημα είναι αυτό της κατ’ εξοχήν Εξόρυξης. Πολλές από τις μεθόδους εξόρυξης έχουν σχεδιαστεί για συλλογές από ευπαρουσίαστα αντικείμενα σε συμπαγείς δομές πινάκων. Παρόλα αυτά, εκτός από μαζικότατες δομές με μη οργανωμένη και μη σταθμισμένη πληροφορία, όπως κείμενο (Kreuzthaler, 2011) (Holzinger et al., 2012c), κατακλυζόμαστε από μεγάλες συλλογές από συσχετιζόμενα αντικείμενα, των οποίων η φυσική αναπαράσταση είναι ένα συσσωμάτωμα σημείων ή μια εικόνα γραφήματος (π.χ. πρωτεϊνικές δομές, δίκτυα αλληλεπίδρασης κ.α.). Οι πιο εξεζητημένες τεχνικές εξόρυξης αποτελούν:

- Εξόρυξη σε γραφήματα,
- Εξόρυξη Βάσει Εντροπίας και
- Εξόρυξη Βάσει Τοπολογίας.

Οι τεχνικές αυτές εμπεριέχουν στην μεθοδολογία τους γνώση από άλλους κλάδους. Συγκεκριμένα, η Εξόρυξη Βάσει Εντροπίας είναι θεμελιωμένη στην Θεωρία Πληροφορίας και την Θεωρία Γραφημάτων. Γενικά, η Θεωρία Πληροφορίας (Shannon & Weaver, 1949) σχετίζεται με την ποσοτικοποίηση της πληροφορίας και την διερεύνηση της επικοινωνίας ως διεργασία. Η μετάφραση αυτού του σχήματος στην Θεωρία Γραφημάτων είναι πολύπλοκη. Ως αποτέλεσμα, πολλά γραφήματα εντροπίας έχουν αναπτυχθεί, αλλά η βιβλιογραφία στερείται από λοιπές μετρήσεις δικτύων (Dehmer & Mowshowitz, 2011). Συμπερασματικά, αρκετή δουλειά ακόμα μένει για το μέλλον.

Τελικώς, τα αποτέλεσμα που απέδωσε η εφαρμογή εξειδικευμένων αλγορίθμων, σε χώρους πολλαπλών διαστάσεων, στο πεδίο της κατ’ εξοχήν εξόρυξης, θα πρέπει να μπορεί να παρασταθεί απεικονιστικά σε δισδιάστατο

χώρο, που να είναι αντιληπτός απ’ τον ευκλείδειο νου. Φαίνεται ότι καθώς ο κόσμος είναι σε μεγάλο βαθμό πολυδιάστατος, εμείς είτε επιλέγουμε είτε παραμορφώνουμε τα ευρήματα, για να τα απεικονίσουμε στις λίγες διαστάσεις που γίνονται αντιληπτά. Το γεγονός αυτό οδηγεί στον ορισμό της απεικονιστικής διαδικασίας, ως την χαρτογράφηση ή τον ισομορφισμό μιας εικόνας, από ένα χώρο με πολλές διαστάσεις σε λιγότερες, μία κίνηση που αναγκαστικά ενέχει τον κίνδυνο να δημιουργεί πλασματικά ευρήματα. Παρόλο που η απεικόνιση δεδομένων αποτελεί μία επιστημονική εμπειρία με υπόβαθρο δεκαετιών, υφίστανται ακόμα πολλές προκλήσεις και ανοιχτά ζητήματα για την έρευνα, ειδικά στα πλαίσια της διαδραστικής εξόρυξης. Μεγάλο είναι το κενό που δημιουργείται από την απουσία εργαλειοθήκης, που να υποστηρίζει λειτουργίες ανάλυσης στην βιοϊατρική διαδικασία (Jeanquartier & Holzinger, 2013).

Ενδιαφέρον παρουσιάζει η παρατήρηση ότι παρόλο που υπάρχει διαθέσιμη πληθώρα εξειδικευμένων τεχνικών απεικόνισης, οι υπάρχουσες σπάνια χρησιμοποιούνται. Κάποιες πρωτόλειες απόπειρες απεικόνισης καταλήγουν κατά κύριο λόγο αναποτελεσματικές ή παραπλανητικές, καθώς η αποτελεσματική οπτικοποίηση προϋποθέτει την κατανόηση που ο ανθρώπινος εγκέφαλος επεξεργάζεται την πληροφορία (Tory & Moller, 2004).

Τέλος, δεν μπορούμε να παραβλέψουμε την πολιτική πραγματικότητα η οποία συνοδεύει την επιστημονική διαδικασία. Τα βιοϊατρικά δεδομένα, που παράγονται από ασθενείς και αποτελούν κατά κάποιο τρόπο ιδιοκτησία τους, προκύπτουν ζητήματα ιδιωτικότητας, προστασίας αυτής, καθώς και πλαισίου δίκαιης διαχείρισης (Weippl et al., 2006). Αντιμετωπίζουμε ένα εύρος από ερευνητικές προκλήσεις στην ανάπτυξη μεθόδων Εξόρυξης, που να μπορούν να χειρίζονται, όπως αρμόζει, όπως δεδομένα που αφορούν σε παθολογικές ή όχι για το άτομο ιδιαιτερότητες.

3. Εφαρμογές Εξόρυξης από Δεδομένα στη Βιοτεχνολογία και Βιοπληροφορική

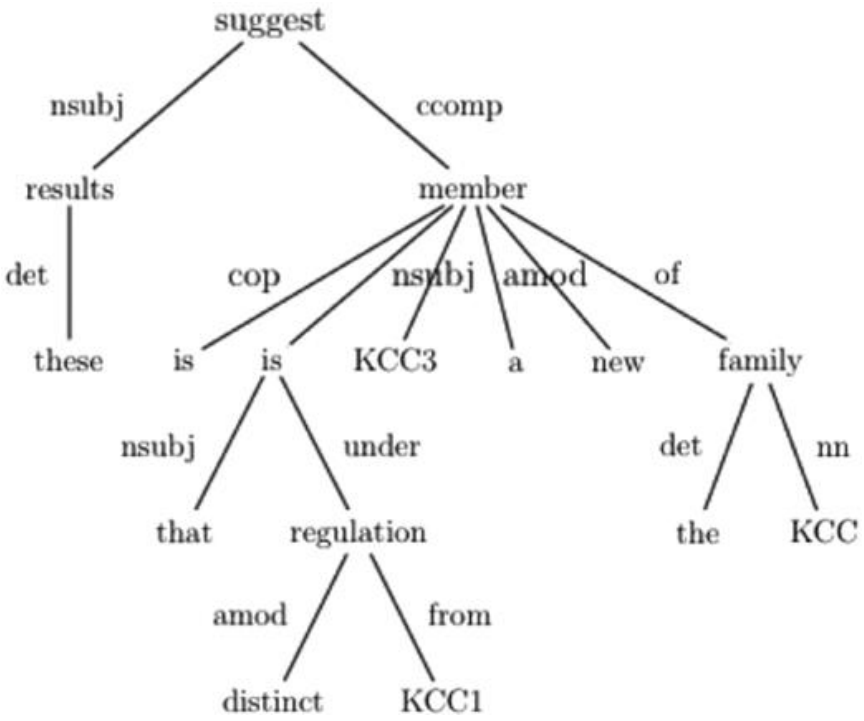
Οι εφαρμογές της Εξόρυξης στα πλαίσια της Βιοτεχνολογίας και Βιοπληροφορικής εκτείνονται σε ένα μεγάλο φάσμα απ' την Φαρμακολογία και την εύρεση συγκεκριμένων γονιδίων στόχων έως την επιτάχυνση εργαστηριακών διαδικασιών. Οι δημοσιεύσεις που αφορούν αποκλειστικά στην Εξόρυξη από Δεδομένα είναι λίγες στο σύνολό τους. Τις περισσότερες φορές μια δημοσίευση επιστρατεύει την Εξόρυξη ως μέσο μιας γενικότερης διαδικασίας, πολλές φορές για να ισχυροποιήσει τα αποτελέσματά της ή την χρησιμοποιεί ως έναυσμα για την έρευνα που περιγράφει. Παρόλα αυτά, η Εξόρυξη ως διαδικασία δεν παύει να αποτελεί όλο και περισσότερο ανάγκη για μια σύγχρονη έρευνα, προσφέροντας την συνεχώς αυξανόμενη εργαλειοθήκη της.

3.1. Ταυτοποίηση οντοτήτων που σχετίζονται με ασθένειες

Η Εξόρυξη από Κείμενο έχει επανειλημμένως εφαρμοστεί στην ταυτοποίηση οντοτήτων (γονίδια ή πρωτεΐνες) που σχετίζονται με ασθένειες και την κατανόηση του ρόλου τους σε αυτές (εικόνα 16). Προσφάτως, ο Özgür και οι συνεργάτες του περιέγραψαν μια νέα μέθοδο αναγνώρισης, που ανακτά και ιεραρχεί τα υποψήφια γονίδια που συσχετίζονται με τον καρκίνο του προστάτη (Özgür et al., 2008). Αρχικά, δομείται μια λίστα από 15 αρχικά γονίδια εκκίνησης, τα οποία είναι γνωστά και σεσημασμένα, προερχόμενα από βιβλιοθήκες, όπως η Online Mendelian Inheritance in Man (OMIM). Η λίστα των αρχικών γονιδίων (seed genes), αρχική λίστα, χρησιμοποιείται ως πρώτη ύλη για την κατασκευή ενός δικτύου με αλληλεπιδράσεις γονιδίων για την συγκεκριμένη ασθένεια, το οποίο έχει εξορυχθεί από ολόκληρα άρθρα του PubMed Central (PMC) και χρησιμοποιεί τις μεθόδους ανάλυσης εξάρτησης (Dependency Parsing) και Support Vector Machines (SVM). Τέτοιες μέθοδοι δημιουργούν δένδρα εξάρτησης ή συσχέτισης με της λέξεις κλειδιά και τα υπό

αναζήτηση γονίδια. Οι φράσεις αποτελούν χωρία της βιβλιογραφίας και προκειμένου να αναλυθεί η πληροφορία που ενέχεται στις φράσεις αυτές, από τον υπολογιστή πρέπει με κάποιο τρόπο να αλλάξει η μορφή. Αυτή είναι η μορφή που παίρνει σε δενδρικό διάγραμμα (**Εικόνα 16**).

Η εκτεταμένη λίστα γονιδίων στο δίκτυο των αλληλεπιδράσεων γονιδίων βαθμονομήθηκε και ιεραρχήθηκε σύμφωνα με την βέλτιστη χωροθέτηση ως προς το κέντρο του δικτύου βιβλιογραφίας. Εντύπωση κάνει το γεγονός ότι το 95% των 20 κορυφαίων στην διαλογή γονιδίων, που προτάθηκαν από αυτή τη μέθοδο, είναι ήδη επιβεβαιωμένο ότι συσχετίζεται με τον καρκίνο του προστάτη.



Εικόνα 16. *Παράδειγμα εξάρτησης δενδρικού διαγράμματος απ' την φράση: "These results suggest KCC3 is a new member of the KCC family that is under distinct regulation from KCC1".*
(Η εικόνα αποτελεί μέρος της δημοσίευσης: Özgür et al., 2008.)

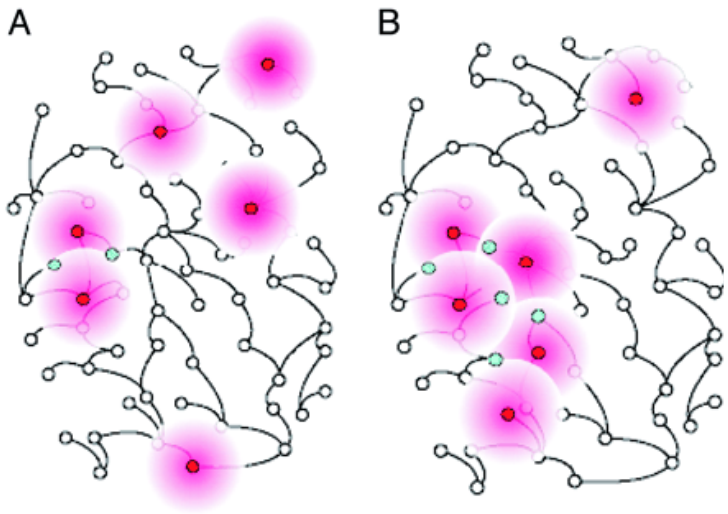
Παρομοίως, άλλη ομάδα υιοθετώντας μια άλλη προσέγγιση, που συνδυάζει κειμενική και δομική αναζήτηση προκειμένου να ανακτήσει πιθανά ένζυμα στόχους στον εξωκυττάριο χώρο καρκινικών κυττάρων, για έξι κοινούς και θανάσιμους καρκίνους στον άνθρωπο. Η αναζήτηση έγινε πάνω σε περιλήψεις δημοσιεύσεων από το PybMed, τη UniProt, την InterPro και φυσικά την NCBI (Pospisil et al., 2006). Πρώτα, το εργαλείο εξόρυξης βιβλιογραφίας LSGraph

χρησιμοποιήθηκε για να εξαχθούν οντότητες από τις επιμελημένες βιβλιοθήκες που προαναφέρθηκαν, με αίτημα λέξεως κλειδί (keyword based searching) και όρους Γονιδιακής Οντολογίας (gene ontology terms, GO). Αυτές οι οντότητες εμπλουτίζονται με σχετικούς σχολιασμούς (annotations) που χαρακτηρίζουν τη λειτουργία και ομαδοποιούνται στη συνέχεια βάσει της χωροθέτησής τους στο κύτταρο και την βιοχημική λειτουργία με την οποία είναι καταχωρημένες στα αρχεία της Ingenuity knowledge-base. Τελικά, η μέθοδος οδηγεί στην ταυτοποίηση μιας σειράς από υδρολάσες (hydrolases) συντεταγμένες σε λίστα, που σχετίζονται με κάθε διαφορετικό τύπο καρκίνο. Μεταξύ αυτών βρίσκονται η φωσφατάση προστατικού οξέως (Prostatic Acid Phosphatase, ACPP), το αντιγόνο του προστάτη (Prostate-Specific Antigen, PSA) και η θειοφατάση (Sulfatase_1, SULF1), άρα επιλέγονται ως κατάλληλοι στόχοι για τη διαμεσολαβούμενη από ένζυμα θεραπεία καρκίνου (Pospisil et al., 2007).

Ένα χαρακτηριστικό παράδειγμα εφαρμογής της Κειμενικής Εξόρυξης, για την ταυτοποίηση δικτύων που οι κόμβοι τους αποτελούν οντότητες που σχετίζονται με ασθένειες, είναι αποτέλεσμα της ερευνητικής εργασίας του Krauthammer και του συνεργάτη του (Krauthammer & Kaufmann, 2004). Στα πλαίσια αυτής δημιουργείται το εξορυκτικό εργαλείο GeneWays, που αυτοματοποιημένα εξετάζει μεγάλο αριθμό από πλήρη κείμενα και προβλέπει πιθανές φυσικές αλληλεπιδράσεις, ακμές μεταξύ των κόμβων του δικτύου, που συμβολίζουν τα υποψήφια γονίδια που λανθάνουν στη βιβλιογραφία.

Πρώτα, δοκιμάζεται η εξορυκτική ικανότητα του GeneWays σε 25 επιστημονικά περιοδικά, με αποτέλεσμα ένα δίκτυο αλληλεπίδρασης βασισμένο στην βιβλιογραφία, το οποίο περιγράφει τις άμεσες συσχετίσεις, όπως δεσμός ή φωσφορυλίωση (**Εικόνα 17**). Στη συνέχεια, καταγράφεται ένας κατάλογος από 60 υποψήφια γονίδια που σχετίζονται με τη νόσο του Alzheimer, σχηματισμένος από έναν ειδικό στο πεδίο. Τα γονίδια αυτά χρησιμοποιούνται για την αναζήτηση σε υπό-δίκτυα στα οποία μπορεί να ενσωματώσουν νέα γονίδια που σχετίζονται με τη νόσο του Alzheimer. Αυτά τα γονίδια διαλέγονται απ' την μέθοδο του μοριακού τριγωνισμού και στη συνέχεια συμπληρώνεται μια λίστα με 60 από αυτά. Η ανάλυση ξεκινά με ένα πιθανώς θορυβώδες σύνολο γονιδίων εκκίνησης (seed genes) που έχουν ταυτοποιηθεί ότι φέρουν πληροφορίες σχετικά με το μοριακό υποσύστημα το οποίο παρουσιάζει σχετική διαταραχή. Όταν πλέον κατασκευαστεί το δίκτυο που απεικονίζει την σχετική

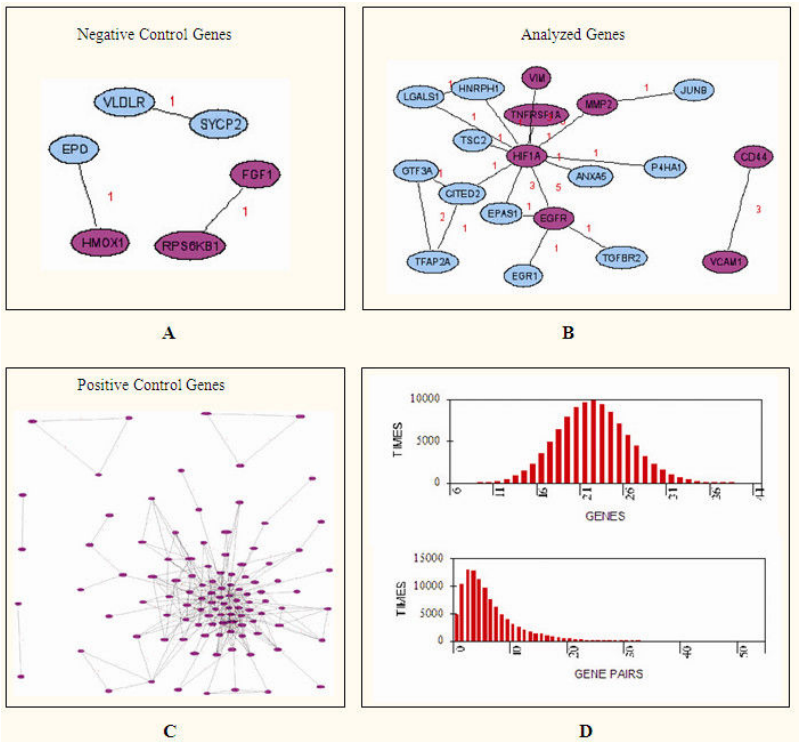
έκφραση γονιδίων, προκύπτει μια δεύτερη κατηγορία γονιδίων που είναι αυτά που ενώνονται άμεσα στο δίκτυο με δύο γονίδια εκκίνησης. Εάν τα γονίδια των εκκίνησης κατανέμονται τυχαία τείνουν να απέχουν πολύ από το ένα το άλλο. Επομένως, λίγοι κόμβοι βρίσκονται σε άμεση γειτνίαση με περισσότερα από ένα τέτοια γονίδια. Από την άλλη πλευρά, τα γονίδια εκκίνησης συσσωματώνονται σε μια συμπαγή μοριακή γειτονία, υποδεικνύοντας έτσι το ελαττωματικό υποσύστημα, τείνοντας να έχουν επικαλυπτόμενες γειτνιάσεις (**Εικόνα 17**). Στο σχήμα φαίνονται τα κομβικά γονίδια τα οποία είναι χρωματισμένα, με κόκκινο τα γονίδια εκκίνησης και με κυανό αυτά που συνορεύουν με τουλάχιστον δύο κόκκινα. (A) Τυχαία κατανομή κόκκινων γονιδίων, άρα και λίγα κυανά. (B) Κατανομή κόκκινων γονιδίων σε συστάδες, άρα πολλά κυανά γονίδια και δημιουργία παθογενών περιοχών πάνω στο δίκτυο. Αυτή η μέθοδος έχει πολύ καλή συμπεριφορά στην πρόβλεψη παθογόνων κόμβων ενός δικτύου που ενώνονται με κάποιο απ' τα υποψήφια γονίδια, γεγονός που επιβεβαιώνεται από πολλούς και αξιόλογους ερευνητές.



Εικόνα 17. *Στο σχήμα φαίνονται τα κομβικά γονίδια (κόκκινο) και γονίδια εκκίνησης (κυανό). Το σχήμα παρήχθη απ' το πρωτότυπο λογισμικό GeneWays.*
(Η εικόνα αποτελεί μέρος της δημοσίευσης: Krauthammer et al., 2004.)

Προσφάτως, σημαντικές προσπάθειες έχουν καταβληθεί για την ανάπτυξη εργαλείων εξόρυξης, που εξειδικεύονται στην αναγνώριση αλληλεπίδρασης μεταξύ δικτύων που αφορούν στην ίδια ασθένεια, μέσω βιβλιογραφίας. Για παράδειγμα, PolySearch (Cheng et al., 2008) είναι ένα νέο-αναπτυχθέν

διαδικτυακό εργαλείο, που ταυτοποιεί συσχετισμούς και δίκτυα, από περιλήψεις δημοσιεύσεων και πληθώρα επιμελώς σχολιασμένων βάσεων δεδομένων. Παρομοίως, το GenCLip (Huang et al., 2008) είναι ένα εργαλείο εξόρυξης από βιβλιογραφία, που αναπτύχθηκε με σκοπό να ανακαλύψει συστάδες (clusters) και δίκτυα σχετικά με την εκάστοτε παθολογία. Σε μία διαδικασία αναζήτησης γονιδιακής σύμπτωσης με την αναζήτηση των όρων «hypoxia» και «fibroblast», η οποία ακολουθείται από την αναζήτηση για το ποια γονίδια σχετίζονται επιπλέον με τον όρο «keloid» (A), τα οποία είναι 232 αρνητικώς συσχετιζόμενα στο σύνολο. Στο (B) επιδεικνύονται τα γονίδια που αναλύθηκαν. Στο (C) τα θετικώς συσχετιζόμενα. Τα γονίδια που είναι χρωματισμένα με μοβ αναπαριστούν συσχέτιση με τον όρο «keloid». Στο (D) εμφανίζεται ο συσχετισμός των κατανομών των συσχετιζόμενων γονιδίων και των αναμενόμενων από κανονική κατανομή (**Εικόνα 18**).



Εικόνα 18. *Συστάδες και δίκτυα που σχετίζονται με παθογόνα γονίδια.*
(Η εικόνα αποτελεί μέρος της δημοσίευσης: Huanget al., 2008.

Παρόλο που με την εξόρυξη διευκολύνεται κατά πολύ η συγκομιδή οντοτήτων και γνώσεων, από έναν αστρονομικό αριθμό ερευνητικών άρθρων, ακόμα παραμένουν μερικά δομημένα προβλήματα. Το πρώτο είναι με την

ιδιότητα της ποικιλομορφίας και ασάφειας των βιοϊατρικών οντοτήτων (Desany & Zhang, 2004). Η ποικιλομορφία του όρου συμβαίνει όταν μια έννοια μπορεί να υποδηλωθεί από διάφορες εκφάνσεις. Για παράδειγμα, «προστάτης» και «προστατικό» μπορούν να χρησιμοποιηθούν ως είσοδοι αναζήτησης. Αντίστροφα, η λέξη ασάφεια προκύπτει όταν ο ίδιος όρος μπορεί να αναφέρεται σε πολλές βιοϊατρικές έννοιες, όπως για παράδειγμα, η λέξη "λίπος", που είναι αρκετά κοινή (Cohen & Hunter, 2008). Αυτές οι ασάφειες μπορούν να οδηγήσουν σε λανθασμένους συσχετισμούς μεταξύ της Μοριακής Βιολογίας και ανθρώπινων ασθενειών.

Για να ξεπεραστεί αυτό το πρόβλημα, έχουν προταθεί μέθοδοι για την ταχεία ανάπτυξη ελεγχόμενων λεξιλογίων στην εξόρυξη κειμένου. Για παράδειγμα, η χρήση των όρων GO (επίσης γνωστός ως λεξικό οντολογίας ελεγχόμενου γονιδίου, αγγ.: controlled gene ontology vocabulary) που υποδεικνύει την υποκυτταρική θέση, τη μοριακή λειτουργία και τη βιολογική διαδικασία επέτρεψε τον πιο κατάλληλο σχολιασμό για τις οντότητες και την ενισχυμένη ανάκτηση. Ένας δεύτερος περιορισμός είναι πρόσβαση στο πλήρες κείμενο των εγγράφων και στις παραπομπές. Πιο ολοκληρωμένες, συγκεκριμένες και λεπτομερείς πληροφορίες κρύβονται στον κορμό του κειμένου ενός άρθρου, παρά στις περιλήψεις. Έτσι, ο αριθμός των οντοτήτων που εντοπίζονται από εξόρυξη κειμένου υποβιβάζεται σε μεγάλο βαθμό, εξαιτίας της συμπυκνωμένης φύσης των περιλήψεων της βιβλιογραφίας. Τέλος, είναι σημαντικό οι ερευνητές να γνωρίζουν τα επίπεδα αξιοπιστίας και ακρίβειας των διαφόρων μεθόδων εξόρυξης και των εργαλείων τους. Η γεφύρωση των κενών μεταξύ βιολόγων και υπολογιστικών επιστημόνων μοιάζει δύσκολο έργο. Επομένως, ενώ οι βιολόγοι θα πρέπει να γνωρίζουν την καινοτομία της εξόρυξης κειμένων για την ανακάλυψη βιοϊατρικών στόχων, οι υπολογιστικοί ερευνητές πρέπει να ενθαρρύνονται να αναπτύξουν φιλικότερες, προς το χρήστη, μεθόδους και εργαλεία, για την διευκόλυνση των πειραματικών συνεργατών τους.

3.2. Εξόρυξη από δεδομένα μικρο-συστοιχιών (microarrays)

Η Εξόρυξη από Δεδομένα μικρο-συστοιχιών (microarrays) αναφέρεται στην εφαρμογή βίο-πληροφορικών προσεγγίσεων στα δεδομένα, για την ανακάλυψη οντοτήτων και μονοπατιών, που ορίζουν φαινοτύπους π.χ. ασθενειών. Αξίζει να σημειωθεί ότι όπου υπάρχει ο όρος «μονοπάτι» δεν ορίζεται ως «σηματοδοτικό» μονοπάτι, δεν συμβολίζει τον όρο που αντιλαμβάνονται οι βιολόγοι ως μονοπάτι, με την χρονική αλληλουχία των βιολογικών διεργασιών (καταρρακτών, αγγ: cascades). «Μονοπάτι» ορίζεται η απόσταση μεταξύ δύο κόμβων, επί ενός δικτύου και αποτελεί μετρική της απεικονιστικής τεχνικής αυτής. Για την αποφυγή παρεξηγήσεων, η βιολογική έννοια θα δηλώνεται ρητά ως «σηματοδοτικό μονοπάτι».

Εδώ δύο είναι οι προσεγγίσεις, οι οποίες χρησιμοποιούνται ευρέως: Η Αναγνώριση Συστάδων χωρίς επίβλεψη και η επιβλεπόμενη ταξινόμηση (Mount & Pandey, 2005). Στην προηγούμενη προσέγγιση μια ομάδα γονιδίων συντονισμένης έκφρασης σε ένα υποσύνολο συνθηκών, προσδιορίζεται χρησιμοποιώντας μεθόδους ομαδοποίησης όπως η Ιεραρχική Συσσωμάτωση (Hierarchical Clustering), της Ανάλυσης Κύριων Συστατικών (Principal Component Analysis, PCA) και των χαρτών με δυνατότητα αυτό-οργάνωσης (self organizing custering, SOC) (Mount & Pandey, 2005). Για παράδειγμα, η μέθοδος SOM (Self-Organized Mapping: αυτό-οργανωμένη απεικόνιση) βρίσκει ένα βέλτιστο σύνολο από κεντροειδή που χαρακτηρίζουν συστάδες, γύρω απ’ τα οποία συναθροίζονται τα δεδομένα της γονιδιακής έκφρασης. Ύστερα, δείγματα ιστών ή κυττάρων χωρίζονται σε ομάδες, με κάθε κεντροειδές να ορίζει μια Συστάδα (Cluster), η οποία βασίζει την συνοχή της στις μετρικές ευκλείδειας απόστασης και στον συντελεστή συσχέτισης κατά Pearson χωρίς επίβλεψη (D' Haeseleer, 2005).

Αντιθέτως, στην εποπτευόμενη ταξινόμηση, αναζητούμε γονίδια που μπορούν να διακρίνουν τα γνωστά δείγματα και τις συνθήκες κάτω απ’ τις οποίες ελήφθησαν. Στα πλαίσια μιας τυπικής εποπτευόμενης ταξινόμησης, τα γενικά αρχεία γονιδιακής έκφρασης από ασθενείς ιστούς ή εκκρίματα, θα συγκριθούν με τα φυσιολογικά. Από την διαδικασία αυτή θα προσδιοριστεί μια λίστα με γονίδια ή μονοπάτια στόχους, τα οποία σχετίζονται ισχυρά με την ασθένεια. Επίσης, χρησιμοποιήθηκαν και μέθοδοι εποπτευόμενης ταξινόμησης, όπως η Γραμμικά Διακριτοποιημένη Ανάλυση (Linear Discriminant Analysis), η

Εύρεση Εγγύτατου Γείτονα (Nearest Neiborhood Search) και οι Γενετικοί Αλγόριθμοι (Mount & Pandey, 2005).

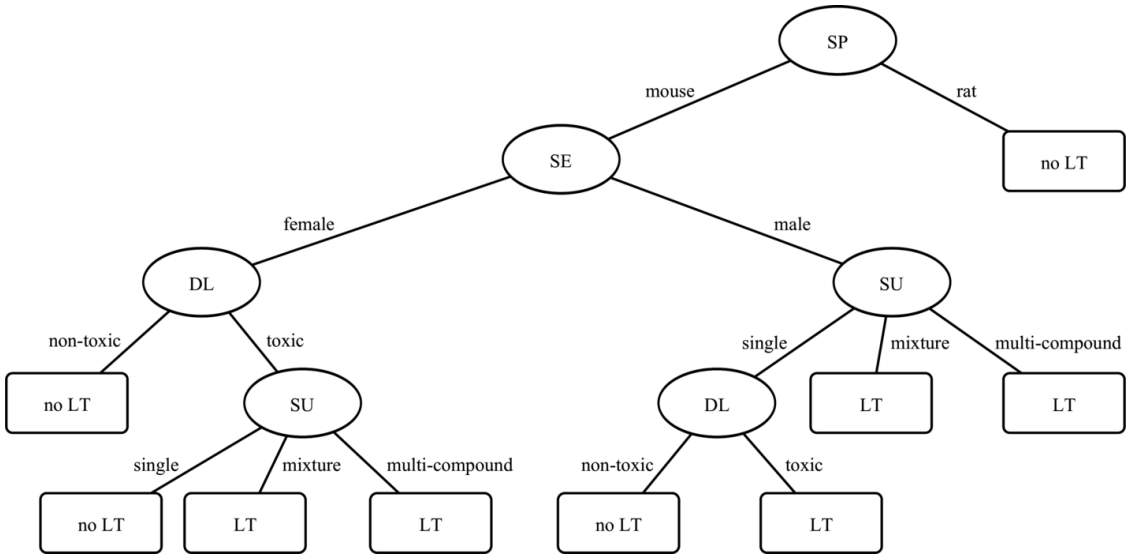
Σε σχέση με την εκθετική αύξηση των δεδομένων από μικρο-συστοιχίες τα τελευταία χρόνια, μια υπολογίσιμη προσπάθεια έγκειται στην οργάνωση βιβλιοθηκών από τέτοια δεδομένα. Οι βιβλιοθήκες θα χρησιμοποιούνται και θα υποστηρίζονται απ' αυτή, πάντα με στόχο την στοχευμένη αναζήτηση. Συνεπώς, παρέχεται η δυνατότητα μετά-ανάλυσης πολλαπλών πακέτων από δεδομένα μικρο-συστοιχιών, που απευθύνεται σε παρόμοιες βιολογικές υποθέσεις (Rhodes & Chinnaiyan, 2004). Η αξία τέτοιων μετά-αναλυτικών μεθόδων είναι ότι η στατιστικές μετρήσεις μπορούν πάντα να συγκριθούν μεταξύ τους και έτσι να υπάρχει ταυτόχρονη αξιολόγηση των αποτελεσμάτων. Επιπροσθέτως, άλλες στρατηγικές ανακάλυψης γονιδίων, όπως η MPSS (Massive Paralel Signature Sequencing), η SAGE (Serial Analysis of Gene Expression) και η EST (Expressed Sequence Tags), επίσης αποδεικνύονται ουσιαστικές στην ταυτοποίηση στόχων και δεικτών (markers) (Narayanan, 2007).

3.2.1 Ταυτοποίηση Θεραπευτικών Στόχων

Η εξόρυξη από δεδομένα μικρο-συστοιχιών είναι αποδεδειγμένα αποτελεσματική στην ανακάλυψη γονιδίων στόχων. Για παράδειγμα, το γονίδιο IGFBP3 έχει ταυτοποιηθεί ως στόχος υπέρ-μεθυλίωσης για τον καρκίνο του προστάτη (Perry & Loftus et al., 2007). Συνοπτικά, τα γονίδια που υφίστανται σημαντική καταστολή σε καρκινικό προστάτη σε σχέση με ένα κανονικό, αναγνωρίστηκαν από την βάση δεδομένων του Άτλαντα Γονιδιακής Έκφρασης. Τα ανακτηθέντα γονίδια οργανώθηκαν με τη βοήθεια του GeneCards (http://www.genecards.org/). Μεταξύ ενός καταλόγου 631 ανακτηθέντων γονιδίων, 16 από αυτά ταυτοχρόνως ταυτοποιήθηκαν από παλαιότερες μελέτες και, τέλος, η IGFBP3 επιλέχθηκε και επαληθεύτηκε ως στόχος υπέρ-μεθυλίωσης του καρκίνου του προστάτη.

Ένα άλλο αξιόλογο παράδειγμα είναι του Ryu και των συνεργατών του, όπου προσφάτως κατέβαλαν προσπάθεια στην ταυτοποίηση νέων μοριακών υπογραφών ως θεραπευτικούς στόχους, σε σχέση με το Καλπάζων Μελάνωμα,

μια απ' τις πιο ραγδαία αυξανόμενες μορφές καρκίνου στις ΗΠΑ (Ryu et al., 2007). Πρώτα, συνέκριναν και ανέλυσαν μια μεγάλη μάζα από προφίλ γονιδιακών εκφράσεων από μια σειρά καρκινικών κυττάρων, που αντιπροσωπεύουν διακεκριμένα στάδια κακοήθους εξέλιξης, καθώς και τα πρωτογενή ανθρώπινα μελανοκύτταρα μέσω μεθόδων ιεραρχικής ομαδοποίησης χωρίς εποπτεία, που εφαρμόζονται στο GeneCluster. Η ομαδοποίησή τους επέτρεψε να αναγνωρίσουν δύο διακριτές ομάδες κυτταρικών σειρών, μία κύρια και μία επιθετική ομάδα μελανώματος. Στη συνέχεια επιστρατεύτηκε μια πλατφόρμα εποπτευόμενης εξόρυξης σε δεδομένα μικρο-συστοιχιών (SAM), σε συνδυασμό με μια ανάλυση των σχολίων που αφορούν στις λειτουργίες, με σκοπό να ταυτοποιηθεί ένα σύνολο από ιδιαίτερα επιθετικά γονίδια, μεταξύ των οποίων τα NF-kB, CXCL1, CXCL2, IL-8, MMP1 και IGFBP3, που έχουν προηγουμένως εμπλακεί στην προαγωγή της αγγειογένεσης στους όγκους, ενός βασικού χαρακτηριστικού των επιθετικών όγκων.



Εικόνα 19. *Το παρόν δενδρικό διάγραμμα εκπαιδεύτηκε με τον αλγόριθμο C4.5 και επιτελεί ταξινόμηση όγκων στο ήπαρ (liver tumors, LT).*
(Η εικόνα είναι μέρος της δημοσίευσης: Ring & Eskofier 2015.)

Η εργασία των Ring και Eskofier (2015) ως σκοπό έχει να αποδείξει συστηματικά σφάλματα στην εμφάνιση καρκίνου του ήπατος (liver tumors, LT) σε τρωκτικά, τα οποία διακρίνει με μία διαδικασία δενδρικής ταξινόμησης βάσει πληροφορίων για το είδος (spiece, SP) και το φύλο (sex, SE), για την τοξική

ουσία που παράγεται (substance, SU) και την δόση τοξικότητας (dose level, DL) (**Εικόνα 19**). Τα δεδομένα πάνω στα οποία γίνεται η εξόρυξη ανήκουν στην Βάση Δεδομένων του NTP (U.S. National Toxicology Program). Η ταξινόμηση των περιστατικών των πειραματόζωων που εμφανίζουν καρκίνο στο ήπαρ, γίνεται από ένα Δέντρο Λήψης Απόφασης. Οι μεταβλητές είναι το είδος (αν είναι ποντικός ή αρουραίος), το φύλο, το υπόστρωμα, και το κατά πόσο η δόση είναι τοξική ή όχι. Η μελέτη τους χρησιμοποιεί τεχνικές του Data Mining σε δεδομένα από διαδικτυακές τράπεζες. Δείχθηκε ότι τα περιστατικά που εμφάνισαν καρκίνο στο ήπαρ, μπορούν να προβλεφθούν βάσει της δόσης που τους χορηγήθηκε. Η ίδια ακριβώς μέθοδος εξόρυξης μπορεί να χρησιμοποιηθεί εν γένει, σε διαφορετικές ασθένειες, όργανα και οργανισμούς.

3.2.2 Ταυτοποίηση Δεικτών για Διάγνωση ή Πρόγνωση

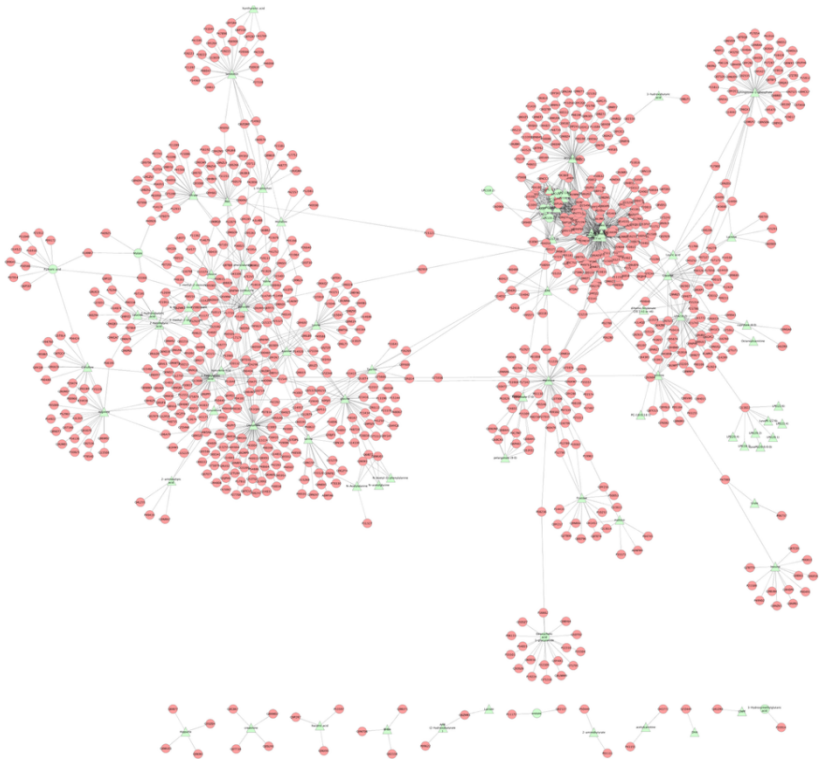
Οι βίο-δείκτες (biomarkers) είναι μόρια που χρησιμοποιούνται για την αναγνώριση της φυσιολογικής κατάστασης, αποτελώντας σήμα κατατεθέν για οποιαδήποτε αλλαγή της φυσιολογίας, τόσο σε επίπεδο ιστού όσο σε επίπεδο χυμών κατά την εξέλιξη της ασθένειας (Campagne & Skrabanek, 2006). Με τις σημερινές αυξανόμενες ανάγκες για εύρεση νέων βίο-δεικτών, η εξόρυξη σε δεδομένα από μικρο-συστοιχίες γνωρίζει αυξητική τάση στον εντοπισμό διαγνωστικών και προγνωστικών γονιδίων δεικτών. Για παράδειγμα, ο Kim και οι συνεργάτες του αναφέρουν ότι η εξόρυξη σε δημόσια δεδομένα γονιδιακής έκφρασης από τις βιβλιοθήκες CGAP και GEO, έχουν σκοπό να βρεθούν υποψήφιοι δείκτες για τον καρκίνο του πνεύμονα (Kim et al., 2007). Πρώτα, ανέκτησαν μερικές εκατοντάδες εκφραζόμενων γονιδίων στο καρκίνο του πνεύμονα, μέσω μιας μετά-ανάλυσης των δύο βιβλιοθηκών, με την χρήση της μεθόδου Fisher. Στη συνέχεια, μέσω μιας συστηματικής εξέτασης, που βασίζεται στις αναφερόμενες στο σχολιασμό ιδιότητες των γονιδίων και στις στατιστικές τιμές P-value, κατέληξαν σε ένα σύνολο 20 υποψήφιων γονιδίων, που δευτερογενώς υπόκεινται σε πειραματική αξιολόγηση. Τελικώς, 7 από τα υπέρ-παραγόμενα επιλέγονται ως πιθανοί διαγνωστικοί δείκτες.

Παρομοίως, άλλη ομάδα κατέγραψε σε έναν κατάλογο βίο-δείκτες στο αίμα και σχετίζονται με έξι κοινούς τύπους καρκίνου στον άνθρωπο, με την

εφαρμογή εξορυκτικής στρατηγικής στην πλατφόρμα μικρο-συστοιχιών Oncomine και μιας επιμελημένης γνωσιακής βάσης δεδομένων για μονοπάτια (Yang et al., 2008). Πρώτο, όλα τα γονίδια με σημαντική αύξηση στην έκφραση και οντολογικά καθορισμένη (σε GO) κυτταρική λειτουργία στον καρκίνο, συλλέχθηκαν με τον κανόνα των ελάχιστων ψευδώς θετικών (false discovery rate cut-off). Τα ανακεκτημένα γονίδια στη συνέχεια υπόκεινται σε ανάλυση των μονοπατιών στα οποία συμμετέχουν και τελικά παραμένουν στη λίστα μόνο όσα κωδικοποιούν κρυφές πρωτεΐνες-δείκτες σε αίμα ή ορό ή πλάσμα. Περαιτέρω, μια μελέτη σύγκρισης των γονιδίων ανακτηθέντων δεικτών σε διάφορους τύπους όγκων έχει οδηγήσει στην ταυτοποίηση κοινών και μοναδικών δεικτών σε έξι όγκους, μεταξύ των οποίων ErbB2, BRCA1 / BRCA2, PSA, HABP2 και IGF-II έχουν επίσης επιλεγεί από άλλες μελέτες ως υποψήφιοι δείκτες όγκου και ήδη χρησιμοποιούνται κλινικά. Αξίζει να σημειωθεί ότι, μετά από χειροκίνητη διασταύρωση δεδομένων με τη βάση δεδομένων iHOP και άλλες επιμελημένες βάσεις δεδομένων, 13 από τους συνήθεις 35 δείκτες (περίπου το 1/3) σε όγκους προστάτη, μαστού και πνεύμονα έχουν επιβεβαιωθεί από τη βιβλιογραφία ότι χρησιμεύουν ως προγνωστικοί δείκτες στην εξέλιξη και την επιθετικότητα των ανθρώπινων όγκων (Yang et al., 2008). Επιπροσθέτως, τα MMP1, CD44, CP και NOTCH4 επιλέχθηκαν και ιεραρχήθηκαν ως υποσχόμενοι δείκτες στο αίμα, σύμφωνα με την κανονικοποιημένη τιμή κατωφλίου από την RT-PCR.

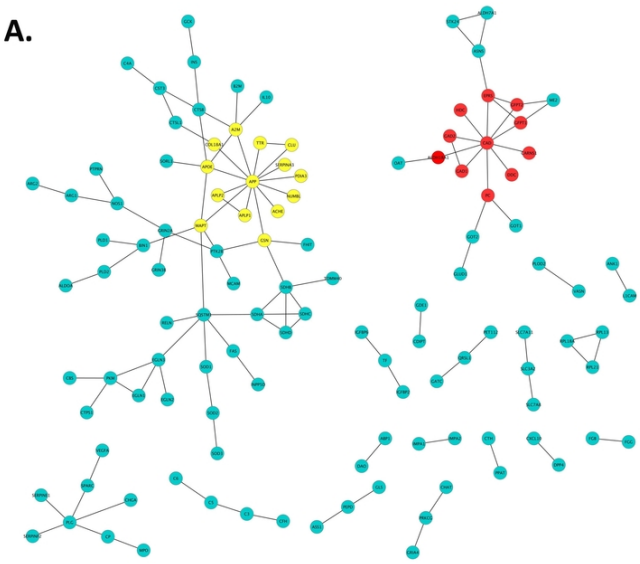
Εδώ και καιρό έχει αναγνωριστεί ότι ο παραδοσιακός τρόπος κατασκευής εμπορεύσιμων φαρμάκων, είναι μία επίπονη, χρονοβόρα και δαπανηρή διαδικασία. Υπολογίζεται ότι μια μελέτη, ώστε το φάρμακο να βγει στην αγορά, διαρκεί 10-17 χρόνια και λιγότερο από 10% των μελετών κατορθώνει να κατασκευάσει ένα φάρμακο για ανθρώπινη χρήση. Ένα απλό σύστημα είναι η εξέταση ήδη εγκεκριμένων φαρμάκων, έτσι ώστε να επαναπροσδιοριστούν οι δράσεις τους σε άλλες παρεμφερείς ή όχι ασθένειες. Πρόκειται, για την εκ νέου στόχευση φαρμάκων που ήδη έχουν περάσει τον κλινικό έλεγχο, μειώνοντας σημαντικά τον κίνδυνο, τα κόστη και την αναμονή. Στην περίπτωση ο στόχος είναι ο επαναπροσδιορισμός των δυνατοτήτων ήδη εγκεκριμένων φαρμάκων για την θεραπεία του διαβήτη. Στην προσπάθεια αυτή αναλύθηκαν δεδομένα από γονιδιωματικές (GWAS, genome wide assosiation studies), πρωτεωμικές και μεταβολωμικές μελέτες. Απ’ τις μελέτες αποκαλύπτονται 992 πρωτεΐνες με

πιθανή αντί-διαβητική δράση στον άνθρωπο και με την κατασκευή του δικτύου συσχέτισης μεταξύ μεταβολιτών και πρωτεϊνών που σχετίζονται την νόσο του διαβήτη. Συνολικά απεικονίζονται 1.660 ζεύγη (ακμές) από πρωτεΐνες και μεταβολίτες που σχετίζονται με την νόσο. Στο δίκτυο απεικονίζονται τόσο οι μεταβολείτες όσο και οι πρωτεΐνες που σχετίζονται με αυτούς, σύμφωνα με την βάση δεδομένων HMDB. (**Εικόνα 20**).



Εικόνα 20. *Δίκτυο συσχέτισης μεταξύ μεταβολιτών και πρωτεϊνών που σχετίζονται την νόσο του διαβήτη. Τα πράσινα τρίγωνα αναπαριστούν μεταβολίτες, ενώ οι κόκκινοι κύκλοι πρωτεΐνες που σχετίζονται με τους μεταβολίτες. Το δίκτυο κατασκευάστηκε απ' το λογισμικό CytoScape.*
(Η εικόνα αποτελεί μέρος της δημοσίευσης: Zhang et al., 2015)

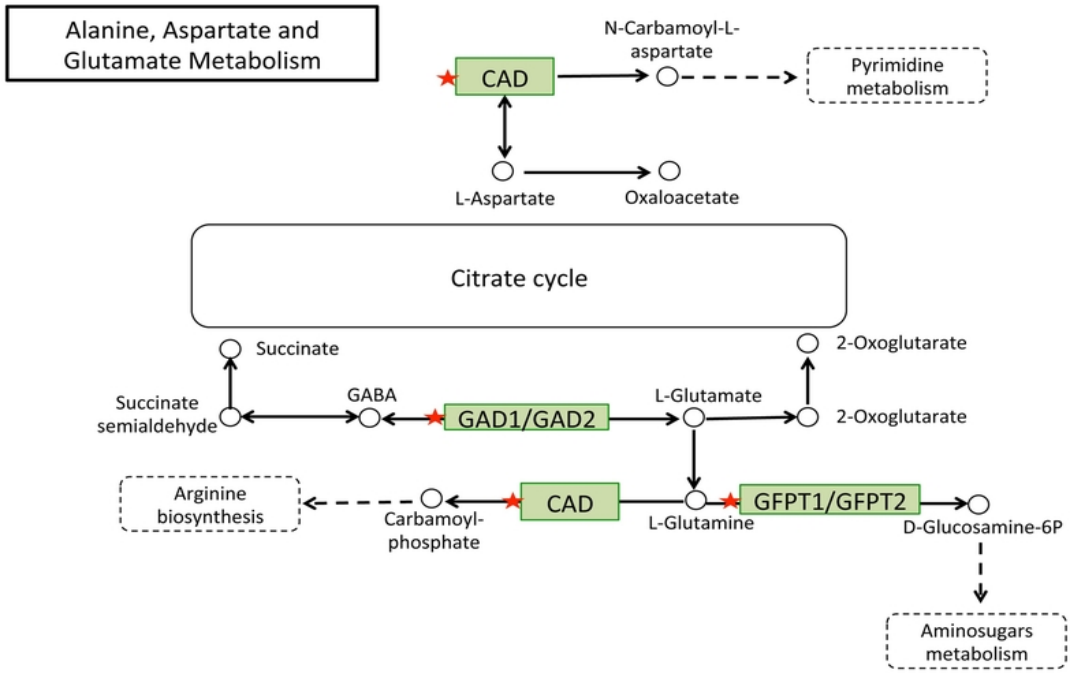
Ύστερα σε συνδυασμό με πληροφορίες για το ποια φάρμακα στοχεύουν σε τέτοιες πρωτεΐνες, βρέθηκε ότι οι 108 από αυτές είναι στόχοι ήδη εμπορεύσιμων φαρμάκων. Οι μέθοδοι που χρησιμοποιούνται είναι η κατασκευή και η ανάλυση χαρτών συγγένειας που απεικονίζονται σε δίκτυα, καθώς και η εξόρυξη από κείμενο. Τελικά, 9 από αυτά δείχνεται ότι μπορούν να επαναπροσδιορίσουν την στόχευση τους προς την αντιμετώπιση του διαβήτη (Zhang et al., 2015).



Εικόνα 21. *Μια ανάλυση διαπρωτεϊνικής αλληλεπίδρασης 524 σχετιζόμενων πρωτεϊνών αποκαλύπτει δύο μεγάλες πρωτεϊνικές συστάδες: το δίκτυο APP (14 κίτρινοι κόμβοι) και το δίκτυο CAD (11 κόκκινοι κόμβοι).*
(Η εικόνα αποτελεί μέρος της δημοσίευσης: Zhang et al., 2016)

Σε μια πολύ παρόμοια εργασία της οποίας ηγείται και πάλι ο Zhang (Zhang et al., 2016), αναζητούνται πιθανά φάρμακα κλινικά ελεγμένα που να στοχεύουν στην θεραπεία της νόσου του Alzheimer. Δημιουργήθηκε μια λίστα με πρωτεΐνες που σχετίζονται με την ασθένεια. Η λίστα αυτή προέκυψε από την ανάλυση ομικών δεδομένων κάθε μορφής, πρωτεομικά, επιγονιδιωματικά, γονιδιωματικά και μεταβολομικά από της βιβλιοθήκες PubMed και OMIM. Ενώ οι πληροφορίες σχετικά με τα φάρμακα πάρθηκαν από την DrugBank και την Therapeutic Target Database. Η λίστα πρωτεϊνών που σχετίζονται με την ασθένεια αποτελείται από 524 πρωτεΐνες, 18 εκ των οποίων είναι στόχοι 75 ήδη υπαρχόντων φαρμάκων που χρησιμοποιούνται για άλλη ασθένεια (**Εικόνα 21**). Ένας αλγόριθμος βαθμονόμησης που ιεραρχεί τα υποψήφια φάρμακα και αποκαλύπτει τα επικρατέστερα και δραστικότερα. Επίσης, ανακαλύπτονται και 7 νέα φάρμακα που ενεργοποιούν μονοπάτια με ανασταλτική δράση ως προς την ασθένεια (Zhang et al., 2016). Οι μέθοδοι που χρησιμοποιήθηκαν αφορούν στην αναζήτηση σε βιβλιοθήκες για πιθανούς στόχους με ανασταλτική δράση, η χαρτογράφηση και απεικόνιση των σχέσεων μεταξύ μεταβολιτών και πρωτεϊνών, η χαρτογράφηση και η απεικόνιση των σχέσεων μεταξύ πρωτεϊνών

και υπαρχόντων φαρμάκων (**Εικόνα 22**), καθώς και μία πρώτη προσπάθεια να δειχθεί ότι τα φάρμακα που κρίθηκαν υποψήφια για να επαναστοχεύσουν την χρήση τους στην θεραπεία της νόσου του Alzheimer, μπορούν όντως να είναι άξια της φήμης τους.



Εικόνα 22. *Η ανάλυση με επιπλέον εμπλουτισμό των λειτουργιών που επιτελούν αυτές οι πρωτεΐνες προκύπτουν νέα πιο εξειδικευμένα στοιχεία. Για την δημιουργία της εικόνας χρησιμοποιήθηκαν από κοινού δεδομένα απ' το διαδικτυακό εργαλείο David και την βάση δεδομένων KEGG.*

(Η εικόνα αποτελεί μέρος της δημοσίευσης: Zhang et al., 2016)

3.3. Εξόρυξη σε Πρωτεομικά Δεδομένα

Με την έλευση της μετά-γονιδιωματικής εποχής η πρωτεομική, εμφανίστηκε σαν μια νέα τεχνολογία που βασίζεται στην ανάλυση φασματομετρίας μάζας υψηλής απόδοσης (Mass Spectrometry, MS) (Siepen et al., 2008). Συνεπώς, απαιτείται πρωτεομική εξόρυξη για την ανάλυση και εξαγωγή χρήσιμων πληροφοριών από δεδομένα φασματομετρίας μάζας. Δεδομένου ότι μπορούν να συμπεριληφθούν 1 έως 2 εκατομμύρια δεδομένα ανά δείγμα, για ένα φασματοσκοπικό όργανο υψηλής ευκρίνειας, η πρωτεομική εξόρυξη είναι δύσκολη λόγω μεγέθους και διαστάσεων των δεδομένων (Gerling et al., 2006).

Για παράδειγμα, μια τυπική ανάλυση τέτοιων δεδομένων, για ένα δείγμα αίματος ασθενούς θα μπορούσε να οδηγήσει στη δημιουργία 350.000-400.000 σημείων. Σε πρακτικό επίπεδο είναι απαγορευτικό να αναλυθούν αυτά τα πολλά σύνολα δεδομένων με παραδοσιακά εργαλεία. Συνεπώς, υπάρχει ανάγκη να αναπτυχθούν νέα εργαλεία και μέθοδοι εξόρυξης, για να υπερπηδηθεί το πρόβλημα στη βάση μιας πρωτεομικής προσέγγισης.

Στην εργασία του Zafar (2001) επιχειρείται μια ανάλυση της πρωτεϊνικής ομοιότητας σε αλληλουχίες απ' το BLASTP. Μεταξύ της προς αναζήτηση ακολουθίας και της αλληλουχίας αναφοράς, η πρωτεΐνη συνδυάζεται ξεχωριστά με κάθε ένα στοιχείο της πρωτεϊνικής Βάσης Δεδομένων. Αν ο βαθμός ομολογίας τους τείνει ή ξεπερνάει έναν αριθμό κατωφλίου, θεωρείται οι πρωτεΐνες σχετίζονται και οι αριθμοί των γονιδίων τους αποθηκεύονται από το σύστημα. Ο χρήστης μπορεί είτε να ορίσει τρία διαφορετικά κατώφλια ομοιότητας ή να χρησιμοποιήσει τα σταθμισμένα, ήδη υπάρχοντα ως προεπιλογές. Όταν ολοκληρωθεί η διαδικασία, τα σκορ συγκεντρώνονται σε ένα διάγραμμα γενετικής τάξης. Αυτά τα δισδιάστατα γραφήματα έχουν σημειακές αναπαραστάσεις των πρωτεϊνών. Παρ' όλα αυτά, τα σημεία σε αυτό το διάγραμμα αναπαριστούν γονίδια με πρωτεϊνική ομολογία, που ταυτοποιούνται μεταξύ των δύο γονιδιωμάτων. Ο τρόπος με τον οποίον είναι στοιχισμένες είναι σύμφωνα με τις θέσεις ταύτισης. Η διάταξη των σημείων περί της κεντρικής διαγωνίου υποδεικνύει την συγγραμμικότητα των γονιδίων. Η αντιμετάθεση γονιδίων και η τακτική αναδιοργάνωση εμφανίζονται ως μοτίβα από εικονιζόμενα σημεία ή ευθείες που απέχουν της διαγωνίου.

Αυτό το γραφικό αποτέλεσμα, με διαφοροποιημένα χρώματα και σύμβολα αναπαριστά διάφορες σε επίπεδα ομοιότητας μεταξύ γονιδίων, εκτείνει το εύρος χρήσεων του ήδη υπάρχοντος προγράμματος GeneOrder2.0. Αυτοί οι διαφορετικοί βαθμοί ομοιότητας μπορούν να αναλυθούν για δύο αλληλουχίες, επιτρέποντας την ανάλυση ταξινομημένων γονιδίων που ενέχουν την πιθανότητα να διαφέρουν λόγω εξέλιξης (Zafar et al., 2001). Επί παραδείγματι, υπάρχουν δύο δείγματα γονιδίων στην στοίχιση των γονιδίων του νουκλεοπολυεδρικού ιού, που φαίνεται να εμπλέκονται σε ένα γρηγορότερο ρυθμό εξέλιξης απ' τα γειτονικά γονίδια. Ο πίνακας που περιέχει όλες αυτές τις ομολογίες πρωτεϊνών, είναι διαθέσιμος στο διαδίκτυο μέσω της GenBank.

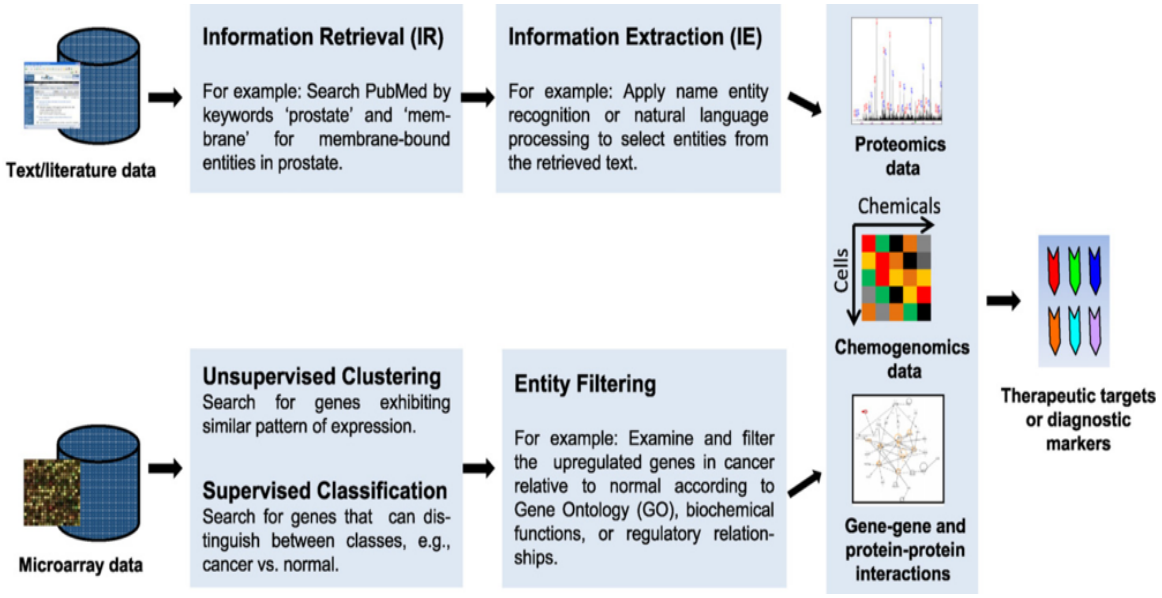
Πρόσφατα, η Ανοιχτή Πρωτεομική Βιβλιοθήκη (Open Proteomic Database, OPD) και η βάση δεδομένων EMBL Proteomic (PRIDE) έγιναν διαθέσιμα στο κοινό και έχουν προταθεί μεθόδους εξόρυξης όπως η ανάλυση κατά Bayes, η Βασισμένη σε Κανόνες Ανάλυση (Rule-Based Analysis) και η βαθμονόμηση χαρακτηριστικών υπογραφών (Siepen et al., 2008). Οι προηγμένες υπολογιστικές μέθοδοι, ωστόσο, απαιτούνται ακόμη για την ολοκλήρωση, την εξόρυξη, τη συγκριτική ανάλυση και τη λειτουργική ερμηνεία των πρωτεομικών δεδομένων υψηλής απόδοσης, χαρακτηριστικό παράδειγμα είναι αυτό της ταυτόχρονης απεικόνισης δεδομένων γονιδιακής έκφρασης, που επιπλέον έχουν υποστεί ανάλυση Συσταδοποίησης.

3.4. Εξόρυξη σε Χημικό-Γονιδιακά Δεδομένα

Μια άλλη αναδυόμενη προσέγγιση, εξόρυξη σε χημικό-γενετικά δεδομένα, ερμηνεύει τα δεδομένα από τη χημική γονιδιωματική, μια νέα τεχνολογία που εξετάζει τους ενδιαφέροντες φαινοτύπους (όπως βιωσιμότητα, μορφολογία κυττάρων, προφίλ συμπεριφοράς και γονιδιακής έκφρασης) με τρόπο παράλληλο, προσαρμόζοντας μικρά μόρια από χημικές βιβλιοθήκες σε μια βιβλιοθήκη κυττάρων (Wuster & Babu, 2008). Στον διδιάστατο πίνακα ή χώρο, που προκύπτει από την εξέταση χημικό-γενετικής, η μία διάσταση είναι η χημική βιβλιοθήκη και η άλλη διάσταση είναι η βιβλιοθήκη διαφορετικών κυτταρικών τύπων.

Αυτό μπορεί να δημιουργήσει νέους τρόπους εντοπισμού κυτταρικών στόχων φαρμάκων και μονοπατιών ασθενειών. Η ερμηνεία και το φιλτράρισμα των πολυδιάστατων χημικό-γενετικών δεδομένων είναι δύσκολο έργο. Η πρόκληση σχετίζεται με την εξόρυξη σε κειμενικά δεδομένα και μικρο-συστοιχίες όταν αυτά μπορούν να συνδυαστούν με πρωτεομικά ή χημικό-γενετικά για την ανακάλυψη στόχων. Διαφορετικές πηγές και δομές δεδομένων μπορούν να χαρτογραφηθούν ή να απεικονιστούν βάσει των μονοπατιών αλληλεπίδρασης γονιδίου με γονίδιο και πρωτεΐνης με πρωτεΐνη (**Εικόνα 23**), με αποτέλεσμα την ανάπτυξη εργαλείων και μεθόδων εξόρυξης για ανάλυση δεδομένων με συστηματικό τρόπο (Kwon, 2006). Συγκεκριμένα, έχουν προταθεί διάφοροι αλγόριθμοι ομαδοποίησης, εποπτευόμενοι ή μη, για να

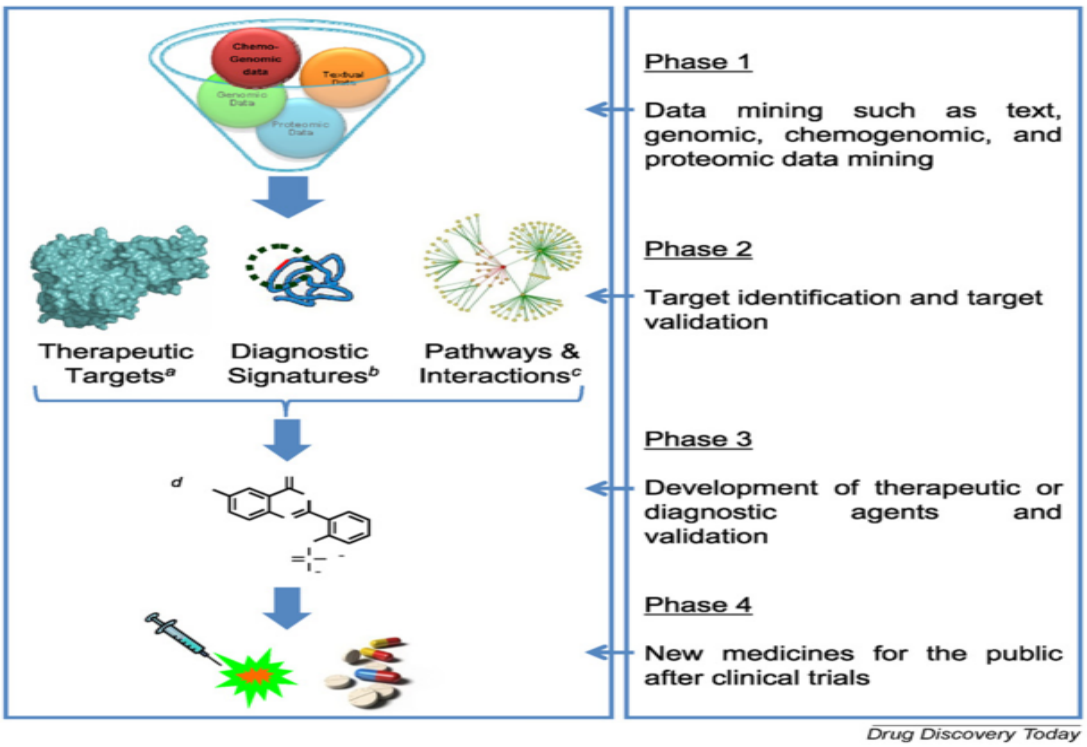
κατακρατηθεί το υποσύνολο γονιδίων με τις ενδιαφέρουσες ιδιότητες. Τέτοιοι αλγόριθμοι υπάγονται σε μεθόδους όπως η Ιεραρχική Συσταδοποίηση (Hierarchical Clustering), η μέθοδος των k-means, οι αυτό-οργανωμένοι χάρτες (self-organizing maps), βίο-Συσταδοποίηση (bioclustering) ή βελτιστοποίησης πράξεων μεταξύ πινάκων (matrix operation) (Wuster & Babu, 2008).



Εικόνα 23. *Η ακολουθία εργασιών κατά την Εξόρυξη από κειμενικά δεδομένα και δεδομένα μικρο-συστοιχιών ενοποιημένα με υψηλής απόδοσης και δεδομένα αλληλεπιδράσεων, με σκοπό την ανακάλυψη θεραπευτικών στόχων.*
(Η εικόνα αποτελεί μέρος της δημοσίευσης: Yang et al., 2009)

Προκειμένου να υποστηριχθεί και να μορφοποιηθεί μαθηματικά ο συσχετισμός γνωστών ή διερμηνευμένων λειτουργικών χαρακτηριστικών απ' το γονίδιο στο φαινότυπο, στα πλαίσια της εργασίας των Perez-Iratxeta et al., 2002, έχει αναπτυχθεί ένα σύστημα που υπακούει στις αρχές της εξόρυξης από δεδομένα βασιζόμενο σε ένα σύνολο σύμφωνα με την Θεωρία Ασαφών Συνόλων (Fuzzy Set Theory). Εφαρμόστηκε το σύστημα ιεράρχησης των υποψήφιων γονιδίων, για 455 γενετικά κληρονομούμενες ασθένειες, με τις οποίες ακόμα δεν σχετίζεται με κανένα γονίδιο. Η πρώτη φάση της διαδικασίας εξόρυξης ασχολείται με τον συνδυασμό πληροφορίας από μια πρωτεϊνική βάση δεδομένων και απ' το MEDLINE, μια βάση με αρχειοθετημένες της δημοσιεύσεις, τις παραθέσεις τους και ακόμα τις περιλήψεις αυτών ως λεκτικό

απόθεμα, με θεματολογία κυρίως από τον τομέα την βιοϊατρικής βιβλιογραφίας, με πάνω από 11 εκατομμύρια καταχωρήσεις.



Εικόνα 24. *Σχεδιάγραμμα της διαδικασίας ανακάλυψης φαρμάκων στην εποχή των omics. (a) Η απεικόνιση της αλκαλικής φωσφατάσης στον ανθρώπινο πλακούντα ως στόχο προς θεραπεία. (b) Οι διαγνωστικές υπογραφές εγχύονται στο ανθρώπινο αίμα. (c) Το αναπτυξιακό και πολλαπλασιαστικό κυτταρικό μονοπάτι ενός όγκου στον προστάτη. (d) Δομή από την ενζυμική ενεργοποίηση του φαρμάκου.*

(Η εικόνα αποτελεί μέρος της δημοσίευσης: Yang et al., 2009)

Η διαδικασία αποτελείται από τρία στάδια (**Εικόνα 24**): (i) Υπολογίζεται ο συσχετισμός μεταξύ παθολογικής συνθήκης και χημικών σχέσεων με τη βοήθεια του MEDLINE. (ii) Ακολούθως, υπολογίζεται η σχέση μεταξύ χημικών όρων (chemical terms) και όρων που περιγράφουν την πρωτεϊνική λειτουργία. Με τη Βάση Δεδομένων NCBI RefSeq, που περιέχει πάνω από δέκα χιλιάδες γονίδια, των οποίων η λειτουργία είναι πλήρως χαρακτηρισμένη, και σημειώνονται με όρους από ένα ελεγχόμενο λειτουργικό λεξιλόγιο. Πειραματικά στοιχεία αποδεικνύουν κάθε συσχέτιση μεταξύ πρωτεΐνης και λειτουργίας, με την βοήθεια ενός δείκτη απ' το MEDLINE. Ενώ, θεωρείται ότι τα σεσημασμένα γονίδια σχετίζουν τους λειτουργικούς τους όρους (οντολογικούς όρους) με τους

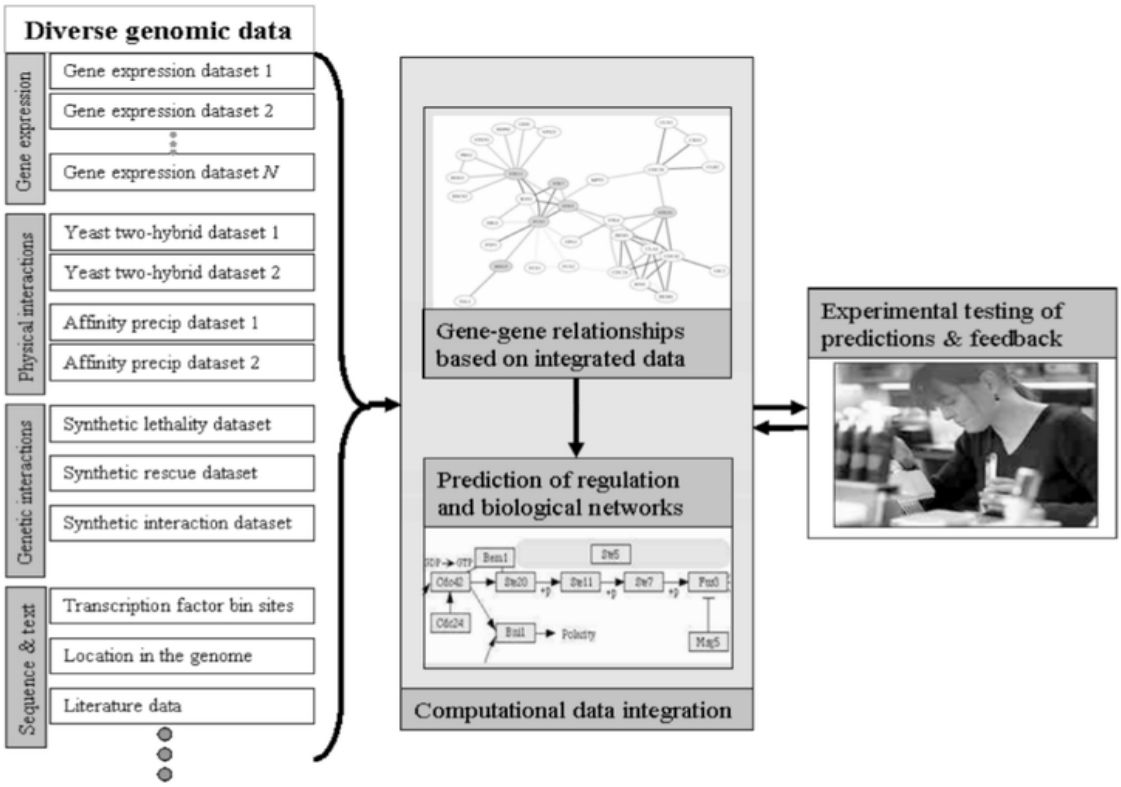
χημικούς όρους που βρέθηκαν στην προσαρμοσμένη βιβλιογραφία. (iii) Από τον συνδυασμό των συσχετισμών μεταξύ λειτουργικών και χημικών όρων, με τους ήδη εδραιωμένους συσχετισμούς παθολογίας και χημικών όρων, θα προκύψουν οι προαναφερθείσες σχέσεις μεταξύ παθολογικών συνθηκών και πρωτεϊνικών λειτουργικών όρων.

3.5. Ενοποιημένη ή Ενσωματωμένη Εξόρυξη

Η ανακάλυψη στόχων είναι μια επίπονη διαδικασία, κυρίως λόγω της πολυπλοκότητας των ανθρώπινων ασθενειών και της ανομοιογένειας διαφόρων βιολογικών δεδομένων. Δεν υπάρχει μια ενιαία επαρκής προσέγγιση εξόρυξης δεδομένων για την κατανόηση των κυτταρικών μηχανισμών και την ανασυγκρότηση των βιολογικών δικτύων (Chen & Chen, 2008). Για την ανάκτηση και προτεραιότητα των βιολογικά σημαντικών στόχων, πρέπει να ενσωματωθεί και να αναλυθεί ένας πλούτος δεδομένων σε πολλά διαφορετικά επιστημονικά πεδία, όπως οι αναλύσεις γονιδιακών ή πρωτεομικών δεδομένων (Hu et al., 2003) και δεδομένων μικρο-συστοιχιών. Πολλαπλά δεδομένα γονιδιακής έκφρασης και ποικίλα γονιδιωματικά δεδομένα μπορούν να συγχωνευτούν με υπολογιστικές μεθόδους και να δημιουργήσουν μια καθολική εικόνα των λειτουργικών σχέσεων μεταξύ των γονιδίων. Τα συγχωνευμένα δεδομένα είναι έτοιμα να χρησιμοποιηθούν στην πρόγνωση βιολογικών λειτουργιών, πρωτεϊνικών ρυθμίσεων ή στην κατασκευή και μοντελοποίηση δικτύων αλληλεπίδρασης μεταξύ τους (**Εικόνα 25**). Εναλλακτικά, μοντελοποιήσεις βιολογικών δικτύων χρησιμοποιούνται στην απευθείας ανάλυση γενομικών δεδομένων (Troyanskaya, 2005).

Οι προσεγγίσεις Βιοπληροφορικής που ενσωματώνουν διαφορετικές πηγές δεδομένων, λαμβάνοντας υπόψη τα πλεονεκτήματα και τα μειονεκτήματα της κάθε μιας, αναμένεται να ενισχύσουν σημαντικά την ανακάλυψη πολύτιμων στόχων (Kim et al., 2007). Συγκεκριμένα, ο συνδυασμός ή η ενσωμάτωση της Εξόρυξης σε Κείμενο και της Ανάλυσης Δεδομένων Υψηλής Απόδοσης (όπως γονιδιωματικά, πρωτεομικά ή χημικό-γενετικά δεδομένα) χρησιμοποιείται όλο και περισσότερο για την αναζήτηση δεικτών νόσου και στόχων φαρμάκων.

Αντιθέτως, με την εμφάνιση της Βιολογίας Συστημάτων, τη συνεχιζόμενη ανάπτυξη δεδομένων αλληλεπίδρασης μεταξύ γονιδίων ή πρωτεϊνών, επέτρεψε στους επιστήμονες να αναλύουν και να απεικονίζουν μια ποικιλία δεδομένων στο πλαίσιο των βιολογικών δικτύων ή μονοπατιών, κυρίως με μια Βάση Γνώσεων, όπως η KEGG (Kyoto Encyclopedia of Genes and Genomes), άλλα και βάσεις δεδομένων από πειραματικά επιβεβαιωμένες αλληλεπιδράσεις, όπως το UniHI (Unified Human Interactome). Για παράδειγμα, το PathwayExplorer είναι ένα εργαλείο που εξορύσσει από δεδομένα υψηλής απόδοσης από βάσεις γνώσεων όπως η KEGG και η GenMAPP. Επιπροσθέτως, επιτρέπει τη χαρτογράφηση του προφίλ έκφρασης γονιδίων ή πρωτεϊνών ταυτόχρονα σε κύριες ρυθμιστικές, μεταβολικές και κυτταρικές οδούς.



Εικόνα 25. *Εποπτική επίδειξη της διαδικασίας της καθολικής ανάλυσης πολλαπλών δεδομένων.*

(Η εικόνα αποτελεί μέρος της δημοσίευσης: Troyanskaya 2005)

Προκειμένου να βρεθούν νέα γονίδια που να σχετίζονται με μυοπάθειες, επιχειρείται από τον Neto και την ομάδα του (Neto et al., 2014), μια προσέγγιση ενσωματωμένου Data Mining. Σκοπός είναι· πρώτον η εξαγωγή των

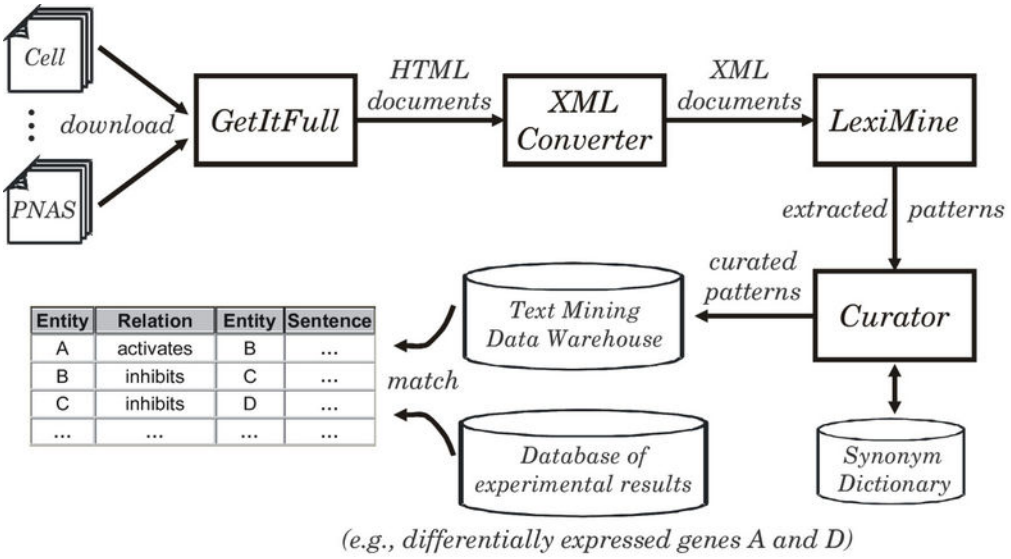
συγκεκριμένων γονιδιακών υπογραφών που χαρακτηρίζουν τις ομάδες μυοπαθειών, από γονίδια που είναι ήδη γνωστό ότι σχετίζονται. Ύστερα οι χαρακτηριστικές υπογραφές αυτές χρησιμεύουν στην αναζήτηση επιπλέον γονιδίων στο ανθρώπινο γονιδίωμα και στην οντολογία φαινοτύπου ποντικών. Ο καθορισμός των ομάδων 9 ασθενειών, στις οποίες πρόκειται να ταξινομηθούν τα γονίδια, γίνεται βάσει του Γονιδιακού Πίνακα Νευρωμυϊκών Διαταραχών (Gene Table of Neuromuscular Disorders). Όλα τα γνωστά γονίδια για τις διάφορες μυοπάθειες απεικονίζονται σε 4 δίκτυα βάσει των οντολογικών συσχετίσεων που έχουν μεταξύ τους. Εκτός απ' τις ακμές που ενώνουν τους κόμβους (γονίδια), επί των δικτύων σημειώνονται και οι ομάδες των γονιδίων σύμφωνα με την ταξινόμηση των μυοπαθειών. Στη συνέχεια γίνεται επιλογή των γονιδίων που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου. Ο τρόπος επιλογής των γονιδίων λαμβάνει υπ' όψιν το γράφημα συσχετίσεων, έτσι ώστε για την εκπαίδευση να χρησιμοποιηθούν γονίδια που να συγχέουν όσο το δυνατόν λιγότερο τις κλάσεις ασθενειών. Η εργασία πετυχαίνει την ανάκτηση τόσο συγκεκριμένων υπογραφών για κάθε κατηγορία μυοπάθειας, όσο και να αποκαλύψει νέα γονίδια που μπορεί σχετίζονται με τις μυοπάθειες. Οι χαρακτηριστικές υπογραφές ασθενειών καθώς και τα υποψήφια γονίδια, αναδεικνύουν την προοπτική κοινών παθολογικών μηχανισμών και συσχετισμών μεταξύ των διαφόρων ασθενειών (Neto et al., 2014).

3.5.1 Ενσωμάτωση της Κειμενικής Εξόρυξης στην Ανάλυση Δεδομένων Υψηλής Απόδοσης (High-throughput Data Analysis)

Οι βιολογικές διεργασίες εκ φύσεως επηρεάζονται απ' τις σύνθετες αλληλεπιδράσεις μεταξύ βιομορίων. Η ενσωματωμένη υψηλών αποδόσεων ομική μελέτη, επιτρέπει διεισδυτική και πολυεπίπεδη θέαση των κυττάρων, οργανισμών ή αποικιών. Με την έλευση των μετά-γονομικών τεχνολογιών, οι ομικές μελέτες γίνονται όλο και πιο κυρίαρχες στο χώρο της συστηματικής μελέτης. Το πλήρες όφελος από τις μεθοδολογίες αυτές δεν έχει ακόμα προσδιοριστεί, όσο εκκρεμεί ο εναρμονισμός των δεδομένων, ο διαμοιρασμός τους, η μετά-ανάλυση και η δόμησή τους ώστε να υποστηρίζεται η

ενσωματωμένη αναζήτηση. Αυτά τα καθοριστικά βήματα απαιτούν δημιουργία, συλλογή και αναδιανομή μετά-δεδομένων (Kolker et al., 2014).

Στα πλαίσια της εργασίας (Natarajan et al., 2006) συνδυάστηκε επιτυχώς η εξόρυξη αντικειμένων από πλήρες κείμενο με την ανάλυση γονιδιωματικών δεδομένων, με στόχο να αποκαλύψει την επίδραση της φωσφορικής σφιγγοσίνης I (SIP), μιας διέγερσης λυσοφωσφολιπιδίων, που εμπλέκεται στην απόπτωση, τον πολλαπλασιασμό και τη μετανάστευση των κυττάρων, στο επεμβατικό ανθρώπινο γλοιοβλάστωμα και την καταστολή των αλλεπάλληλων γεγονότων. Τα πλήρη άρθρα των δημοσιεύσεων ανακτήθηκαν και επεξεργάστηκαν απ' το GetItRight. Τα τύπου HTML αρχεία που προέκυψαν μετατράπηκαν σε τύπου XML. Οι βιολογικές οντότητες (γονίδια και πρωτεΐνες) καθώς και οι συσχετισμοί τους (ενεργοποίηση, καταστολή ή άλλου είδους ρύθμιση) εξήχθησαν απ' το LexiQuest Mine (SPSS, Chicago, IL). Τα μοτίβα που προέκυψαν αποθηκεύτηκαν εκ νέου σε αποθήκες κειμενικής εξόρυξης, για να αποτελέσουν έτοιμη γνώση στο μέλλον. Ενώ τα αποτελέσματα διασταυρώθηκαν με άλλα πειράματα γονιδιακής έκφρασης (**Εικόνα 26**).



Εικόνα 26. *Σχηματικό διάγραμμα της διαδικασίας που ακολουθήθηκε στην ανάλυση των δεδομένων από την εξόρυξη κειμένου.*

(Η εικόνα αποτελεί μέρος της εργασίας: Natarajan et al., 2006)

Αρχικά, ταυτοποιήθηκαν ένα σύνολο 72 διαφορετικά εκπεφρασμένων γονιδίων από την ανάλυση δεδομένων μικροσυστοιχιών ως μοναδική

απάντηση στο SIP, συγκρίνοντάς τα με τα προφίλ έκφρασης υπό την επίδραση του επιδερμικού αυξητικού παράγοντα (epidermal growth factor, EGF). Αυτό το σύνολο γονιδίων στη συνέχεια χρησιμοποιήθηκε για να εξαγάγει τα δίκτυα αλληλεπίδρασης μεταξύ γονιδίων, που εξήχθησαν από εξόρυξη σε πλήρη άρθρα 20 δημοφιλών επιστημονικών περιοδικών στον τομέα της έρευνας για τον καρκίνο σε μια πενταετία (1999-2003). Η μεθόδος αφορά στην Επεξεργασία Φυσικής Γλώσσας (Natural Language Proccess, NLP). Μεταξύ των προερχόμενων δικτύων αλληλεπίδρασης γονιδίων, έχουν χαρτογραφήσει ένα ιδιαίτερα ενδιαφέρον δίκτυο, που ενεργοποιείται από το SIP, όπου ο πίνακας της μεταλλοπρωτεϊνάσης-9 (MMP-9), αναγνωρίστηκε ως βασικός παίκτης στα επεμβατικά γλοιοβλαστώματα.

Παρομοίως, άλλη ομάδα εφάρμοσε την προσέγγιση συνδυασμένης βιβλιογραφικής εξόρυξης και ανάλυσης μικρο-συστοιχιών (Literature Mining and Microarray Analysis, LMMA), για να κατασκευάσει ένα δίκτυο-στόχο, που θα απεικονίζει την αγγειογένεση, μια διαδικασία δημιουργίας νέων τριχοειδών αιμοφόρων αγγείων για ένα θεμελιώδες βήμα στη μετάβαση όγκων από αδρανή κατάσταση σε κακοήθη κατάσταση (Li et al., 2006). Αυτή η προσέγγιση είναι ιδιαίτερα ενδιαφέρουσα διότι έχει συνοψίσει και ενσωματώσει συστηματικά μεγάλες ποσότητες σχετικών στοιχείων βιβλιογραφίας και δεδομένων μικρο-συστοιχιών. Πρώτα, συνέλεξαν όλες τις σχετικές δημοσιεύσεις του PubMed χρησιμοποιώντας την αγγειογένεση ως λέξη-κλειδί, από την οποία τα 1929 γονίδια με σύμβολα HUGO και 9514 παραπομπές ανακτήθηκαν για να κατασκευαστεί ένα δίκτυο γονιδίων, που εκφράζονται ταυτόχρονα κατά την αγγειογένεση. Στη συνέχεια, τα προφίλ έκφρασης γονιδίων σχετιζόμενα με την αγγειογένεση των ενδοθηλιακών κυττάρων (Endothelial Cells, EC) και των συμπαγών όγκων (Solid Tumors, ST) συλλέχθηκαν από τη βάση δεδομένων του Stanford (Stanford Microarray Database, SMD).

Περαιτέρω, το δίκτυο αγγειογένεσης βασισμένο στη βιβλιογραφία, διυλίστηκε με βάση τα προφίλ έκφρασης γονιδίων, που ανακτήθηκαν μέσω μίας της διαδικασίας μεταβλητής επιλογής, βάσει της υπόθεσης ότι: Τα ζευγάρια γονιδίων που συναντώνται σε κοινή βιβλιογραφία, θα αλληλεπιδρούν πράγματι μεταξύ τους στον φυσικό κόσμο. Παρότι, η υπόθεση παρουσιάζει κενά, δίνει πολύ καλά αποτελέσματα στα πλαίσια αυτής της νέας επιστήμης της Εξόρυξης. Τέλος, προέκυψε ένα βελτιωμένο δίκτυο αγγειογένεσης, στο οποίο πολλά

κομβικά γονίδια θα μπορούσαν να χρησιμοποιηθούν ως στόχοι για την αναστολή της αγγειογένεσης του όγκου, όπως ο παράγοντας νέκρωσης όγκων (Tumor Necrosis Factor, TNF), η ιντερλευκίνη (IL) -1, -6 και ο αγγειακός ενδοθηλιακός παράγοντας ανάπτυξης (Vascular Endothelial Growth Factor, VEGF).

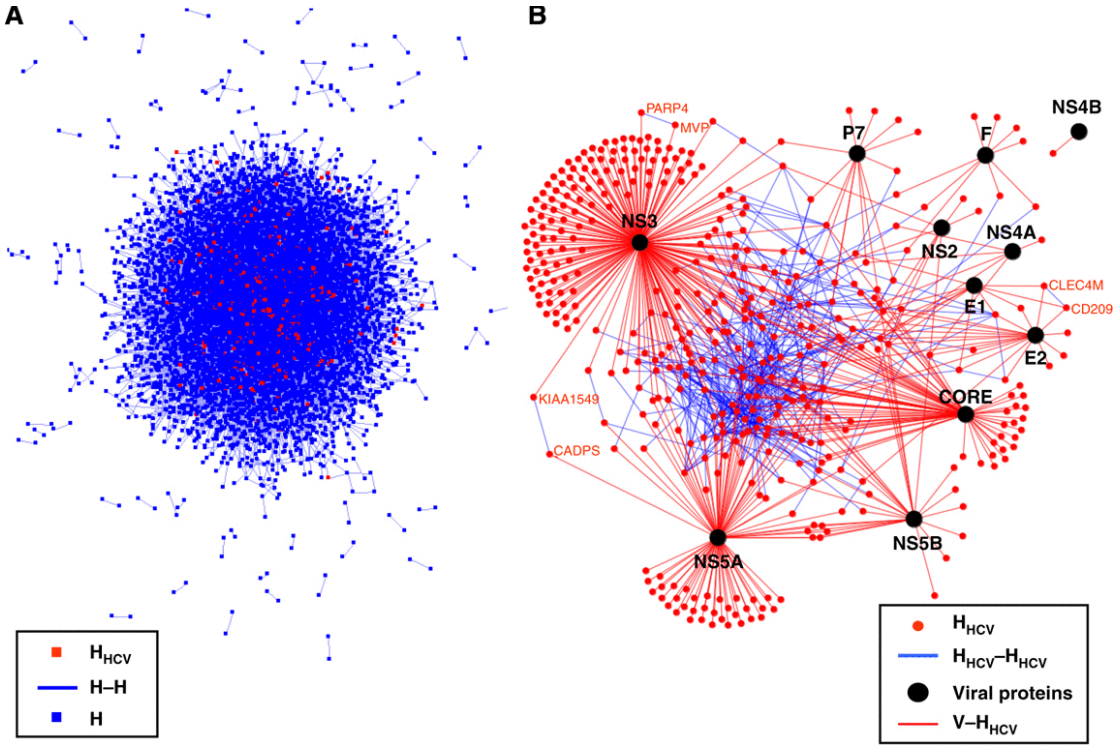
Στα πλαίσια της εργασίας του Kolker (Kolker et al., 2014), προκειμένου να βεβαιωθεί η διαφάνεια επί των δεδομένων, η ικανότητα να μη χαθεί η πληροφορία κατά τη σύνθεση υπερδομών, επιβλήθηκεο εναρμονισμός των δεδομένων και η μεγιστοποίησητης αναπαραγωγικής και χρηστικής ικανότητας μιας μελέτης στα πλαίσια των επιστημών υγείας, προτείνοντας μια λίστα επιβεβαίωσης με κοινά ομικά μετά-δεδομένα. Η λίστα μπορεί να χρησιμεύσει σαν μέτρο για την καθοδήγηση του πειραματικού σχεδιασμού. Ακόμα μπορεί να καταδεικνύει σημαντικές παραμέτρους και να αποτελεί κανόνα για δημοσιεύσεις με ασυσχέτιστα αποτελέσματα. Στόχος είναι η ισχύς της λίστας να επιβεβαιώνεται συνεχώς καθώς και να βελτιστοποιείται από σύγκριση με την σύγχρονη βιβλιογραφία.

3.5.2 Ενσωμάτωση Εξόρυξης με Βιβλιοθήκες Μονοπατιών

Ως εξαιρετικό παράδειγμα, κατασκευάστηκε επιτυχώς ένα δίκτυο-στόχων με κόμβους πρωτεΐνες, που σχετίζονται με την λοίμωξη από τον ιό της ηπατίτιδας C (HCV). Στο δίκτυο συμπεριλήφθηκαν και τα αποτελέσματα από μια διαδικασία διερευνήσεων διαπρωτεϊνικής αλληλεπίδρασης της βιβλιογραφίας και οκτώ επιλεγμένων βιβλιοθηκών αλληλεπίδρασης, συμπεριλαμβανομένων των BIND, BioGRID, DIP, GeneRIF, HPRD, IntAct, MINT και Reactome (De Chassey et al., 2008). Πρώτον, ταυτοποιήθηκαν 314 αλληλεπιδράσεις μεταξύ των HCV και ανθρώπινων πρωτεϊνών με πειράματα τύπου yeast two-hybrid και 170 με εξόρυξη κειμένου.

Για την προσέγγιση κειμενικής εξόρυξης ανακτήθηκαν όλες οι περιλήψεις που σχετίζονται με τις αλληλεπιδράσεις HCV και τις πρωτεΐνες των λέξεων-κλειδιών, όπως στην γραφική αναπαράσταση του δικτύου H-H, μεταξύ ιικών και ανθρώπινων πρωτεϊνών. Κάθε κόμβος συμβολίζει μια πρωτεΐνη και κάθε ακμή μια αλληλεπίδραση. Οι κόμβοι είναι χρωματισμένοι κόκκινοι αν αποτελούν

πρωτεΐνες της οικογένειας του HCV–human, ενώ οι μπλε ανθρώπινες πρωτεΐνες (**Εικόνα 27 Α**).



Εικόνα 27. *Γραφική αναπαράσταση του δικτύου αλληλεπίδρασής HCV–human (Hepatitis C virus, που προσβάλει τον άνθρωπο).*
(Η εικόνα αποτελεί μέρος της δημοσίευσης: De Chassey et al., 2008)

Αυτές οι διαπρωτεϊνικές αλληλεπιδράσεις χρησιμοποιήθηκαν ως σπόροι δεδομένα (input seeds) και ενσωματώθηκαν στις οκτώ βιβλιοθήκες,με σκοπό την ανασύσταση ενός δικτύου αλληλεπίδρασης HCV. Οι πρωτεΐνες CORE, NS3 και NS5A, ταυτοποιήθηκαν ως κύριοι στόχοι-παραγωγοί αντί-ιικών μορίων καθώς η θέση τους στο δίκτυο ήταν μεταξύ των κεντρικών ιικών και των ανθρώπινων πρωτεϊνών. Με μαύρο χρωματίζονται οι ιικές πρωτεΐνες, με κόκκινο οι ανθρώπινες. Οι ακμές που χρωματίζονται με κόκκινο συμβολίζουν την αλληλεπίδραση μεταξύ ιικών και ανθρωπίνων, ενώ με μπλε οι η αλληλεπιδράσεις μεταξύ των ανθρωπίνων πρωτεϊνών. Το μεγαλύτερο δομικό στοιχείο που αποτελείται από 196 πρωτεΐνες αναπαριστάται στο κέντρο (**Εικόνα 27 Β**). Εκτός από τις επιμέρους βιβλιοθήκες, χρησιμοποιήθηκαν

εντατικά και πειραματικές βάσεις δεδομένων αλληλεπίδρασης, για τον εντοπισμό πιθανών στόχων.

Πρόσφατα, επιτεύχθηκε η κατασκευή ενός δικτύου στόχου για το αντικαρκινικό φάρμακο Ganoderic Acid D (GAD), ένα σημαντικό συστατικό σε παραδοσιακό κινεζικό φυτικό φάρμακο. Αυτό συνέβει την ανάλυση των πρωτεομικών δεδομένων, με εξόρυξη επί της βάσης δεδομένων UniHI, μιας πειραματικής βάσης δια-πρωτεϊνικών αλληλεπιδράσεων ο Yue (Yue et al., 2008). Εν συντομία, 21 διαφορικά εκφρασμένες πρωτεΐνες αναγνωρίστηκαν αρχικά ως κυτταρικοί στόχοι της GAD, μέσω ανάλυσης πρωτεομικών δεδομένων. Αυτές οι 21 πρωτεΐνες στη συνέχεια χρησιμοποιήθηκαν ως γονίδια εκκίνησης για να αλιεύσουν αλληλεπιδρούσες πρωτεΐνες στη UniHI. Η επαναληπτική έρευνα τέτοιων πρωτεϊνών-εταίρων οδήγησε σε ένα διευρυμένο δίκτυο, που περιλαμβάνει και τις 21 πρωτεΐνες, γνωστές απ’ τις πειραματικές μελέτες. Τέλος, έχουν αναγνωρίσει την οικογένεια πρωτεϊνών 14-3-3, ως μείζονα παράγοντα στον μηχανισμό κυτταροτοξικότητας του GAD.

Σχετικά με τον καρκίνο του παχέος εντέρου (CRC, colorectal cancer) έχουν ταυτοποιηθεί αρκετά ευπαθή γονίδια σε μεμονωμένα εργαστηριακά πειράματα. Δυστυχώς, η πληροφορία που καρποφόρησε δεν έχει συστηματοποιηθεί ώστε να αποτελεί μια ολότητα με ενιαία σύνταξη και ερμηνεία. Έτσι, κατασκευάστηκε η γονιδιακή βάση δεδομένων με το όνομα gbCRC. Μια πρώτη γονιδιακή βάση δεδομένων που συλλέγει τα γονίδια που σχετίζονται με τον καρκίνο του παχέος εντέρου. Η πληροφορίες της βιβλιοθήκης αυτής πηγάζουν απ’ την βιβλιογραφία και σχετικές δημοσιεύσεις. Τα χαρακτηριστικά της είναι τόσο ο μη-αυτοματοποιημένος χειρισμός των πειραματικά ενδεδειγμένων γονιδίων που αναφέρονται στην βιβλιογραφία, όσο και η καθολική διάδραση, μέσω αυτής της βιβλιοθήκης, πέντε μεγάλων βιβλιοθηκών με δεδομένα, όπως OMIM (Online Mendelian Inheritance in Man), GAD (The Genetic Association database), GeneRif, GWASCatalog και CRCGene database. Επίσης, βασικό χαρακτηριστικό είναι η δυνατότητα υπολογισμού των ρυθμιστικών μοτίβων που εμπλέκουν μεταγραφικούς παράγοντες, miRNA, και μεγάλα τμήματα μη εκφραζόμενων RNA. Συνολικά 2067 γονίδια συσχετίζονται με 2819 δημοσιεύσεις απ’ το PubMed (Zhao et al., 2016).

4. Σχόλια και Συμπεράσματα

Με την πληθώρα των βιοϊατρικών δεδομένων και πληροφοριών, που παράγονται από μια ποικιλία καινοτόμων τεχνολογιών, βρισκόμαστε στα όρια μιας συναρπαστικής εποχής ανακάλυψης διεργασιών, ιδιοτήτων, βιοδεικτών και φαρμάκων. Αναπόφευκτα, οι προσεγγίσεις Εξόρυξης θα αποτελέσουν την πρώτη φάση των μελλοντικών διόδων ανακάλυψης, βοηθώντας στην επιλογή κατάλληλων στόχων και στην καλύτερη κατανόηση των κυτταρικών μηχανισμών ή φαινοτύπων των ανθρώπινων λειτουργιών και ασθενειών. Πράγματι, η Εξόρυξη από Δεδομένων έχει ήδη εφαρμοστεί ευρέως για τον προσδιορισμό στόχων για θεραπευτική εφεύρεση και έγκαιρη διάγνωση.

Οι προσεγγίσεις συνίστανται σε Εξόρυξη Κειμένου, Εξόρυξη από δεδομένα Μικρο-Συστοιχιών και άλλες δύο αναδυόμενες προσεγγίσεις Εξόρυξης: Εξόρυξη Πρωτεομικών Δεδομένων και Εξόρυξη Χημικό-Γενετικών Δεδομένων. Ευτυχώς, ένας μεγάλος αριθμός βάσεων, που αποθηκεύουν μια ποικιλία δεδομένων, αξιόπιστα εργαλεία και μέθοδοι εξόρυξης βρίσκονται ήδη εν ενεργεία. Λόγω των εγγενών περιορισμών των διαφόρων προσεγγίσεων Εξόρυξης, ωστόσο, προτείνεται να εφαρμοστεί ένας συνδυασμός ή ενσωμάτωση διαφορετικών προσεγγίσεων εξόρυξης προκειμένου να ξεπεραστούν τα μειονεκτήματα μιας μεμονωμένης μεθόδου. Συνεπώς, οι μελλοντικές εργασίες θα πρέπει να κατευθύνονται προς την ανάπτυξη ολοκληρωμένων Βιβλιοθηκών, με ομοιομορφία δομής και λογισμικού, καθώς και εργαλείων φιλικών προς τους βιολόγους, που θα επιτρέψουν την επιτάχυνση της ανακάλυψης στόχων. Αυτό είναι δύσκολο γιατί οι βιολογικές λειτουργίες είναι εξαιρετικά πολύπλοκες διαδικασίες και τα δεδομένα είναι σε μεγάλο βαθμό ετερογενή και ακαθόριστα.

Η διαδικασία της Εξόρυξης ενώ τροφοδοτεί την Βιολογία με εξαιρετικά εργαλεία, εμφανίζει κάποια πρωτογενή προβλήματα. Για την αντιμετώπιση των οποίων θα χρειαστούν ριζικές αλλαγές, τόσο στην επιστημονική μεθοδολογία, όσο και στην αντίληψη των επιστημόνων. Το πιο κομβικό παράδειγμα είναι το πώς θα κατορθώσουν επιστήμονες από διαφορετικά πεδία να συνεργαστούν εποικοδομητικά.

Η σύνθεση μιας ερευνητικής κοινοπραξίας από ειδικούς, με συμπληρωματικές γνώσεις αλλά κοινά ενδιαφέροντα, είναι αρχικώς αποτελεσματικά φερέγγυα. Η διατμηματική ένωση ειδικών από διαφορετικές επιστήμες ή από διαφορετικούς τομείς της ίδιας επιστήμης, ιδίως υπό το πρίσμα της αλληλεξάρτησής τους, είναι μια ακόμα πρόκληση για την επίτευξη δύσκολων τεχνικά ερευνητικών διαδικασιών, αλλά και ευκαιρία έμπνευσης ή διεύρυνσης των ερεθισμάτων (Holzinger, 2011). Για παράδειγμα, η εφαρμογή των αρχών της συνεργασίας ανθρώπου και υπολογιστή (HCI), στις μελέτες που υπαγορεύονται απ' τα ήδη γνωστά δεδομένα, γνωστά και ως data-driven projects, έχει ολοένα και μεγαλύτερη συχνότητα. Τέτοιες μελέτες, που αναφορικά αποτελούν την επιτομή του Τέταρτου Παραδείγματος (Fourth Paradigm) μελέτης της Φύσης απ' τον άνθρωπο, έχουν μια ιδιαιτερότητα: η παρατήρηση, η οποία διεγείρει ερωτήματα και θεωρήσεις, οι οποίες με την σειρά τους θα μορφώσουν το πλαίσιο πειραματισμού, δεν πηγάζει απ' την ανθρώπινη, αλλά απ' την μηχανική νόηση. **Το υποκείμενο παρατήρησης δεν είναι ο άνθρωπος, αλλά ο υπολογιστής**, γεγονός ιστορικά πρωτοφανές. Στις βιοεπιστήμες, οι ερευνητές είναι ταυτοχρόνως παραγωγοί και τελικοί χρήστες δεδομένων, μηχανικοί της πληροφορίας και οι αναλυτές, που βοηθούν στην οργάνωση, στην συγχώνευση, στην οπτικοποίηση, ανάλυση και αξιολόγηση των δεδομένων. Επί παραδείγματι, στην «Βιολογία Συστημάτων» η συνύφανση των δύο λειτουργιών, παραγωγής και επεξεργασίας, μηχανικών και αναλυτών, έχει αποδεδειγμένα οδηγήσει στην βελτίωση τόσο των θεωρητικών μοντέλων, όσο και των πειραματικών αποτελεσμάτων. Σε σύνθετους τομείς όπως η Βιοϊατρική ή η Βιοπληροφορική, απαιτούνται ειδήμονες που κατανοούν το περιβάλλον του συστήματος, το πρόβλημα και το σύνολο των δεδομένων, οπότε τελικά το πλήρες πλαίσιο αναζήτησης (Fawcett, 2006).

Επίσης, λόγω της αναγκαιότητας του υπολογιστή τίθεται το ζήτημα κατά πόσο τα αποτελέσματα είναι δυνατόν να ερμηνευθούν. Όσο η ροή εργασιών Εξόρυξης διευρύνεται, πρέπει να βελτιωθούν οι μετρητικές μέθοδοι που χρησιμοποιούνται για την αξιολόγηση των αποτελεσμάτων. Πλέον δεν είναι επαρκής η εστίαση σε μετρικές απόδοσης, όπως η ROC (Fawcett, 2006), ακρίβειας και ανάκλησης, χρειάζεται κανείς να αναλογιστεί πως μη λειτουργικές απαιτήσεις μπορούν να ικανοποιηθούν, όπως η Δυνατότητα Ερμηνείας. Στη Βιολογία είναι απαραίτητο να αιτιολογούνται και διαπιστεύονται επιχειρήματα

της απόφασης, η Εξόρυξη από μόνη της είναι ανεπαρκής. Το δόγμα είναι πως τα αποτελέσματα θα πρέπει να είναι κατανοητά και από άλλους (Hirsh, 2008). Πώς ποιοτικά και ποσοτικά μπορεί να αξιολογηθεί η δυνατότητα ερμηνείας; Σε ομοιότητα με τις έννοιες του ενδιαφέροντος ή της ομορφιάς, η ερμηνευτική δυνατότητα ίσως αποτελεί υποκειμενική μετρική του παρατηρητή και όχι ιδιότητα του αποτελέσματος. Μέχρι στιγμής, η ερμηνεία είναι κάτι που θεωρείται ότι βασίζεται σε προηγούμενη γνώση, στον βαθμό εξειδίκευσης του ερευνητή ή γενικότερα, του υποκειμένου παρατήρησης, οπότε απαιτείται, κατ' ελάχιστον, προσαρμοστικά εργαλεία για την ικανοποίηση αρχαρίων και ειδικών (Holzinger et al., 2009).

Πέρα από αυτά τα πιο λεπτά ζητήματα υπάρχει και το απλό τεχνικό της υπολογιστικής ισχύος. Καθώς η μηχανική των υπολογιστών εξελίσσεται, είναι αναγκαίο να βελτιστοποιηθούν οι αλγόριθμοι επεξεργασίας δεδομένων και η ροή εργασίας (Workflow), ώστε να συνεργάζονται καλύτερα με το ανεπτυγμένο υπολογιστικό σύστημα (hardware). Η ανοιχτή επικοινωνία συνδεδεμένων υπολογιστών, χρησιμοποιείται ήδη στην έρευνα (Garg et al., 2011). Παρόλα αυτά, μεγάλο μέρος του όγκου δεδομένων τις βιοϊατρικής υπακούν σε αυστηρούς κανονισμούς προστασίας, ασφάλειας και ιδιωτικότητας. Οι μεγάλες διαδικτυακές εταιρίες, προσφέρουν ήδη υπηρεσίες για την εντατική επεξεργασία δεδομένων, και παρόμοιες πολιτικές οδηγούν στην ανάπτυξη μεγάλων υπολογιστικών συμπλεγμάτων, που προορίζονται για μαζική ανάλυση δεδομένων (http://www.worldcommunitygrid.org).

Το νεαρόν της ηλικίας της Εξόρυξης από Δεδομένα αποτελεί, προς το παρόν, εμπόδιο στην καθορισμό Δεικτών Αξιολόγησης (Benchmarks). Προκειμένου να αξιολογηθείη ικανότητα των μεθόδων Εξόρυξης, η σύγκριση με αποτελέσματα άλλων μεθόδων που αντιμετωπίζουν το ίδιο πρόβλημα είναι καθοριστικής σημασίας. Οι Δείκτες αυτοί επιτρέπουν να συγκριθούν αποτελέσματα ανταγωνιστικών μεθόδων και έτσι να βαθμονομηθείη ικανότητα παραγωγής πληροφορίας της μεθόδου. Ακόμα όμως λίγα είναι αποδεκτά και διαδεδομένα (Hall & Holmes, 2003).

Τέλος, μεγάλη προσπάθεια γίνεται για να σταθμιστούν οι απαραίτητοι παράγοντες ώστε οι μέθοδοι Εξόρυξης να είναι παγκοσμίως αναπαράξιμοι. Πολύ διαδεδομένο πρόβλημα μεταξύ των μοντέρνων επιστημών, συμπεριλαμβανομένης και την Εξόρυξης από Δεδομένα. Συχνά δεν είναι καν

δυνατόν να επαληθευτεί ή να επαναληφθεί η πειραματική διαδικασία, διότι η διαδικασία ή οι αλγόριθμοι αποτελούν πνευματική ιδιοκτησία και δεν μπορούν να είναι ανοιχτοί στο ευρύ κοινό. Αυτό είναι ένα απ’ τα βασικότερα ζητήματα με όλο και αυξανόμενη αναφορά στις δημοσιεύσεις (Pastrello et al., 2014). Έτσι, η μεγαλύτερη πρόκληση παραμένει, το πως θα υπάρξει η δυνατότητα παραγωγής αντικειμενικών αποτελεσμάτων, στα οποία να μπορεί να συμφωνεί οποιοσδήποτε, οπουδήποτε.

Βιβλιογραφία

Δημοσιεύσεις:

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining association rules. In Proc. 20th int. conf. very large data bases, VLDB (Vol. 1215, pp. 487-499).

Ananiadou, S., Kell, D. B., & Tsujii, J. (2006). Text mining and its potential applications in systems biology. Trends in Biotechnology, 24(12), 571–579.

Baker CJ, Butler G, Jurisica I: Data Integration in the Life Sciences: 9th International Conference, DILS 2013, Montreal, Canada, July 11-12, 2013, Proceedings. Springer Publishing Company, Incorporated; 2013.

Bleiholder J, Naumann F: Data fusion. ACM Computing Surveys (CSUR) 2008, 41(1):1.

Campagne, F., & Skrabanek, L. (2006). Mining expressed sequence tags identifies cancer markers of clinical interest. BMC bioinformatics, 7(1), 481.

Carpenter, G. A. (1989). Neural network models for pattern recognition and associative memory. Neural networks, 2(4), 243-257.

Catchpoole, D. R., Kennedy, P., Skillicorn, D. B., & Simoff, S. (2010). The curse of dimensionality: a blessing to personalized medicine. Journal of Clinical Oncology, 28(34), e723-e724.

Charaniya, S., Hu, W. S., & Karypis, G. (2008). Mining bioprocess data: opportunities and challenges. Trends in Biotechnology.

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., & Wishart, D. S. (2008). PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic acids research, 36(suppl_2), W399-W405.

Cohen, K. B., & Hunter, L. (2008). Getting started in text mining. PLoS computational biology, 4(1), e20.

D'haeseleer, P. (2005). How does gene expression clustering work?. Nature biotechnology, 23(12), 1499-1501.

De Chassey, B., Navratil, V., Tafforeau, L., Hiet, M. S., Aublin-Gex, A., Agaugue, S., ... & Le Breton, M. (2008). Hepatitis C virus infection protein network. Molecular systems biology, 4(1), 230.

Dehmer M, Mowshowitz A: A history of graph entropy measures. Inf Sci 2011, 181(1):57-78.

Desany, B., & Zhang, Z. (2004). Bioinformatics and cancer target discovery. Drug discovery today, 9(18), 795-802.

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nature protocols, 4(8), 1184-1191.

Elloumi, M., & Zomaya, A. Y. (2013). Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data (Vol. 23). John Wiley & Sons.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. ACM SIGKDD explorations newsletter, 4(1), 65-75.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), 861-874.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996a). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 39(11), 27-34.

Fernández-Suárez, X. M., & Birney, E. (2008). Advanced Genomic Data Mining. PLoS Computational Biology, 4(9), 1–7.

Garg, V., Arora, S., & Gupta, C. (2011). Cloud computing approaches to accelerate drug discovery value chain. Combinatorial chemistry & high throughput screening, 14(10), 861-871.

Gerling, I. C., Singh, S., Lenchik, N. I., Marshall, D. R., & Wu, J. (2006). New data analysis and mining approaches identify unique proteome and transcriptome markers of susceptibility to autoimmune diabetes. Molecular & Cellular Proteomics, 5(2), 293-305.

Giudici, P., & Figini, S. (2009). Market basket analysis. Applied Data Mining for Business and Industry, Second Edition, 175-191.

Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data engineering, 15(6), 1437-1447.

Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. Data mining and knowledge discovery, 2(1), 9-37.

Hirschman, L., Park, J. C., Tsujii, J., Wong, L., & Wu, C. H. (2002). Accomplishments and challenges in literature data mining for biology. Bioinformatics, 18(12), 1553-1561.

Hirsh, H. (2008). Data mining research: Current status and future opportunities. Statistical Analysis and Data Mining: The ASA Data Science Journal, 1(2), 104-107.

Holzinger A, Simonic KM, Yildirim P (2012). Disease-Disease Relationships for Rheumatic Diseases: Web-Based Biomedical Textmining an Knowledge Discovery to Assist Medical Decision Making. IEEE 36th Annual Computer Software and Applications Conference (COMPSAC): 16-20 July 2012 2012; Izmir, Turkey 573-580.

Holzinger, A. (2011). Successful management of research & development. BoD–Books on Demand.

Holzinger, A. (2013, September). Human-Computer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together?. In International Conference on Availability, Reliability, and Security (pp. 319-328). Springer, Berlin, Heidelberg.

Holzinger, A., Dehmer, M., & Jurisica, I. (2014). Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions. BMC bioinformatics, 15(6), I1.

Holzinger, A., Kickmeier-Rust, M. D., Wassertheurer, S., & Hessinger, M. (2009). Learning performance with interactive simulations in medical education: Lessons learned from results of learning complex physiological models with the HAEMOdynamics SIMulator. Computers & Education, 52(2), 292-301.

Holzinger, A., Scherer, R., Seeber, M., Wagner, J., & Müller-Putz, G. (2012). Computational sensemaking on examples of knowledge discovery from neuroscience data: towards enhancing stroke rehabilitation. Information Technology in Bio-and Medical Informatics, 166-168.

Hu, Y., Hines, L. M., Weng, H., Zuo, D., Rivera, M., Richardson, A., & LaBaer, J. (2003). Analysis of genomic and proteomic data using advanced literature mining. Journal of proteome research, 2(4), 405-412.

Huang, Z. X., Tian, H. Y., Hu, Z. F., Zhou, Y. B., Zhao, J., & Yao, K. T. (2008). GenCLiP: a software program for clustering gene lists by literature profiling and constructing gene co-occurrence networks related to custom keywords. BMC bioinformatics, 9(1), 308.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern recognition letters, 31(8), 651-666.

Jarke, M., Jeusfeld, M. A., Quix, C., & Vassiliadis, P. (2013). Architecture and Quality in Data Warehouses. In Seminal Contributions to Information Systems

Engineering (pp. 161-181). Springer Berlin Heidelberg.

Jeanquartier, F., & Holzinger, A. (2013, September). On visual analytics and evaluation in cell physiology: a case study. In International Conference on Availability, Reliability, and Security (pp. 495-502). Springer, Berlin, Heidelberg.

Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. Nature reviews. Genetics, 7(2), 119.

Jurisica, I., Mylopoulos, J., Glasgow, J., Shapiro, H., & Casper, R. F. (1998). Case-based reasoning in IVF: Prediction and knowledge mining. Artificial intelligence in medicine, 12(1), 1-24.

Kim, B., Lee, H. J., Choi, H. Y., Shin, Y., Nam, S., Seo, G., ... & Kim, J. (2007). Clinical validity of the lung cancer biomarkers identified by bioinformatics analysis of public expression data. Cancer research, 67(15), 7431-7438.

Kolker, E., Özdemir, Martens Lennart, Hancock William, Anderson Gordon, Anderson Nathaniel, Aynacioglu Sukru, Baranova Ancha, Campagna Shawn R., Chen Rui, Choiniere John, Dearth Stephen P., Feng Wu-Chun, Ferguson Lynnette, Fox Geoffrey, Frishman Dmitrij, Grossman Robert, Heath Allison, Higdon Roger, Hutz Mara H., Janko Imre, Jiang Lihua, Joshi Sanjay, Kel Alexander, Kemnitz Joseph W., Kohane Isaac S., Kolker Natali, Lancet Doron, Lee Elaine, Li Weizhong, Lisitsa Andrey, Llerena Adrian, MacNealy-Koch Courtney, Marshall Jean-Claude, Masuzzo Paola, May Amanda, Mias George, Monroe Matthew, Montague Elizabeth, Mooney Sean, Nesvizhskii Alexey, Noronha Santosh, Omenn Gilbert, Rajasimha Harsha, Ramamoorthy Preveen, Sheehan Jerry, Smarr Larry, Smith Charles V., Smith Todd, Snyder Michael, Rapole Srikanth, Srivastava Sanjeeva, Stanberry Larissa, Stewart Elizabeth, Toppo Stefano, Uetz Peter, Verheggen Kenneth, Voy Brynn H., Warnich Louise, Wilhelm Steven W. & Yandl Gregory (2014). Toward more transparent and reproducible omics studies through a common metadata checklist and data publications. Omics: a journal of integrative biology, 18(1), 10-14.

Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., & Rzhetsky, A. (2004). Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. Proceedings of the National Academy of Sciences of the United States of America, 101(42), 15148-15153.

Kreuzthaler, M., Bloice, M., Simonic, K. M., & Holzinger, A. (2011). Navigating through very large sets of medical records: an information retrieval evaluation architecture for non-standardized text. Information Quality in e-Health, 455-470.

Kumar, K., & Abhishek, B. (2012). Artificial neural networks for diagnosis of kidney stones disease.

Kwon, H. J. (2006). Discovery of new small molecules and targets towards angiogenesis via chemical genomics approach. Current drug targets, 7(4), 397-405.

Lee, M. L., Lu, H., Ling, T. W., & Ko, Y. T. (1999, August). Cleansing data for mining and warehousing. In DEXA (Vol. 99, pp. 751-760).

Li, S., Wu, L., & Zhang, Z. (2006). Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. Bioinformatics, 22(17), 2143-2150.

Liew, A. C., Yan, H., & Yang, M. (2005). Data mining for Bioinformatics. In Bioinformatics technologies (pp. 63-116). Springer Berlin Heidelberg.

Liu Lin, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, & Maggie Law (2012). Comparison of next-generation sequencing systems. BioMed Research International, 2012.

Mamitsuka, H. (2013). Data mining for systems biology. New York, NJ: Humana Press.

Manning, C., Raghavan, P., & Schütze, H. (2009). Introduction to information retrieval/Christopher D.

McShan, D. C., Rao, S., & Shah, I. (2003). PathMiner: predicting metabolic pathways by heuristic search. Bioinformatics (Oxford, England), 19(13), 1692–1698.

Mount, D. W., & Pandey, R. (2005). Using bioinformatics and genome analysis for new therapeutic interventions. Molecular cancer therapeutics, 4(10), 1636-1643.

Müller, H., & Freytag, J. C. (2005). Problems, methods, and challenges in comprehensive data cleansing. Professoren des Inst. Für Informatik.

Narayanan, R. (2007). Bioinformatics approaches to cancer gene discovery. Target Discovery and Validation Reviews and Protocols: Volume 1, Emerging Strategies for Targets and Biomarker Discovery, 13-31.

Natarajan Jeyakumar, Daniel Berrar, Werner DubitzkyEmail, Catherine Hack, Yonghong Zhang, Catherine DeSesa, James R. Van Brocklyn & Eric G. Bremer (2006). Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. BMC bioinformatics, 7(1), 373.

Neto, O. A., Tassy, O., Biancalana, V., Zanoteli, E., Pourquie, O., & Laporte, J. (2014). Integrative data mining highlights candidate genes for monogenic myopathies. PloS one, 9(10), e110888.

Özgür, A., Vu, T., Erkan, G., & Radev, D. R. (2008). Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics, 24(13), i277-i285.

Papalexakis, E. E., & Faloutsos, C. (2016). Unsupervised Tensor Mining for Big Data Practitioners. Big data, 4(3), 179-191.

Pastrello, C., Pasini, E., Kotlyar, M., Otasek, D., Wong, S., Sangrar, W., Rahmati S. & Jurisica, I. (2014). Integration, visualization and analysis of human interactome. Biochemical and biophysical research communications, 445(4), 757-773.

Perez-Iratxeta, C., Bork, P., & Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. Nature genetics, 31(3), 316.

Perry, A. S., Loftus, B., Moroose, R., Lynch, T. H., Hollywood, D., Watson, R. W. G., Woodson K. & Lawler, M. (2007). In silico mining identifies IGFBP3 as a novel target of methylation in prostate cancer. British journal of cancer, 96(10), 1587.

Phoebe Chen, Y. P., & Chen, F. (2008). Identifying targets for drug discovery using bioinformatics. Expert opinion on therapeutic targets, 12(4), 383-389.

Pospisil, P., Iyer, L. K., Adelstein, S. J., & Kassis, A. I. (2006). A combined approach to data mining of textual and structured data to identify cancer-related targets. BMC bioinformatics, 7(1), 354.

Pospisil, P., Wang, K., Al Aowad, A. F., Iyer, L. K., Adelstein, S. J., & Kassis, A. I. (2007). Computational modeling and experimental evaluation of a novel prodrug for targeting the extracellular space of prostate tumors. Cancer research, 67(5), 2197-2205.

Rebholz-Schuhmann, D., Kirsch, H., & Couto, F. (2005). Facts from text—is text mining ready to deliver?. PLoS biology, 3(2), e65.

Rhodes, D. R., & Chinnaiyan, A. M. (2004). Bioinformatics Strategies for Translating Genome-Wide Expression Analyses into Clinically Useful Cancer Markers. Annals of the New York Academy of Sciences, 1020(1), 32-40.

Ring, M., & Eskofier, B. M. (2015). Data mining in the US National Toxicology Program (NTP) database reveals a potential bias regarding liver tumors in rodents irrespective of the test agent. PloS one, 10(2), e0116488.

Ryu, B., Kim, D. S., DeLuca, A. M., & Alani, R. M. (2007). Comprehensive expression profiling of tumor cell lines identifies molecular signatures of melanoma progression. PloS one, 2(7), e594.

Shannon, C. E., & Weaver, W. (1949). The Mathematical Theory of

Communication (Champaign, IL. Urbana: University of Illinois Press.

Siepen, J. A., Selley, J. N., & Hubbard, S. J. (2008). PepSeeker: mining information from proteomic data. Functional Proteomics: Methods and Protocols, 319-332.

Su, G., Morris, J. H., Demchak, B., & Bader, G. D. (2014). Biological network exploration with cytoscape 3. *Current protocols in bioinformatics*, 8-13.

Tansley, S., & Tolle, K. M. (2009). The fourth paradigm: data-intensive scientific discovery (Vol. 1). T. Hey (Ed.). Redmond, WA: Microsoft research.

Tory, M., & Moller, T. (2004). Human factors in visualization research. IEEE transactions on visualization and computer graphics, 10(1), 72-84.

Troyanskaya, O. G. (2005). Putting microarrays in a context: integrated analysis of diverse biological data. Briefings in bioinformatics, 6(1), 34-43.

Uchiyama, T., & Arbib, M. A. (1994). Color image segmentation using competitive learning. IEEE Transactions on pattern analysis and machine intelligence, 16(12), 1197-1206.

Viceconti, M., Taddei, F., Montanari, L., Testi, D., Leardini, A., Clapworthy, G., & Jan, S. V. S. (2007). Multimod Data Manager: A tool for data fusion. Computer methods and programs in biomedicine, 87(2), 148-159.

Von Bertalanffy, L. (1968). General system theory. New York, 41973(1968), 40.

Weippl, E., Holzinger, A., & Tjoa, A. M. (2006). Security aspects of ubiquitous computing in health care. e & i Elektrotechnik und Informationstechnik, 123(4), 156-161.

Wiltgen, M., Holzinger, A., Groell, R., Wolf, G., & Habermann, W. (2006). Usability of image fusion: optimal opacification of vessels and squamous cell carcinoma in CT scans. e & i Elektrotechnik und Informationstechnik, 123(4), 144-147.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Wuster, A., & Babu, M. M. (2008). Chemogenomics and biotechnology. Trends in biotechnology, 26(5), 252-258.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.

Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." IEEE Transactions on neural networks 16.3 (2005): 645-678.

Yang, Y., Adelstein, S. J., & Kassis, A. I. (2009). Target discovery from data mining approaches. Drug discovery today, 14(3), 147-154.

Yang, Y., Iyer, L. K., Adelstein, S. J., & Kassis, A. I. (2008). Integrative genomic data mining for discovery of potential blood-borne biomarkers for early diagnosis of cancer. PloS one, 3(11), e3661.

Yue, Q. X., Cao, Z. W., Guan, S. H., Liu, X. H., Tao, L., Wu, W. Y., Yi-Xue Li, Peng-Yuan Yang, Liu X. & Guo, D. A. (2008). Proteomics characterization of the cytotoxicity mechanism of ganoderic acid D and computer-automated estimation of the possible drug target network. Molecular & Cellular Proteomics, 7(5), 949-961.

Zadeh, L. A. (1965). Fuzzy sets. Information and control, 8(3), 338-353.

Zafar, N., Mazumder, R., & Seto, D. (2001). Comparisons of gene colinearity in genomes using GeneOrder2. 0. Trends in biochemical sciences, 26(8), 514-516.

Zhang, M., Luo, H., Xi, Z., & Rogaeva, E. (2015). Drug repositioning for diabetes based on'omics' data mining. PloS one, 10(5).

Zhang, M., Schmitt-Ulms, G., Sato, C., Xi, Z., Zhang, Y., Zhou, Y., Peter St George-Hyslop & Rogaeva, E. (2016). Drug Repositioning for Alzheimer's Disease Based on Systematic 'omics' Data Mining. PloS one, 11(12),

e0168812.

Zhao, M., Liu, Y., Huang, F., & Qu, H. (2016). A gene browser of colorectal cancer with literature evidence and pre-computed regulatory information to identify key tumor suppressors and oncogenes. Scientific reports, 6, 30624.

Αργυράκης, Π. (2001). Νευρωνικά Δίκτυα και Εφαρμογές. Ελληνικό Ανοικτό Πανεπιστήμιο, Σχολή θετικών Επιστημών και Τεχνολογίας. ISBN, 960-538.

Ιστοσελίδες:

Arthur M. Lesk (2009), Bioinformatics, (https://www.britannica.com/science/bioinformatics, τελευταία πρόσβαση στις 15/9/2017).

Fumihito Miyatake, Kazunori Iwabuchi (2005), Effect of high compost temperature on enzymatic activity and species diversity of culturable bacteria in cattle manure compost, Bioresource Technology, Volume 96, Issue 16, 2005, Pages 1821-1825, ISSN 0960-8524, (http://dx.doi.org/10.1016/j.biortech.2005.01.005, τελευταία πρόσβαση στις 20/10/2017).

Michael A. Nielsen (2015). Neural Networks and Deep Learning, (http://neuralnetworksanddeeplearning.com/, τελευταία πρόσβαση στις 11/11/2017).

Murtagh, F. (1983). A Survey of Recent Advances in Hierarchical Clustering Algorithms, (https://www.researchgate.net/profile/Fionn_Murtagh/publication/220459555_A_Survey_of_Recent_Advances_in_Hierarchical_Clustering_Algorithms/links/0046351425ae488419000000/A-Survey-of-Recent-Advances-in-Hierarchical-Clustering-Algorithms.pdf, τελευταία πρόσβαση στις 25/9/2017).

Rachel Nuwer (2017), What if the internet stopped working for a day?, (http://www.bbc.com/future/story/20170207-what-if-the-internet-stopped-for-a-

day, τελευταία πρόσβαση στις 12/9/2017)