

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Θέμα: Χρήση R-εργαλείου για την αντιστοίχιση γονιδίων με τους κλινικούς φαινοτύπους τους μέσω PubMed και HPO

Ονοματεπώνυμο Συγγραφέα: Άρτεμις Ντούσκου **ΑΜ:** 214

Επιβλέπων Καθηγητής: Ιωάννης Γεωργίου, Καθηγητής Ιατρικής Γενετικής και Υποβοηθούμενης Αναπαραγωγής



Ιωάννινα, 2026

Ευχαριστίες

Η παρούσα διπλωματική εργασία, με θέμα «Χρήση R-εργαλείου για την αντιστοίχιση γονιδίων με τους κλινικούς φαινοτύπους τους μέσω PubMed και HPO», πραγματοποιήθηκε στο Εργαστήριο Ιατρικής Γενετικής του Τμήματος Ιατρικής του Πανεπιστημίου Ιωαννίνων, εν όψει του προγράμματος μεταπτυχιακών σπουδών «Βασικές Βιοϊατρικές Επιστήμες» και κατεύθυνσης «Κυτταρογενετική».

Με αυτή την ευκαιρία θα ήθελα να εκφράσω την ευγνωμοσύνη μου και να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή της παρούσας εργασίας, τον καθηγητή κ. Ιωάννη Γεωργίου, για την εμπιστοσύνη που μου έδειξε από την αρχή της συνεργασίας μας, για τις πολύτιμες συμβουλές και την συνεχή καθοδήγησή του κατά την υλοποίηση αυτού του έργου, για την εξαιρετικά σημαντική και ταχεία βοήθεια που μου έδινε σε δύσκολα προβλήματα καθώς και για την εύθυμη και προσιτή ατμόσφαιρα που δημιουργούσε. Ο πολύτιμος χρόνος που αφιέρωσε αποτέλεσε πολύ σημαντική παράμετρο αυτής της εργασίας και η καθοδήγησή του υπήρξε καθοριστική τόσο για την ακαδημαϊκή και επαγγελματική μου πορεία όσο και για τη διαμόρφωση της σκέψης και της προσωπικότητάς μου.

Θα ήθελα επίσης να ευχαριστήσω την κα. Αικατερίνη Νάκα, διευθύντρια καθηγήτρια Καρδιολογίας και υπεύθυνη του παρόντος προγράμματος μεταπτυχιακών σπουδών, την ομότιμη καθηγήτρια κα. Πατρώνα Βεζυράκη και όλους τους διδάσκοντες και μη που έλαβαν μέρος στην υλοποίηση αυτού του προγράμματος, οι οποίοι με τη συμβολή και τη στήριξή τους κατέστησαν δυνατή την ολοκλήρωση αυτού του σημαντικού ακαδημαϊκού κύκλου, συμβάλλοντας ουσιαστικά στη διεύρυνση των γνώσεων και δεξιοτήτων μου και εν γένει της ακαδημαϊκής μου εξέλιξης.

Δεν θα μπορούσα να μην ευχαριστήσω εκ βάθους καρδίας, την υποψήφια διδάκτορα Αθανασία Σεσσέ, για την βοήθεια και στήριξη που μου προσέφερε, τον χρόνο που αφιέρωσε για την ένταξη και εξοικείωσή μου στο εργαστήριο σε συνδυασμό με το ευχάριστο, φιλικό κλίμα που επικρατούσε ανάμεσά μας.

Ευχαριστώ θερμά όλα τα μέλη του εργαστηρίου και μεταξύ αυτών την κα Ιωάννα Μπούμπα, ΕΔΙΠ Ιατρικής Γενετικής και τον κ. Χάρη Κωστούλα, ακαδημαϊκό υπότροφο και επιστημονικό συνεργάτη του εργαστηρίου, οι οποίοι ήταν πάντα πολύ συνεργάσιμοι και πρόθυμοι να βοηθήσουν.

Τέλος, ευχαριστώ ολόψυχα τις πολύ καλές μου φίλες Ευαγγελία Παπακωνσταντίνου, Ιωάννα Κιοσέ και Κατερίνα Τσιμπούλη, καθώς και την οικογένειά μου για την αγάπη τους, την εμψύχωσή τους, την στήριξή τους και την αμέριστη βοήθεια και συμπαράσταση που μου προσέφεραν με κάθε τρόπο καθ' όλη την διάρκεια των σπουδών μου.

Πίνακας Περιεχομένων

Περίληψη 8
Abstract 9
Εισαγωγή 10
1.1 Γονιδιακή πληροφορία και κλινικοί φαινότυποι 10
1.2 Ανάγκη αυτοματοποιημένης αντιστοίχισης γονιδίων-φαινοτύπων 10
1.3 Στόχοι και συνεισφορά της διπλωματικής 11
Θεωρητικό Υπόβαθρο 11
2.1 Βασικές έννοιες Μοριακής Γενετικής 11
2.1.1 Γονίδια και DNA 11
2.1.2 Μεταλλάξεις 12
2.1.3 Γονιδιακή έκφραση 12
2.2 Κλινικοί φαινότυποι και οντολογίες 12
2.2.1 Τι είναι φαινότυπος 12
2.2.2 Σχέση γονότυπου-φαινοτύπου 13
2.2.3 Συσχέτιση γονιδίων, ασθενειών και φαινοτύπων 13
2.2.4 Τι είναι οντολογία 13
2.2.5 Πλεονεκτήματα τυποποίησης 14
2.3 Human Phenotype Ontology (HPO) 14
2.3.1 Δομή 14
2.3.2 HPO IDs 14
2.3.3 Συσχετίσεις γονιδίων - φαινοτύπων 14
2.3.4 Χρήση στην κλινική γενετική 15
2.4 PubMed και βιοϊατρική βιβλιογραφία 15
2.4.1 Δομή και οργάνωση της PubMed 15
2.4.2 MeSH terms 15
2.4.3. Αναζήτηση και περιορισμοί 15
2.4.4 Πρόσβαση μέσω NCBI Entrez API 15
Τεχνολογίες και Εργαλεία 16

3.1 Γλώσσα R στη βιοπληροφορική 16
3.2 R πακέτα που χρησιμοποιήθηκαν 16
3.3 APIs και δεδομένα 17
Μεθοδολογία 18
4.1 Γενική αρχιτεκτονική συστήματος 18
4.2 Ροή εργασίας (Workflow) – Γενική περιγραφή και Αναπαράσταση 19
4.3 Εισαγωγή και Προεπεξεργασία δεδομένων γονιδίων 21
4.4 Ανάκτηση βιβλιογραφικών δεδομένων από την PubMed 21
4.5 Ανάκτηση και αντιστοίχιση φαινοτυπικών όρων από την HPO 22
4.6 Επεξεργασία και οργάνωση βιβλιογραφικών και φαινοτυπικών δεδομένων 23
4.7 Παραγωγή και παρουσίαση αποτελεσμάτων 23
4.8 Περιορισμοί της μεθοδολογίας 23
Υλοποίηση του R Εργαλείου 24
5.1 Δομή του κώδικα 24
5.2 Περιγραφή βασικών συναρτήσεων 25
5.2.1 Συναρτήσεις εισόδου και προεπεξεργασίας 25
5.2.2 Αναζήτηση βιβλιογραφίας στο PubMed 25
5.2.3 Αντιστοίχιση γονιδίων με HPO όρους 25
5.2.4 Επεξεργασία και οργάνωση αποτελεσμάτων 26
5.2.5 Κύρια συνάρτηση ανάλυσης 26
5.2.6 Μορφοποίηση και εξαγωγή αποτελεσμάτων 26
5.3 Παραγωγή εξόδων 26
5.4 Παράδειγμα χρήσης του εργαλείου 27
Αποτελέσματα 27
6.1 Περιγραφικά αποτελέσματα 27
6.1.1 Αριθμός φαινοτύπων ανά γονίδιο 28
6.1.2 Πλήθος άρθρων PubMed ανά γονίδιο 28
6.2 Παραδείγματα περιπτώσεων (case studies) 28
6.3 Αξιολόγηση αποτελεσμάτων 32

Συζήτηση 33
7.1 Πλεονεκτήματα του εργαλείου 33
7.2 Περιορισμοί 35
7.3 Σύγκριση με υπάρχοντα εργαλεία 36
Συμπεράσματα και Μελλοντική Εργασία 37
8.1 Συμπεράσματα 37
8.2 Μελλοντικές επεκτάσεις 37
8.2.1 Ενσωμάτωση NLP (Natural Language Processing) 37
8.2.2 Machine Learning για gene–phenotype συσχετίσεις 38
8.2.3 Ανάπτυξη Web εφαρμογής 38
8.2.4 Ενσωμάτωση κλινικών δεδομένων 38
8.2.5 Οπτικοποίηση αποτελεσμάτων 39
8.2.6 Ενσωμάτωση δεδομένων ασθενειών 39
8.2.7 Σύνθετα queries με MeSH όρους 39
8.2.8 Λοιπές βελτιώσεις που προκύπτουν από τεχνικούς περιορισμούς του R-εργαλείου 39
Βιβλιογραφικές αναφορές 40
Παραρτήματα 43
Παράρτημα Α: Κώδικας R 43
Α.1 Κύριο αρχείο εκτέλεσης (run_tool.R) 44
Α.2 Συναρτήσεις εισόδου (input_functions.R) 44
Α.3 Συναρτήσεις ανάκτησης PubMed (pubmed_functions.R) 45
Α.4 Συναρτήσεις ανάκτησης HPO (hpo_functions.R) 45
Α.5 Συναρτήσεις επεξεργασίας (processing_functions.R) 47
Α.6 Κύριο pipeline (pipeline.R) 48
Α.7 Μορφοποίηση αποτελεσμάτων (output_formatting.R) 48
Παράρτημα Β: Οδηγός χρήσης εργαλείου 49
Β.1 Απαιτήσεις συστήματος 49
Β.2 Απαραίτητα αρχεία 49
Β.3 Προετοιμασία αρχείου εισόδου 50

Β.4 Εκτέλεση στο RStudio 50
Β.5 Εκτέλεση μέσω γραμμής εντολών 50
Β.6 Παραγόμενα αρχεία εξόδου 51
Β.7 Ενδεικτικό μήνυμα ολοκλήρωσης 51
Β.8 Πιθανά προβλήματα κατά την εκτέλεση 52

Η υλοποίηση του εργαλείου πραγματοποιήθηκε στη γλώσσα R χρησιμοποιώντας το περιβάλλον ανάπτυξης RStudio, το οποίο παρέχει δυνατότητες συγγραφής, εκτέλεσης και διαχείρισης κώδικα. Το εργαλείο μπορεί να εκτελεστεί είτε μέσω του περιβάλλοντος RStudio είτε μέσω γραμμής εντολών χρησιμοποιώντας την εντολή Rscript, επιτρέποντας την αυτοματοποιημένη εκτέλεση του pipeline.

Περίληψη

Η παρούσα διπλωματική εργασία επικεντρώνεται στην ανάπτυξη ενός υπολογιστικού εργαλείου σε γλώσσα R για την αυτοματοποιημένη αντιστοίχιση γονιδίων με τους αντίστοιχους κλινικούς φαινοτύπους και τη σχετική βιοϊατρική βιβλιογραφία. Η ανάγκη για την ανάπτυξη ενός τέτοιου εργαλείου προκύπτει από τον μεγάλο όγκο διαθέσιμων γενετικών και φαινοτυπικών δεδομένων, καθώς και από τη δυσκολία χειροκίνητης επεξεργασίας.

Το εργαλείο βασίζεται σε δύο κύριες βάσεις δεδομένων: τη βιβλιογραφική βάση PubMed και την οντολογία Human Phenotype Ontology (HPO). Η μεθοδολογία που ακολουθείται υλοποιείται ως υπολογιστικό pipeline, το οποίο περιλαμβάνει την εισαγωγή και προεπεξεργασία γονιδιακών δεδομένων, την αυτοματοποιημένη αναζήτηση βιβλιογραφίας μέσω του Entrez API, την ανάκτηση φαινοτυπικών όρων από τα αρχεία συσχέτισης της HPO και την εξαγωγή των αποτελεσμάτων σε δομημένη μορφή.

Τα αποτελέσματα του εργαλείου περιλαμβάνουν πίνακες αντιστοίχισης γονιδίων–φαινοτύπων και βιβλιογραφικών μεταδεδομένων, οι οποίοι εξάγονται σε μορφή CSV, επιτρέποντας την περαιτέρω ανάλυση και αξιοποίησή τους. Η αξιολόγηση του εργαλείου πραγματοποιήθηκε μέσω εκτέλεσής του με ενδεικτικά γονίδια, αναδεικνύοντας τη δυνατότητά του να αποτυπώνει γνωστές βιολογικές συσχετίσεις.

Το εργαλείο παρουσιάζει σημαντικά πλεονεκτήματα ως προς την αυτοματοποίηση, την επεκτασιμότητα και την παράλληλη αξιοποίηση ετερογενών δεδομένων, ωστόσο συνοδεύεται από περιορισμούς που σχετίζονται με τη στρατηγική αναζήτησης, την πληρότητα των δεδομένων και την απουσία σημασιολογικής ανάλυσης. Παρά τους περιορισμούς, αποτελεί ένα χρήσιμο εργαλείο αρχικής διερεύνησης για τη μελέτη των σχέσεων μεταξύ γονιδίων και φαινοτύπων.

Λέξεις-κλειδιά: Γονίδια, φαινότυποι, Human Phenotype Ontology (HPO), PubMed, βιοπληροφορική, υπολογιστικό pipeline, αντιστοίχιση γονιδίων-φαινοτύπων, συσχέτιση δεδομένων

Abstract

This thesis focuses on the development of a computational tool implemented in the R programming language for the automated mapping of genes to their corresponding clinical phenotypes and related biomedical literature. The need for such a tool arises from the large volume of available genomic and phenotypic data, as well as the limitations associated with manual data processing.

The tool is based on two primary data sources: the PubMed bibliographic database and the Human Phenotype Ontology (HPO). The proposed methodology is implemented as a computational pipeline, which includes gene data input and pre-processing, automated literature retrieval via the Entrez API, retrieval of phenotypic terms from HPO, and the generation of structured outputs.

The results of the tool consist of gene–phenotype association tables and bibliographic metadata, which are exported in CSV format, enabling further analysis and utilization. The evaluation of the tool was performed through its application to representative genes, demonstrating its ability to capture known biological associations.

The tool presents significant advantages in terms of automation, scalability, and the parallel use of heterogeneous data sources. However, it is subject to certain limitations related to the search strategy, data completeness, and the absence of semantic analysis. Despite these limitations, it constitutes a useful exploratory tool for the study of gene–phenotype relationships.

Keywords: Genes, phenotypes, Human Phenotype Ontology (HPO), PubMed, bioinformatics, computational pipeline, gene–phenotype mapping, data linking

Εισαγωγή

Η ραγδαία ανάπτυξη της βιοϊατρικής έρευνας και των τεχνολογιών αλληλούχισης έχει οδηγήσει στη συσσώρευση μεγάλου όγκου γενετικών και φαινοτυπικών δεδομένων. Η κατανόηση της σχέσης μεταξύ γονιδίων και κλινικών φαινοτύπων αποτελεί βασικό στόχο της σύγχρονης βιοπληροφορικής και της ιατρικής έρευνας, καθώς συμβάλλει στην κατανόηση των ασθενειών και στην υποστήριξη της διάγνωσης. (Köhler et al., 2021)

Παράλληλα, η συνεχής αύξηση της βιοϊατρικής βιβλιογραφίας και των διαθέσιμων βάσεων δεδομένων καθιστά δύσκολη τη χειροκίνητη ανάλυση. Για τον λόγο αυτό, αναδεικνύεται η ανάγκη ανάπτυξης αυτοματοποιημένων υπολογιστικών εργαλείων που μπορούν να αξιοποιούν συνδυαστικά δεδομένα από διαφορετικές πηγές και να υποστηρίζουν την εξαγωγή χρήσιμων συσχετίσεων.

1.1 Γονιδιακή πληροφορία και κλινικοί φαινότυποι

Η σχέση μεταξύ γονιδίων και φαινοτύπων αποτελεί θεμελιώδη έννοια στη μοριακή γενετική, καθώς γενετικές μεταβολές μπορούν να οδηγήσουν στην εμφάνιση συγκεκριμένων κλινικών χαρακτηριστικών ή ασθενειών. (Strachan & Read, 2018)

Στο πλαίσιο της ιατρικής ακριβείας, η συσχέτιση γενετικών και φαινοτυπικών δεδομένων επιτρέπει την εξατομικευμένη κατανόηση των ασθενειών και την ανάπτυξη στοχευμένων θεραπευτικών προσεγγίσεων. (Ashley, 2016)

Η τυποποιημένη αναπαράσταση των φαινοτύπων μέσω οντολογιών, όπως η Human Phenotype Ontology (HPO), επιτρέπει τη συστηματική καταγραφή και ανάλυση των φαινοτυπικών χαρακτηριστικών και τη σύνδεσή τους με γενετικά δεδομένα. (Köhler et al., 2021)

1.2 Ανάγκη αυτοματοποιημένης αντιστοίχισης γονιδίων-φαινοτύπων

Η βάση δεδομένων PubMed περιλαμβάνει εκατομμύρια βιοϊατρικές δημοσιεύσεις, καθιστώντας δύσκολη τη χειροκίνητη αναζήτηση σχετικής πληροφορίας (Sayers et al., 2023). Παράλληλα, η φαινοτυπική πληροφορία είναι κατανεμημένη σε διαφορετικές πηγές, γεγονός που δυσχεραίνει τη συνδυαστική ανάλυση.

Η ανάγκη για αυτοματοποιημένη αντιστοίχιση των γονιδίων με τους φαινοτύπους καθίσταται επομένως κρίσιμη, καθώς επιτρέπει την ταχύτερη, οργανωμένη και συστηματική εξαγωγή πληροφορίας.

1.3 Στόχοι και συνεισφορά της διπλωματικής

Στόχος της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη ενός υπολογιστικού εργαλείου σε γλώσσα R για την αυτοματοποιημένη αντιστοίχιση γονιδίων με τους αντίστοιχους φαινοτυπικούς όρους και τη σχετική βιβλιογραφία.

Πιο συγκεκριμένα, το εργαλείο αυτό στοχεύει στην εισαγωγή και προεπεξεργασία δεδομένων γονιδίων, στην αυτοματοποιημένη αναζήτηση βιβλιογραφίας στη PubMed, στην αντιστοίχιση γονιδίων με φαινοτυπικούς όρους της HPO, στην οργάνωση των δεδομένων και την παραγωγή δομημένων αποτελεσμάτων.

Η συνεισφορά της εργασίας έγκειται στην ανάπτυξη ενός απλού και επεκτάσιμου υπολογιστικού pipeline, το οποίο επιτρέπει την αυτοματοποιημένη ανάλυση και οργάνωση βιοϊατρικών δεδομένων, υποστηρίζοντας τη μελέτη των σχέσεων μεταξύ γονιδίων και φαινοτύπων.

Θεωρητικό Υπόβαθρο

Στο κεφάλαιο αυτό θα αναπτυχθούν έννοιες από τον κλάδο της Μοριακής Γενετικής και της Βιολογίας καθώς και από τις διαδεδομένες μηχανές βιολογικής/ιατρικής αναζήτησης PubMed και HPO, στις οποίες και βασίστηκε η ανάπτυξη του εν λόγω R - εργαλείου.

2.1 Βασικές έννοιες Μοριακής Γενετικής

2.1.1 Γονίδια και DNA

Η μοριακή γενετική αποτελεί θεμελιώδη κλάδο της βιολογίας που μελετά τη δομή, τη λειτουργία και τη ρύθμιση των γονιδίων σε μοριακό επίπεδο καθώς και τους μηχανισμούς μέσω των οποίων τα γονίδια επηρεάζουν τη φυσιολογία και την ανάπτυξη των οργανισμών. Τα γονίδια ορίζονται ως συγκεκριμένες αλληλουχίες DNA που περιέχουν την απαραίτητη πληροφορία για τη σύνθεση λειτουργικών μορίων, κυρίως πρωτεϊνών. Το DNA αποτελείται από νουκλεοτίδια οργανωμένα σε διπλή έλικα, όπου η αλληλουχία των βάσεων καθορίζει τη γενετική πληροφορία του οργανισμού.

Η ακριβής αντιγραφή και μετάδοση του DNA είναι απαραίτητη για τη διατήρηση της γενετικής πληροφορίας στα κύτταρα και στους οργανισμούς. (Alberts et al., 2022; Watson et al., 2014)

2.1.2 Μεταλλάξεις

Οι μεταλλάξεις ορίζονται ως αλλαγές στην αλληλουχία του DNA και αποτελούν θεμελιώδη μηχανισμό δημιουργίας γενετικής ποικιλότητας. Μπορούν να προκύψουν είτε λόγω σφαλμάτων κατά την αντιγραφή του DNA είτε λόγω εξωτερικών παραγόντων όπως ακτινοβολία ή χημικές ουσίες. Οι μεταλλάξεις διακρίνονται σε διάφορες κατηγορίες, όπως για παράδειγμα σημειακές μεταλλάξεις, ενθέσεις και ελλείψεις κ.λπ., και μπορούν να έχουν ουδέτερες, επιβλαβείς ή ευεργετικές επιπτώσεις στον οργανισμό. Σε πολλές περιπτώσεις, οι μεταλλάξεις επηρεάζουν τη δομή και τη λειτουργία των πρωτεϊνών, οδηγώντας στην εμφάνιση γενετικών ασθενειών ή παθολογικών καταστάσεων. (Griffiths et al., 2020; Lodish et al., 2021)

2.1.3 Γονιδιακή έκφραση

Η γονιδιακή έκφραση αναφέρεται στη διαδικασία μέσω της οποίας η πληροφορία που περιέχεται στο DNA μετατρέπεται σε λειτουργικό προϊόν, συνήθως πρωτεΐνη. Η διαδικασία αυτή περιλαμβάνει δύο βασικά στάδια: τη μεταγραφή του DNA σε RNA (DNA → RNA) και τη μετάφραση του RNA σε πρωτεΐνη (RNA → πρωτεΐνη), σύμφωνα με το «κεντρικό δόγμα» της μοριακής βιολογίας. Η ρύθμιση της γονιδιακής έκφρασης είναι ιδιαίτερα πολύπλοκη και εξαρτάται από διάφορους μηχανισμούς, όπως μεταγραφικούς παράγοντες, επιγενετικές τροποποιήσεις και τη δομή της χρωματίνης, οι οποίοι επηρεάζουν την ενεργοποίηση ή καταστολή των γονιδίων. Οι διαφοροποιήσεις στην έκφραση των γονιδίων κατέχουν καθοριστικό ρόλο στη διαφοροποίηση των κυττάρων και στην ανάπτυξη των οργανισμών. (Alberts et al., 2022; Lodish et al., 2021)

2.2 Κλινικοί φαινότυποι και οντολογίες

2.2.1 Τι είναι φαινότυπος

Ο φαινότυπος αναφέρεται στο σύνολο των παρατηρήσιμων χαρακτηριστικών ενός οργανισμού, όπως μορφολογικά, φυσιολογικά ή συμπεριφορικά γνωρίσματα. Ο φαινότυπος προκύπτει από την αλληλεπίδραση του γονότυπου (δηλαδή της γενετικής σύστασης) με ποικίλους παράγοντες (περιβαλλοντικοί, επιγενετικοί κ.λπ.) και δεν καθορίζεται αποκλειστικά από τη γενετική πληροφορία. (Griffiths et al., 2020)

2.2.2 Σχέση γονότυπου-φαινοτύπου

Η σχέση μεταξύ γονότυπου και φαινοτύπου είναι σύνθετη και συχνά μη γραμμική. Ένα γονίδιο μπορεί να σχετίζεται με πολλαπλούς φαινότυπους, ενώ ένας φαινότυπος μπορεί να επηρεάζεται από πολλά γονίδια. Ταυτόχρονα, περιβαλλοντικοί παράγοντες και ρυθμιστικοί μηχανισμοί επηρεάζουν σημαντικά την έκφραση των χαρακτηριστικών. Η πολυπλοκότητα αυτή αποτελεί βασικό αντικείμενο μελέτης στη σύγχρονη βιοπληροφορική και ιατρική γενετική καθώς έχει αναδειχθεί ιδιαιτέρως σε σύγχρονες μελέτες σύνθετων χαρακτηριστικών (complex traits). (Boyle et al., 2017)

2.2.3 Συσχέτιση γονιδίων, ασθενειών και φαινοτύπων

Η σχέση μεταξύ γονιδίων, ασθενειών και φαινοτύπων αποτελεί βασικό αντικείμενο μελέτης στη γενετική και τη βιοπληροφορική. Γενετικές μεταβολές -όπως μεταλλάξεις ή αλλαγές στην έκφραση γονιδίων- μπορούν να οδηγήσουν σε διαταραχές φυσιολογικών λειτουργιών που εκδηλώνονται μέσω συγκεκριμένων κλινικών φαινοτύπων (εμφάνιση ασθενειών). Ωστόσο, όπως ήδη αναφέρθηκε, οι συσχετίσεις αυτές δεν είναι πάντα άμεσες, καθώς συχνά μεσολαβούν πολύπλοκα μοριακά και βιολογικά δίκτυα. Ένα γονίδιο μπορεί να σχετίζεται με περισσότερες από μία ασθένειες ενώ κάθε ασθένεια μπορεί να εκδηλώνεται μέσω πολλαπλών φαινοτύπων, γεγονός που αντικατοπτρίζει την πολυπλοκότητα των gene–disease–phenotype σχέσεων. Η σύγχρονη έρευνα στη γονιδιωματική και τη βιοπληροφορική επιδιώκει τη χαρτογράφηση αυτών των σχέσεων, συμβάλλοντας στην κατανόηση των μηχανισμών των ασθενειών και στην ανάπτυξη εξατομικευμένων θεραπευτικών προσεγγίσεων. Βάσεις δεδομένων όπως η OMIM και η HPO είναι γνωστές στο να αποτυπώνουν τέτοιες σχέσεις. (Robinson et al., 2014; Amberger et al., 2019)

2.2.4 Τι είναι οντολογία

Στο πλαίσιο της βιοπληροφορικής, μια οντολογία αποτελεί ένα τυποποιημένο σύστημα αναπαράστασης γνώσης, το οποίο οργανώνει έννοιες και τις μεταξύ τους σχέσεις με τρόπο που επιτρέπει τη συνεπή περιγραφή και την υπολογιστική επεξεργασία δεδομένων. Οι οντολογίες χρησιμοποιούνται για τη συστηματική αναπαράσταση βιοϊατρικών δεδομένων και επιτρέπουν τη διαλειτουργικότητα μεταξύ διαφορετικών βάσεων δεδομένων, εφαρμογών και υπολογιστικών συστημάτων. Συνεπώς, διευκολύνουν την οργάνωση δεδομένων από διαφορετικές πηγές και υποστηρίζουν την ανάπτυξη βιοϊατρικών εφαρμογών. (Smith et al., 2007)

2.2.5 Πλεονεκτήματα τυποποίησης

Η χρήση οντολογιών προσφέρει σημαντικά πλεονεκτήματα, όπως:

- τυποποίηση της ορολογίας,
- βελτιωμένη αναζήτηση και ανάκτηση δεδομένων,
- δυνατότητα οργάνωσης δεδομένων από πολυάριθμες πηγές,
- διευκόλυνση υπολογιστικής ανάλυσης και διαλειτουργικότητας μεταξύ διαφορετικών συστημάτων.

Η τυποποίηση είναι ιδιαίτερα σημαντική στη βιοϊατρική έρευνα, όπου τα δεδομένα είναι συχνά ετερογενή και πολύπλοκα. (Bodenreider, 2008)

2.3 Human Phenotype Ontology (HPO)

Η Human Phenotype Ontology (HPO) αποτελεί ένα ευρέως χρησιμοποιούμενο σύστημα οντολογίας για την περιγραφή ανθρώπινων φαινοτύπων. Περιλαμβάνει χιλιάδες τυποποιημένους όρους που οργανώνονται σε ιεραρχική δομή, επιτρέποντας τη συστηματική αναπαράσταση φαινοτυπικών χαρακτηριστικών. (Robinson et al., 2008)

2.3.1 Δομή

Η HPO οργανώνεται ως κατευθυνόμενο γράφημα (directed acyclic graph), στο οποίο κάθε όρος μπορεί να έχει έναν ή περισσότερους «γονικούς» όρους. Αυτό επιτρέπει την αναπαράσταση γενικών και εξειδικευμένων φαινοτύπων και τη μοντελοποίηση πολύπλοκων σχέσεων μεταξύ τους.

2.3.2 HPO IDs

Κάθε φαινοτυπικός όρος αντιστοιχεί σε ένα μοναδικό αναγνωριστικό (HPO ID), π.χ. HP:0001250, το οποίο επιτρέπει την ακριβή αναφορά σε συγκεκριμένο φαινότυπο.

2.3.3 Συσχετίσεις γονιδίων - φαινοτύπων

Η HPO περιλαμβάνει επιμελημένες συσχετίσεις μεταξύ γονιδίων και φαινοτυπικών όρων, γνωστές ως annotations. Οι συσχετίσεις αυτές βασίζονται σε σχέσεις gene–disease–phenotype και επιτρέπουν την αντιστοίχιση ενός γονιδίου με πολλαπλούς φαινότυπους. Είναι σημαντικό να σημειωθεί ότι οι συσχετίσεις αυτές δεν υποδηλώνουν απαραίτητα άμεση αιτιολογική σχέση. (Robinson et al., 2014)

2.3.4 Χρήση στην κλινική γενετική

Η HPO χρησιμοποιείται ευρέως στην κλινική γενετική για τη διάγνωση και την κατηγοριοποίηση ασθενειών, καθώς και για την ανάλυση γενετικών δεδομένων. Επιτρέπει την τυποποιημένη καταγραφή φαινοτύπων και τη σύγκριση μεταξύ ασθενών και μελετών. (Köhler et al., 2021)

2.4 PubMed και βιοϊατρική βιβλιογραφία

Η PubMed αποτελεί μία από τις σημαντικότερες βάσεις δεδομένων βιοϊατρικής βιβλιογραφίας, παρέχοντας πρόσβαση σε εκατομμύρια επιστημονικά άρθρα. Διαχειρίζεται από το National Center for Biotechnology Information (NCBI) και περιλαμβάνει κυρίως εγγραφές από τη βάση MEDLINE. (Sayers et al., 2023)

2.4.1 Δομή και οργάνωση της PubMed

Η PubMed οργανώνει τα άρθρα με βάση βιβλιογραφικά μεταδεδομένα, όπως τίτλο, συγγραφείς, περίληψη (abstract) και λέξεις-κλειδιά. Η δομή αυτή επιτρέπει την αποτελεσματική αναζήτηση και ανάκτηση επιστημονικής πληροφορίας.

2.4.2 MeSH terms

Τα MeSH (Medical Subject Headings) αποτελούν ελεγχόμενο λεξιλόγιο που χρησιμοποιείται για την εύρεση των άρθρων στην PubMed. Μέσω της ιεραρχικής τους δομής επιτρέπουν την ακριβέστερη και πιο στοχευμένη αναζήτηση επιστημονικής βιβλιογραφίας μειώνοντας την ασάφεια των όρων. (Lipscomb, 2000, Sayers et al., 2023)

2.4.3. Αναζήτηση και περιορισμοί

Η αναζήτηση στην PubMed μπορεί να πραγματοποιηθεί με διάφορους τρόπους, όπως μέσω λέξεων-κλειδιών ή MeSH όρων. Ωστόσο, η απλή αναζήτηση με βάση λέξεις-κλειδιά μπορεί να οδηγήσει σε μη ειδικά ή ελλιπή αποτελέσματα, ιδιαίτερα όταν οι όροι είναι αμφίσημοι ή γενικοί.

2.4.4 Πρόσβαση μέσω NCBI Entrez API

Η PubMed παρέχει προγραμματιστική πρόσβαση μέσω του NCBI Entrez API, το οποίο επιτρέπει την αυτοματοποιημένη αναζήτηση και ανάκτηση βιβλιογραφικών δεδομένων. Η δυνατότητα αυτή είναι ιδιαίτερα σημαντική για εφαρμογές βιοπληροφορικής, καθώς επιτρέπει την ενσωμάτωση βιβλιογραφικών δεδομένων σε υπολογιστικά pipelines. (Sayers et al., 2023)

Τεχνολογίες και Εργαλεία

Στο παρόν κεφάλαιο παρουσιάζονται οι τεχνολογίες και τα εργαλεία που χρησιμοποιήθηκαν για την υλοποίηση του R-εργαλείου. Η επιλογή των συγκεκριμένων τεχνολογιών βασίστηκε στην ευρεία χρήση τους στη βιοπληροφορική, στη διαθεσιμότητα εξειδικευμένων βιβλιοθηκών και στη δυνατότητα διασύνδεσης με βιοϊατρικές βάσεις δεδομένων.

3.1 Γλώσσα R στη βιοπληροφορική

Η R αποτελεί μία από τις πιο διαδεδομένες γλώσσες προγραμματισμού στη βιοπληροφορική και τη βιοστατιστική, καθώς προσφέρει ισχυρές δυνατότητες για ανάλυση και επεξεργασία δεδομένων υψηλής πολυπλοκότητας (R Core Team, 2023). Ένα σημαντικό πλεονέκτημα της R είναι η υποστήριξη αναπαραγώγιμων αναλύσεων (reproducibility), κάτι που αποτελεί βασική απαίτηση στη σύγχρονη επιστημονική έρευνα. Επιπλέον, επιτρέπει την εύκολη ενσωμάτωση δεδομένων από εξωτερικές πηγές μέσω APIs, γεγονός που την καθιστά ιδιαίτερα κατάλληλη για εφαρμογές που συνδέονται με βάσεις δεδομένων όπως η PubMed (Gentleman et al., 2004).

Η επιλογή της R στην παρούσα εργασία βασίστηκε στην ευρεία χρήση της στη βιοπληροφορική, στη διαθεσιμότητα εξειδικευμένων πακέτων και στη δυνατότητα ανάπτυξης αυτοματοποιημένων pipelines για την ανάλυση δεδομένων.

3.2 R πακέτα που χρησιμοποιήθηκαν

Για την υλοποίηση του εργαλείου χρησιμοποιήθηκαν συγκεκριμένα πακέτα της R, τα οποία καλύπτουν τις βασικές ανάγκες ανάκτησης, επεξεργασίας και οργάνωσης των δεδομένων.

rentrez

Το πακέτο rentrez παρέχει διεπαφή με το Entrez API του NCBI και επιτρέπει την αυτοματοποιημένη αναζήτηση και ανάκτηση δεδομένων από βάσεις δεδομένων όπως η PubMed (Winter, 2017). Μέσω αυτού, είναι δυνατή η εκτέλεση ερωτημάτων και η λήψη βασικών μεταδεδομένων άρθρων, όπως το PubMed ID και ο τίτλος.

Στο πλαίσιο της παρούσας εργασίας, το rentrez χρησιμοποιείται για την αναζήτηση βιβλιογραφικών δεδομένων με βάση το gene symbol.

dplyr

Το πακέτο dplyr χρησιμοποιείται για τον μετασχηματισμό και τη διαχείριση δεδομένων, προσφέροντας ένα σύνολο εργαλείων για φιλτράρισμα, επιλογή και τροποποίηση δεδομένων (Wickham et al., 2023). Η χρήση του συμβάλλει στη δημιουργία καθαρού και ευανάγνωστου κώδικα, καθώς και στην αποτελεσματική επεξεργασία των αποτελεσμάτων.

Στο εργαλείο αξιοποιείται για:

- κανονικοποίηση των δεδομένων,
- φιλτράρισμα αποτελεσμάτων,
- απομάκρυνση διπλότυπων εγγραφών
- και μετασχηματισμό και αναδιάρθρωση πινάκων δεδομένων

readr

Το πακέτο readr χρησιμοποιείται για την αποδοτική ανάγνωση δεδομένων από αρχεία κειμένου και αποτελεί βασικό εργαλείο για τη διαχείριση δομημένων δεδομένων (Wickham et al., 2023). Σε σύγκριση με βασικές συναρτήσεις της R, προσφέρει καλύτερη απόδοση και πιο αξιόπιστη διαχείριση τύπων δεδομένων.

Στην παρούσα εργασία χρησιμοποιείται για την ανάγνωση του αρχείου genes_to_phenotype.txt, το οποίο περιλαμβάνει τις συσχετίσεις γονιδίων–φαινοτύπων.

3.3 APIs και δεδομένα

HPO annotations

Η φαινοτυπική πληροφορία που χρησιμοποιείται στο εργαλείο προέρχεται από την Human Phenotype Ontology (HPO), η οποία παρέχει επιμελημένες συσχετίσεις μεταξύ γονιδίων και φαινοτυπικών όρων. Οι συσχετίσεις αυτές βασίζονται σε σχέσεις gene–disease–phenotype και χρησιμοποιούνται ευρέως για την ανάλυση γενετικών και κλινικών δεδομένων. (Köhler et al., 2021)

Στην παρούσα υλοποίηση, τα δεδομένα αυτά αξιοποιούνται μέσω του αρχείου genes_to_phenotype.txt, το οποίο περιλαμβάνει τις κατάλληλες αντιστοιχίσεις σε δομημένη μορφή.

PubMed Entrez API

Η πρόσβαση στη βιβλιογραφική βάση PubMed πραγματοποιείται μέσω του Entrez API του National Center for Biotechnology Information (NCBI), το οποίο επιτρέπει την αυτοματοποιημένη αναζήτηση και ανάκτηση βιβλιογραφικών δεδομένων. (Sayers et al., 2023)

Στο εργαλείο, το API χρησιμοποιείται μέσω του πακέτου rentrez για την εκτέλεση ερωτημάτων και την ανάκτηση βασικών μεταδεδομένων άρθρων.

Μορφή δεδομένων (Data Format)

Τα δεδομένα που ανακτώνται μέσω του Entrez API επιστρέφονται σε δομημένες μορφές όπως XML και JSON, οι οποίες επιτρέπουν την εύκολη εξαγωγή συγκεκριμένων πληροφοριών τους από υπολογιστικά συστήματα (π.χ. έτος δημοσίευσης). Οι μορφές αυτές διευκολύνουν την επεξεργασία και την ενσωμάτωση των δεδομένων σε υπολογιστικές ροές ανάλυσης. (Sayers et al., 2023)

Μεθοδολογία

Στο συγκεκριμένο κεφάλαιο περιγράφεται ο τρόπος λειτουργίας του παρόντος R-εργαλείου.

4.1 Γενική αρχιτεκτονική συστήματος

Η παρούσα διπλωματική εργασία ακολουθεί μια υπολογιστική και αυτοματοποιημένη μεθοδολογική προσέγγιση με στόχο την αντιστοίχιση γονιδίων με τους κλινικούς φαινοτύπους τους. Η μεθοδολογία βασίζεται στην ανάκτηση, επεξεργασία και οργάνωση δεδομένων από δύο κύριες πηγές βιοϊατρικής πληροφορίας:

- τη βιβλιογραφική βάση PubMed και
- την οντολογία Human Phenotype Ontology (HPO)

Η προτεινόμενη προσέγγιση υλοποιείται ως υπολογιστικό pipeline, το οποίο ξεκινά με την εισαγωγή γονιδίων από αρχείο εισόδου και καταλήγει στην παραγωγή δομημένων αποτελεσμάτων που περιλαμβάνουν αντιστοιχίσεις γονιδίων με φαινοτύπους και σχετική βιβλιογραφία. Η συνολική ροή της διαδικασίας παρουσιάζεται στο **Σχήμα 1**.

4.2 Ροή εργασίας (Workflow) – Γενική περιγραφή και Αναπαράσταση

Η ροή εργασίας του συστήματος αποτελείται από διαδοχικά στάδια επεξεργασίας δεδομένων, τα οποία εκτελούνται με προκαθορισμένη σειρά ώστε να διασφαλίζεται η ορθότητα και η επαναληψιμότητα της ανάλυσης. Το προτεινόμενο R-εργαλείο υλοποιεί μια αυτοματοποιημένη ροή εργασίας για την αντιστοίχιση γονιδίων με τους κλινικούς φαινοτύπους τους, αξιοποιώντας δεδομένα από τη βιβλιογραφική βάση PubMed και την οντολογία HPO.

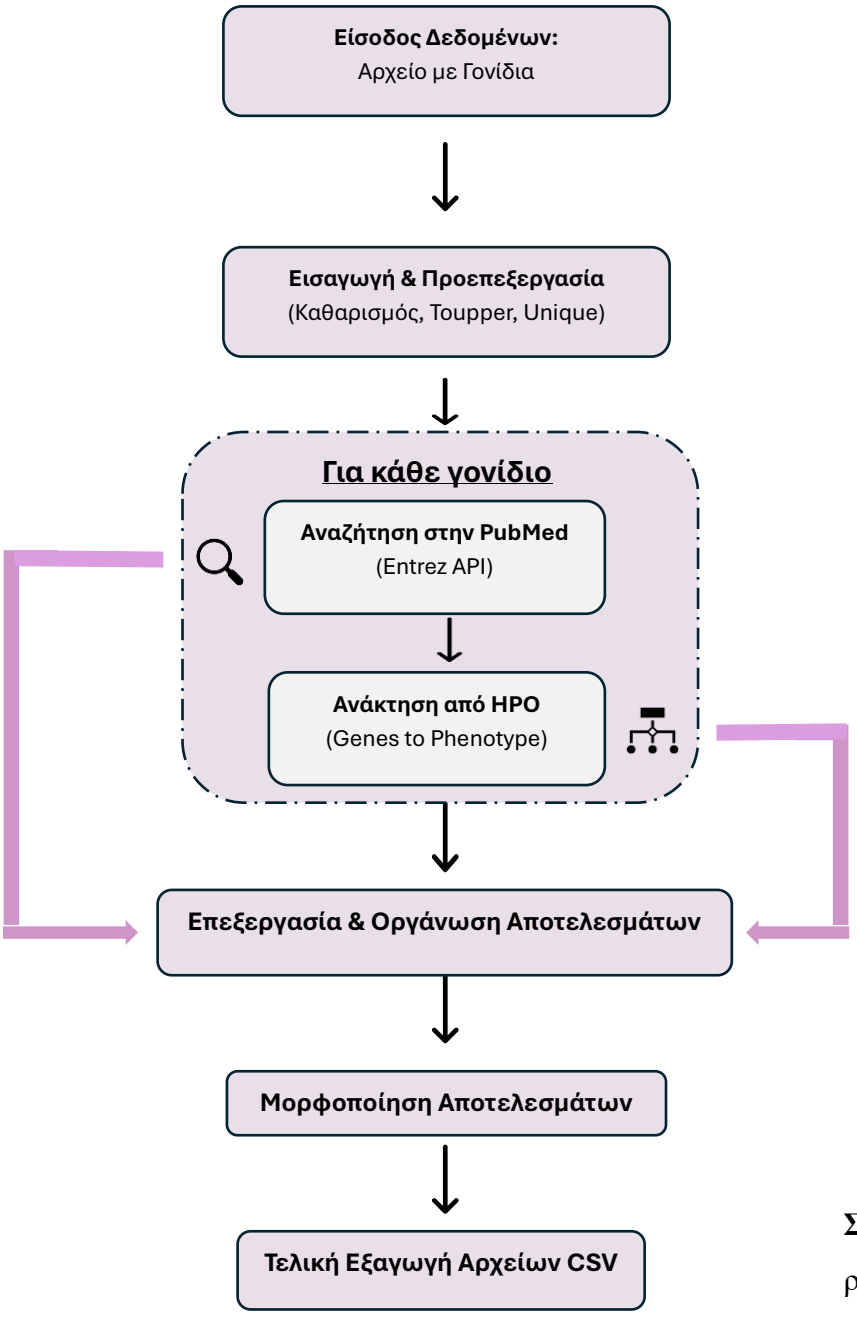
Η ροή ξεκινά από την εισαγωγή ενός ή περισσότερων γονιδίων μέσω αρχείου εισόδου και καταλήγει στην παραγωγή δομημένων αποτελεσμάτων που περιλαμβάνουν συσχετισμένους HPO όρους και σχετική βιβλιογραφία.

Η διαδικασία είναι πλήρως αυτοματοποιημένη και επαναλήψιμη, επιτρέποντας την ανάλυση τόσο μεμονωμένων γονιδίων όσο και πολλαπλών γονιδίων.

Η διαδικασία περιλαμβάνει:

1. Εισαγωγή και προεπεξεργασία (καθαρισμός και τυποποίηση) γονιδίων
2. Αναζήτηση σχετικής βιβλιογραφίας στη βάση PubMed
3. Ανάκτηση φαινοτυπικών όρων από την HPO
4. Επεξεργασία και οργάνωση βιβλιογραφικών και φαινοτυπικών δεδομένων
5. Μορφοποίηση αποτελεσμάτων
6. Εξαγωγή αποτελεσμάτων σε αρχεία CSV

Κάθε στάδιο της ροής εργασίας αντιστοιχεί σε συγκεκριμένες λειτουργικές ενότητες του εργαλείου, οι οποίες περιγράφονται αναλυτικά στο **Κεφάλαιο 5** (Υλοποίηση).



Σχήμα 1. Σχηματική αναπαράσταση ροής εργασίας (Workflow).

Η συνολική ροή εργασίας του συστήματος παρουσιάζεται στο **Σχήμα 1** και ακολουθεί μια γραμμική και αυτοματοποιημένη διαδικασία επεξεργασίας δεδομένων. Η ανάλυση ξεκινά από την εισαγωγή και προεπεξεργασία των γονιδίων εισόδου και συνεχίζεται με την ανεξάρτητη ανάκτηση

βιβλιογραφικών δεδομένων από τη PubMed και φαινοτυπικών όρων από την HPO για κάθε γονίδιο. Στη συνέχεια, τα δεδομένα οργανώνονται και μορφοποιούνται κατάλληλα, οδηγώντας στην παραγωγή δομημένων αποτελεσμάτων σε μορφή αρχείων CSV.

4.3 Εισαγωγή και Προεπεξεργασία δεδομένων γονιδίων

Η μεθοδολογία υποστηρίζει την ανάλυση τόσο μεμονωμένων γονιδίων όσο και συνόλων γονιδίων. Τα δεδομένα εισόδου παρέχονται μέσω αρχείου κειμένου, στο οποίο κάθε γραμμή αντιστοιχεί σε ένα gene symbol.

Κατά το στάδιο της προεπεξεργασίας, το οποίο υλοποιείται μέσω της συνάρτησης load_genes(), πραγματοποιείται βασικός καθαρισμός και τυποποίηση των δεδομένων εισόδου. Πιο συγκεκριμένα γίνεται:

- αφαίρεση περιττών κενών χαρακτήρων (trimws()),
- απομάκρυνση κενών γραμμών,
- αφαίρεση διπλότυπων γονιδίων (unique()),
- μετατροπή των gene symbols σε κεφαλαία γράμματα (toupper()).

Το στάδιο αυτό είναι κρίσιμο, καθώς προσφέρει ομοιόμορφη μορφή των δεδομένων εισόδου και έτσι διασφαλίζει τη σωστή αντιστοίχιση των γονιδίων με τις εξωτερικές βάσεις δεδομένων που χρησιμοποιούνται στα επόμενα βήματα της ανάλυσης.

4.4 Ανάκτηση βιβλιογραφικών δεδομένων από την PubMed

Για κάθε γονίδιο πραγματοποιείται αυτοματοποιημένη αναζήτηση στη βιβλιογραφική βάση PubMed. Η αναζήτηση βασίζεται στη δημιουργία δυναμικών ερωτημάτων που στοχεύουν στον εντοπισμό επιστημονικών άρθρων στα οποία το γονίδιο αναφέρεται στον τίτλο ή στο abstract. Η κατασκευή των ερωτημάτων αναζήτησης (query construction) πραγματοποιείται με βάση το σύμβολο του γονιδίου. Πιο συγκεκριμένα, για κάθε γονίδιο, η συνάρτηση search_pubmed() κατασκευάζει ερώτημα αναζήτησης της μορφής gene[Title/Abstract], ώστε να εντοπίζονται άρθρα στα οποία το γονίδιο αναφέρεται στον τίτλο ή στο abstract.

Η ανάκτηση των βιβλιογραφικών δεδομένων πραγματοποιείται μέσω του NCBI Entrez API με τη χρήση του πακέτου rentrez της γλώσσας R.

Για κάθε γονίδιο εκτελείται αναζήτηση στη βάση PubMed και ανακτώνται βασικά μεταδεδομένα από τα σχετικά άρθρα, όπως:

- το αναγνωριστικό PubMed (PMID),
- ο τίτλος του άρθρου,
- η ημερομηνία δημοσίευσης.

Τα δεδομένα αυτά χρησιμοποιούνται ως σχετική βιβλιογραφική πληροφορία για το υπό μελέτη γονίδιο και οργανώνονται παράλληλα με τα φαινοτυπικά δεδομένα που ανακτώνται από την HPO.

4.5 Ανάκτηση και αντιστοίχιση φαινοτυπικών όρων από την HPO

Η φαινοτυπική πληροφορία ανακτάται από την Human Phenotype Ontology (HPO), η οποία παρέχει τυποποιημένους και επιμελημένους όρους για την περιγραφή ανθρώπινων φαινοτύπων. Στο πλαίσιο της παρούσας εργασίας, η αντιστοίχιση γονιδίων με φαινοτυπικούς όρους πραγματοποιείται μέσω του αρχείου genes_to_phenotype.txt, το οποίο προέρχεται από την HPO και περιλαμβάνει επιμελημένες συσχετίσεις γονιδίων με HPO όρους.

Για κάθε γονίδιο πραγματοποιείται HPO mapping μέσω αναζήτησης των αντίστοιχων φαινοτυπικών όρων, οι οποίοι περιγράφουν κλινικά χαρακτηριστικά και παθολογικές εκδηλώσεις που έχουν συσχετιστεί με αυτό. Οι όροι αυτοί αναπαρίστανται μέσω μοναδικών αναγνωριστικών (HPO IDs) και συνοδεύονται από την ονομασία του αντίστοιχου φαινοτύπου.

Η φύση της συσχέτισης είναι τύπου one-to-many, καθώς κάθε γονίδιο μπορεί να αντιστοιχεί σε πολλαπλούς κλινικούς φαινότυπους. Το σύστημα διατηρεί κάθε αντιστοίχιση γονιδίου–φαινοτύπου ως ανεξάρτητη εγγραφή, επιτρέποντας την καταγραφή πολλαπλών φαινοτυπικών συσχετίσεων για το ίδιο γονίδιο.

Σε αυτό το σημείο, είναι σημαντικό να αναφερθεί ότι οι συσχετίσεις γονιδίων-φαινοτύπων βασίζονται σε HPO annotations που προκύπτουν από σχέσεις gene–disease–phenotype και δεν υποδηλώνουν απαραίτητα άμεση αιτιολογική σχέση για κάθε επιμέρους φαινότυπο.

4.6 Επεξεργασία και οργάνωση βιβλιογραφικών και φαινοτυπικών δεδομένων

Στο στάδιο αυτό πραγματοποιείται η επεξεργασία και οργάνωση των δεδομένων που προέρχονται από την PubMed και την HPO.

Η επεξεργασία περιλαμβάνει αρχικά βασική κανονικοποίηση της μορφής των δεδομένων ώστε να εξασφαλίζεται ομοιομορφία στην αναπαράσταση των γονιδίων και των φαινοτυπικών όρων. Ειδικότερα, εφαρμόζεται τυποποίηση των συμβόλων των γονιδίων, ενιαία μορφή των αναγνωριστικών HPO και σταθερή δομή στα επιμέρους πεδία των αποτελεσμάτων. Έπειτα γίνεται αφαίρεση διπλότυπων εγγραφών, έτσι ώστε κάθε μοναδική καταγραφή να διατηρείται μία φορά χωρίς περιττή επανάληψη.

Τα αποτελέσματα δεν συγχωνεύονται σε έναν κοινό πίνακα, αλλά οργανώνονται σε κοινή λογική δομή ανά γονίδιο η οποία περιλαμβάνει:

- τους αντίστοιχους φαινοτυπικούς όρους από την HPO,
- τη σχετική επιστημονική βιβλιογραφία από την PubMed.

Στη συνέχεια, εξάγονται σε καθαρή και κανονικοποιημένη μορφή, κατάλληλη για εξαγωγή και παρουσίαση σε δύο διακριτούς πίνακες αποτελεσμάτων.

4.7 Παραγωγή και παρουσίαση αποτελεσμάτων

Τα τελικά αποτελέσματα της παρούσας μεθοδολογικής προσέγγισης παρουσιάζονται σε δομημένη μορφή και αποθηκεύονται σε δύο αρχεία εξόδου τύπου CSV. Το αρχείο gene_hpo_results.csv περιλαμβάνει τις αντιστοιχίσεις γονιδίων με φαινοτυπικούς όρους της HPO, ενώ το αρχείο pubmed_results.csv περιλαμβάνει τα βιβλιογραφικά δεδομένα που ανακτήθηκαν από την PubMed.

Όπως αναφέρθηκε και προηγουμένως, η μορφή αυτή επιτρέπει την περαιτέρω ανάλυση, παρουσίαση και αξιοποίηση των αποτελεσμάτων σε ερευνητικά ή κλινικά πλαίσια.

4.8 Περιορισμοί της μεθοδολογίας

Η προτεινόμενη μεθοδολογία παρουσιάζει ορισμένους περιορισμούς οι οποίοι σχετίζονται κυρίως με τη φύση των δεδομένων και τη στρατηγική ανάκτησής τους. Συγκεκριμένα, η ακρίβεια της ανάλυσης επηρεάζεται από την πληρότητα και την επικαιρότητα των δεδομένων της Human

Phenotype Ontology (HPO), καθώς και από τους περιορισμούς της βιβλιογραφικής αναζήτησης στη βάση PubMed.

Επιπλέον, οι συσχετίσεις γονιδίων–φαινοτύπων που προέρχονται από την HPO βασίζονται σε επιμεληημένα annotations και δεν υποδηλώνουν απαραίτητα άμεση αιτιολογική σχέση για κάθε επιμέρους φαινότυπο ενώ η βιβλιογραφική αναζήτηση περιορίζεται σε τίτλους και abstracts, χωρίς να πραγματοποιείται ανάλυση του πλήρους κειμένου των άρθρων. Τέλος, εφαρμόζεται ανώτατο όριο στον αριθμό των ανακτηθέντων άρθρων ανά γονίδιο, το οποίο μπορεί να επηρεάσει την πληρότητα των βιβλιογραφικών αποτελεσμάτων.

Οι περιορισμοί αυτοί λαμβάνονται υπόψη κατά την ερμηνεία των αποτελεσμάτων και συζητούνται περαιτέρω στο κεφάλαιο της Συζήτησης (βλ. **Κεφάλαιο 7**).

Υλοποίηση του R Εργαλείου

Σε αυτή την ενότητα παρουσιάζεται το κυρίως έργο της παρούσας διπλωματικής με τον κώδικα του R-εργαλείου αναλυτικά. Το εργαλείο υλοποιήθηκε και δοκιμάστηκε σε περιβάλλον Windows με έκδοση 11 Pro 24H2 και επιβεβαιώθηκε η ορθή λειτουργία του για σύνολα δοκιμαστικών γονιδίων. Στο **Παράρτημα Α** περιγράφεται επιπλέον η modular δομή και τα αρχεία που διαθέτει.

5.1 Δομή του κώδικα

Το R εργαλείο που αναπτύχθηκε ακολουθεί αρθρωτή (modular) δομή, έτσι ώστε κάθε στάδιο του workflow (βλ. **Κεφάλαιο 4**) να αντιστοιχεί σε σαφώς ορισμένες συναρτήσεις. Η δομή αυτή επιτρέπει τον διαχωρισμό της ανάλυσης, της επεξεργασίας και της παρουσίασης των αποτελεσμάτων, διευκολύνοντας τη συντήρηση και τη μελλοντική επέκταση του εργαλείου.

Ο κώδικας οργανώνεται σε διακριτά αρχεία R, τα οποία ομαδοποιούν συναρτήσεις σχετικές με:

- την είσοδο και προεπεξεργασία των γονιδίων
- την αναζήτηση βιβλιογραφικής πληροφορίας από την PubMed και την ανάκτηση φαινοτυπικών δεδομένων από αρχείο HPO
- τον συντονισμό της συνολικής ανάλυσης μέσω του κύριου pipeline

- τη μορφοποίηση των τελικών αποτελεσμάτων

Η εκτέλεση του εργαλείου πραγματοποιείται μέσω του κύριου script (run_tool.R), το οποίο καλεί τις επιμέρους συναρτήσεις και ενεργοποιεί το pipeline ανάλυσης.

Η ιεραρχική αυτή διάρθρωση αντικατοπτρίζει άμεσα τα βήματα του workflow (βλ. **Κεφάλαιο 4.2**) και διασφαλίζει την λογική ροή από την είσοδο των γονιδίων έως την παραγωγή των τελικών αποτελεσμάτων.

5.2 Περιγραφή βασικών συναρτήσεων

5.2.1 Συναρτήσεις εισόδου και προεπεξεργασίας

Η συνάρτηση load_genes() αναλαμβάνει την ανάγνωση των γονιδίων από αρχείο εισόδου και την προεπεξεργασία τους. Στο στάδιο αυτό πραγματοποιείται βασικός καθαρισμός, απομάκρυνση κενών τιμών και διπλοτύπων, καθώς και τυποποίηση των συμβόλων των γονιδίων σε κεφαλαία μορφή, όπως περιγράφηκε στο **Κεφάλαιο 4.3**.

5.2.2 Αναζήτηση βιβλιογραφίας στο PubMed

Η συνάρτηση search_pubmed() υλοποιεί την αναζήτηση βιβλιογραφικής πληροφορίας στη βάση δεδομένων PubMed για κάθε γονίδιο. Η αναζήτηση βασίζεται σε δυναμικά ερωτήματα της μορφής gene[Title/Abstract], ώστε να εντοπίζονται άρθρα στα οποία το γονίδιο αναφέρεται στον τίτλο ή στο abstract. Στη συνέχεια, επιστρέφει βασικά μεταδεδομένα των άρθρων, όπως το αναγνωριστικό PubMed (PMID), τον τίτλο και το έτος δημοσίευσης.

Η προσέγγιση αυτή είναι σύμφωνη με το **Κεφάλαιο 4.4** (workflow) και επιλέχθηκε ώστε να επιτυγχάνεται ισορροπία μεταξύ πληρότητας και υπολογιστικής απόδοσης.

5.2.3 Αντιστοίχιση γονιδίων με HPO όρους

Η συνάρτηση get_hpo_terms() πραγματοποιεί την αντιστοίχιση κάθε γονιδίου με τους σχετικούς φαινοτυπικούς όρους της Human Phenotype Ontology. Η αντιστοίχιση πραγματοποιείται μέσω του αρχείου genes_to_phenotype.txt, το οποίο περιλαμβάνει επιμελημένες συσχετίσεις γονιδίων με HPO όρους. Για κάθε γονίδιο είναι δυνατή η ανάκτηση πολλαπλών HPO όρων, αντανακλώντας τη σύνθετη φύση των κλινικών φαινοτύπων, όπως περιγράφηκε στο **Κεφάλαιο 4.5**.

5.2.4 Επεξεργασία και οργάνωση αποτελεσμάτων

Η συνάρτηση process_results() συνδυάζει τα δεδομένα που προκύπτουν από το PubMed και την HPO και υλοποιεί:

- κανονικοποίηση της μορφής των δεδομένων
- αφαίρεση διπλότυπων εγγραφών
- οργάνωση των αποτελεσμάτων σε δομημένη μορφή

Τα δεδομένα δεν συγχωνεύονται σε ενιαίο πίνακα, αλλά διατηρούνται ως διακριτά σύνολα ανά πηγή.

Το στάδιο αυτό όπως αναφέρθηκε και στο βήμα 4.6 του workflow, αποτελεί κρίσιμο σημείο για τη διασφάλιση της ποιότητας των δεδομένων.

5.2.5 Κύρια συνάρτηση ανάλυσης

Η συνάρτηση gene_phenotype_pipeline() αποτελεί τον κεντρικό μηχανισμό του εργαλείου και υλοποιεί την επαναληπτική εφαρμογή της ανάλυσης για κάθε γονίδιο εισόδου μέσω της συνάρτησης analyze_gene(). Πρακτικά, για κάθε γονίδιο καλεί τη συνάρτηση analyze_gene(), η οποία συντονίζει την αναζήτηση PubMed, την αντιστοίχιση HPO και την επεξεργασία των αποτελεσμάτων. Η έξοδος της συνάρτησης είναι μια δομημένη λίστα αποτελεσμάτων για όλα τα γονίδια εισόδου.

5.2.6 Μορφοποίηση και εξαγωγή αποτελεσμάτων

Η συνάρτηση format_results() αναλαμβάνει τη συγκέντρωση των επιμέρους αποτελεσμάτων και τη δημιουργία των τελικών πινάκων. Πιο συγκεκριμένα, συγκεντρώνει τα δεδομένα HPO και PubMed από όλα τα γονίδια σε δύο ξεχωριστούς πίνακες, οι οποίοι στη συνέχεια αποθηκεύονται σε αρχεία CSV.

5.3 Παραγωγή εξόδων

Μετά την ολοκλήρωση του κύριου pipeline, τα αποτελέσματα της ανάλυσης εξάγονται σε δομημένα αρχεία μορφής CSV, διευκολύνοντας την περαιτέρω ανάλυση και αξιοποίηση τους. Πιο συγκεκριμένα, μορφοποιούνται σε τελικούς πίνακες μέσω της συνάρτησης format_results(). Οι πίνακες που παράγονται περιλαμβάνουν:

◊ πίνακα αντιστοίχισης γονιδίων - HPO φαινοτύπων,

◊ πίνακα βιβλιογραφικής πληροφορίας, με δεδομένα από την PubMed

Η διάκριση της παραγωγής εξόδων από την κύρια ανάλυση επιτρέπει ευελιξία στην ανάλυση και επαναχρησιμοποίηση των δεδομένων και αντιστοιχεί στο βήμα 4.7 του **Κεφαλαίου 4**.

5.4 Παράδειγμα χρήσης του εργαλείου

Το δεδομένο R εργαλείο μπορεί να χρησιμοποιηθεί ως εξής:

Ο χρήστης παρέχει ένα αρχείο εισόδου (genes.txt) που περιέχει μία λίστα γονιδίων. Στη συνέχεια εκτελείται το κύριο script του εργαλείου (run_tool.R), το οποίο ενεργοποιεί το pipeline ανάλυσης και καλεί τη συνάρτηση gene_phenotype_pipeline().

Κατά την εκτέλεση, παράγεται μία δομημένη λίστα αποτελεσμάτων, η οποία περιλαμβάνει για κάθε γονίδιο τη σχετική βιβλιογραφική πληροφορία από την PubMed και τους αντίστοιχους φαινοτυπικούς όρους από την HPO. Στη συνέχεια, τα αποτελέσματα μορφοποιούνται σε τελικούς πίνακες και αποθηκεύονται σε αρχεία CSV.

Η προσέγγιση αυτή επιτρέπει την αυτοματοποιημένη ανάλυση πολλαπλών γονιδίων χωρίς την ανάγκη χειροκίνητης παρέμβασης.

Αποτελέσματα

Στο παρόν κεφάλαιο παρουσιάζονται τα αποτελέσματα που προκύπτουν από την εκτέλεση του αναπτυγμένου R-εργαλείου. Η ανάλυση βασίζεται στη σύνδεση γονιδίων με φαινοτυπικούς όρους της HPO και στη σχετική βιβλιογραφία από την PubMed. Τα αποτελέσματα παρουσιάζονται τόσο σε συνοπτική μορφή όσο και μέσω ενδεικτικών παραδειγμάτων.

6.1 Περιγραφικά αποτελέσματα

Τα αποτελέσματα του εργαλείου περιλαμβάνουν δύο βασικά σύνολα δεδομένων:

- έναν πίνακα αντιστοίχισης γονιδίων–φαινοτύπων (HPO),
- έναν πίνακα βιβλιογραφικών δεδομένων από την PubMed.

Η ανάλυση των δεδομένων αυτών επιτρέπει την εξαγωγή βασικών περιγραφικών στατιστικών.

6.1.1 Αριθμός φαινοτύπων ανά γονίδιο

Για κάθε γονίδιο, το εργαλείο ανακτά έναν αριθμό φαινοτυπικών όρων από την HPO. Παρατηρείται ότι ο αριθμός αυτός διαφέρει σημαντικά μεταξύ των γονιδίων, γεγονός που αντανακλά την πολυπλοκότητα των βιολογικών λειτουργιών και τη διαφορετική έκταση της βιβλιογραφικής γνώσης για κάθε γονίδιο.

Γονίδια που σχετίζονται με καλά μελετημένες ασθένειες τείνουν να εμφανίζουν μεγαλύτερο αριθμό φαινοτυπικών συσχετίσεων, ενώ λιγότερο μελετημένα γονίδια συνοδεύονται από περιορισμένο αριθμό φαινοτύπων.

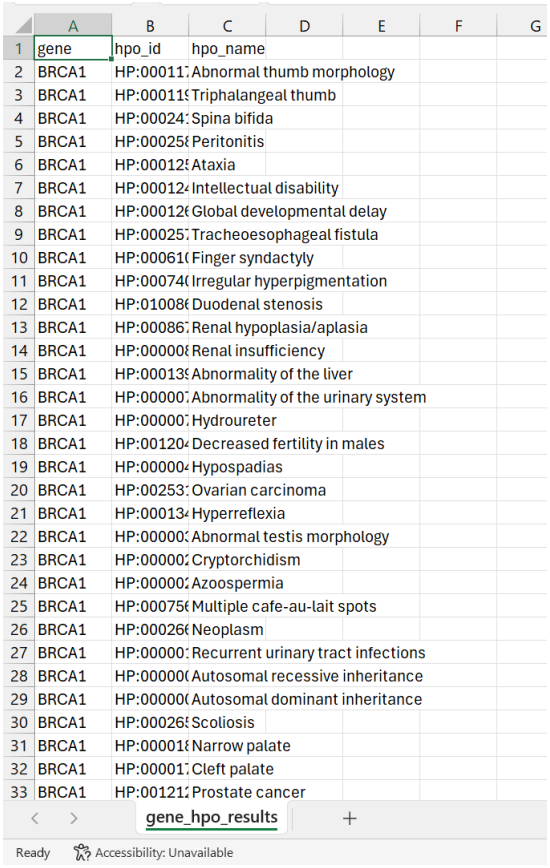
6.1.2 Πλήθος άρθρων PubMed ανά γονίδιο

Ο πραγματικός αριθμός των διαθέσιμων άρθρων στην PubMed μπορεί να διαφέρει ανά γονίδιο. Ωστόσο, στο παρόν εργαλείο έχει τεθεί ανώτατο όριο ανάκτησης 20 άρθρων ανά γονίδιο, για λόγους απόδοσης και διαχειρισιμότητας των αποτελεσμάτων.

Γονίδια με υψηλό ερευνητικό ενδιαφέρον εμφανίζουν μεγάλο αριθμό δημοσιεύσεων, ενώ άλλα γονίδια εμφανίζουν περιορισμένη βιβλιογραφική κάλυψη. Η παρατήρηση αυτή αναδεικνύει τη σημασία της βιβλιογραφικής βάσης ως συμπληρωματικής πηγής πληροφορίας, αλλά και τους περιορισμούς της ως προς την πληρότητα και την ομοιομορφία των δεδομένων.

6.2 Παραδείγματα περιπτώσεων (case studies)

Για την αξιολόγηση της λειτουργίας του εργαλείου χρησιμοποιήθηκαν ως ενδεικτικά παραδείγματα τα γονίδια BRCA1, TP53 και CFTR, τα οποία αποτελούν καλά μελετημένα γονίδια με εκτεταμένη βιβλιογραφική τεκμηρίωση. Η ακριβής διαδικασία περιγράφεται στο **Παράρτημα Β**. Τα αποτελέσματα στις **Εικόνα 1** και **Εικόνα 2** καθώς και στους **Πίνακας 1** και **Σχήμα 2**, έδειξαν ότι το εργαλείο ανακτά μεγάλο αριθμό φαινοτυπικών όρων από την HPO, καθώς και σχετική βιβλιογραφία από την PubMed.

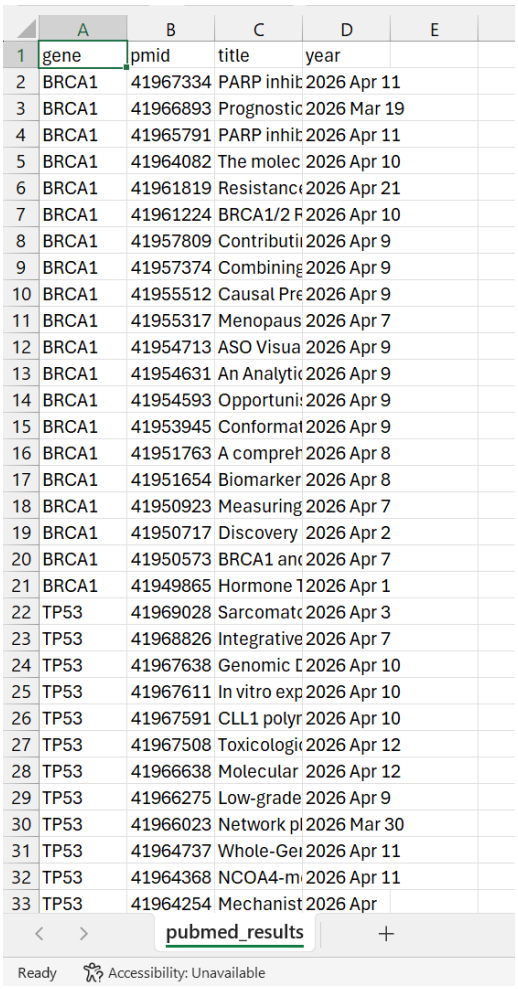


	A	B	C	D	E	F	G
1	gene	hpo_id	hpo_name				
2	BRCA1	HP:00011	Abnormal thumb morphology				
3	BRCA1	HP:00011	Triphalangeal thumb				
4	BRCA1	HP:00024	Spina bifida				
5	BRCA1	HP:00025	Peritonitis				
6	BRCA1	HP:00012	Ataxia				
7	BRCA1	HP:00012	Intellectual disability				
8	BRCA1	HP:00012	Global developmental delay				
9	BRCA1	HP:00025	Tracheoesophageal fistula				
10	BRCA1	HP:00061	Finger syndactyly				
11	BRCA1	HP:00074	Irregular hyperpigmentation				
12	BRCA1	HP:01008	Duodenal stenosis				
13	BRCA1	HP:00086	Renal hypoplasia/aplasia				
14	BRCA1	HP:00000	Renal insufficiency				
15	BRCA1	HP:00013	Abnormality of the liver				
16	BRCA1	HP:00000	Abnormality of the urinary system				
17	BRCA1	HP:00000	Hydroureter				
18	BRCA1	HP:00120	Decreased fertility in males				
19	BRCA1	HP:00000	Hypospadias				
20	BRCA1	HP:00253	Ovarian carcinoma				
21	BRCA1	HP:00013	Hyperreflexia				
22	BRCA1	HP:00000	Abnormal testis morphology				
23	BRCA1	HP:00000	Cryptorchidism				
24	BRCA1	HP:00000	Azoospermia				
25	BRCA1	HP:00075	Multiple cafe-au-lait spots				
26	BRCA1	HP:00026	Neoplasm				
27	BRCA1	HP:00000	Recurrent urinary tract infections				
28	BRCA1	HP:00000	Autosomal recessive inheritance				
29	BRCA1	HP:00000	Autosomal dominant inheritance				
30	BRCA1	HP:00026	Scoliosis				
31	BRCA1	HP:00001	Narrow palate				
32	BRCA1	HP:00001	Cleft palate				
33	BRCA1	HP:00121	Prostate cancer				

gene_hpo_results

Ready Accessibility: Unavailable

Εικόνα 1. Το αρχείο gene_hpo_results.csv περιλαμβάνει τις αντιστοιχίσεις γονιδίων με φαινοτυπικούς όρους της HPO. Κάθε γραμμή αντιστοιχεί σε μία συσχέτιση γονιδίου–φαινοτύπου.



	A	B	C	D	E
1	gene	pmid	title	year	
2	BRCA1	41967334	PARP inhit	2026 Apr 11	
3	BRCA1	41966893	Prognostic	2026 Mar 19	
4	BRCA1	41965791	PARP inhit	2026 Apr 11	
5	BRCA1	41964082	The molec	2026 Apr 10	
6	BRCA1	41961819	Resistanc	2026 Apr 21	
7	BRCA1	41961224	BRCA1/2 F	2026 Apr 10	
8	BRCA1	41957809	Contributi	2026 Apr 9	
9	BRCA1	41957374	Combining	2026 Apr 9	
10	BRCA1	41955512	Causal Pre	2026 Apr 9	
11	BRCA1	41955317	Menopaus	2026 Apr 7	
12	BRCA1	41954713	ASO Visua	2026 Apr 9	
13	BRCA1	41954631	An Analytic	2026 Apr 9	
14	BRCA1	41954593	Opportuni	2026 Apr 9	
15	BRCA1	41953945	Conformat	2026 Apr 9	
16	BRCA1	41951763	A compreh	2026 Apr 8	
17	BRCA1	41951654	Biomarker	2026 Apr 8	
18	BRCA1	41950923	Measuring	2026 Apr 7	
19	BRCA1	41950717	Discovery	2026 Apr 2	
20	BRCA1	41950573	BRCA1 and	2026 Apr 7	
21	BRCA1	41949865	Hormone T	2026 Apr 1	
22	TP53	41969028	Sarcomatc	2026 Apr 3	
23	TP53	41968826	Integrative	2026 Apr 7	
24	TP53	41967638	Genomic D	2026 Apr 10	
25	TP53	41967611	In vitro exp	2026 Apr 10	
26	TP53	41967591	CLL1 polyr	2026 Apr 10	
27	TP53	41967508	Toxicologic	2026 Apr 12	
28	TP53	41966638	Molecular	2026 Apr 12	
29	TP53	41966275	Low-grade	2026 Apr 9	
30	TP53	41966023	Network pl	2026 Mar 30	
31	TP53	41964737	Whole-Ger	2026 Apr 11	
32	TP53	41964368	NCOA4-m	2026 Apr 11	
33	TP53	41964254	Mechanist	2026 Apr	

pubmed_results

Ready Accessibility: Unavailable

Εικόνα 2. Το αρχείο pubmed_results.csv περιλαμβάνει βιβλιογραφικά μεταδεδομένα για κάθε γονίδιο. Ενδεικτικά περιλαμβάνει το gene symbol, το PMID, τον τίτλο του άρθρου και το έτος δημοσίευσης.

Πιο συγκεκριμένα, το γονίδιο BRCA1 αποτελεί ένα από τα πιο μελετημένα γονίδια που σχετίζονται με καρκινικές διαταραχές, ιδιαίτερα με τον καρκίνο του μαστού και των ωοθηκών. Τα αποτελέσματα του εργαλείου έδειξαν μεγάλο αριθμό φαινοτυπικών όρων από την HPO, όπως neoplasm, ovarian carcinoma και prostate cancer, οι οποίοι είναι συμβατοί με τη γνωστή βιολογική του λειτουργία. Παράλληλα, ανακτήθηκε σημαντικός αριθμός άρθρων από την PubMed, γεγονός που αντανακλά το υψηλό ερευνητικό ενδιαφέρον για το συγκεκριμένο γονίδιο. Η παράλληλη παρουσία φαινοτυπικών και βιβλιογραφικών δεδομένων υποδεικνύει ότι το εργαλείο λειτουργεί αποτελεσματικά σε καλά μελετημένες περιπτώσεις.

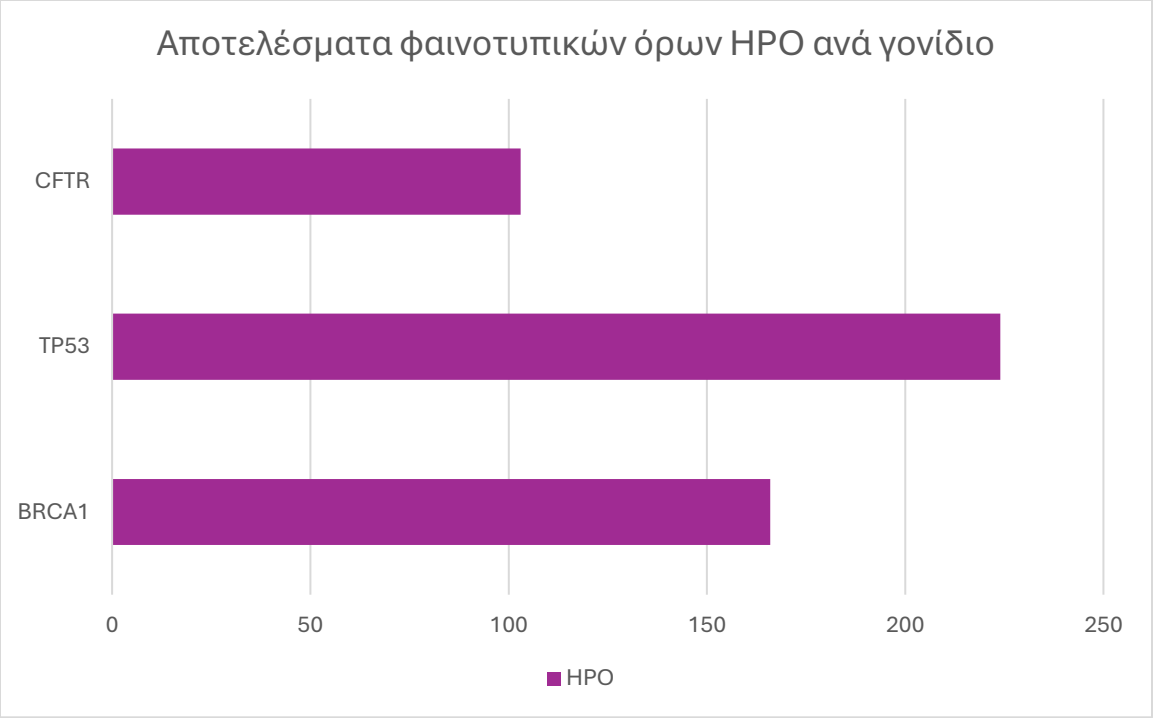
Στην περίπτωση του TP53 το οποίο εμπλέκεται σε πλήθος κυτταρικών μηχανισμών, όπως η απόπτωση και η επιδιόρθωση DNA, τα αποτελέσματα έδειξαν ότι το εργαλείο ανακτά ποικιλία φαινοτυπικών όρων, οι οποίοι σχετίζονται κυρίως με νεοπλασματικές καταστάσεις. Επιπλέον, η

ανάκτηση σημαντικού αριθμού άρθρων από την PubMed επιβεβαιώνει τη σημαντική παρουσία του γονιδίου στη σύγχρονη επιστημονική βιβλιογραφία. Η περίπτωση του TP53 αναδεικνύει τη δυνατότητα του εργαλείου να διαχειρίζεται γονίδια με πολύπλοκες και πολλαπλές φαινοτυπικές συσχετίσεις.

Τέλος, το γονίδιο CFTR που σχετίζεται κυρίως με την κυστική ίνωση και άλλες διαταραχές του αναπνευστικού και πεπτικού συστήματος παρουσιάζει -σε αντίθεση με τα προηγούμενα γονίδια- πιο εξειδικευμένο φαινοτυπικό προφίλ. Τα αποτελέσματα του εργαλείου ανέδειξαν σχετικούς φαινοτυπικούς όρους, ενώ η βιβλιογραφική ανάκτηση από την PubMed παρέχει σχετική πληροφορία για τη σύνδεση του γονιδίου με συγκεκριμένες παθολογικές καταστάσεις. Η περίπτωση αυτή δείχνει ότι το εργαλείο είναι ικανό να αποτυπώσει τόσο γενικευμένα όσο και εξειδικευμένα πρότυπα φαινοτυπικών συσχετίσεων.

Γονίδιο	Αριθμός φαινοτύπων (HPO)	Αριθμός άρθρων (PubMed)	Παρατηρήσεις
BRCA1	Υψηλός (166)	20	Σχετίζεται με πολλαπλές μορφές καρκίνου/πολλοί φαινότυποι
TP53	Υψηλός (224)	20	Πολύπλοκες και εκτεταμένες φαινοτυπικές συσχετίσεις
CFTR	Μέτριος (103)	20	Πιο εξειδικευμένο φαινοτυπικό προφίλ

Πίνακας 1. Παρουσίαση αποτελεσμάτων εκτέλεσης R-εργαλείου για ενδεικτικά γονίδια BRCA1, TP53 και CFTR.



Σχήμα 2. Αναπαράσταση αποτελεσμάτων για τα γονίδια BRCA1, TP53 και CFTR.

Το **Σχήμα 2** παρουσιάζει τη σύγκριση του αριθμού φαινοτυπικών όρων και βιβλιογραφικών αναφορών για τα εξεταζόμενα γονίδια. Παρατηρείται σημαντική διαφοροποίηση στον αριθμό των φαινοτύπων, με τα γονίδια BRCA1 και TP53 να εμφανίζουν αυξημένο αριθμό συσχετίσεων, γεγονός που αντανακλά την εκτεταμένη μελέτη τους στη βιβλιογραφία, ενώ το CFTR παρουσιάζει πιο περιορισμένο αλλά πιο εξειδικευμένο προφίλ.

Αντίθετα, ο αριθμός των βιβλιογραφικών αναφορών παραμένει σταθερός, καθώς έχει τεθεί ανώτατο όριο ανάκτησης (20 άρθρα ανά γονίδιο), γεγονός που δεν επιτρέπει την άμεση σύγκριση της βιβλιογραφικής κάλυψης μεταξύ των γονιδίων.

6.3 Αξιολόγηση αποτελεσμάτων

Πληρότητα

Η πληρότητα των αποτελεσμάτων εξαρτάται από την κάλυψη των εξωτερικών πηγών δεδομένων. Η HPO παρέχει επιμελημένες συσχετίσεις γονιδίων - φαινοτύπων, ωστόσο δεν περιλαμβάνει όλες τις πιθανές φαινοτυπικές εκδηλώσεις για κάθε γονίδιο. Αντίστοιχα, η PubMed περιλαμβάνει

μεγάλο όγκο βιβλιογραφίας, αλλά η ανάκτηση σχετικών άρθρων εξαρτάται από τη στρατηγική αναζήτησης και τον περιορισμό του αριθμού των επιστρεφόμενων αποτελεσμάτων.

Ακρίβεια

Η ακρίβεια των αποτελεσμάτων επηρεάζεται από την ποιότητα των δεδομένων και τη μεθοδολογία αναζήτησης. Οι HPO συσχετίσεις θεωρούνται υψηλής αξιοπιστίας, καθώς βασίζονται σε επιμελημένα δεδομένα, ωστόσο το αρχείο genes_to_phenotype.txt απαιτεί τακτική ενημέρωση ώστε να αντανακλά την πιο πρόσφατη διαθέσιμη γνώση. Αντίθετα, τα αποτελέσματα από την PubMed μπορεί να περιλαμβάνουν γενικές ή μη άμεσα σχετικές αναφορές, λόγω της χρήσης απλών ερωτημάτων αναζήτησης, της απουσίας μηχανισμών αξιολόγησης συνάφειας και του περιορισμένου αριθμού αποτελεσμάτων (ανώτατο όριο 20 άρθρων ανά γονίδιο), γεγονός που μπορεί να επηρεάσει την αντιπροσωπευτικότητα των δεδομένων.

Περιορισμοί

Οι παραπάνω παράγοντες επηρεάζουν τη συνολική ποιότητα των αποτελεσμάτων και συνδέονται άμεσα με τους περιορισμούς της μεθοδολογίας (βλ. **Κεφάλαιο 4.8**), οι οποίοι αναλύονται εκτενέστερα στο **Κεφάλαιο 7.2** (Περιορισμοί), για αυτό θα πρέπει να λαμβάνονται υπόψιν κατά την ερμηνεία των αποτελεσμάτων.

Συζήτηση

Το παρόν κεφάλαιο παρουσιάζει την κριτική ανάλυση των αποτελεσμάτων που προέκυψαν από την εφαρμογή του εργαλείου. Η συζήτηση επικεντρώνεται στα πλεονεκτήματα και τους περιορισμούς της προτεινόμενης μεθοδολογίας, καθώς και στη σύγκρισή της με υπάρχοντα εργαλεία βιοπληροφορικής.

7.1 Πλεονεκτήματα του εργαλείου

Το προτεινόμενο εργαλείο παρουσιάζει σημαντικά πλεονεκτήματα ως προς την αυτοματοποίηση της διαδικασίας αντιστοίχισης γονιδίων με φαινοτυπικούς όρους και βιβλιογραφικά δεδομένα. Η ενσωμάτωση δεδομένων από την Human Phenotype Ontology (HPO) και την PubMed επιτρέπει τον συνδυασμό επιμελημένης γνώσης και βιβλιογραφικής πληροφορίας, κάτι που αποτελεί βασική προσέγγιση στη σύγχρονη βιοπληροφορική. (Köhler et al., 2021; Sayers et al., 2023)

Ένα σημαντικό πλεονέκτημα είναι η δυνατότητα αυτοματοποιημένης ανάλυσης πολλαπλών γονιδίων μέσω ενός ενιαίου pipeline, μειώνοντας την ανάγκη για χειροκίνητη αναζήτηση και επεξεργασία δεδομένων. Η προσέγγιση αυτή ευθυγραμμίζεται με πρακτικές που χρησιμοποιούνται σε μεγάλης κλίμακας βιοϊατρικές αναλύσεις. (Huber et al., 2015)

Επιπλέον, η χρήση τυποποιημένων φαινοτυπικών όρων μέσω της HPO διασφαλίζει συνέπεια και διαλειτουργικότητα των δεδομένων, διευκολύνοντας την ερμηνεία και τη σύγκριση των αποτελεσμάτων. (Robinson et al., 2014)

Άλλα προτερήματα της χρήσης του εργαλείου στην αναζήτηση από τον ερευνητή είναι η συγκέντρωση γνώσης σε ένα σημείο (παράγει αρχεία δομημένων αποτελεσμάτων CSV).

Το εργαλείο μπορεί να αξιοποιηθεί ως μέσο αρχικής διερεύνησης (exploratory analysis), υποστηρίζοντας τον ερευνητή στον εντοπισμό πιθανών σχέσεων και στην κατανόηση της βιολογικής σημασίας ενός γονιδίου, χωρίς να αντικαθιστά εξειδικευμένες αναλυτικές μεθόδους.

Όσον αφορά τα τεχνικά του πλεονεκτήματα, το εργαλείο υλοποιήθηκε με modular αρχιτεκτονική, όπου κάθε λειτουργικό στάδιο αντιστοιχεί σε ξεχωριστές συναρτήσεις και αρχεία κώδικα. Η προσέγγιση αυτή διευκολύνει την κατανόηση, τη συντήρηση και την περαιτέρω ανάπτυξη του λογισμικού καθώς επιτρέπει την εύκολη ενσωμάτωση επιπλέον πηγών δεδομένων.

Ακόμη, βασίζεται σε pipeline αρχιτεκτονική, επιτρέποντας την αυτοματοποιημένη επεξεργασία πολλαπλών γονιδίων μέσω της συνάρτησης lapply της R. Με τον τρόπο αυτό, το εργαλείο μπορεί να εφαρμοστεί σε σύνολα μεγάλου μεγέθους χωρίς τροποποίηση του κώδικα.

Για τη βελτίωση της υπολογιστικής απόδοσης, χρησιμοποιείται μηχανισμός προσωρινής αποθήκευσης (caching), ώστε δεδομένα όπως οι HPO συσχετίσεις να φορτώνονται μία φορά στη μνήμη και να επαναχρησιμοποιούνται κατά την ανάλυση. Αυτό μειώνει σημαντικά τον χρόνο εκτέλεσης.

Τέλος, η δομή του εργαλείου επιτρέπει την εύκολη επέκτασή του (extensibility), καθώς νέες βάσεις δεδομένων μπορούν να ενσωματωθούν μέσω ανεξάρτητων συναρτήσεων χωρίς σημαντικές αλλαγές στη βασική λογική του pipeline.

7.2 Περιορισμοί

Παρά τα πλεονεκτήματα, το εργαλείο παρουσιάζει ορισμένους περιορισμούς (οι οποίοι αναφέρθηκαν και στα **Κεφάλαια 4.8 και 6**) που σχετίζονται κυρίως με τη φύση των δεδομένων και τη μεθοδολογία ανάλυσης.

Ένας βασικός περιορισμός αφορά τη βιβλιογραφική αναζήτηση στην PubMed. Η χρήση απλών ερωτημάτων όπως για παράδειγμα η αναζήτηση με βάση το gene symbol σε τίτλο και abstract (gene[Title/Abstract]) μπορεί να οδηγήσει σε μη ειδικά ή γενικά αποτελέσματα. Η μη αξιοποίηση προηγμένων τεχνικών αναζήτησης, όπως οι MeSH όροι και οι λογικοί τελεστές (Boolean operators), περιορίζει την ακρίβεια των ανακτηθέντων άρθρων. Επιπλέον, η επιβολή ανώτατου ορίου στον αριθμό των επιστρεφόμενων αποτελεσμάτων (στο παρόν R – εργαλείο, 20) ενδέχεται να επηρεάσει την αντιπροσωπευτικότητα της βιβλιογραφίας. (Sayers et al., 2023)

Ένας δεύτερος περιορισμός αφορά την ελλιπή κάλυψη της διαθέσιμης βιβλιογραφίας. Παρότι η PubMed αποτελεί μία από τις μεγαλύτερες βάσεις βιοϊατρικής πληροφορίας, δεν περιλαμβάνει το σύνολο της παγκόσμιας επιστημονικής παραγωγής, γεγονός που μπορεί να οδηγήσει σε ελλιπή αναπαράσταση ορισμένων γονιδίων. (Lu, 2011)

Επιπλέον, υπάρχουν γλωσσικοί περιορισμοί, καθώς η πλειονότητα των άρθρων στη PubMed είναι στην αγγλική γλώσσα, γεγονός που μπορεί να περιορίζει την πρόσβαση σε μη αγγλόφωνη βιβλιογραφία και να εισάγει πιθανή μεροληψία στα αποτελέσματα. (Lu, 2011)

Ακόμη, οι συσχετίσεις γονιδίων–φαινοτύπων που προέρχονται από την HPO βασίζονται σε επιμελημένα annotations και δεν υποδηλώνουν απαραίτητα άμεση αιτιολογική σχέση για κάθε επιμέρους φαινότυπο, ενώ η πληρότητα και επικαιρότητα των δεδομένων εξαρτάται από τη συνεχή ενημέρωση της βάσης.

Εξίσου σημαντικοί περιορισμοί είναι πως το εργαλείο δεν πραγματοποιεί σημασιολογική ανάλυση του περιεχομένου των άρθρων (text mining), ούτε συνδέει άμεσα φαινοτυπικούς όρους με συγκεκριμένες βιβλιογραφικές αναφορές. Η ανάκτηση των δεδομένων βασίζεται αποκλειστικά σε λέξεις-κλειδιά, γεγονός που περιορίζει την ικανότητα εντοπισμού πιο σύνθετων σχέσεων μεταξύ γονιδίων και φαινοτύπων· ενώ οι δύο πηγές δεδομένων (HPO και PubMed) λειτουργούν συμπληρωματικά και ανεξάρτητα αλλά όχι συνδυαστικά σε επίπεδο άμεσης αντιστοίχισης της βιβλιογραφίας με τους HPO όρους. Με απλά λόγια, η HPO χρησιμοποιείται για την ανάκτηση

δομημένων φαινοτυπικών συσχετίσεων, ενώ η PubMed για την ανάκτηση σχετικής βιβλιογραφίας που αφορά το συγκεκριμένο γονίδιο, χωρίς να υπάρχει μηχανισμός άμεσης σύνδεσης συγκεκριμένων άρθρων με συγκεκριμένους φαινοτυπικούς όρους.

Ακόμη, το παρόν εργαλείο δεν αποτελεί διαγνωστικό ή κλινικό εργαλείο ούτε εντοπίζει νέες βιολογικές συσχετίσεις παρά μόνο οργανώνει ήδη υπάρχουσα γνώση, ενώ επίσης δεν ενοποιεί τα αποτελέσματα σε έναν ενιαίο τελικό πίνακα.

7.3 Σύγκριση με υπάρχοντα εργαλεία

Το προτεινόμενο εργαλείο παρουσιάζει ομοιότητες αλλά και σημαντικές διαφορές σε σχέση με υπάρχοντα εργαλεία βιοπληροφορικής.

Η βάση δεδομένων OMIM (Online Mendelian Inheritance in Man) παρέχει επιμελημένες και αξιόπιστες πληροφορίες για τη σχέση γονιδίων και ασθενειών, βασισμένες σε εκτενή βιβλιογραφική τεκμηρίωση. Ωστόσο, η πρόσβαση σε πλήρη δεδομένα μπορεί να είναι περιορισμένη και δεν παρέχεται άμεση δυνατότητα αυτοματοποιημένης ανάλυσης μέσω pipeline, γεγονός που περιορίζει τη χρήση της σε μεγάλης κλίμακας επεξεργασία δεδομένων. (Amberger et al., 2019)

Το DisGeNET αποτελεί μια ολοκληρωμένη πλατφόρμα που ενσωματώνει δεδομένα από πολλαπλές πηγές και προσφέρει εκτεταμένες συσχετίσεις μεταξύ γονιδίων και ασθενειών. Παρόλα αυτά, η πολυπλοκότητα της δομής του και η ανάγκη για εξειδικευμένη επεξεργασία μπορεί να περιορίζουν την απλή, άμεση χρήση του σε απλούστερα pipelines για γρήγορη διερεύνηση. (Piñero et al., 2020)

Το εργαλείο Phenolyzer χρησιμοποιεί προηγμένες τεχνικές, όπως text mining και δικτυακή ανάλυση, για την πρόβλεψη γονιδίων με βάση φαινοτυπικές πληροφορίες. Σε αντίθεση με αυτό, το προτεινόμενο εργαλείο υιοθετεί μια απλούστερη και πιο διαφανή προσέγγιση, η οποία βασίζεται σε άμεση ανάκτηση και οργάνωση δεδομένων χωρίς σύνθετα υπολογιστικά μοντέλα. (Yang et al., 2015)

Συνολικά, το προτεινόμενο εργαλείο δεν στοχεύει να αντικαταστήσει τα παραπάνω συστήματα, αλλά να προσφέρει μια απλή, επαναλήψιμη, αυτοματοποιημένη, κατανοητή και εύχρηστη προσέγγιση για την αρχική διερεύνηση (exploratory analysis) σχέσεων γονιδίων και φαινοτύπων.

Συμπεράσματα και Μελλοντική Εργασία

8.1 Συμπεράσματα

Στην παρούσα εργασία αναπτύχθηκε ένα υπολογιστικό εργαλείο σε γλώσσα R για την αυτοματοποιημένη αντιστοίχιση γονιδίων με φαινοτυπικούς όρους και σχετική βιβλιογραφία. Το εργαλείο βασίζεται στην ανάκτηση δεδομένων από την Human Phenotype Ontology (HPO) και τη βάση PubMed, επιτρέποντας την ταχεία και οργανωμένη πρόσβαση σε πληροφορία που είναι κατανεμημένη σε διαφορετικές πηγές. Η προσέγγιση αυτή ευθυγραμμίζεται με σύγχρονες πρακτικές στη βιοπληροφορική, όπου η ενσωμάτωση ετερογενών δεδομένων αποτελεί βασική πρόκληση. (Köhler et al., 2021; Sayers et al., 2023)

Το εργαλείο επιτυγχάνει την αυτοματοποίηση της διαδικασίας ανάλυσης, επιτρέποντας την επεξεργασία πολλαπλών γονιδίων μέσω ενός επαναλήψιμου pipeline. Η δυνατότητα αυτή μειώνει σημαντικά τον χρόνο και την προσπάθεια που απαιτείται για χειροκίνητη αναζήτηση και επεξεργασία δεδομένων, ενώ ταυτόχρονα διασφαλίζει τη συνέπεια των αποτελεσμάτων. Αντίστοιχες προσεγγίσεις αυτοματοποιημένης ανάλυσης χρησιμοποιούνται ευρέως σε σύγχρονες βιοπληροφορικές εφαρμογές. (Huber et al., 2015)

Η κύρια συνεισφορά της εργασίας έγκειται στην ανάπτυξη ενός απλού, επεκτάσιμου και αναπαραγώγιμου εργαλείου που μπορεί να χρησιμοποιηθεί για αρχική διερεύνηση (exploratory analysis) σχέσεων γονιδίων και φαινοτύπων. Σε αντίθεση με πιο σύνθετα εργαλεία, η προτεινόμενη προσέγγιση δίνει έμφαση στη διαφάνεια και την ευκολία χρήσης, στοιχεία που θεωρούνται σημαντικά για την ευρεία αξιοποίηση υπολογιστικών εργαλείων στη βιοϊατρική έρευνα. (Gentleman et al., 2004)

8.2 Μελλοντικές επεκτάσεις

Παρά τη λειτουργικότητα του εργαλείου, υπάρχουν σημαντικές δυνατότητες επέκτασης που θα μπορούσαν να ενισχύσουν περαιτέρω τη χρησιμότητά του.

8.2.1 Ενσωμάτωση NLP (Natural Language Processing)

Μια σημαντική κατεύθυνση αφορά την ενσωμάτωση τεχνικών επεξεργασίας φυσικής γλώσσας (NLP) για την ανάλυση του περιεχομένου των abstracts. Σύγχρονες μελέτες δείχνουν ότι μοντέλα βασισμένα σε large language models μπορούν να χρησιμοποιηθούν για την αναγνώριση και

εξαγωγή βιοϊατρικών οντοτήτων από επιστημονικά κείμενα με υψηλή ακρίβεια, ακόμη και πέρα από τους ήδη καταγεγραμμένους όρους της HPO. Η ενσωμάτωση τέτοιων τεχνικών θα μπορούσε να επιτρέψει την πιο άμεση συσχέτιση βιβλιογραφικών δεδομένων με φαινοτυπικές πληροφορίες. (Yang et al., 2024)

8.2.2 Machine Learning για gene–phenotype συσχετίσεις

Η ενσωμάτωση μεθόδων μηχανικής μάθησης (Machine Learning) αποτελεί μια ακόμη πιθανή μελλοντική επέκταση του εργαλείου. Σύγχρονες προσεγγίσεις αξιοποιούν συνδυασμό γονιδιωματικών και φαινοτυπικών δεδομένων για την υποστήριξη της αναγνώρισης γονιδίων που σχετίζονται με συγκεκριμένες ασθένειες και φαινοτυπικά πρότυπα. Ιδιαίτερα, πρόσφατες μελέτες έχουν δείξει ότι μοντέλα βασισμένα σε machine learning και πιο συγκεκριμένα deep learning, μπορούν να αξιοποιήσουν φαινοτυπικές περιγραφές για την ιεράρχηση υποψήφιων γονιδίων και την υποστήριξη της διαδικασίας διάγνωσης σε σπάνιες γενετικές νόσους. Στο πλαίσιο αυτό, η δυνατότητα ενσωμάτωσης τέτοιων μεθόδων πιθανώς θα μπορούσε να ενισχύσει τον αναλυτικό χαρακτήρα του εργαλείου, επεκτείνοντάς το πέρα από την απλή περιγραφή των συσχετίσεων (Alsentzer et al., 2025)

8.2.3 Ανάπτυξη Web εφαρμογής

Η μετατροπή του εργαλείου σε διαδικτυακή εφαρμογή θα μπορούσε να διευκολύνει σημαντικά τη χρήση του από μη εξειδικευμένους χρήστες και ερευνητές χωρίς προγραμματιστικές γνώσεις. Στη βιοπληροφορική, αναπτύσσονται και χρησιμοποιούνται Web-based εφαρμογές που εκτελούνται απευθείας μέσω φυλλομετρητή (Web browser), χωρίς την ανάγκη τοπικής εγκατάστασης λογισμικού. Η προσέγγιση αυτή διευκολύνει την πρόσβαση και τη χρήση υπολογιστικών εργαλείων από τους ερευνητές (Yachdav et al., 2016).

8.2.4 Ενσωμάτωση κλινικών δεδομένων

Η ενσωμάτωση δεδομένων από ηλεκτρονικούς ιατρικούς φακέλους (Electronic Health Records – EHRs) αποτελεί σημαντική κατεύθυνση για την υποστήριξη της ανάλυσης φαινοτυπικών δεδομένων. Σύγχρονες μελέτες δείχνουν ότι η αξιοποίηση κλινικών δεδομένων σε συνδυασμό με μεθόδους μηχανικής μάθησης μπορεί να συμβάλει στον πιο ακριβή ορισμό και την αναγνώριση σύνθετων φαινοτύπων, βασισμένων σε πραγματικά κλινικά δεδομένα (Binkheder et al., 2025). Η ενσωμάτωση τέτοιων δεδομένων πιθανώς θα μπορούσε να ενισχύσει την αναλυτική δυνατότητα του εργαλείου ως προς τη φαινοτυπική περιγραφή.

Επιπροσθέτως, η ενσωμάτωση κλινικών δεδομένων θα μπορούσε να ενισχύσει τη χρησιμότητα του εργαλείου σε εφαρμογές ιατρικής ακριβείας, όπου ο συνδυασμός γενετικών και φαινοτυπικών δεδομένων είναι κρίσιμος για την κατανόηση των ασθενειών. (Ashley, 2016)

8.2.5 Οπτικοποίηση αποτελεσμάτων

Η προσθήκη δυνατοτήτων οπτικοποίησης των αποτελεσμάτων θα μπορούσε να βελτιώσει την ερμηνεία των δεδομένων, καθώς η οπτικοποίηση αποτελεί βασικό εργαλείο για την ανάλυση πολύπλοκων δεδομένων και την ανάδειξη προτύπων (Heer et al., 2010), γεγονός που την καθιστά ιδιαίτερα χρήσιμη και σε βιοπληροφορικές εφαρμογές.

8.2.6 Ενσωμάτωση δεδομένων ασθενειών

Ακόμη, μια σημαντική επέκταση θα ήταν η ενσωμάτωση δεδομένων ασθενειών, ώστε να επιτευχθεί η σύνδεση gene–disease–phenotype. Η προσέγγιση αυτή χρησιμοποιείται ευρέως στη βιοπληροφορική για την κατανόηση των μοριακών μηχανισμών των ασθενειών και των σχέσεων τους με γονίδια και φαινότυπους. (Köhler et al., 2021, Piñero et al., 2020)

8.2.7 Σύνθετα queries με MeSH όρους

Η βελτίωση της στρατηγικής αναζήτησης στη PubMed, μέσω χρήσης MeSH όρων ή σύνθετων ερωτημάτων αναζήτησης (queries), θα μπορούσε να αυξήσει την ακρίβεια και τη συνάφεια των βιβλιογραφικών αποτελεσμάτων (Sayers et al., 2023). Ακόμη, η παραμετροποίηση των αναζητήσεων ως προς χρονικά ή θεματικά κριτήρια, με μικρή αλλαγή στον κώδικα, θα μπορούσε να ενισχύσει την ευελιξία του εργαλείου και να προσαρμόσει τα αποτελέσματα στις ανάγκες του χρήστη.

8.2.8 Λοιπές βελτιώσεις που προκύπτουν από τεχνικούς περιορισμούς του R-εργαλείου

Άλλη εξίσου σημαντική βελτίωση θα μπορούσε να είναι η εφαρμογή **μεθόδων αξιολόγησης συνάφειας (relevance scoring)**, ώστε να ιεραρχούνται οι φαινοτυπικές συσχετίσεις και τα βιβλιογραφικά αποτελέσματα με βάση τη σημασία τους.

Επιπλέον, η υλοποίηση **δυναμικής ενημέρωσης των δεδομένων από τις αντίστοιχες βάσεις** (π.χ. HPO) θα μπορούσε να διασφαλίσει ότι τα αποτελέσματα παραμένουν επικαιροποιημένα, δεδομένου ότι στο παρόν εργαλείο χρησιμοποιούνται στατικά αρχεία συσχετίσεων.

Μια ακόμη πιθανή επέκταση θα ήταν η παραμετροποίηση του ορίου ανάκτησης βιβλιογραφικών αποτελεσμάτων ή η υλοποίηση μηχανισμού διαδοχικής ανάκτησης (pagination), ώστε να αυξηθεί η πληρότητα των βιβλιογραφικών δεδομένων.

Ενώ τέλος, για τη βελτίωση της απόδοσης σε μεγάλα σύνολα γονιδίων, θα μπορούσε να προστεθεί υποστήριξη **παράλληλης εκτέλεσης** σε αντίθεση με την τρέχουσα σειριακή εκτέλεση όπου έχει μεγαλύτερη καθυστέρηση και πιο αργή εκτέλεση.

Βιβλιογραφικές αναφορές

- Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: From polygenic to omnigenic. *Cell, 169*(7), 1177–1186. https://doi.org/10.1016/j.cell.2017.05.038
- Robinson, P. N., Köhler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The Human Phenotype Ontology: A tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics, 83*(5), 610–615. https://doi.org/10.1016/j.ajhg.2008.09.017
- Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., & Losick, R. (2014). *Molecular Biology of the Gene* (7th ed.). Pearson.
- Robinson, P. N., Köhler, S., Oellrich, A., Wang, K., Mungall, C. J., Lewis, S. E., et al. (2014). Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Research, 24*(2), 340–348. https://doi.org/10.1101/gr.160325.113
- Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L. C., Lewis-Smith, D., Vasilevsky, N. A., et al. (2021). The Human Phenotype Ontology in 2021. *Nucleic Acids Research, 49*(D1), D1207–D1217. https://doi.org/10.1093/nar/gkaa1043
- Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., et al. (2023). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research, 51*(D1), D29–D38. https://doi.org/10.1093/nar/gkac1032
- Lipscomb, C. E. (2000). Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association, 88*(3), 265–266.

- Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Bretscher, A., Ploegh, H., Amon, A., & Scott, M. P. (2021). *Molecular Cell Biology* (9th ed.). W. H. Freeman.
- Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), R80.
- R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
- Winter DJ. 2017. rentrez: An R package for the NCBI eUtils API. *PeerJ Preprints* 5:e3179v2 https://doi.org/10.7287/peerj.preprints.3179v2
- Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2022). *Molecular Biology of the Cell* (7th ed.). Garland Science.
- Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods, 12*(2), 115–121. https://doi.org/10.1038/nmeth.3252
- Lu, Z. (2011). PubMed and beyond: a survey of web tools for searching biomedical literature. *Database*, *2011*, baq036.
- Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, *48*(D1), D845-D855.
- Strachan, T. and Read, A., (2018). *Human Molecular Genetics*. 5th ed. New York: Garland Science. DOI: 10.4324/9780203833544-1
- Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics*, *17*(9), 507-522.
- Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Communications of the ACM, 53*(6), 59–67.
- Yang, H., Robinson, P. N., & Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nature methods*, *12*(9), 841-843.

- Guy Yachdav, Sebastian Wilzbach, et al. MSAViewer: interactive JavaScript visualization of multiple sequence alignments, *Bioinformatics*, Volume 32, Issue 22, November 2016, Pages 3501–3503, https://doi.org/10.1093/bioinformatics/btw474
- Yang, J., Liu, C., Deng, W., et al. (2024). Enhancing phenotype recognition in clinical notes using large language models: PhenoBCBERT and PhenoGPT. *Patterns*, *5*(1).
- Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., & Doebley, J. (2020). *Introduction to Genetic Analysis* (12th ed.). W. H. Freeman.
- Alsentzer, E., Li, M. M., Kobren, et al. (2025). Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. *npj Digital Medicine*, *8*(1), 380.
- Binkheder, S., Liu, X., Wu, et al. (2025). Biomedical literature-based clinical phenotype definition discovery using large language models. *Database*, *Volume 2025, 2025,* baaf047, https://doi.org/10.1093/database/baaf047
- Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: Leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Research, 47*(D1), D1038–D1043. https://doi.org/10.1093/nar/gky1151
- Bodenreider, O. (2008). Biomedical ontologies in action: Role in knowledge management, data integration and decision support. *Yearbook of Medical Informatics, 17*(01), 67–79. https://doi.org/10.1055/s-0038-1638582
- Smith, B., Ashburner, M., et al. (2007). The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology, 25*(11), 1251–1255. https://doi.org/10.1038/nbt1346
- Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: A grammar of data manipulation (Version 1.2.0) [Computer software]. https://CRAN.R-project.org/package=dplyr
- Wickham, H., Hester, J., & Bryan, J. (2023). readr: Read rectangular text data (Version 2.2.0) [Computer software]. https://CRAN.R-project.org/package=readr

Παραρτήματα

Παράρτημα Α: Κώδικας R

Στο παράρτημα αυτό παρατίθεται ο κώδικας του εργαλείου που αναπτύχθηκε στην παρούσα διπλωματική εργασία. Ο κώδικας έχει οργανωθεί σε επιμέρους αρχεία, σύμφωνα με αρθρωτή (modular) αρχιτεκτονική, ώστε κάθε λειτουργικό στάδιο του pipeline να υλοποιείται από ξεχωριστές συναρτήσεις.

Η συνολική δομή του κώδικα περιλαμβάνει τα ακόλουθα αρχεία:

- input_functions.R: περιλαμβάνει συναρτήσεις για την ανάγνωση και προεπεξεργασία των γονιδίων εισόδου.
- pubmed_functions.R: περιλαμβάνει συναρτήσεις για την αναζήτηση βιβλιογραφικών δεδομένων στη PubMed μέσω του Entrez API.
- hpo_functions.R: περιλαμβάνει συναρτήσεις για την ανάκτηση και αντιστοίχιση φαινοτυπικών όρων από τα αρχεία συσχέτισης της HPO.
- processing_functions.R: περιλαμβάνει συναρτήσεις για την κανονικοποίηση, οργάνωση και απομάκρυνση διπλότυπων εγγραφών.
- pipeline.R: περιλαμβάνει την κεντρική λογική του εργαλείου και τις συναρτήσεις ανάλυσης ανά γονίδιο.
- output_formatting.R: περιλαμβάνει συναρτήσεις για τη μορφοποίηση και εξαγωγή των αποτελεσμάτων.
- run_tool.R: αποτελεί το κύριο αρχείο εκτέλεσης του εργαλείου.

Ο κώδικας του εργαλείου παρατίθεται στις επόμενες ενότητες του παραρτήματος, με τη σειρά που εκτελείται από το κύριο script.

Α.1 Κύριο αρχείο εκτέλεσης (run_tool.R)

```
source("input_functions.R")
source("pubmed_functions.R")
source("hpo_functions.R")
source("processing_functions.R")
source("pipeline.R")
source("output_formatting.R")

results <- gene_phenotype_pipeline("genes.txt")

final_tables <- format_results(results)

write.csv(final_tables$gene_hpo, "gene_hpo_results.csv", row.names = FALSE)
write.csv(final_tables$pubmed,   "pubmed_results.csv",   row.names = FALSE)

cat(" Ολοκληρώθηκε. \n")
cat("-> Δημιουργήθηκαν:\n")
cat("   gene_hpo_results.csv\n")
cat("   pubmed_results.csv\n")
cat("-> Γραμμές:\n")
cat("   HPO:", nrow(final_tables$gene_hpo), "\n")
cat("   PubMed:", nrow(final_tables$pubmed), "\n")
```

Α.2 Συναρτήσεις εισόδου (input_functions.R)

```
load_genes <- function(input_file) {
  genes <- readLines(input_file, warn = FALSE)

  genes <- trimws(genes)
  genes <- genes[genes !=""]
  genes <- toupper(genes)

  unique(genes)
}
```

Α.3 Συναρτήσεις ανάκτησης PubMed (pubmed_functions.R)

```
library(rentrez)
library(dplyr)

search_pubmed <- function(gene, retmax = 20) {
 query <- paste0(gene, "[Title/Abstract]")
 search <- entrez_search(
  db = "pubmed",
  term = query,
  retmax = retmax
 )
 if(length(search$ids) ==0) {
  return(data.frame())
 }
 summaries <- entrez_summary(db = "pubmed", id=search$ids)

 data.frame(
  gene=gene,
  pmid = sapply(summaries, function(x) x$uid),
  title = sapply(summaries, function(x) x$title),
  year = sapply(summaries, function(x) x$pubdate),
  stringsAsFactors = FALSE
 )
}
```

Α.4 Συναρτήσεις ανάκτησης HPO (hpo_functions.R)

```
library(dplyr)
library(readr)

# Φόρτωση mapping μία φορά (cache)
.load_hpo_map <- local({
 cache <- NULL
 function(filepath = "genes_to_phenotype.txt") {
  if (!is.null(cache)) return(cache)

  df <- readr::read_tsv(filepath, comment = "#", show_col_types = FALSE)
  names(df) <- tolower(names(df))
```

```
# Προσαρμογή σε συνηθισμένα ονόματα στηλών
  if (!("gene" %in% names(df))) {
   if ("gene_symbol" %in% names(df)) df <- df %>% rename(gene = gene_symbol)
   if ("symbol" %in% names(df))     df <- df %>% rename(gene = symbol)
  }
  if (!("hpo_id" %in% names(df))) {
   if ("hpo" %in% names(df))       df <- df %>% rename(hpo_id = hpo)
   if ("hpoid" %in% names(df))     df <- df %>% rename(hpo_id = hpoid)
  }
  if (!("hpo_name" %in% names(df))) {
   if ("term" %in% names(df))      df <- df %>% rename(hpo_name = term)
   if ("phenotype" %in% names(df))  df <- df %>% rename(hpo_name = phenotype)
   if ("name" %in% names(df))      df <- df %>% rename(hpo_name = name)
  }

  cache <<- df %>%
   transmute(
    gene = toupper(as.character(gene)),
    hpo_id = as.character(hpo_id),
    hpo_name = as.character(hpo_name)
   ) %>%
   distinct()

  cache
 }
})


get_hpo_terms <- function(gene, filepath = "genes_to_phenotype.txt") {
 gene <- toupper(gene)
 hpo_map <- .load_hpo_map(filepath)

 out <- hpo_map %>% filter(gene == !!gene)

 if (nrow(out) == 0) {
  return(data.frame(
   gene = gene,
```

```
    hpo_id = character(),
    hpo_name = character(),
    stringsAsFactors = FALSE
   ))
 }
 out
}
```

A.5 Συναρτήσεις επεξεργασίας (processing_functions.R)

```
library(dplyr)
process_results <- function(gene, pubmed_data, hpo_data) {
 gene <- toupper(gene)
 if(nrow(hpo_data) > 0) {
  hpo_data <- hpo_data %>%
   mutate(
    gene = gene,
    hpo_id = as.character(hpo_id),
    hpo_name = as.character(hpo_name)
   ) %>%
   distinct()
 }
 if(nrow(pubmed_data) > 0) {
  pubmed_data <- pubmed_data %>%
   mutate(gene = gene) %>%
   distinct()
 }
 list(
  gene = gene,
  hpo = hpo_data,
  pubmed = pubmed_data
 )
}
```

A.6 Κύριο pipeline (pipeline.R)

```
analyze_gene <- function(gene) {
 pubmed_data <- search_pubmed(gene)
 hpo_data <- get_hpo_terms(gene)

 process_results(gene, pubmed_data, hpo_data)
}
gene_phenotype_pipeline <- function(input_genes) {
 genes <- load_genes(input_genes)

 results <- lapply(genes, analyze_gene)
 names(results) <- genes

 results
}
```

A.7 Μορφοποίηση αποτελεσμάτων (output_formatting.R)

```
library(dplyr)

format_results <- function(results) {
 gene_hpo_table <- bind_rows(
  lapply(results, function(x) x$hpo)
 )
 pubmed_table <- bind_rows(
  lapply(results, function(x) x$pubmed)
 )

 list(
  gene_hpo = gene_hpo_table,
  pubmed = pubmed_table
 )
}
```

Παράρτημα Β: Οδηγός χρήσης εργαλείου

Το παρόν παράρτημα περιγράφει συνοπτικά τα βήματα που απαιτούνται για την εκτέλεση του εργαλείου και την παραγωγή των αποτελεσμάτων του.

Β.1 Απαιτήσεις συστήματος

Για την εκτέλεση του εργαλείου απαιτούνται:

- εγκατεστημένη η γλώσσα R,
- περιβάλλον ανάπτυξης RStudio (προαιρετικά, αλλά προτείνεται),
- τα απαραίτητα πακέτα R:
    - rentrez
    - dplyr
    - readr

Η εγκατάσταση των πακέτων πραγματοποιείται με τις εντολές:

- `install.packages("rentrez")`
- `install.packages("dplyr")`
- `install.packages("readr")`

Β.2 Απαραίτητα αρχεία

Στον ίδιο φάκελο πρέπει να βρίσκονται τα εξής αρχεία:

- run_tool.R
- input_functions.R
- pubmed_functions.R
- hpo_functions.R
- processing_functions.R
- pipeline.R
- output_formatting.R

- genes.txt
- genes_to_phenotype.txt

B.3 Προετοιμασία αρχείου εισόδου

Ο χρήστης δημιουργεί ένα αρχείο κειμένου με όνομα genes.txt, στο οποίο καταχωρίζεται ένα gene symbol ανά γραμμή. Ενδεικτικό παράδειγμα (**Εικόνα 3**):

```
BRCA1
TP53
CFTR
```

Εικόνα 3. Αρχείο εισόδου για τα γονίδια BRCA1, TP53 και CFTR. Το εργαλείο διαβάζει το αρχείο αυτό και εφαρμόζει βασικά βήματα προεπεξεργασίας, όπως αφαίρεση περιττών κενών, απομάκρυνση κενών γραμμών, μετατροπή σε κεφαλαία γράμματα και αφαίρεση διπλότυπων εγγραφών.

B.4 Εκτέλεση στο RStudio

Η εκτέλεση του εργαλείου μπορεί να πραγματοποιηθεί μέσα από το RStudio.

Βήματα:

1. Ο χρήστης ανοίγει το RStudio.
2. Ορίζει ως working directory τον φάκελο όπου βρίσκονται τα αρχεία του εργαλείου.
3. Εκτελεί στο Console την εντολή: source("run_tool.R")

Με την εντολή αυτή ενεργοποιείται το κύριο pipeline και παράγονται τα αρχεία αποτελεσμάτων.

B.5 Εκτέλεση μέσω γραμμής εντολών

Εναλλακτικά, το εργαλείο μπορεί να εκτελεστεί μέσω terminal ή command prompt με την εντολή: Rscript run_tool.R

Η προσέγγιση αυτή επιτρέπει την αυτοματοποιημένη εκτέλεση του εργαλείου χωρίς τη χρήση του περιβάλλοντος RStudio.

Β.6 Παραγόμενα αρχεία εξόδου

Μετά την επιτυχή ολοκλήρωση της εκτέλεσης, παράγονται δύο αρχεία αποτελεσμάτων:

- gene_hpo_results.csv
- pubmed_results.csv

Το πρώτο περιλαμβάνει τις αντιστοιχίσεις γονιδίων–φαινοτύπων και το δεύτερο τη σχετική βιβλιογραφική πληροφορία από την PubMed.

Β.7 Ενδεικτικό μήνυμα ολοκλήρωσης

Μετά την εκτέλεση του εργαλείου, ο χρήστης μπορεί να δει μήνυμα επιτυχούς ολοκλήρωσης στο Console, όπως (**Εικόνα 4**):

Ολοκληρώθηκε.

Δημιουργήθηκαν:

gene_hpo_results.csv

pubmed_results.csv

```
Ολοκληρώθηκε.
Δημιουργήθηκαν:
gene_hpo_results.csv
pubmed_results.csv
Γραμμές:
HPO: 493
PubMed: 60
```

Εικόνα 4. Μήνυμα επιτυχούς ολοκλήρωσης για τα γονίδια BRCA1, TP53 και CFTR.

Β.8 Πιθανά προβλήματα κατά την εκτέλεση

Κατά την εκτέλεση ενδέχεται να εμφανιστούν προβλήματα όπως:

- λανθασμένος φάκελος εργασίας,
- απουσία απαιτούμενων πακέτων,
- λανθασμένη κωδικοποίηση αρχείων,
- κενό ή λανθασμένο αρχείο εισόδου,
- απουσία του αρχείου genes_to_phenotype.txt.

Η σωστή οργάνωση των αρχείων και η εγκατάσταση των πακέτων εξασφαλίζουν την ομαλή λειτουργία του εργαλείου.