**ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ**

**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ**



**Βιοπληροφορική**

**Θέμα Εργασίας: Δεδομένα - Ρυθμιστικά δίκτυα – δείκτες στον καρκίνο του πνεύμονα**

**Ονοματεπώνυμο: Σταμάτης Καλλιπόσης**

**Αριθμός μητρώου: inf2021071**

**e-mail: inf2021071@ionio.gr**

**Ακαδημαϊκό έτος: 2024 - 2025**

**Περιεχόμενα**

[Ενότητα 1: Εισαγωγή 3](#_Toc193690570)

[Ενότητα 2: Περιγραφή των Ανοιχτών Βάσεων Δεδομένων 5](#_Toc193690571)

[Ενότητα 3 : Ανασκόπηση Μεθοδολογιών 6](#_Toc193690572)

[Ενότητα 4: Συνοπτική Παρουσίαση Σημαντικών Βιοδεικτών στον Καρκίνο του Πνεύμονα 7](#_Toc193690573)

[Βιβλιογραφία 9](#_Toc193690574)

# 

# 

# Ενότητα 1: Εισαγωγή

Ο καρκίνος του πνεύμονα αποτελεί μία από τις πιο συχνές και θανατηφόρες μορφές καρκίνου σε παγκόσμιο επίπεδο. Σύμφωνα με πρόσφατα δεδομένα, κάθε χρόνο καταγράφονται εκατομμύρια νέες διαγνώσεις και θάνατοι από καρκίνο του πνεύμονα, γεγονός που τον καθιστά κορυφαία αιτία θανάτου από καρκίνο τόσο στους άνδρες όσο και στις γυναίκες (Siegel et al., 2023). Η υψηλή θνησιμότητα οφείλεται κυρίως στην καθυστερημένη διάγνωση και στις περιορισμένες επιλογές αποτελεσματικής θεραπείας, ειδικά στα προχωρημένα στάδια της νόσου.

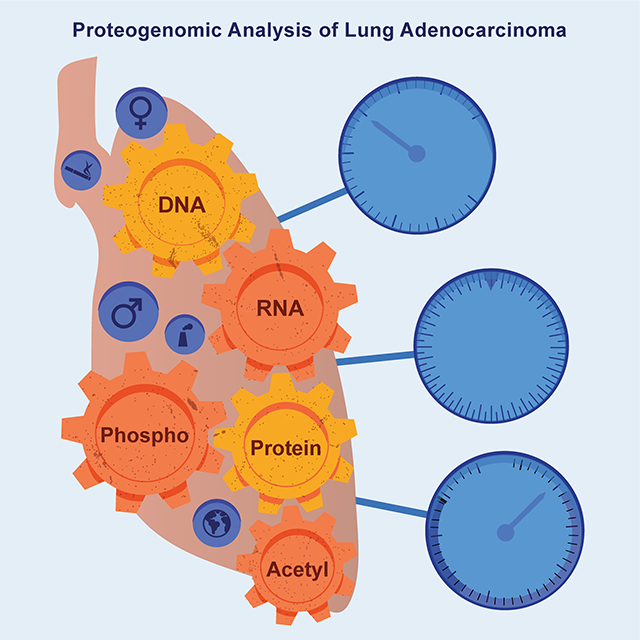
Τα τελευταία χρόνια, η εξέλιξη της τεχνολογίας έχει προσφέρει νέα εργαλεία για την καλύτερη κατανόηση της μοριακής βάσης του καρκίνου του πνεύμονα. Μέθοδοι όπως η γονιδιωματική ανάλυση, η μεταγραφωματική και η πρωτεομική ανάλυση, έχουν συμβάλει στην ανακάλυψη βιοδεικτών (biomarkers) και στην αποκάλυψη ρυθμιστικών δικτύων που είναι κρίσιμα για την ανάπτυξη και εξέλιξη του όγκου (Chen et al., 2021).

Οι βιοδείκτες αποτελούν μοριακά χαρακτηριστικά (γονίδια, πρωτεΐνες ή RNA) που μπορούν να παρέχουν πληροφορίες σχετικά με τη διάγνωση, την πρόγνωση ή την ανταπόκριση ενός ασθενή στη θεραπεία. Παράλληλα, τα ρυθμιστικά δίκτυα αποτελούν ένα σύνολο από γονίδια ή πρωτεΐνες που αλληλεπιδρούν μεταξύ τους, επηρεάζοντας καθοριστικά τις βιολογικές λειτουργίες του κυττάρου. Η κατανόηση αυτών των δικτύων είναι σημαντική, διότι μπορεί να αναδείξει στόχους για νέες θεραπείες ή να βελτιώσει τις ήδη υπάρχουσες (Zhang et al. (2023).

Ένα ιδιαίτερα σημαντικό σημείο της σύγχρονης έρευνας είναι η ενσωμάτωση πολλαπλών ομικών επιπέδων (multi-omics integration), όπου συνδυάζονται δεδομένα από διαφορετικά βιολογικά επίπεδα (DNA, RNA, πρωτεΐνες) με στόχο τη δημιουργία ολοκληρωμένων προφίλ των όγκων. Αυτή η πολυδιάστατη ανάλυση επιτρέπει την ακριβέστερη κατανόηση των μηχανισμών που οδηγούν στην καρκινογένεση, αλλά και τον εντοπισμό νέων βιοδεικτών και θεραπευτικών στόχων που μπορούν να χρησιμοποιηθούν στην εξατομικευμένη ιατρική (Chunchao Zhang).

Η παρούσα εργασία επικεντρώνεται στην ανασκόπηση των πρόσφατων επιστημονικών δεδομένων για τα ρυθμιστικά δίκτυα και τους βιοδείκτες στον καρκίνο του πνεύμονα, αλλά και στην εφαρμογή μεθόδων βιοπληροφορικής για την ανάλυση δεδομένων από ανοιχτές βάσεις, με στόχο την ανάδειξη και επικύρωση μοριακών στόχων με πιθανή κλινική εφαρμογή.

**Εικόνα 1:**

****

(Η εικόνα 1 περιγράφει ένα ολοκληρωμένο σχήμα που δείχνει τη διαδικασία πρωτεογονιδιωματικής ανάλυσης, που συνδυάζει δεδομένα από το γονιδίωμα, μεταγράφημα, πρωτέωμα και μετα-μεταφραστικές τροποποιήσεις για την εύρεση βιοδεικτών και μοριακών μονοπατιών.)

# Ενότητα 2: Περιγραφή των Ανοιχτών Βάσεων Δεδομένων

Η έρευνα στον τομέα του καρκίνου του πνεύμονα, ειδικότερα όσον αφορά τη γονιδιωματική, τη μεταγραφωματική και την πρωτεομική ανάλυση, στηρίζεται σε μεγάλο βαθμό σε δεδομένα υψηλής διαθεσιμότητας. Οι πιο σημαντικές ανοιχτές βάσεις δεδομένων που παρέχουν τέτοιου είδους πληροφορίες είναι το TCGA, το GEO και το CPTAC.

**The Cancer Genome Atlas (TCGA)**

Το TCGA είναι ένα από τα πιο μεγάλα και αναγνωρισμένα προγράμματα, που παρέχει εκτενή δεδομένα πολλαπλών ομικών επιπέδων (γονιδιωματικά, μεταγραφωματικά, επιγενετικά, πρωτεϊνικά) για πάνω από 30 διαφορετικούς τύπους καρκίνου, συμπεριλαμβανομένου του καρκίνου του πνεύμονα. Τα δεδομένα περιλαμβάνουν μεταλλάξεις, αλλαγές στον αριθμό αντιγράφων γονιδίων (copy number variations - CNVs), δεδομένα έκφρασης RNA (RNA-seq), καθώς και πρωτεϊνικά δεδομένα (RPPA - Reverse Phase Protein Array). Το TCGA διαθέτει ειδικά σύνολα δεδομένων για το αδενοκαρκίνωμα πνεύμονα (LUAD) και τον ισχωματοκυτταρικό καρκίνο του πνεύμονα (LUSC), τα οποία είναι διαθέσιμα μέσω του Genomic Data Commons (GDC) [TCGA Portal](https://portal.gdc.cancer.gov/).

**Gene Expression Omnibus (GEO)**

Το GEO είναι μια διεθνώς αναγνωρισμένη βάση δεδομένων του NCBI, η οποία περιέχει δεδομένα έκφρασης γονιδίων από πειράματα μικροσυστοιχιών (microarray) και RNA-seq. Περιλαμβάνει πλήθος από μελέτες στον καρκίνο του πνεύμονα, προσφέροντας πρόσβαση σε μεγάλο αριθμό δειγμάτων και σχετικές κλινικές πληροφορίες (π.χ., τύπος ιστού, στάδιο καρκίνου, θεραπευτική προσέγγιση). Δημοφιλείς σειρές δεδομένων όπως οι GSE31210 και GSE43458 έχουν χρησιμοποιηθεί εκτενώς για τη δημιουργία προγνωστικών υπογραφών και την κατασκευή ρυθμιστικών δικτύων γονιδίων. Τα δεδομένα αυτά είναι προσβάσιμα στην ιστοσελίδα του GEO [GEO Database](https://www.ncbi.nlm.nih.gov/geo/).

**Clinical Proteomic Tumor Analysis Consortium (CPTAC)**

Το CPTAC παρέχει εξειδικευμένα πρωτεομικά δεδομένα που έχουν παραχθεί με προηγμένες τεχνικές μάζας-φασματομετρίας. Έχει εστιάσει στη λεπτομερή πρωτεομική ανάλυση πολλών καρκινικών τύπων, συμπεριλαμβανομένου και του καρκίνου του πνεύμονα. Τα δεδομένα του CPTAC συνδυάζουν γονιδιωματικά και πρωτεομικά προφίλ, συμβάλλοντας έτσι σε μια πιο ολοκληρωμένη κατανόηση των μηχανισμών του καρκίνου. Συγκεκριμένα, για τον καρκίνο του πνεύμονα, παρέχει εκτενείς πρωτεογονιδιωματικές αναλύσεις με στόχο την ανακάλυψη νέων βιοδεικτών και τη βελτίωση της εξατομικευμένης θεραπείας. Τα δεδομένα είναι διαθέσιμα στην ιστοσελίδα του CPTAC [CPTAC Data Portal](https://proteomics.cancer.gov/data-portal).

Η χρήση των παραπάνω βάσεων δεδομένων είναι ζωτικής σημασίας για τη σύγχρονη βιοπληροφορική έρευνα, καθώς επιτρέπει στους ερευνητές να αξιοποιούν εκτεταμένα και καλά οργανωμένα δεδομένα για τη μελέτη του καρκίνου του πνεύμονα, προωθώντας έτσι την ανάπτυξη αποτελεσματικότερων διαγνωστικών και θεραπευτικών προσεγγίσεων.

# Ενότητα 3 : Ανασκόπηση Μεθοδολογιών

Η βιοπληροφορική προσφέρει ένα μεγάλο εύρος τεχνικών και μεθόδων που αξιοποιούνται ευρέως για τη μελέτη των μοριακών μηχανισμών και την ανακάλυψη βιοδεικτών στον καρκίνο του πνεύμονα. Οι κυριότερες από αυτές τις μεθοδολογίες είναι:

**Ανάλυση Διαφορικής Έκφρασης (Differential Expression Analysis)**

Η ανάλυση διαφορικής έκφρασης επιτρέπει τον εντοπισμό γονιδίων που παρουσιάζουν σημαντικές διαφορές στην έκφρασή τους μεταξύ δύο ή περισσότερων ομάδων δειγμάτων (π.χ., καρκινικός έναντι φυσιολογικού ιστού). Δημοφιλή εργαλεία που χρησιμοποιούνται είναι τα DESeq2 για δεδομένα RNA-seq και το limma για δεδομένα μικροσυστοιχιών (microarrays). Αυτές οι μέθοδοι παρέχουν τη βάση για περαιτέρω αναλύσεις, όπως η κατασκευή δικτύων και η αναγνώριση πιθανών βιοδεικτών ([DESeq2](https://bioconductor.org/packages/release/bioc/html/DESeq2.html), [limma](https://bioconductor.org/packages/release/bioc/html/limma.html)).

**Ανάλυση Δικτύων Συν-έκφρασης (Co-expression Networks, WGCNA)**

Η μέθοδος Weighted Gene Co-expression Network Analysis (WGCNA) χρησιμοποιείται για την οργάνωση των γονιδίων σε ομάδες (modules) βάσει παρόμοιων μοτίβων έκφρασης. Αυτές οι ομάδες συχνά σχετίζονται με συγκεκριμένα βιολογικά χαρακτηριστικά, όπως την επιβίωση ή το στάδιο της νόσου. Το WGCNA επιτρέπει επίσης την ταυτοποίηση γονιδίων-κόμβων (hub genes), που έχουν κρίσιμο ρόλο στη ρύθμιση των δικτύων ([WGCNA](https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/)).

**Κατασκευή Δικτύων Αλληλεπίδρασης Πρωτεϊνών (PPI Networks, Cytoscape)**

Η ανάλυση των δικτύων αλληλεπίδρασης πρωτεϊνών είναι ιδιαίτερα χρήσιμη για τον εντοπισμό σημαντικών πρωτεϊνικών αλληλεπιδράσεων που μπορεί να επηρεάζουν κρίσιμα μοριακά μονοπάτια στον καρκίνο του πνεύμονα. Δημοφιλή εργαλεία όπως το Cytoscape χρησιμοποιούνται ευρέως για την κατασκευή, την ανάλυση και την οπτικοποίηση αυτών των δικτύων ([Cytoscape](https://cytoscape.org/)).

**Λειτουργικός Εμπλουτισμός (Functional Enrichment Analysis, DAVID, Enrichr)**

Η ανάλυση λειτουργικού εμπλουτισμού χρησιμοποιείται για να προσδιορίσει τις βιολογικές λειτουργίες και τα μονοπάτια στα οποία συμμετέχουν γονίδια που έχουν αναδειχθεί από προηγούμενες αναλύσεις. Εργαλεία όπως το DAVID και το Enrichr παρέχουν τη δυνατότητα να εντοπίζονται σημαντικά εμπλουτισμένες βιολογικές διεργασίες, μοριακές λειτουργίες και μονοπάτια ([DAVID](https://david.ncifcrf.gov/), [Enrichr](https://maayanlab.cloud/Enrichr/)).

**Ανάλυση Επιβίωσης (Survival Analysis)**

Η στατιστική ανάλυση επιβίωσης είναι σημαντική για την αξιολόγηση της προγνωστικής αξίας συγκεκριμένων βιοδεικτών ή γονιδιακών υπογραφών. Μέθοδοι όπως η Kaplan-Meier ανάλυση και τα μοντέλα Cox proportional hazards χρησιμοποιούνται για να καθοριστεί αν η έκφραση ενός γονιδίου ή μιας ομάδας γονιδίων συνδέεται με την επιβίωση των ασθενών ([Kaplan-Meier & Cox Models](https://lifelines.readthedocs.io/en/latest/)).

**Μηχανική Μάθηση για Ανακάλυψη Βιοδεικτών**

Η εφαρμογή αλγορίθμων μηχανικής μάθησης προσφέρει τη δυνατότητα ανακάλυψης και επικύρωσης βιοδεικτών από μεγάλα και πολύπλοκα σύνολα δεδομένων. Τεχνικές όπως η παλινδρόμηση LASSO, οι ταξινομητές τυχαίου δάσους (Random Forest), και τα νευρωνικά δίκτυα είναι δημοφιλείς στην ανακάλυψη βιοδεικτών και μπορούν να προσφέρουν βελτιωμένες προβλέψεις όσον αφορά τη διάγνωση και την πρόγνωση της νόσου ([scikit-learn](https://scikit-learn.org/stable/)).

Η χρήση αυτών των μεθοδολογιών συνδυαστικά επιτρέπει μια ολοκληρωμένη ανάλυση και μπορεί να προσφέρει πολύτιμες πληροφορίες για την κατανόηση της μοριακής βάσης του καρκίνου του πνεύμονα και την ανάπτυξη νέων θεραπευτικών και διαγνωστικών εργαλείων.

# Ενότητα 4: Συνοπτική Παρουσίαση Σημαντικών Βιοδεικτών στον Καρκίνο του Πνεύμονα

Η έρευνα στον καρκίνο του πνεύμονα έχει αναδείξει αρκετούς σημαντικούς βιοδείκτες, οι οποίοι έχουν αξιολογηθεί για τη διαγνωστική, προγνωστική ή θεραπευτική τους σημασία. Παρακάτω συνοψίζονται μερικοί από τους πιο ευρέως αποδεκτούς και μελετημένους βιοδείκτες που έχουν προταθεί στη βιβλιογραφία:

**EGFR (Epidermal Growth Factor Receptor)**

Οι μεταλλάξεις στο γονίδιο EGFR είναι από τους πιο σημαντικούς βιοδείκτες στον καρκίνο του πνεύμονα, ειδικά στο μη μικροκυτταρικό καρκίνο (NSCLC). Αυτές οι μεταλλάξεις είναι συνήθως προγνωστικές για την αποτελεσματικότητα συγκεκριμένων θεραπειών με αναστολείς τυροσινικής κινάσης του EGFR (π.χ., erlotinib και osimertinib) ([EGFR mutations in NSCLC](https://pubmed.ncbi.nlm.nih.gov/29401005/)).

**ALK (Anaplastic Lymphoma Kinase)**

Η συγχώνευση του γονιδίου ALK, όπως η γνωστή EML4-ALK, έχει καθιερωθεί ως ένας κρίσιμος προγνωστικός και θεραπευτικός βιοδείκτης στο NSCLC. Η παρουσία αυτής της συγχώνευσης είναι δείκτης θετικής ανταπόκρισης σε στοχευμένες θεραπείες όπως crizotinib και alectinib ([ALK fusion in lung cancer](https://pubmed.ncbi.nlm.nih.gov/31217343/)).

**PD-L1 (Programmed Death-Ligand 1)**

Η έκφραση της πρωτεΐνης PD-L1 στα καρκινικά κύτταρα έχει γίνει σημαντικός βιοδείκτης στην επιλογή ασθενών για ανοσοθεραπεία. Υψηλά επίπεδα έκφρασης PD-L1 συχνά προβλέπουν καλύτερη ανταπόκριση σε ανοσοθεραπευτικά φάρμακα που στοχεύουν το μονοπάτι PD-1/PD-L1, όπως το pembrolizumab ([PD-L1 expression as biomarker](https://pubmed.ncbi.nlm.nih.gov/31711696/)).

**KRAS (Kirsten Rat Sarcoma Virus)**

Οι μεταλλάξεις στο γονίδιο KRAS είναι συχνές στον καρκίνο του πνεύμονα και έχουν συσχετιστεί με φτωχή πρόγνωση και αντίσταση σε διάφορες στοχευμένες θεραπείες. Πρόσφατα αναπτύχθηκαν νέοι θεραπευτικοί παράγοντες που στοχεύουν συγκεκριμένες μεταλλάξεις KRAS (π.χ., KRAS G12C), ανοίγοντας νέους δρόμους για θεραπεία ([KRAS mutations](https://pubmed.ncbi.nlm.nih.gov/34151404/)).

**ROS1**

Οι συγχωνεύσεις ROS1 είναι σπάνιες αλλά σημαντικές στο NSCLC, καθώς σχετίζονται με εξαιρετική ανταπόκριση σε στοχευμένες θεραπείες όπως το crizotinib. Αυτό καθιστά το ROS1 έναν πολύτιμο θεραπευτικό βιοδείκτη ([ROS1 fusion-positive lung cancer](https://pubmed.ncbi.nlm.nih.gov/32086372/)).

**MET**

Η ενίσχυση ή μετάλλαξη του MET σχετίζεται με αντοχή στις τρέχουσες θεραπείες και φτωχότερη πρόγνωση. Η στοχευμένη θεραπεία με αναστολείς MET έχει δείξει αποτελεσματικότητα σε ασθενείς με συγκεκριμένες μεταλλάξεις ή ενισχύσεις MET ([MET alterations](https://pubmed.ncbi.nlm.nih.gov/33029068/)).

**BRAF**

Οι μεταλλάξεις BRAF, ιδιαίτερα η μετάλλαξη V600E, αν και σχετικά σπάνιες, είναι σημαντικές καθώς ανταποκρίνονται καλά σε θεραπεία με αναστολείς BRAF και MEK, προσφέροντας θεραπευτική εναλλακτική ([BRAF mutations in lung cancer](https://pubmed.ncbi.nlm.nih.gov/31736317/)).

Η συνεχής έρευνα και αξιολόγηση αυτών των βιοδεικτών είναι ζωτικής σημασίας για τη βελτίωση της διάγνωσης, της πρόγνωσης και της θεραπευτικής προσέγγισης του καρκίνου του πνεύμονα, συμβάλλοντας καθοριστικά στην ανάπτυξη εξατομικευμένης ιατρικής.

# Ενότητα 5: Εφαρμογή Βιοπληροφορικής Ανάλυσης σε Ανοιχτά Δεδομένα

Στην παρούσα ενότητα περιγράφεται η εφαρμογή βιοπληροφορικής ανάλυσης σε πραγματικά δεδομένα από ανοικτές βάσεις του GEO (Gene Expression Omnibus). Ο στόχος είναι να αναλυθεί η έκφραση γονιδίων μεταξύ φυσιολογικών και καρκινικών δειγμάτων στον καρκίνο του πνεύμονα. Για την ανάλυση χρησιμοποιήθηκαν δύο datasets: GSE31210 και GSE19804.

Το GSE31210 περιλαμβάνει δεδομένα έκφρασης από 226 δείγματα καρκίνου του πνεύμονα. Το GSE19804 περιέχει επίσης δείγματα από καρκινικούς ιστούς και φυσιολογικούς πνεύμονες, με πιο στοχευμένο σχεδιασμό για διαφοροποίηση μεταξύ καρκινικών και μη δειγμάτων. Αμφότερα τα datasets βασίζονται στην πλατφόρμα Affymetrix U133 Plus 2.0 (GPL570), για την οποία διατίθεται αρχείο συσχέτισης probes με γονίδια.

Η επεξεργασία των αρχείων πραγματοποιήθηκε σε Python. Αρχικά, τα δεδομένα διαβάστηκαν και οργανώθηκαν σε dataframes. Από το annotation αρχείο προέκυψε αντιστοίχιση των probes (ID\_REF) σε ονόματα γονιδίων (Gene Symbols). Η μετατροπή αυτή ήταν κρίσιμη για να μπορέσουν να συνενωθούν τα datasets με συνεπή τρόπο, και να μειωθεί η πολυπλοκότητα των δεδομένων. Στη συνέχεια, έγινε ομαδοποίηση των γραμμών βάσει Gene Symbol και υπολογισμός του μέσου όρου για τις επαναλαμβανόμενες τιμές.

Από τα δύο datasets απομονώθηκαν τα δείγματα που ήταν χρήσιμα για την ανάλυση. Η πληροφορία αυτή εξήχθη αυτόματα από τα metadata των GEO αρχείων και καταγράφηκε σε δύο αρχεία κειμένου: cancer.txt και normal.txt. Μέσα από αυτά τα αρχεία κατασκευάστηκαν οι ετικέτες κατηγοριοποίησης (labels), που υποδεικνύουν ποια δείγματα ανήκουν στην κατηγορία "cancer" και ποια στην "normal".

Ένα τελικό dataframe κατασκευάστηκε, όπου οι στήλες αντιστοιχούν σε γονίδια και οι γραμμές σε δείγματα. Προστέθηκε επιπλέον η στήλη "label" ώστε να είναι δυνατή η επίβλεψη της ανάλυσης. Το τελικό σχήμα των δεδομένων είναι 366 δείγματα (286 καρκινικά και 80 φυσιολογικά) και 23.521 γονίδια.

Παρακάτω παρουσιάζεται screenshotes από τον πίνακα expression\_data όπως εμφανίζεται στο Jupyter Notebook, το οποίο επιβεβαιώνει την επιτυχή οργάνωση των δεδομένων:

Εικόνα 2,3:

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, λογισμικό, οθόνη

Το περιεχόμενο που δημιουργείται από τεχνολογία AI ενδέχεται να είναι εσφαλμένο.

Εικόνα που περιέχει κείμενο, στιγμιότυπο οθόνης, γραμματοσειρά, μαύρο

Το περιεχόμενο που δημιουργείται από τεχνολογία AI ενδέχεται να είναι εσφαλμένο.

# Βιβλιογραφία