Biological Big Data Mining

Λευτέρης Χατζηευφραιμίδης

Εξόρυξη Δεδομένων Μεγάλου Όγκου 2021-22
Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
Πανεπιστήμιο Θεσσαλίας, Βόλος
echatzief@e-ce.uth.gr

Περίληψη Σε αυτήν την αναφορά θα περιγράψουμε το ρόλο του data mining στο χώρο της βιολογίας και της ιατρικής. Τα δεδομένα στον χώρο της ιατρικής και βιολογίας αυξάνονται με ραγδαίους ρυθμούς και παρατηρούμε ότι πολλοί ερευνητές και ερευνητικά προγράμματα ασχολούνται με την συλλογή, την επεξεργασία και την χρήση αυτών των δεδομένων για την εξαγωγή συμπερασμάτων και την δημιουργία μοντέλων που βοηθούν την ανθρωπότητα. Αυτές οι ενεργειες είναι απόρροια της τεχνολογικής εξέλιξης που ωθεί τους επιστήμονες να ασχολούνται με την απάντηση ολοένα και περισσότερων ερωτημάτων μέσω πειραμάτων. Βέβαια, ακόμα υπάρχουν διάφορες δυσκολίες που εμποδίζουν την όλη διαδικασία και κάποιες από αυτές θα τις συναντήσουμε στην παρούσα αναφορά.

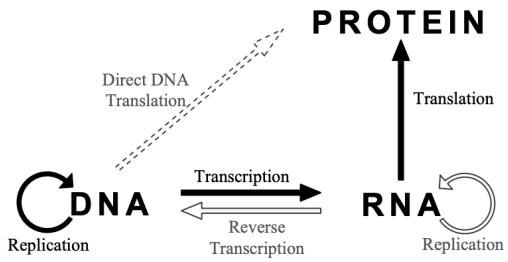
Λέξεις Κλειδιά: big data analytics, machine learning, bioinformatics, health care, biological databases, biology

1 Εισαγωγή στο θέμα

Η συλλογή και η ανάλυση δεδομένων είχε ξεκινήσει από πολύ παλιά ωστόσο δεν υπήρχαν τα απαραίτητα μέσα για να γίνεται όλη η διαδικασία πιο αυτοματοποιημένη και πιο γρήγορη. Ακόμα από τα αρχαία χρόνια ορισμένοι πολιτισμοί παρατηρούσαν και κατέγραφαν τα άστρα και τους πλανήτες. Με βάση αυτές τις παρατηρήσεις μπορούσαν να μετράνε το χρόνο και να προβλέπουν τις αλλαγές στις εποχές και στον καιρό.Οι προβλέψεις ηταν καθοριστικές γιατί επηρέαζαν πολλαπλές πτυχές της καθημερινότητάς τους όπως, η παραγωγή τροφής μέσω της καλλιέργειας της γης. Αργότερα αυτές οι παρατηρήσεις γίνονταν σε πιο μεγάλη κλίμακα και με πιο συνθέτες διαδικασίες που οδήγησαν σε πολλές από τις πιο γνωστές επιστημονικές επαναστάσεις, οι οποίες άλλαξαν καθοριστικά τις ζωές των ανθρώπων και όλη την κοινωνία και την καθημερινότητα.

Μαζί με όλες αυτές τις επαναστάσεις και τις μεταβολές τον 20ο αιώνα ήρθε και η ανακάλυψη του μοντέλου του DNA σαν μια αλληλουχία από 2 στήλες από τους Watson και Crick και αυτό εδραίωσε την μοριακή βιολογία σαν ένα κύριο πεδίο έρευνας της βιολογίας.Τραβώντας το ενδιαφέρον πολλών επιστμόνων και αλλάζοντας καθοριστικά τον τρόπο που αντιμετώπιζαν την βιολογία και την ιατρική μέχρι τότε. Βέβαια όλα αυτά δεν θα είχαν γίνει αν δεν υπήρχε διαρκή συλλογή και ανάλυση όλο και περισσότερων δεδομένων, μέσω σύνθετων τεχνικών ανάλυσης και εξαγωγής συμπερασμάτων με την βοήθεια ορισμένων εργαλείων και εφευρέσεων. Για παράδειγμα, αν δεν είχε ανακαλυφθεί το τηλεσκόπιο δεν θα είχαν παρατηρηθεί και δεν θα είχαν ανακαλυφθεί οι πλανήτες

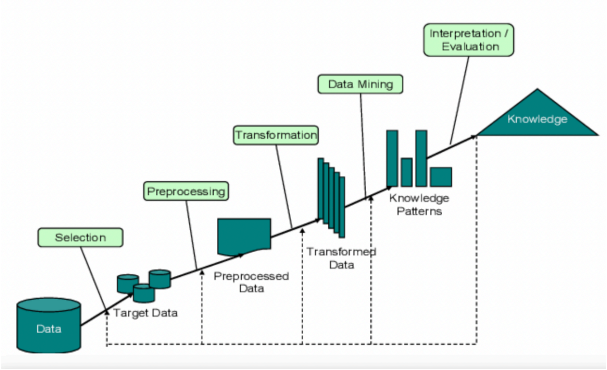
και τα ποικίλα αντικείμενα που υπάρχουν στο σύμπαν μας. Το ίδιο έγινε και στον τομέα της βιολογίας με την ανακάλυψη του μικροσκοπίου που συνέβαλε στην μελέτη μικροοργανισμών όπου υπάρχουν και στον άνθρωπο και στο γενικότερο περιβάλλον του. Βέβαια σήμερα με την σύγχρονη τεχνολογία και τα μέσα που έχουμε συλλέγουμε μέγαλου όγκου δεδομένα κατά την διάρκεια πειραμάτων που λαμβάνουν χώρα από επιστήμονες. Βέβαια πρώτα για να δούμε πως συνδέεται το data mining με το κομμάτι της βιολογίας θα πρέπει να εξηγήσουμε πώς λειτουργεί το βασικό δόγμα για τον κύκλο της βιολογικής πληροφορίας. Οι βασικές μεταφορές στους περισσότερους οργανισμούς περιγράφονται με το παρακάτω διάγραμμα.



Εικ. 1. Διαγράμμα μετασχηματισμού του DNA σε πρωτείνη.

Δηλαδή το DNA μεταγράφεται στό RNA και έτσι δημιουργείται η εκάστοτε πρωτείνη. Ο κυκλικός βρόχος γύρω από το DNA δηλώνει την επαναληπτικότητα της διαδικασίας. Πέρα από αυτή την διαδικασία υπάρχουν κάποιες επιπλέον μεταφορές που αναφέρονται σε εξειδικευμένους ιούς, όπως την μετατροπή του RNA στο DNA μέσω μολυσμένων κυττάρων και γενικότερα υπάρχουν αρκετοι τρόποι για να μεταφέρεται η πληροφορία από μορφή σε μορφή. Η εισαγωγή των νέων τεχνολογιών και τεχνικών, όπως είναι και το data mining μας οδήγησε να μελετάμε αυτές τις παραπάνω ενέργειες και να εξάγουμε μαζικά δεδομένα που θα μας κατευθύνουν σε νέους δρόμους και πεδία έρευνας.Όπως αναφέραμε, ένας από τους τρόπους συλλογής και ανάλυσης των δεδομένων είναι και το data mining που ήρθε να δώσει λύση στα προβλήματα που είχαν οι παλιότερες μέθοδοι ανάλυσης δεδομένων εξαιτίας του μεγάλου όγκου της πληροφορίας. Πρακτικά το data mining είναι το κύριο βήμα στην διαδικασία του Knowledge Discovery των βάσεων. Το Knowledge Discovery είναι μια ρουτίνα που διαχωρίζει τα χρήσιμα και έγκυρα δεδομένα από τα υπόλοιπα μέσω ανίχνευσης συγκεκριμένων μοτίβων. Αυτά το μοτίβα εντοπίζονται με την βοήθεια ποικίλων αλγορίθμων data mining που εφαρμόζονται και περιλαμβάνουν τα παρακάτω βήματα. Αρχικά, έχουμε το preprocessing των δεδομένων που περιλαμβάνει τον διαχωρισμό της ποσότητας των δεδομένων που είναι έγκυρα, τον καθαρισμό αυτών και το μετασχηματισμό τους ανάλογα με την περίπτωση που μελετάμε. Στην συνέχεια, ακολουθεί το data mining, όπου εφαρμόζονται ορισμένοι αλγόριθμοι κατηγοριοποίησης, συσταδοποίησης ή αναδρομής για να εξάγωγουμε πληροφορίες από τα δεδομένα του πρώτου βήματος. Τέλος, από την προηγούμενη διαδικασία παράγονται ορισμένα μοντέλα και μοτίβα που αξιολογούνται από

την δοκιμή τους σε νέα δεδομένα που αντλούμε. Τα παραπάνω βήματα περιγράφονται αναλυτικά στο παρακάτω διάγραμμα.



Εικ. 2. Στάδια Data Mining

2 Ανασκόπηση σχετικής βιβλιογραφίας και αιτιολόγηση επιλογής πηγών που θα χρησιμοποιηθούν στη συνέχεια

Μιας και το θέμα της βιολογίας σε συνδυασμό με το data mining είναι αρκετά επίκαιρο και υπάρχει αρκετή έρευνα γύρω από τις τεχνικές και τα εμπόδια που αντιμετωπίζουν οι ερευνητές στην πορεία τους να αναπτύξουν συστήματα με σκοπό να βελτιώσουν το data mining, θεώρησα σωστό να καλύψω και τις δύο πλευρές αυτού του θέματος. Δηλαδή, θα αναλύσω μια πηγή που μελετά τις αρκετές από τις μεθόδους και τις πρακτικές που χρησιμοποιούν για να εξάγουν πληροφορία από τις εφαρμογές της βιολογίας καθώς και τις δυσκολίες που υπάρχουν και εμποδίζουν τους ερευνητές πολλές φορές να φτάσουν σε συμπεράσματα. Η πρώτη πηγη που θα αναλύσουμε είναι μια δημοσίευση του Khalid Raza με τίτλο “Application of Data Mining in Bioinformatics” [1] που αναφέρεται στην συσχέτιση της βιοπληροφορικής και του data mining και αναλύει ορισμένες από τις εφαρμογές του data mining στο κομμάτι της βιοπληροφορικής. Στην συνέχεια, θα αναλύσουμε μια άλλη δημοσίευση που αφορά τις ευκαιρίες και τις δυσκολίες που έχουν οι ερευνητές όταν χρησιμοποιούν στην βιολογία το data mining, η οποία γράφτηκε από τους Yixue Li και Luoman Chen με τίτλο “Big Biological Data: Challenges and Opportunuties” [2]. Έτσι, θα έχουμε μια πλήρη εικόνα για το κομμάτι των biological data, τις εφαρμογές που έχει στην σημερινή κοινωνία και τι εμπόδια υπάρχουν που δυσχεραίνουν την εξέλιξη αυτού του τομέα.

3 Παρουσίαση 2-3 πηγών επί του θέματος της εργασίας

Στην πρώτη δημοσίευση αναφέρει ότι, οι γρήγορες εξελίξεις στην βιολογία οδήγησαν στην παραγωγή μεγάλου όγκου δεδομένων. Βέβαια για να εξάγουμε συμπεράσματα από αυτά τα δεδομένα χρειαζόμαστε αρκετή επεξεργασία και υπολογιστική ανάλυση. Για αυτό έχει αναπτυχθεί ένα νέο πεδίο έρευνας που αναλύει αυτά τα μεγάλα datasets και εξάγει λογικά ευρήματα από τα δεδομένα. Το πεδίο αυτό ονομάζεται βιοπληροφορικη και συνδυάζει την βιολογία και ένα κομμάτι της στατιστικής καθώς και παραδείγματα ανάλυσης αυτού του τομέα είναι η πρόβλεψη των δομών της πρωτεΐνης, την κατηγοριοποίηση των γονιδίων, την συσταδοποίηση των γονιδιακών συνδυασμών και αρκετές ακόμα προσεγγίσεις. Στην συνέχεια, εξηγεί τον όρο της βιοπληροφορικής, τις πτυχές που την απαρτίζουν και τους τομείς που έχει εφαρμογή

Μια από τις πιο γνωστές εφαρμογές ονομάζεται sequence analysis και στοχεύει στην εύρεση βιολογικών ακολουθιών, οι οποίες είναι όμοιες μεταξύ τους και άλλες που είναι διαφορετικές και σε ποια κομμάτια διαφέρουν. Η μέθοδος sequence analysis ξεκινάει με την υποβολή ενός DNA σε δοκιμές, όπως το sequence alignment και άλλες αναζητήσεις με σκοπό να βρούμε τις όμοιες ακολουθίες και τις διαφορετικές. Η συγκεκριμένη διαδικασία επιτυγχάνεται με το data mining χρησιμοποιώντας τεχνικές κατηγοριοποίησης και συσταδοποίησης. Άλλη μια επιπλέον αλλά εξίσου σημαντική εφαρμογή είναι η ανάλυση των γονιδιακών εκφράσεων. Γνωρίζουμε πως το mRNA και οι πρωτεΐνες μας δίνουν μια ξεκάθαρη εικόνα για την δραστηριότητα των γονιδίων και έτσι μέσω της ανάλυσης με διάφορους data mining αλγορίθμους μπορούμε να βρούμε συσχετίσεις πρωτεϊνών με τα αντίστοιχα γονίδια, τα οποία καθορίζουν σημαντικά την ζωή του κάθε ανθρώπου. Μια άλλη παρόμοια εφαρμογή είναι η ανίχνευση των μεταλλάξεων του καρκίνου, δηλαδή παρατηρώντας τα γονίδια του ατόμου μπορούμε να προβλέψουμε την καρκινογένεση. Η βιοπληροφορική έχει δημιουργήσει πολλαπλά συστήματα για να αναλύει τις ακολουθίες των γονιδίων και μέσω της ανάλυσής τους να εντοπίζει και να προβλέπει αν ένα άτομο θα εμφανίσει καρκίνο και αν θα υπάρξει κάποια μετάλλαξη του. Όλα αυτά δεν θα ήταν δυνατόν να επιτευχθούν αν δεν υπήρχε η τεχνική του data mining διότι θα υπήρχαν τα δεδομένα αλλά δεν θα υπήρχε δυνατότητα αξιοποίησης τους. Τέλος, ακόμη μια τεχνική που είναι ευρέως γνωστή είναι ανάλυση εικόνων για την εξαγωγή παρατηρήσεων και συμπερασμάτων, δηλαδή αναπτύσσονται συστήματα για ανάλυση εικόνας που αντικαθιστούν πλήρως έναν παρατηρητή. Ο σκοπός τους είναι να εκτελούν κλινικές αναλύσεις σε εικόνες, σε διαγράμματα DNA και διάφορα βιοιατρικά φαινόμενα. Αυτό βοηθά να εντοπίζονται αυτοματοποιημένα και να προβλέπονται ασθένειες σε κάθε περίπτωση ατόμου απλώς δίνοντας τις κατάλληλες εικόνες για ανάλυση. Το δεύτερο κομμάτι της δημοσίευσης αναφέρεται στις δυνατότητες και τις ενέργειες του data mining πάνω στο κομμάτι της βιοπληροφορικής που είναι η κατηγοριοποίηση, η πρόβλεψη, ο υπολογισμός, οι κανόνες συσχέτισης και η συσταδοποίηση.

Η δεύτερη δημοσίευση αναφέρεται στα εμπόδια και στις δυσκολίες που εμφανίζονται κατα την ανάλυση των δεδομένων με τις διάφορες τεχνικές του data mining. Μια από τις βασικές δυσκολίες είναι ο όγκος των δεδομένων. Με τις πληροφορίες για τους ανθρώπους, τα μικρόβια και όλα όσα αφορούν μια έρευνα αναπτύσσονται μεγάλα datasets με εκατομμύρια εγγραφές. Για παράδειγμα, μια έρευνα για την αναπαραγωγή των σκυλιών παρήγαγε ένα dataset με 3.2 terabytes δεδομένα. Για αυτό τον σκοπό το διαίρεσαν σε 12 sample datasets και τελικά κατάφεραν να πραγματοποιήσουν την ανά-

λυση που ήθελαν και εντόπισαν μια μετάλλαξη στα γονίδια. Σε αυτή την περίπτωση κατάφεραν να εξάγουν κάποιο αποτέλεσμα αλλά σε άλλες περιπτώσεις τα πολλά δεδομένα προκαλούν εμπόδιο στην ανάλυση, μιας και δεν υπάρχει δυνατότητα να βρούν το σημείο εστίασης μέσα στο μεγάλο όγκο δεδομένων. Άλλο ένα μείζον πρόβλημα των δεδομένων στο κομμάτι της βιολογίας είναι το γεγονός ότι είναι ετερογενή. Δηλαδή υπάρχουν πολλαπλα διαφορετικά μοτίβα και αρχές. Επομένως, για να ξεπεραστεί το εμπόδιο αυτό θα πρέπει η έρευνα να είναι hypothesis-driven, δηλαδή να εστιάζουμε σε μια αρχή που υπάρχει και να ερευνούμε τα δεδομένα με βάση αυτή, ώστε να υπάρχει κάποιος στόχος και όχι απλά να ψάχνουμε για γενικά μοτίβα. Αυτό ο τρόπος προσέγγισης μειώνει τον όγκο των δεδομένων και το χρόνο των υπολογισμών, άρα οδηγούμαστε πιο γρήγορα σε κάποιο αποτέλεσμα. Τέλος, αναφέρει ότι η φύση των δεδομένων και οι δυσκολίες διατήρησής τους έχει οδηγήσει στην δημιουργία δομών για την αποθήκευση και την επεξεργασία όπως βάσεις δεδομένων, εργαλεία ανάλυσης και cloud services, τα οποία διευκολύνουν την διαδικασία.

4 Σύγκριση του περιεχομένου / των αποτελεσμάτων των παραπάνω πηγών

Οι δυο δημοσιεύσεις αναφέρονται στις εφαρμογές του data mining στο τομέα της βιολογίας. Γενικά η ύπαρξη του data mining έχει βοηθήσει στην ανάλυση και στην εξαγωγή συμπερασμάτων που έχει ανοίξει πολλαπλούς ορίζοντες και έχει δώσει το έναυσμα σε νέες έρευνες, οι οποίες με τη σειρά τους έχουν βελτιώσει την καθημερινότητα των ανθρώπων. Δηλαδή μέσω τον τεχνικών που έχουν αναπτυχθεί προβλέπονται και υπολογίζονται μεταλλάξεις και ασθένειες και έτσι αυξάνεται το όριο ζωής και η ποιότητα της. Βέβαια για να φτάσει ως εδώ η έρευνα έχει περάσει πολλαπλά εμπόδια και δυσκολίες που κάποια από αυτά αντιμετωπίστηκαν. Τα συγκεκριμένα εμπόδια αναφέρονται και στο κομμάτι της δεύτερης δημοσίευσης αλλά σταδιακά με την βοήθεια της τεχνολογίας αντιμετωπίζονται και υπάρχει μεγαλύτερη εξέλιξη στο κομμάτι της βιολογίας με πολλαπλά θετικά αποτελέσματα στους ανθρώπους.

5 Συμπεράσματα και μελλοντικά σχέδια μελέτης

Όπως παρατηρήσαμε και στις δυο δημοσιεύσεις το θέμα του data mining προυπήρχε από πιο παλιά αλλά λόγω της αναπτυγμένης τεχνολογικής ανάπτυξης έχει εδραιωθεί στις σημερινές έρευνες και μελέτες που διεξάγονται. Προφανώς, η συμβολή του data mining είναι απαραίτητη μιας και διευκολύνει και λύνει λύση σε αναλύσεις μεγάλου όγκου δεδομένων και βοηθά στην διεξαγωγή συμπερασμάτων που κάποια από αυτά αλλάζουν ριζικά την καθημερινότητά μας. Έτσι και ο τομέας της βιοιατρικής και βιοπληροφορικής έχει ενσωματώσει το data mining σαν ένα από τα κυριότερα εργαλεία για να ανακαλύπτει καινούργια πράγματα και να εμβαθύνει σε έννοιες που σε άλλες περιπτώσεις θα ήταν αδύνατο. Συνεπώς, καλό θα ήταν να συνεχίσει να εξελίσσεται ο τομέας του data mining διότι με αυτόν θα βελτιωθούν πολλαπλοί ακόμη τομείς. Σαν μελλοντικά σχέδια μελέτης θα ήταν λογικά να κοιτάξουμε αναλυτικότερα πως θα μπορούσαμε να συνδυάσουμε την μηχανική μάθηση με το data mining, ώστε να δούμε πως

και τα δύο μαζί μπορούν να επηρεάσουν τον χώρο της βιολογίας και πώς θα μπορέσουν μαζί να βοηθήσουν στην εξέλιξή του.

Αναφορές

1. Khaliz Raza: Application of data mining in bioinformatics, Indian Journal of Computer Science and Engineering (2010)
2. Yixue Li, Luonan Chen : Big Biological Data: Challenges and Opportunities, Chinese Academy of Sciences (2014)
3. George Tzanis : Biological and Medical Big Data Mining, Aristotle University of Thessaloniki (2014)