NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

INTERDISCIPLINARY POSTGRADUATE PROGRAM
"INFORMATION TECHNOLOGIES IN MEDICINE AND BIOLOGY"

MSc THESIS

Bioinformatic Analysis of Clinical and Molecular Patient Data with Non-Alcoholic Fatty Liver Disease and Implementation of Statistical Methods in Co-Expression Networks

Vasiliki G. Filippa

Supervisors: **Dr. Ema Anastasiadou,**

Dr. George M. Spyrou,

Dr. George Th. Tsangaris

ATHENS

OCTOBER 2018

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
"ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ"

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βιοπληροφορική ανάλυση κλινικών και μοριακών δεδομένων, ασθενών με μη – αλκοολική λιπώδη νόσο του ήπατος και εφαρμογή στατιστικών μεθόδων για τη δημιουργία δικτύων γονιδιακής συνέκφρασης

Βασιλική Γ. Φίλιππα

Επιβλέποντες: **Δρ. Έμα Αναστασιάδου**

Δρ. Γεώργιος Μ. Σπύρου,

Δρ. Γεώργιος Θ. Τσάγκαρης

ΑΘΗΝΑ

ΣΕΠΤΕΜΒΡΙΟΣ 2017

MSc THESIS

Bioinformatic analysis of clinical and molecular patient data with non-alcoholic fatty liver disease and implementation of statistical methods in co-expression networks

Vasiliki G. Filippa
R.N.: ΠΙΒ0148

SUPERVISORS:

Dr. Ema Anastasiadou, Researcher-Lecturer Level, Bioimedical Research Foundation of the Academy of Athens (BRFAA)

Dr. George M. Spyrou, PhD,Bioinformatics ERA chair, Head of the Bioinformatics Group, The Cyprus Institute of Neurology and Genetics (CING)

EXAMINATION COMMITEE:

Dr. Ema Anastasiadou, Researcher-Lecturer Level, Bioimedical Research Foundation of the Academy of Athens (BRFAA)

Dr. George M. Spyrou, PhD,Bioinformatics ERA chair, Head of the Bioinformatics Group, The Cyprus Institute of Neurology and Genetics (CING)

Dr. George Th.Tsangaris, Staff Research Scientist – Professor Level, Biomedical Research Foundation of the Academy of Athens (BRFAA)

September 2017

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Βιοπληροφορική ανάλυση κλινικών και μοριακών δεδομένων, ασθενών με μη-αλκοολική λιπώδη νόσο του ήπατος και εφαρμογή στατιστικών μεθόδων για τη δημιουργία δικτύων γονιδιακής συνέκφρασης

Βασιλική Γ. Φίλιππα
Α.Μ.: ΠΙΒ0148

ΕΠΙΒΛΕΠΟΝΤΕΣ:

Δρ. Έμα Αναστασιάδου, Ερευνήτρια Δ', Ιδρυμα Ιατρονιολογικών Ερευνών Ακαδημίας Αθηνών (ΙΙΒΕΑΑ)

Δρ. Γεώργιος Μ. Σπύρου, Κάτοχος της έδρας της Βιοπληροφορικής, Επικεφαλής της ομαδας Βιοπληροφορικής, Ινστιτούτο Νευρολογίας και Γενετικής Κύπρου (ΙΝΓΚ)

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ:

Δρ. Έμα Αναστασιάδου, Ερευνήτρια Δ', Ιδρυμα Ιατρονιολογικών Ερευνών Ακαδημίας Αθηνών (ΙΙΒΕΑΑ)

Δρ. Γεώργιος Μ. Σπύρου, Κάτοχος της έδρας της Βιοπληροφορικής, Επικεφαλής της ομαδας Βιοπληροφορικής, Ινστιτούτο Νευρολογίας και Γενετικής Κύπρου (ΙΝΓΚ)

Δρ. Γεώργιος Θ. Τσάγκαρης, Ειδικός Λειτουργικός Επιστήμονας Α', Ιδρυμα Ιατρονιολογικών Ερευνών Ακαδημίας Αθηνών (ΙΙΒΕΑΑ)

Οκτώβρης 2018

ABSTRACT

Non alcoholic fatty liver disease (NAFLD) is currently the most common liver disease worldwide. NAFLD comprises a disease spectrum that ranges from simple steatosis (SS) to non alcoholic steatohepatitis (NASH), which may progress into liver fibrosis and even end stage cirrhosis. Although novel clinical methods advanced tremendously this scientific field, the actual link between SS and the pathogenesis of NASH is still controversial. Therefore, a study focusing on gene expression, interactions and biological pathways while taking into account the clinical data of patients, is considered priceless for the emergence of important genes, biomarkers, etc, related to the above pathological conditions. The proposes thesis is aiming to study and determine the incoherent red lines between SS and NASH by analyzing and combining findings from various parameters such as gene expression in each pathological condition, gene co-expression, clinical features, etc. The methodology regarding the Bioinformatics approach will include differential gene expression analysis of microarray experiment data, biostatistical analysis on patient's clinical data, machine learning methods, gene co-expression network construction, including other techniques. The programming language we use is R, and for the networks visualization we will use the Cytoscape app.

SUBJECT AREA: Molecular Biology, Machine Learning, Biological Networks, Bioinformatics

KEYWORDS: microarray bioinformatic analysis, biomarkers, non-alcoholic fatty liver disease, liver steatosis

ΠΕΡΙΛΗΨΗ

Η μη αλκοολική λιπώδης νόσος του ήπατος (NAFLD) είναι σήμερα η συνηθέστερη νόσος του ήπατος παγκοσμίως. Το φάσμα της ασθένειας είναι εξαιρετικά ευρύ και κυμαίνεται από την απλή στεάτωση (SS) μέχρι τη μη αλκοολική στεατοηπατίτιδα (NASH), η οποία μπορεί να προχωρήσει σε ίνωση του ήπατος, φτάνοντας ακόμη, και στο τελικό στάδιο της κίρρωσης. Παρόλο που νέες κλινικές μέθοδοι έχουν προχωρήσει σημαντικά την έρευνα σε αυτό βιολογικό πεδίο και έχουν ενισχύσει την κατανόησή μας, ως ένα βαθμό, σε σχέση με την εξέλιξη της νόσου, η πραγματική συσχέτιση της απλής στεάτωσης ως προς την μετάβαση της σε μη αλκοολική στεατοηπατίτιδα, παραμένει αμφιλεγόμενη. Ως εκ τούτου, μία μελέτη που να εστιάζει στην ανάλυση της έκφρασης των γονιδίων στις δυο παθολογικές καταστάσεις και στις αλληλεπιδράσεις των βιολογικών μονοπατιών λαμβάνοντας υπόψιν τα κλινικά στοιχεία των ασθενών φαίνεται πως είναι απαραίτητη, για την ανάδειξη σημαντικά εκφρασμένων γονιδίων, βιοδεικτών αλλά και γονιδιακών αλληλεπιδράσεων που σχετίζονται με τις παραπάνω παθολογικές καταστάσεις. Η παρούσα διατριβή αποσκοπεί στη μελέτη αλλά και στον καθορισμό εκείνων των “κόκκινων γραμμών” που μπορούν να διαχωρίσουν την απλή στεάτωση από την μη αλκοολική στεατοηπατίτιδα, μέσω της βιοπληροφορικής ανάλυσης και του συνδυασμού των ευρυμάτων με ποικίλες παραμέτρους. Η μεθοδολογία που αφορά στο πεδίο της βιοπληροφορικής προσέγγισης, περιλαμβάνει την ανάλυση της διαφορικής έκφρασης γονιδίων που προέρχονται από πειράματα μικροσυστιχιών, βιοστατιστική ανάλυση εφαρμοσμένη στα κλινικά στοιχεία των ασθενών, μεθόδους μηχανικής μάθησης, κατασκευή δικτύων γονιδιακής συνέκφρασης, συμπεριλαμβανομένων άλλων τεχνικών. Η γλώσσα προγραμματισμού που χρησιμοποιούμε είναι η R, και για την απεικόνιση των δικτύων χρησιμοποιούμε το λογισμικό Cytoscape.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Μοριακή Βιολογία, Μηχανική Μάθηση, Βιολογικά Δίκτυα, Βιοπληροφορική

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: Βιοπληροφορική ανάλυση μικροσυστοιχιών, βιοδεικτες, μη αλκοολική λιπώδης νόσος του ήπατος, ηπατική στεάτωση, διαφορική έκφραση γονιδίων