BIOMARKERS PREDICTION FOR HEPATOCELLULAR CARCINOMA USING MACHINE LEARNING TECHNIQUES

By

Ola Salah El Din Ayoub Sayed

A Thesis Submitted to the
Faculty of Engineering at Cairo University
In Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
In
BIOMEDICAL ENGINEERING &SYSTEMS

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2016

BIOMARKERS PREDICTION FOR HEPATOCELLULAR CARCINOMA USING MACHINE LEARNING TECHNIQUES

By
Ola Salah El Din Ayoub Sayed

A Thesis Submitted to the
Faculty of Engineering at Cairo University
In Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
In
BIOMEDICAL ENGINEERING &SYSTEMS

Under the Supervision of

Prof. Dr. Yasser M. kadah

Professor of Biomedical Engineer
System and biomedical Engineering
Faculty of Engineering, Cairo University

Dr. Nagwan. M. Abdel Samee

Assistant Professor
Computer Engineering Department,
Faculty of Engineering, Misr University for Science and Technology

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2016

BIOMARKERS PREDICTION FOR HEPATOCELLULAR CARCINOMA USING MACHINE LEARNING TECHNIQUES

By
Ola Salah El Din Ayoub Sayed

A Thesis Submitted to the
Faculty of Engineering at Cairo University
In Partial Fulfillment of the
Requirements for the Degree of
MASTER OF SCIENCE
In
BIOMEDICAL ENGINEERING &SYSTEMS

Approved by the
Examining Committee

_____

Prof. Dr. Yasser M. Kadah, Thesis Main Advisor

_____

Dr. Nagwan M. Abdel Samee, Member

_____

Prof. Dr. Ahmed Sofy AbouTaleb, Internal Examiner

_____

Dr. Mai Mohamed Said Mahmoud Mabrouk, External Examiner

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2016

Engineer's Name: Ola Salah El Din Ayoub Sayed
Date of Birth: 2/10/1987
Nationality: Egyptian
E-mail: ola.salaheldin@gmail.com
Phone: 01200224526
Address: El-Dakhla - New Valley Government.

Registration Date: 1/10/2010
Awarding Date: ./…./…..
Degree: Master of Science
Department: Systems & Biomedical Engineering.

Supervisors:

Prof.Dr.Yasser M. Kadah.
Dr. Nagwan M. AbdelSamee

Examiners:

Porf. Dr.Yasser M. Kadah (Thesis main advisor)
Prof. Ahmed Sofy AbouTaleb (Internal examiner)
Dr. Mai M. Said Mabrook (External examiner)

Title of Thesis:

Biomarkers Prediction for Hepatocellular Carcinoma using Machine Learning Techniques.

Key Words:
Biomarker; Gene selection; Microarray; Multivariate method; Univariate method.

Summary:
Microarray is an effective innovation can recover and study the sub-atomic science of tissues and the quality expression estimations of the entire genome and the estimation of microarrays in comprehension the organic procedures basic a particular ailment which has an awesome part in finding new critical qualities for malignancies in Hepatocellular Carcinoma (HCC). In this work we give a procedure to extract significant genes which have role to understand the identification and characterization of key gene that play a role in the HCC and HCV replication cycle by apply univariate method and multivariate methods. These significant genes can be viewed as cheerful biomarkers in high throughput microarrays of HCC such as provided from univariate method (SPRY1, TXNIP, DDIT4, STC2, COL1A1) and multivariate methods (EEF1A1, FTL, ACTB) and others gene have impact in distinctive kind of cancers. Finally although each method has different procedure it gives me the key genes which have role in HCC.

Acknowledgments

I wish to express my sincere gratitude to Prof. Dr. Yasser M. Kadah who give me support, advices, guidance, valuable comments, suggestions, and provisions that benefited her much in the completion and success of this study and I am lucky to be one of his students.

It is my proud privilege to release the feelings of my gratitude to several to Dr. Nagwan M. Abdel Samee who gives me the way to start this work, gives her love, care, time and shares her knowledge to complete my research. I hope to get your expectation. Thanks Dr. Nagwan.

Dedication

I dedicate my husband and my lovely baby:

Thank you for believing in me; for allowing me to further my studies. Please do not ever doubt my dedication and love for you.

Also I dedicate my work to my lovely Dad and Mum:

Thank you for your unconditional support with my studies. I am honored to have you as my parents. Thank you for giving me a chance to prove and improve myself through all my walks of life. Please do not ever change. I love you.

Finally I dedicate my brother and my sisters:

Hoping that with this research I have proven to you that there is no mountain higher as long as God in our side. Hoping that you will walk again and be able to fulfill your dreams!

Table of Contents

List of Tables

List of Figures

Nomenclature

HCC	Hepatocellular Carcinoma
HCV	Hepatitis C Virus
GEO	Gene Expression Omnibus
PCA	Principal Component Analysis
DAVID	The Database for Annotation, Visualization and Integrated Discovery
CRAN	Comprehensive R Archive Network
KEGG	Kyoto Encyclopedia of Genes and Genome

Abstract

Microarray is an effective innovation can recover and study the sub-atomic science of tissues and the quality expression estimations of the entire genome and the estimation of microarrays in comprehension the organic procedures basic a particular ailment which has an awesome part in finding new critical qualities for malignancies in Hepatocellular Carcinoma (HCC). In this work we give a procedure to extract significant genes which have role to understand the identification and characterization of key gene that play a role in the HCC and HCV replication cycle by apply univariate method such as (F test, T Distribution Formula, Pearson correlation coefficient, Cosine coefficient, Euclidean Distance) and multivariate methods such as Principal Component Analysis). These significant genes can be viewed as cheerful biomarkers in high throughput microarrays of HCC such as provided from univariate method (SPRY1, TXNIP, DDIT4, STC2, COL1A1) and multivariate methods (EEF1A1, FTL, ACTB) and others gene have impact in distinctive kind of cancers. Finally although each method has different procedure it gives me the key genes which have role in HCC.

Chapter 1 : Introduction

1.1. Problem Overview

Molecular biology is an emphasis on genomics and bioinformatics. It is intended for scientists, engineers, computer programmers or anybody with background or strong interest in science, but without background in biology. Then it will subject to the field of bioinformatics and it applications. We capture the gene expressions (mRNA abundance) which describe how the genetic information converted to a functional gene product through the transcription and translation processes. Functional genomics uses microarrays technology to quantify the genes expressions levels under certain conditions and environmental limitations.

Recently it becomes the prevalence of Hepatocellular carcinoma (HCC) and Hepatitis C virus (HCV), where scientific researchers are turning to make scientific trends in research in this area. The scientific researchers are dividing the search in more than one area, such as genetics, image informatics and science of medicine. As we research at genetics, we have process of analyzing and interpreting dataset that have information of genes related to HCC and HCV.

Scientific researchers are interesting in developing treatment for hepatocellular carcinoma (HCC) by studies early drug development and this step it done by target the genes that caused the HCC and then can detect the drug that overcome the genes that has main cause of tumor. So finding new biomarkers on the different period of diagnosis has a great role to analysis what happen and the stages of the disease genes in cancer journey.

In this research we provide to detect the biomarker, we have four period of diagnosis of HCC and HCV replication cycle. We must target the key genes in each period and which one of them has appeared in all period by providing two approach univariate approach and multivariate approach for extracting differential expressed genes.

1.2. Thesis Objective

In this thesis we aims to identify the key genes that play a great role in the replication cycle of C virus (HCV) and Hepatocellular Carcinoma (HCC) by using methods of univariate gene selection and multivariate gene selection and comparing between two approaches. We state the strongest and weakness of two gene selection methods. We apply biological validation by using DAVID and KEGG and Atlas of Genetic and Cytogenetic in Oncology and Hematology, after ranked the highly effective genes.

1.3. Thesis Organization

The thesis is consisting of five chapters and Appendix:

Chapter 1: The first chapter titled "Introduction" and gives a short account of the problem overview, thesis objective and its organization.

Chapter 2: In a second chapter we talk about "Background and Literature Review", which gives a brief introduction of hepatocellular carcinoma, Microarray, Affymetrix Data files, Bioconductor &R, Description of datasets and Data preprocessing. The end of this chapter gives a summary of related work.

Chapter 3: This chapter entitled "Detection of Significant gene for Hepatocellular Carcinoma by using Univariate gene selection methods" provides the univariate methods such as F test, T distribution formula, Pearson correlation coefficient, Cosine Coefficients and Euclidean distance. We follow it result and discussion.

Chapter 4: This chapter entitled "Detection of Significant gene for Hepatocellular Carcinoma by using Multivariate gene selection methods" provides the multivariate method such as principal component analysis and we follow it with results and discussion.

Chapter 5: This chapter entitled "Conclusion and Future Work" which provides the conclusion of all what mention in the thesis and gives comparison between univariate and multivariate approach and which of two methods are applicable for my dataset. Finally give some points for future work.

Chapter 2 : Back ground and Literature Review

2.1. Introduction

Biological data is increasing rapidly. Public databases, for example, Gene Bank and the Protein Data Bank have been becoming exponentially for a long time. With the coming of the internet and fast web affiliations, the data contained in these databases can be gotten too rapidly, easily, and economically from any area in the world. As a result, PC based device now assume an undeniably basic part in the advancement of biological research. Bioinformatics is used for computational tools and procedures to the management and investigation of biological data. The term bioinformatics is generally new, and as characterized here, it infringes on such terms as "computational biology" and others. The utilization of computers in biology research originates before the term bioinformatics by numerous years. Case in point, the determination of 3D protein structure from X-ray crystallographic information has since a long time ago depended on computer investigation. Specifically, bioinformatics is frequently the term utilized when alluding to the information and the methods utilized as a part of vast scale sequencing and analysis whole genomes [1].

2.1.1. Hepatocellular carcinoma

Hepatocellular carcinoma (HCC) is standout amongst the most widely recognized harmful tumor with a high mortality in people [2]. Hepatocellular carcinoma is presently the third driving reason for disease passing around the world, with more than 500,000 individuals influenced. The occurrence of hepatocellular carcinoma is most astounding in Asia and Africa, where the endemic high commonness of hepatitis B and hepatitis C firmly inclines to the advancement of incessant liver malady and ensuing improvement of hepatocellular carcinoma as shown in figure1.1.

The presentation of hepatocellular carcinoma has developed fundamentally in the course of recent decades. While, previously, hepatocellular carcinoma for the most part exhibited at a propelled stage with right upper quadrant torment, weight reduction, and indications of decompensate liver malady, hepatocellular carcinoma is currently progressively perceived at a much before stage as a normal's outcome screening of