GENE ONTOLOGY DATA MINING AND SYSTEMS BIOLOGY OF CANCER

Ph.D. THESIS

by

AJAY SHIV SHARMA

DEPARTMENT OF ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE-247667 (INDIA)

JULY, 2014

GENE ONTOLOGY DATA MINING
AND
SYSTEMS BIOLOGY OF CANCER

A THESIS

Submitted in partial fulfilment of the
requirements for the award of the degree
of

DOCTOR OF PHILOSOPHY

in

ELECTRICAL ENGINEERING

by

AJAY SHIV SHARMA

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE-247667 (INDIA)
JULY, 2014

INDIAN INSTITUTE OF TECHNOLOGY ROORKEE
ROORKEE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **"GENE ONTOLOGY DATA MINING AND SYSTEMS BIOLOGY OF CANCER"** in partial fulfilment of the requirements for the award of the Degree of Doctor of Philosophy and submitted in the Department of Electrical Engineering of the Indian Institute of Technology Roorkee, Roorkee is an authentic record of my own work carried out during a period from July, 2011 to July, 2014 under the supervision of Dr. Hari Om Gupta, Professor and Director, Jaypee Institute of Information Technology-Sector 128, Noida and Dr. Rajendra Prasad, Professor, Department of Electrical Engineering, Indian Institute of Technology Roorkee, Roorkee.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other Institute.

(AJAY SHIV SHARMA)

This is to certify that the above statement made by the candidate is correct to the best of our knowledge.

(Rajendra Prasad)
Supervisor

(Hari Om Gupta)
Supervisor

Date: July, 2014

The Ph.D. Viva-Voce Examination of **Mr. Ajay Shiv Sharma**, Research Scholar, has been held on ……………………………………

Supervisor's

Chairman, SRC

External Examiner

Head of the Department/Chairman, ODC

ABSTRACT

Bioinformatics and computational systems biology fuses several branches of applied sciences and applied engineering and interplay that exploits basic sciences such as mathematics, physics, chemistry, computer science that present a partial picture and biological sciences such as molecular biology, structural biology, and systems biology that present a whole picture. To unravel complicated biological issues, this research work deals with interdisciplinary aspects of contemporary bioinformatics and computational systems biology as defined by the NIHs working definition of bioinformatics and computational biology (http://www.bisti.nih.gov/docs/compubiodef.pdf) in which many science fields come under the umbrella of bioinformatics with numerous applications in the field of life sciences. Contemporary bioinformatics deals with computational tools providing a user-friendly environment for the dissemination of life sciences knowledge through existing biological databases in a particular domain. On similar lines, computational systems biology, which has roots in life sciences, primarily deals with analytical data oriented techniques and mathematical modelling to study and analyze complex biological systems. For this, four different research concerns for progressive biological discoveries are handled in this research work using bioinformatics and systems biology approaches.

At present, all research indications speak in favour of the key challenge for integrative biology: providing physiological models that could facilitate development of novel drugs against diseases such as cancer and Alzheimer's disease against which effective therapeutics currently do not exist. Even though such full physiological models are not always attainable due to inadequate biological data and/or their appropriate integration, functional genomics can be currently considered as a reliable functional basis upon which such models are expected to rely. The research work provides novel insights into how a biological data base, which are essentially descriptive physiological models, can be functionally improved in terms of contemporary bioinformatics depending on the accessibility and integration of data. Most researchers agree that the challenge is data management, data analysis, data interpretation, data modelling and understand all the biological data that are being produced. However, a major issue prevails: all the above-mentioned issues are handled differently at different laboratories throughout the world,

producing plethora of biological data. To fill this research gap, an omics or integrative genomics revolution is need that uses the power of gene ontology (GO). The first concern of this work is to provide theoretical models to achieve this herculean task of integrating biological data by moving from knowledge gained from functional genomics to physiological models. Since a better understanding of many pathological conditions is the ultimate goal of full physiological models, physiology can be understood as the science of the functioning of living systems. To approach a full physiological model, a tremendous amount of biological knowledge contained in various databases needs to be sorted out by discriminating different types of data subjected to double integration: i) vertically - from molecular level, over cell and organ levels, all the way to the level of a whole organism and (ii) horizontally - comprising gene, anatomy and phenotype data. As such, a hypothetical full physiological model is supposed to have its full biological process (BP), full cellular component (CC), full molecular function (MF) and with its specific full ontologies respectively. Connecting individual ontologies from various data resources is a key step leading to a universal full physiological model. As such, the proposed model is supposed to have its full BPCCMF with its specific full ontologies.

After understanding the concept of the full physiology, the illustration using a plants physiological model is implemented in this research work and the same can be extended for other organisms, pathological conditions, etc. The second primary concern in this work is the development of a gene ontology data mining tool using contemporary bioinformatics focusing on the design of a plants physiology database that represents all biological knowledge in a computationally tractable way unambiguously. The idea to serve the plant scientific community by using power of contemporary bioinformatics came from the fact that plants have been the most studied since the advent of classic genetics. Recent studies show that plants are biologically more complex and there are enormous applications to be gained from researching plant genes to progress the reception of nutrients from the earth to enhance plant yields and plant ailments that directly effects the health of humans. This research work focuses on providing a centralized plants physiology database as a new searching and investigating tool after mining plants gene ontological data from GO database. The applications of contemporary database management led to the development of Plants Physiology Database (PPDB), a searching and browsing tool based on the mining of large amounts of gene ontology data currently available. The PPDB is publicly available and freely accessible on-line (http://www.iitr.ernet.in/ajayshiv/) through a user-friendly environment generated by Drupal-6.24.

Another focus of this work is the systems biology of cancer. Last decade has witnessed the emergence of new field of research called systems biology to capture the biological phenomenon with data analysing, modelling, and computational tools. Generations of scientists and physicians have dedicated their life to improving patient care and fighting against cancer. Systems biology offers promising insights to defeat cancer. Cancer is a major health issue responsible for 8.2 million deaths in 2012 and 14.1 million new cancer cases were reported in 2013 worldwide (http://globocan.iarc.fr). It is anticipated that the global yearly number of deaths should reach 17 million in 2030. As such, research progress in cancer treatment is real but insufficient. Cancer is a genetic disease that causes a deregulation of gene networks that control cell growth and dissemination. As a result, methods for modelling gene networks are central to any modern approach of the molecular biology of cancer. Moreover, the sequencing of the human genome and subsequent genomic revolution has impacted cancer research at the molecular level due to high throughput technologies like microarray database (MDB).

As such, this research work focuses on both aspects of systems biology of cancer separating it into different computational approaches dealing with data driven systems biology and model driven systems biology. Data driven models are based on computational statistical tools that can handle high throughput MDB and termed as top-down models. They deal with two types of statistical analysis known as a low level analysis dealing with background correction, normalization using a model based expression index (MBEI) method along with high level analysis dealing with filtering of genes to find interesting genes, hierarchical clustering of filtered genes, genetic association study and gene ontology data mining/enrichment analysis. The central dogma of microarray data analyses is the third research concern in this work. The invaluable information produced after analyses can pave the way for innovative opportunities for early diagnosis of malignancies. This research work can enhance further research in diagnostics, prognostics, disease markers, target validation and targeted therapies using contemporary bioinformatics at a later stage. The list of significant genes or differentially expressed genes helps to find the functional relationships between genes in MDB warehouses by linking it to annotations of GO. For instance, a precautionary double mastectomy on finding the BRCA1 gene with only 87% probable chance of acquiring the disease shows the promising nature of this field.

On the other hand, another approach on how dynamical mathematical models can provide novel insight that cannot be done by doing experiments. Model driven dynamical models or bottom-up models approach is the opposite of a top down model. With the bottom up model, it begins with a hypothesis of a biological mechanism. After having this hypothesis, equations are written down to describe how the components in the biological system interact with one another. Then simulations are run to generate predictions for what would happen under different conditions. Some of the keywords associated with bottom up models are ordinary differential equations, computational tools of dynamical systems to interpret the output and methods for parameter estimation, partial differential equations and stochastic models. The focus of the final research concern deals with developing models consisting of systems of differential equations and using computational tools of dynamical systems in order to interpret the results of these simulations. Therefore, a multi-scale computational approach of tumour growth model is presented. A mathematical model is developed for tumour growth and angiogenesis to simulate the solid tumour growth/progression with chemotherapy drug and anti-angiogenesis drug estimation using partial differential equation (PDE) modelling. The PDE compartmental model incorporated spatiotemporal processes including cellular and tissue-mediated diffusion, cellular transport and migration, cell proliferation, angiogenesis, apoptosis, vessel maturation and formation to model tumour progression and transition from avascular to vascular growth. The angiogenesis process coupled with the solid tumour growth model on a reaction–diffusion kinetics framework portrayed the spatiotemporal development of the generalised functions of a tumour's micro-environment viz., nutrients and growth factors that regulate the tumour's growth during angiogenesis. Most cancers involve an endothelial growth factor receptor/extracellular signal-regulated kinases (EGFR/ERK) signalling pathway that are related to the cell-division cycle promoting tumour cells. Treatment is studied from tyrosine kinase inhibitors (TKI) in EGFR signalling, which are distributed through the blood vessels of a tumour's microvasculature. This showed a huge potential for in-vitro experiments due to the availability of clinical and expression data information, which helps in learning about the responses to treatment. Using ordinary differential equations to model the systems pathway of downstream pathway of EGFR signalling (SOS→RAS→RAF→MEK→ERK→PI3K→AKT), we performed computational simulations to determine the facilitation of glucose, oxygen, tumour angiogenesis factor (TAF), drug (TKI), tumour growth factor alpha (analogue of EGFR) and angiogenesis inhibitor. The simulation results showed signalling pathways of TKI-

EGFR and IGF1R regulation of various active cells, migrating cells, proliferative cells, apoptotic and quiescent cells could be a united behaviour for the entire profile of tumour growth. The results established the dual role behaviour played by angiogenesis as TKI-EGFR and VEGF inhibitors are furnished to diminish tumour incursion. In addition, the neovasculature can transport nutrients to neoplasm cells to continue cell metabolism, thus enhancing the rate of cell endurance. Hence, simulation results suggest that the co-expression of EGFR and IGF1R activates a higher number of ERK receptors compared to down and over-expressions. There is a good agreement between the simulations, an experimental wild type mouse model, and clinical data.

It can be concluded that this work may not be able to solve the numerous convoluted issues in the field of biotechnology, but it can address issues in gene ontology data mining using contemporary bioinformatics taking the example of a plants physiology database and state of the art work related to cancer systems biology.

ACKNOWLEDGEMENTS

I owe a deep sense of gratitude to my Lord who is all sources of prevailing energy and my progenitors who provided me the perseverance, guidance, inspiration, faith, and strength to carry on even when the going got tough.

I express my deep regards and sense of gratitude to my supervisors Dr. Hari Om Gupta, Professor and Director, Jaypee Institute of Information Technology-Sector 128, Noida, India and Dr. Rajendra Prasad, Professor, Department of Electrical Engineering, Indian Institute of Technology, Roorkee, India, for their inspirations, expert guidance, moral support, passing on strong ethical values, and encouragement throughout this work. I also express my deep regards and sense of gratitude to Dr. Petar M. Mitrasinovic, Visiting Professor, Department of Biotechnology, Indian Institute of Technology, Roorkee, India, who was also one of my supervisors up to October, 2013 for invaluable guidance and contribution. It is my lifetime opportunity to work under their expert supervisions.

I express my sincere thanks to Dr. Pramod Agarwal, Dean Academics, IIT Roorkee. I owe my sincere thanks to Dr. G.N. Pillai, Chairman DRC, Department of Electrical Engineering, IIT Roorkee, Dr. Vinod Kumar and Dr. R.S. Anand, Department of Electrical Engineering, IIT Roorkee and Dr. Shailly Tomar, Department of Biotechnology, IIT Roorkee, for being members of my research committee and sparing their valuable time in reviewing and critically examining the work. My sincere thanks are also due to Dr. H.S. Dhaliwal and Dr. Pravindra Kumar, Department of Biotechnology, IIT Roorkee, for extending moral support.

I wish to express my sincere thanks to Dr. Satya P. Srivastava, Professor and Head, Department of Electrical Engineering, for providing the necessary facilities and healthy environment to carry out the research work. I extend my thanks to all members of the Department of Electrical Engineering, with its diverse nature of the faculty that has constituted the strength of the Department. I am thankful to non-teaching staff of Department of Electrical Engineering for their help and support. My sincere thanks are also due to senior Prof. M.K. Vasanta and Dr. Indira Gupta, Department of Electrical Engineering, for showering their continuous blessings and providing conducive environment in the computer laboratory.

I acknowledge my sincere gratitude to the Quality Improvement Program, All India Council for Technical Education, Ministry of Human Resource Development, Government

of India. The research work is carried out under this program. I also acknowledge my appreciation to the coordinator and staff of QIP centre, IIT, Roorkee.

I express my sincere thanks to the Management, Director and HOD (CSE, IT and CC), Guru Nanak Dev Engineering College, Ludhiana (my parent institution), for granting me leave to accomplish the Ph.D. work. I am heartily thankful to my colleagues for sharing my work load at the institute in my absence.

I sincerely acknowledge the Gene Ontology Consortium and open source community as the source and inspiration for Plants Physiology Database (PPDB). I also acknowledge AmiGO for the sourced term images on PPDB website (http://www.iitr.ac.in/ajayshiv/). I am thankful to Head, Institute Computer Center and Information Super Highway, Indian Institute of Technology, Roorkee for providing the facilities to carry out this work. I also acknowledge Affymetrix, Inc. and Broad Institute for providing microarray datasets to enhance further research work.

I am extremely thankful to Dr. V.P. Subramanyam Rallabandi, National Brain Research Center, Gurgaon for the brotherly counsel, motivation and support rendered.

I am indebted to my father Dr. K.N. Sharma and my mother Santosh Sharma for their eternal support and everlasting blessings. I gratefully acknowledge the inspiration and unflagging support of my parents and my mother-in-law Mrs. Kanta Joshi during final days of my thesis writing. I wish to express my deepest gratitude to my sisters, Dr. Aarti Sharma, Er. Shruti, IAS and her husband Mr. Shailesh, IPS along with his family. I wish to express my wholeheartedly thanks to my paternal and maternal family, my brothers, my brothers-in-law and their families.

I express my sincere appreciation and gratitude to my wife Jyoti Sharma and my son Yeemish Shiv Sharma, for bearing joyfully all sorts of inconvenience and hardship during my involvement in this work.

My special thanks are due to all my senior, fellow and junior research scholars at IIT Roorkee, for their cooperation and moral support from time to time.

Last but not the least, I express my sincere thanks to all my friends and well wishers those who have directly or indirectly helped me at various stages of this work.

(Ajay Shiv Sharma)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

The abbreviations used in the text have been defined at appropriate places, however, for easy reference, the list of abbreviations are being given below.

Abbreviation	**Stands for**
A	Adenine
AD	Anno Domini
ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
BC	Before Christ
BP	Biological Process
bp	Base pair
BPCCMF	Biological Process, Cellular Component, Molecular Function
C	Cytosine
CC	Cellular Component
cDNA	complementary DNA
CMS	Content Management System
Cy3	Cyanine green
Cy5	Cyanine red
DAG	Directed Acyclic Graph
DB	Database
DBSF	Drupal Bioinformatic Server Framework
dChip	DNA-Chip Analyzer
de motu cordis	On the Motion of the Heart and Blood
DNA	Deoxyribonucleic acid
EBI	European Bioinformatics Institute
EGFR	Epidermal growth factor receptor
EMBL	European Molecular Biology Laboratory-
ER	Estrogen Receptor
ERK	Extracellular signal-regulated kinases
FDA	U S Food and Drug Administration
G	Guanine

GC-content	guanine-cytosine content
GEO	Gene Expression Omnibus
GO	Gene Ontology
GOC	Gene Ontology Consortium
GOFFA	Gene Ontology for Functional Analysis
GONUTS	the Gene Ontology Normal Usage Tracking System
GSEA	Gene Set Enrichment Analysis
HER1	human epidermal growth factor receptor 1
HER2	Human epidermal growth factor receptor 2
HGP	Human Genome Project
HGU133	Affymetrix Human Genome U133A Array
HGu95	Affymetrix Human Genome U95A Array
HIV	Human immunodeficiency virus
HPC	High Performance Computing
IGF1R	Insulin-like growth factor 1 receptor
IPA	Ingenuity Pathway Analysis
ITALICS	ITeartive and Alternative normalization and Copy number calling for affymetrix Snp arrays
IUPS	International Union of Physiological Sciences
KEGG	Kyoto Encyclopedia of Genes and Genomes
LAMP	Linux,Apache,MySQL,PHP
MANOR	MicroArray NORmalisation
MAS 5.0	Affymetrix®. Microarray Suite.
MBEI	Model Based Expression Index
MDB	Microarray Database Bioinformatics
MeSH	Medical Subject Headings
MF	Molecular Function
MGI	Mouse Genome Informatics
MGI	Mouse Genome Informatics
MIAME	Minimum Information About a Microarray Experiment
MiMI	Michigan Molecular Interactions
MLL	Mixed Lineage Leukemia
MM	Mismatch

mRNA	Messenger RNA
mRNA	Messenger RNA
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
NIH	National Institutes of Health
NSCLC	Non-small cell lung cancer
OBO	The Open Biological and Biomedical Ontologies
ODE	Ordinary differential equation
OMIM	Online Mendelian Inheritance in Man
OSD	The Open Source Definition
PAGE	Parametric Analysis of Gene Set Enrichment
PCR	Polymerase Chain Reaction
PDE	Partial differential equation
PepX	Protein-peptide complexes database
PHP	The Hypertext Preprocessor
Phy-SIM	Physiological Model Simulation, Integration and Modeling
PM	Perfect Match
PM/MM	Perfect Match/Mismatch
PM/MM/BG	Perfect Match/Mismatch/Background
POC	Plant Ontology Consortium
PPDB	Plants Physiology Database
RAID	Redundant Array of Independent Disks
RDBMS	Relational Database Management System
RGD	Rat Genome Database
RNA	Ribonucleic acid
RNA-Seq	RNA Sequencing
SAGE	Serial Analysis of Gene Expression
SGD	Saccharomyces Genome Database
SNP	Single-nucleotide polymorphism
SQL	Structured Query Language
T	Thymine
TAF	Tumour angiogenesis factor
TKI	Tyrosine-kinase inhibitor

VEGF	Vascular endothelial growth factor
XML	Extensible Markup Language
Y2H	Yeast Two-Hybrid System

Chapter – 1
INTRODUCTION

1.1 MOTIVATION

With the completion of the human genome project, followed by the rise in high-throughput technologies like microarray sequencing platforms and scattered biological ontologies clearly necessitates the use of interdisciplinary aspects of contemporary bioinformatics and computational systems biology. The huge information gap requires mining gene based data for interesting associations to develop summaries that directly address the biological questions posed. Contemporary bioinformatics approaches are a major research area that caters to the need for information rich knowledgebase tools through existing biological databases in a particular domain.

As the better understanding of many pathological conditions is the ultimate goal of these specific informatics tools, the key challenge lies in how to integrate and collaborate to build descriptive physiological models. As such, integrative models to achieve the goal of a full physiology are needed after reviewing the overall strategy ranging from functional genomics and its ontologies to functional understanding. Bioinformatics standards and tools for data integration provide the means to enhance cross-software interoperability and data exchanges between laboratories to realize physiology through gene ontology databases. Data integration and collaboration requires the development of unified ontologies precisely to define and categorize biological concepts and data properties.

Recently, there has been a huge interest in the development of a generalized biological framework for gene ontology data mining tools. Recent studies also indicate that plants are biologically more complex and there are enormous applications to be gained from researching plant genes to progress the reception of nutrients from the earth to enhance plant yields and plant ailments that directly affect the health of humans. Therefore, a demonstration of extensible tools like a ‘plants physiology database’ will bridge the gap between bioinformatics and gene ontology data mining. The databases populated with derived information after mining from gene ontology database are of considerable importance for biological discoveries. This facilitates sharing and integration of data in the research community. Web services have also become indispensable tools in bioinformatics. Services provided by host institutes provide data and application access to scientists located anywhere in the world without having to locally install and maintain these resources. These repositories can serve as powerful knowledge resources for a physiology of interest and facilitates further studies. To represent such biological information in an unambiguous and computationally tractable way is one of the ongoing challenges faced by computational systems biology.

Systems biology has been in the focus of intense public and private research in recent years evoking high expectations and hopes with regard to solutions for emerging problems in the health care sector. The notion of systems biology is a rather broad agglomeration of mathematical and computational methods, experimental techniques, and biomedical applications. Systems biology offers promising insights to defeat cancer.

Cancer in its many forms has remained a devastating and pervasive disease for thousands of years. While countless numbers of physicians and scientists through the centuries have been dedicated to battling and understanding cancer, it ultimately remains elusive. Gains have been made, yet more progress is urgently needed. The origin and subsequent growth of cancer is related gene dysfunctions. Since the sequencing of the entire human genome, scientists and researchers can now analyze cancer genes and cells at the molecular level. This has significantly affected cancer research that uses data intensive computational tools to view never before seen biological complexities using data and model driven systems biology of cancer.

Microarray and next generation sequencing technology has greatly influenced our understanding of human diseases as well as clinical practice, such as tumour classification, treatment response and outcome prediction. Gene expression profiling is a promising research area for diagnosis and predicting outcomes in cancer patients after mining the genetic information from the massive data generated from these technologies. Most studies have relied on gene expression microarrays to broadly survey the expression level of many genes and has provided a method for sub grouping heterogeneous cancer types. The heterogeneity within the tumour has a direct consequence on personalized medicine that further needs detailed research. As such, the central dogma of a microarray data analyses can pave the way for innovative opportunities for early diagnosis of malignancies. For instance, a precautionary double mastectomy on finding the BRCA1 gene with only 87% probable chance of acquiring the disease shows the promising nature of this field.

Cancer can be seen as network deregulation of pathologies and processes that govern differentiation, proliferation, and apoptosis in the cellular microenvironment. The questions that can be answered with systems biological approaches are of the following types: What are the cellular pathways involved in the pathology? How to use these pathways to improve predictions? What are the effects of a perturbation on a pathway? Many more questions can be proposed. Mathematical modelling provides tools to capture these issues using model driven systems biology of cancer. Multiscale modelling of cancer and its interaction with drugs is a current research area aimed at resolving various complexities of this disease. These models can help to quantify and recapitulate experimental observations. With the power of model driven or knowledge-based models, hypotheses can be tested in silico before performing experiments in the wet lab.

The great motivating factor is that systems biology of cancer allows us to contemplate cancer cells at the molecular level and to detect phenomena unobservable through the microscope. Moreover, in developing countries like India and the country's food bowl Punjab where cancer incidence rate is the highest and the mortality rate is increasing (http://www.ncrpindia.org/), survival rates are much worse because of late detection and low quality or inexistent healthcare facilities where telemedicine can play a vital role [1]. Thus, cancer systems biology approaches discussed here offer new promising insights to defeat cancer that requires a paradigm shift from the traditional handling of such patients.

1.2 STATEMENT OF PROBLEM

Contemporary bioinformatics and systems biology are an emerging discipline expanding beyond traditional bioinformatics, with a focus on developing computational technologies for real-world biomedical bottlenecks. The problem statement of this thesis can be stated as four concerns:

1) The design of gene ontology (GO) based models to realize physiology is the first primary concern. Although GO annotations may develop into a powerful tool in the future, they are currently limited and incomplete since they depend on database entries by individual investigators. This is particularly problematic, as the translation of biological data into GO terms is highly dependent on a researchers view and scientific background of a given subject. Even more concerning is the limited quality control of the data entries. In addition, just as the initial GO assignments may initially be time consuming and difficult, the problem of keeping results and annotations up to date adds another layer of complexity to the problem. The success of descriptive ontology based full physiological model depends on validated GO databanks.

2) In the recent past, scientists witnessed an overload of GO database ontologies for describing domain specific entities. GO databanks put forward ways to organize information. In addition, each GO based solutions provides a different outlook into the

problem being considered. To advance seamless information discovery, computational tools are required to regularly query across ontologies and develop tailor-made databases from the standard GO databanks. Efficient GO data mining tools are indispensable for representing comprehensive biological systems in an unambiguous and computationally tractable way.

3) With the advent of whole-genome sequencing and high-throughput microarray experimental technologies, transcriptome analysis is an important challenge for analyzing these large-scale data sets and extracting discernible biological information from them. Existing statistical or bioinformatical approaches have several challenges for extracting the differential expressed genes to make a distinction between tumour subtype/phenotypes and integrating heterogeneous genomics data effectively to advance cancer research.

4) In order to promote human health, advances in the microarray database technology and sequencing technologies must be rendered into successful clinical practices. Computational systems biology is reviving a long-standing dream of modelling a whole organism in silico or in other words, formalizing life. For a systematic disease like cancer, the idea is to construct a model of the human physiology at the level of the whole body. Most of the existing models are temporal based using ordinary differential equations (ODEs). In general, most solid tumours have cells that migrate and diffusion properties, therefore spatiotemporal models are necessary (Partial differential equations). Models need to be developed based on the mechanisms of biological processes that utilizes experimental data, validates, and predicts the outcome.

1.3 LITERATURE SURVEY

In recent years, considerable efforts have been devoted to computational systems biology and contemporary bioinformatics methods. The review of selected state-of-the-art for gene ontology data mining and systems biology of cancer and their related issues which have been considered in this thesis are divided into sub-headings as discussed below:

1.3.1 Ontology

The term ontology was originally used to define a philosophical discipline. Ontologies have been studied by philosophers since the time of the ancient Greeks [2] yet today the term ontology is defined as the explicit specification of a conceptualization for use in computer science and information systems [3, 4]. The conceptualization is a representation of a worldly phenomenon that captures all the concepts and requires ontology to satisfy the subsequent conditions [5]: first, all concepts and constraints on the concepts must be explicitly defined. Second, it must capture the state of knowledge from a domain as agreed upon by a group of people. Third, it must be computationally amenable.

Ontology can support a variety of applications including knowledge engineering, artificial intelligence, information retrieval, and integration of databases [6]. The study and use of ontologies has been gaining attention over the past decade due to the exponential growth in the biological space in open domain [7]. Much of the prior research has focused on technical aspects of ontology management. This includes the development of languages for representing ontology [8], yet not much attention has been paid to the investigation of domain based issues related to ontology engineering tools [9]. Creating these ontology based tools takes a lot of effort and time and acts as one of the main obstacles faced by developers of ontologies. Therefore, identifying a relevant knowledge source and organizing it as a part of an ontology are serious challenges [10].

The design of ontology is still more of a craft than an engineering task [11]. Existing ontology creation methodologies provide a range of options, techniques, and guidelines to help ontology construction [12]. Ontology architecture can be classified [13] and ontology creation methodologies can be sorted into top-down and bottom-up views [14]. Top-down methods start with an abstract view of the domain and expand it with detailed specifications [15, 16]. Bottom-up methodologies start from the specification of a certain task and obtain generalizations [17, 18]. Some methods take a middle-out method where the ontology creation starts with key concepts and then generalizations and specializations are created [19]. The key challenges of all these ontology creation methodologies are identifying a relevant knowledge source and a significant manual exertion. Hence, there is increasing interest in the automated creation of ontologies from widely available knowledge sources [20]. If this knowledge could be extracted and organized automatically, it could be effectively used to create domain ontologies. However, prior research on systematic analyzing and using the World Wide Web as a source of knowledge for the creation of domain ontologies is limited.

1.3.2 Biological Ontologies

Ontologies enhance inter-operability between heterogeneous data sources and enable the reuse of data. Ontologies have emerged as the chosen mode of representation of domain-specific concepts and specifications in biology. There is a good review of biological ontologies with examples that describe its creation, applications and future opportunities [21]. The open biological and biomedical ontologies foundry lists over hundred active bio-ontologies (http://www.obofoundry.org/) currently used by the biological and biomedical community [22]. The most widely used of all the computational biology ontologies is Gene Ontology (GO) [23]. The concept of gene ontology was introduced for annotating data in a more relevant and tractable form, so that usable information from the databases can be extracted. Even though more general definitions of GO were previously proposed, an aspect of biological relevance is a tool for the unification of biology [23]. It is in fact a large public database providing a set of controlled gene products vocabulary based on their role within a cell. It also holds data from diverse data resources like GenBank to generate gene annotation data [24]. Scientists deposit their biological findings in GenBank to expedite research, for example RSP-09 lipase gene was submitted to the database with accession number EU414610 [25]. This facility is being extensively exploited by researchers from various backgrounds. Therefore, biological ontologies curation future require cross-linking structure, support and recognition for prosperity of this field [26].

1.3.3 The Gene Ontology

The pioneering works of the Gene Ontology (GO) (http://geneontology.org/) informatics resource transformed genomics biology to a descriptive biology [27]. At the time of its conception, the need for GO was powerful and straightforward: different molecular-biology databases were using different terms to describe important information about gene products [28]. GO started its operation in 1998 as a collaborative attempt for knowledge integration from FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Informatics (MGI) project [29]. Currently, more than fifteen bio-curator groups are working for manual and automatic annotations of GO. GO is the benchmark genomics product ontology based informational database warehouse functionally cross-linked with different species and annotated knowledge databases. This is an extremely useful informational resource for classes of species with restricted experimentally validated biological knowledge where electronic/computer inferred annotations occasionally provide and fill biological information gaps [30].

GO provides a standardized, species-independent representation for the characteristics of genes and gene products where gene products are the biochemical materials produced by gene expression. GO provides a controlled vocabulary for describing characteristics of gene products and is composed of three major components that are the heart of physiology, namely, biological processes (BP), cellular component (CC) and molecular function (MF) [collectively termed as (BPCCMF)]. The process of assigning GO terms to gene products is referred to as annotation [31]. Figure 1.1 and Figure 1.2 illustrate the rapid growth of GO both in terms of the number of GO terms and the total number of GO annotations assigned to gene products. The present and up-to-date GO repository statistics information can be seen at following web link (http://www-test.geneontology.org/page/current-go-statistics).

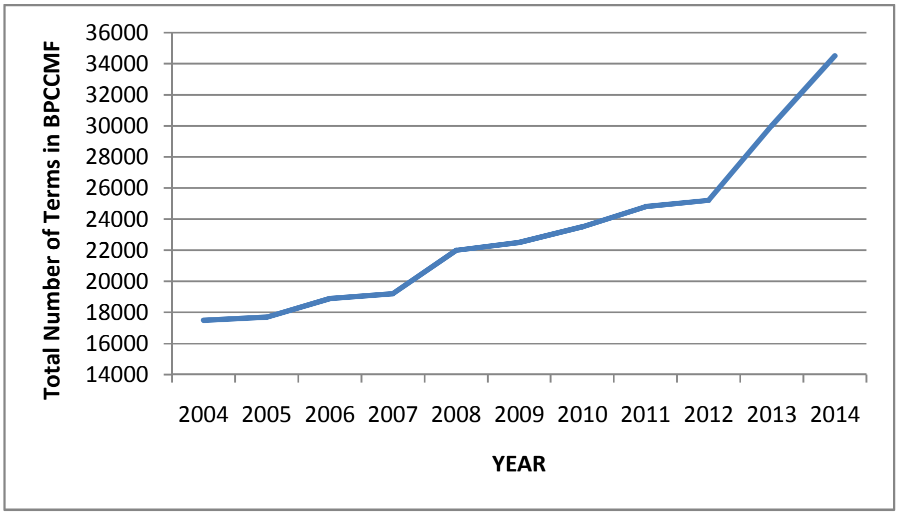


Figure 1.1 Number of terms in Gene Ontology

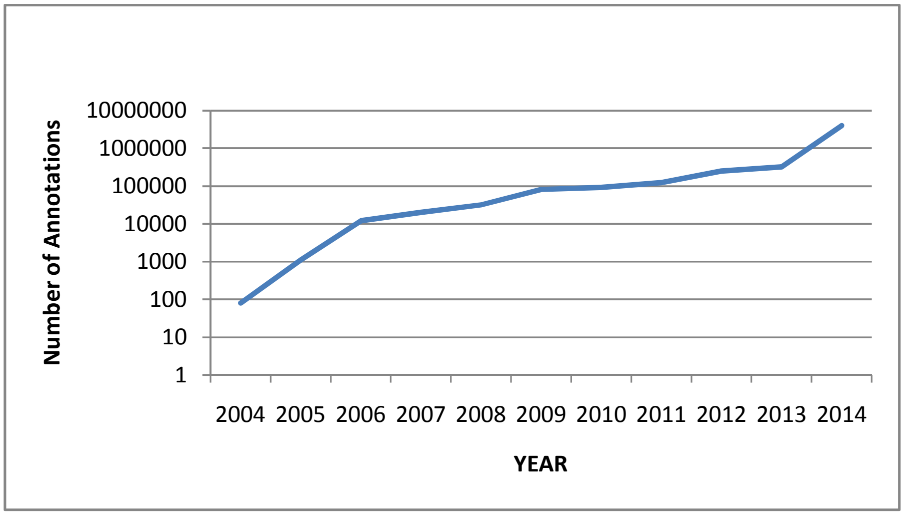


Figure 1.2 Number of Gene Annotations

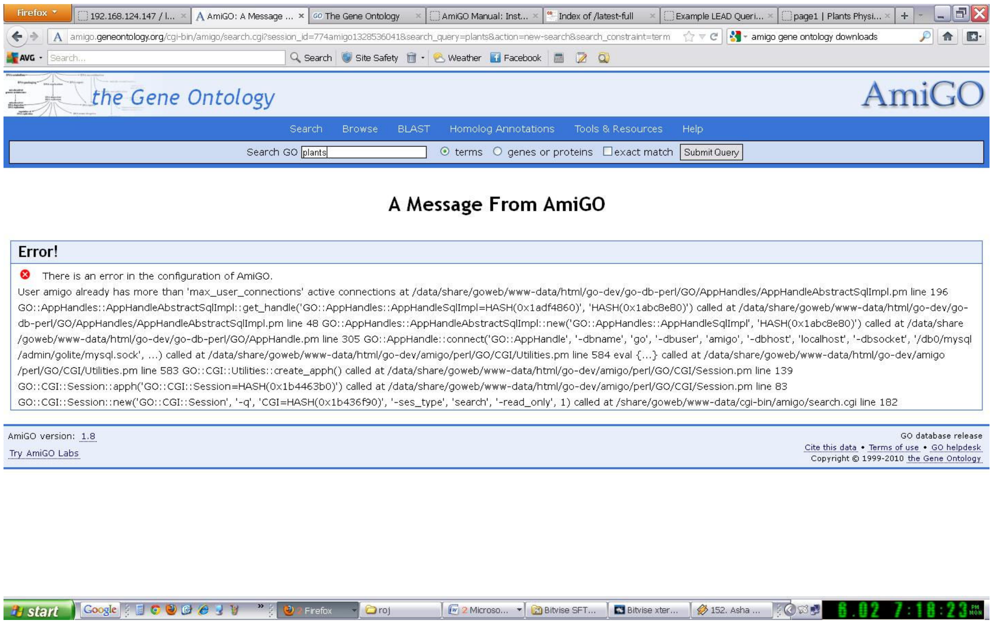
There is increasing research interest in identifying new relations and connections between the three ontologies of the GO [32-34]. Effective data mining techniques are needed to extract value from gene expression data represented by gene based ontologies [35-37].

1.3.4 GO Data Mining

Data mining holds the key for the knowledge discovery process in huge databases by extracting or mining knowledge from colossal quantities of data [38]. The vital aspect of data mining is data retrieval in order to find knowledge in a database warehouse for additional utilization and to present the newly obtained knowledge in a user friendly manner [39]. Data mining is an important component of the knowledge discovery process and employs algorithmic techniques to reveal implicit patterns and relationships from data [40] by means such as user-centric model architecture [41]. Data mining algorithms are routinely applied to bioinformatics data to convert the data into meaningful information that can be of value to researchers [42, 43]. [44] presented an outline of the field of knowledge discovery and data mining (KDD) with some related issues.

Many researchers have attempted to create ontology from target oriented data mining to work out specific goals. Presently, the majority of all gene ontology developments workflow are aimed at computerized data mining [45-47] and are capable of providing knowledge discovery services on the grid/cloud computing environment [48, 49]. Other directions about possible interactions among ontology have aimed at modelling concepts in different fields like the semantic web [50] and information science [51].

The GO database repository is continually being developed by collaborations, which has created a huge knowledge repository [33] that has made it very difficult to find relevant components of knowledge from it. When queried with keywords, this data warehouse may provide a large amount of knowledge that often includes hundreds of irrelevant terms along with other technical issues like data manageability as illustrated in Figure 1.3. Thus, collecting conceptually consumable information from large ontologies has proven to be a very difficult task. GO data mining based on ontology pruning is a solution that removes irrelevant concepts or useless elements to create a subset ontology with relevant elements [52]. Earlier research indicated that pruning is effective in building specific domain based ontologies in medicine [53, 54]. Also, there is claim that pruning increases the effectiveness of a sub-ontology by retaining only relevant concepts [55]. This need for a more enhanced version of GO, recently being developed at the following web link (http://beta.geneontology.org/) and provides a better user experience and integrated data handling capabilities that will ultimately advance the understanding of physiology. In addition, all gene ontological data will be released on a periodical basis, freely offered and can be downloaded in varied formats from the updated GO web link (http://www-test.geneontology.org/page/download-ontology). This will inspire researcher's in particular scientific communities to use this gene ontological database warehouse to build specific tailor-made databases to fulfil their needs from the standard GO ontology repository.



A Message From AmiGO

Error!

```
There is an error in the configuration of AmiGO.
User amigo already has more than 'max_user_connections' active connections at /data/share/goweb/www-data/html/go-dev/go-db-perl/GO/AppHandles/AppHandleAbstractSqlImpl.pm line 196
GO::AppHandles::AppHandleAbstractSqlImpl::get_handle('GO::AppHandles::AppHandleSqlImpl=HASH(0x1adf4860)', 'HASH(0x1abc8e80)') called at /data/share/goweb/www-data/html/go-dev/go-db-perl/GO/AppHandles/AppHandleAbstractSqlImpl.pm line 48 GO::AppHandles::AppHandleAbstractSqlImpl::new('GO::AppHandles::AppHandleSqlImpl', 'HASH(0x1abc8e80)') called at /data/share/goweb/www-data/html/go-dev/go-db-perl/GO/AppHandle.pm line 305 GO::AppHandle::connect('GO::AppHandle', '-dbname', 'go', '-dbuser', 'amigo', '-dbhost', 'localhost', '-dbsocket', '/db0/mysql/admin/golite/mysql.sock', ...) called at /data/share/goweb/www-data/html/go-dev/amigo/perl/GO/CGI/Utilities.pm line 584 eval {...} called at /data/share/goweb/www-data/html/go-dev/amigo/perl/GO/CGI/Utilities.pm line 583 GO::CGI::Utilities::create_apph() called at /data/share/goweb/www-data/html/go-dev/amigo/perl/GO/CGI/Session.pm line 139
GO::CGI::Session::apph('GO::CGI::Session=HASH(0x1b4463b0)') called at /data/share/goweb/www-data/html/go-dev/amigo/perl/GO/CGI/Session.pm line 83
GO::CGI::Session::new('GO::CGI::Session', '-q', 'CGI=HASH(0x1b436f90)', '-ses_type', 'search', '-read_only', 1) called at /share/goweb/www-data/cgi-bin/amigo/search.cgi line 182
```

AmiGO version: 1.8
Try AmiGO Labs

GO database release
Cite this data • Terms of use • GO helpdesk
Copyright © 1999-2010 the Gene Ontology

Figure 1.3 Error message when GO was searched for ‘plants’

These tailor-made databases or specific domain ontologies tools can be built using data mining in GO. One of the benefits of GO data mining is the generation of user-centric association rules focused on patterns of interest to the user [56]. Mining patterns involving user specified concepts reduces the search space by eliminating items that are not of the user’s interest [57]. The most important benefit of GO data mining is the ability to generate multi-level association rules by shifting the abstraction level in the dataset.

Efficient data mining algorithms are needed to mine the wealth of explicit and implicit information embedded in data annotated using ontologies. GO data mining takes advantage of the structure, semantics and relations of the ontology [58]. Another perspective of GO data mining as shown in Figure 1.5 is its wide usage in a variety of ways to provide a functional perspective on the analysis of computationally predicted gene sets. The analysis of microarray results through analyzing the over-representation of GO terms within the differentially represented genes [59, 60] is a common practice.

Although gene based data mining and knowledge discovery have experienced increased recognition and success recently [61], a generalized GO framework that describes and unifies data mining and bioinformatics is still lacking. Also, there is a need for intrusion detection in database systems to be developed as web GO ontologies to protect data from misuse [62, 63]. This requires implementation of rewriting SQL queries for exposure of data infringement [64]. Such a unifying bioinformatics solution [65] can lead to further development of the specific field knowledge through the power of open source technologies.

1.3.5 Untangling Biology through Open Source Bioinformatics

Most of the biological solutions on the web are promoted using open source technologies also follows the principle of open source initiative available at web link (http://opensource.org/osd). GO usage adheres to the creative commons policy and licensing (http://wiki.creativecommons.org/Frequently_Asked_Questions) and has been developed using open source technologies. A vital aspect of GO data mining is data retrieval in order to find knowledge in a database warehouse for additional utilization and to present the newly obtained knowledge in a user-friendly manner. There are huge opportunities in semantic web mining for biological discoveries through development of knowledge extraction management system [66].

Drupal (https://drupal.org/) is popular among many of the open source web content management system frameworks available [67]. It offers a versatile approach to style, management, and organizing content maintained in a dynamic fashion rather than static web pages. Recently, various frontends for biological ontologies like Tripal [68], DBSF [67], EMBRACE [69], RNA-Seq Atlas [70], CASIMIR [71], PepX [72] uses the power of Drupal. This growing list of Drupal users makes it effortless for integration and expansion into comparative genomics oriented database in the near future.

Recently, the Plant Ontology Consortium is considering shifting the base to Drupal 6 (http://wiki.plantontology.org/index.php/Plant_Ontology_Web_Site_Update:_2013) as shown in wikipage. A very good review for overcoming the challenges of DNA barcoding in plants with bioinformatics solutions is presented along the lines of popular support for mitochondrial cytochrome c oxidase 1 gene (COI) barcoding in animals [73]. There is also great need for plant ontological databases as per the assessment report given in late 2005 [74] where bioinformatics can play a major role [75]. Such requirements can be dealt with designing of effective biological database management solutions in a particular domain [76]. Efficient development of extensible prototypes helps in seamless retrieval of information which is necessity for biological discoveries exploiting semantic web [77]. There is also need for integrative analysis approaches such that all the data can be accessible and queried using unified tools.

Gene ontological solutions are descriptive sciences based on integrative and collaborative approaches [78] which will bridge the gap between bioinformatics and gene ontology data mining using open source (http://en.wikipedia.org/wiki/Open_source) software power. The various bioinformatics challenges with clear objective function [79] enable progress in gene identification that can be solved with computational intelligence methods [80]. Computational tools like Scigress Explorer Ultra 7.7 dcoking software [81], Swiss PDB viewer tool [82] are playing a vital role in modelling approach with research in chemical/mechanical/molecular studies [81, 83] of Cytochrome P450 [84], heme enzymes etc. These studies are contributing significantly to scientific community from different perspectives [85-87]. The major research concerns like cross-boundary decision support systems and their proposed conceptual models like CaDHealth [78] for clinical work practices can also be implemented using open source bioinformatics.

1.3.6 Historical Background of Cancer

The first report of metastatic cancer was found in Edmontosaurus fossils (Cretaceous) and neoplasms were discovered in a Neanderthal skull (35000 BC), Egyptian and Incan mummies [88]. The oldest account of cancer in humans can be found in Egyptian papyri written between 3000-1500 BC. Several of the papyrus scrolls including the Georg Ebers papyrus, the Edwin Smith papyrus (circa 1600 BC) and the Kahun papyri (circa 1825 BC) provide detailed information of diseases and conditions that are consistent with modern descriptions of cancer. In Greece, the father of medicine Hippocrates de Cos (460-370 BC), wrote about cancer in his Corpus Hippocraticum. He used the terms carcinos and carcinoma to reference to chronic ulcers or growths that seemed to be malignant tumours and scirrhus for a cancer with a hard consistency. In Greek, carcinos means crayfish, canker, cancer, tumour, while skirros means solid tumours. The word cancer comes from a Latin translation by the Roman doctor Celsus (28 BC-50 AD) of the Greek word carcinos. In Latin is means crab, crayfish, dunce and cancer, canker and was inspired by cancerous lesions that resembles a crab [89]. Galien (131-201) used the Greek term oncos, meaning mass and referred to a growth or tumour that looked malignant [90]. Evidence of cancer has also been made in art history from several works by Rubens and Rembrandt. This allowed physicians to discover physical changes that suggested tumour in the breast of models they painted [91]. The next 2000 years witnessed several key events that helped to refine further the still ongoing main areas of cancer investigation and treatment. The journey to present age systems biology of cancer comes into picture in the last 60 years with the decoding of DNA code that leaded to further developments as illustrated in the timeline Figure 1.4.

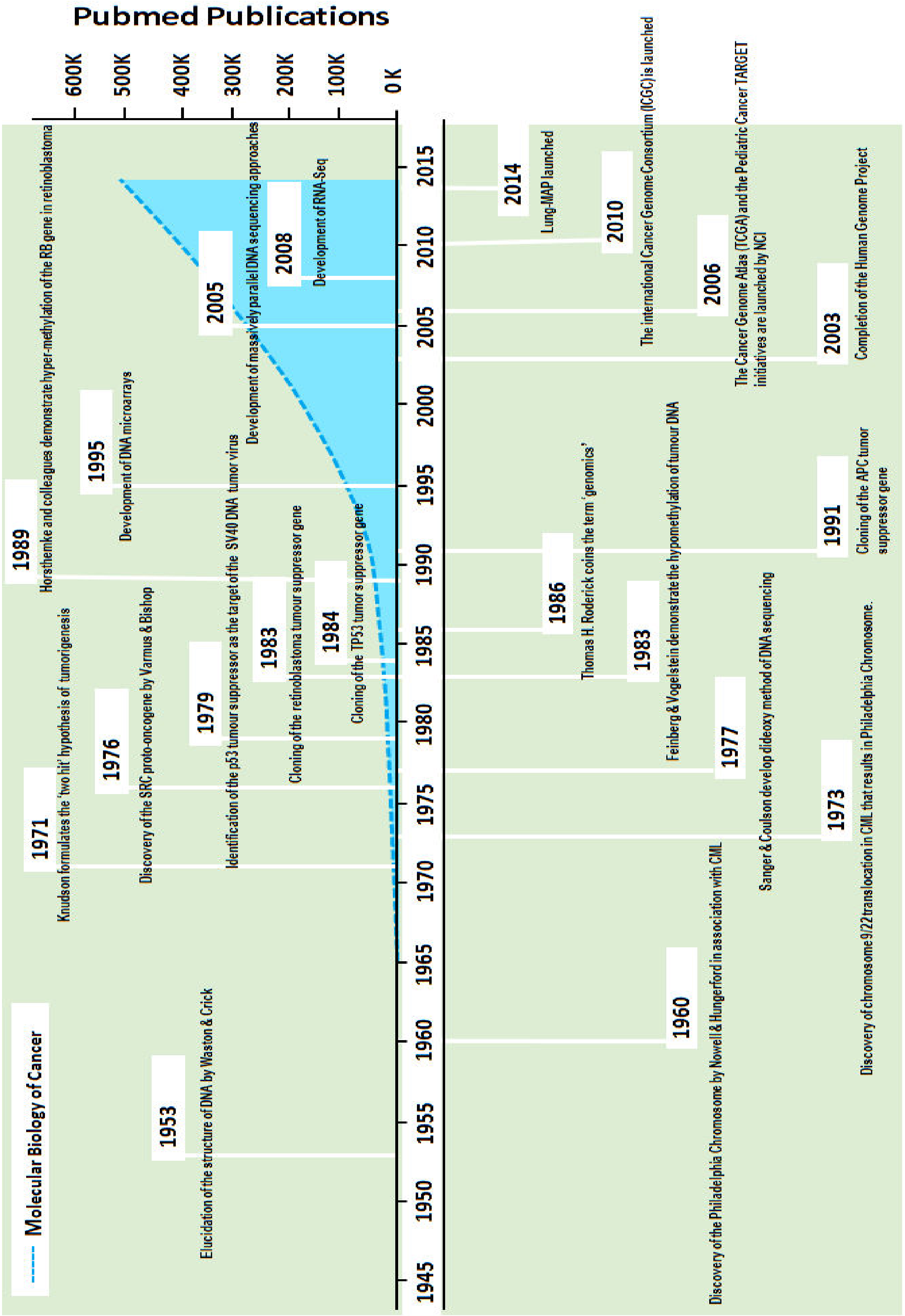


Figure 1.4 Key milestones in Cancer Genomics depicted over a time graph of total number in Pubmed Publications (Adapted from [67])

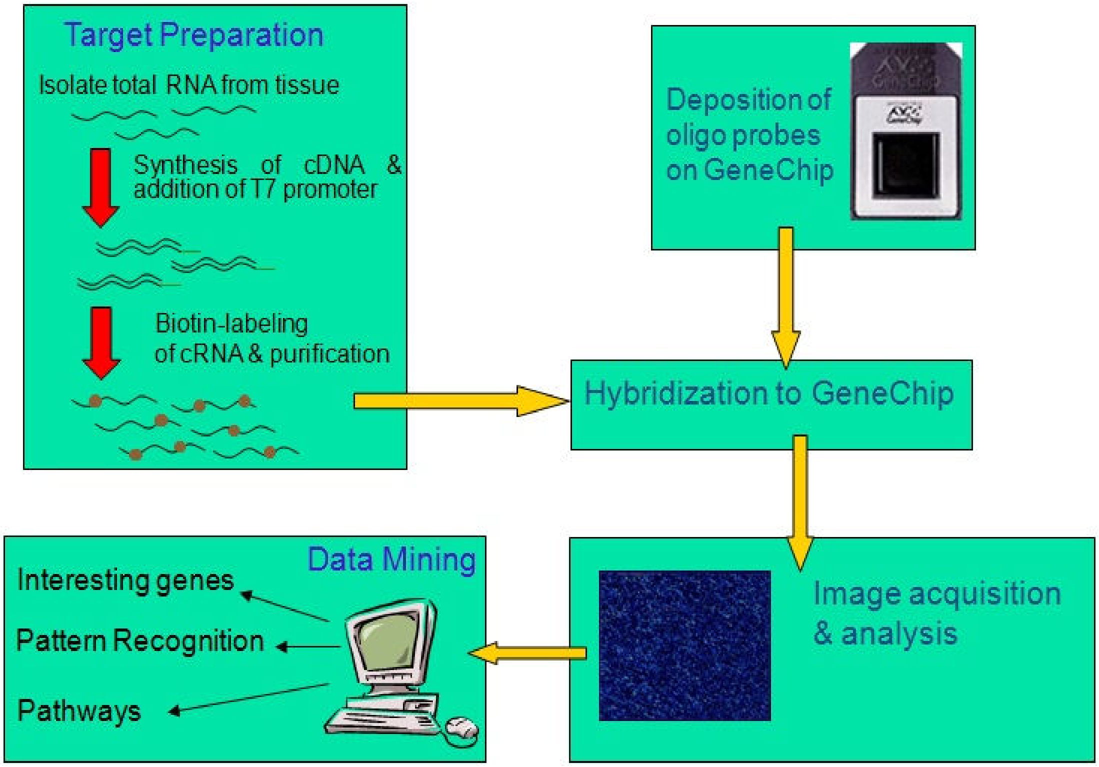
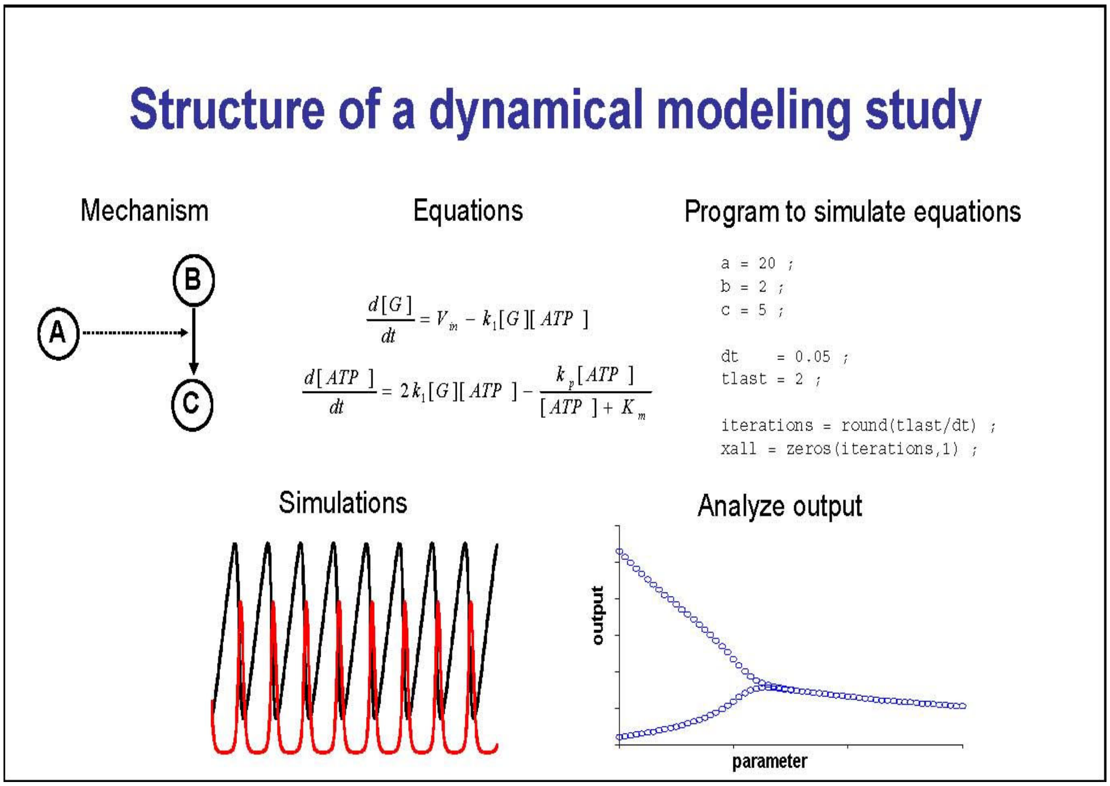

Target Preparation
Isolate total RNA from tissue
Synthesis of cDNA & addition of T7 promoter
Biotin-labeling of cRNA & purification
Deposition of oligo probes on GeneChip
Hybridization to GeneChip
Image acquisition & analysis
Data Mining
Interesting genes
Pattern Recognition
Pathways

Figure 1.5 Illustration of data driven systems biology workflow



Structure of a dynamical modeling study

Mechanism

Equations

$$\frac{d[G]}{dt} = V_{in} - k_1[G][ATP]$$

$$\frac{d[ATP]}{dt} = 2k_1[G][ATP] - \frac{k_p[ATP]}{[ATP]+K_m}$$

Program to simulate equations

```
a = 20 ;
b = 2 ;
c = 5 ;

dt    = 0.05 ;
tlast = 2 ;

iterations = round(tlast/dt) ;
xall = zeros(iterations,1) ;
```

Simulations

Analyze output

Figure 1.6 Illustration of model driven systems biology workflow

1.3.7 Systems Biology of Cancer

Systems biology is transforming cancer research [91, 92].There are several definitions of systems biology proposed till date [93-98] which can be differentiated into two schools of thought. As such, one branch of systems biology is based on data-driven methods where as the other one is based on model-driven methods. In the past, computational systems biology [99] research has been dominated by bioinformatics methods which aim at sequence alignments, finding patterns using data mining approaches and being overtaken by sophisticated statistical tools these days [100] handling large datasets produced from sequencing technologies like microarrays as shown in Figure 1.5. Various stochastic approaches [101, 102] for systems biology of cancer based on data driven approach is currently underway. The second school of thought is based on models which simulate the dynamics of biological system and studies the dynamical behaviour of biological systems by focusing on the interaction of their components [103] as shown in Figure 1.6. The idea is that these behaviours and in particular biological functions are intrinsic properties of the systems that emerge from the interactions between components and cannot be revealed by the study of individual components [104]. Since cancer is a disease of interactions, the systems biology cancer research deals with interactions within signaling pathways and between signalling pathways, interaction between cells and their microenvironment [105]. The most famous review on hallmarks of cancer systems biology recapitulated a quarter century of molecular oncology research and anticipated a major change in our paradigm of cancer research [88, 90] which are the backbone of data driven and model driven systems biology of cancer.

1.3.8 Data Driven Systems Biology of Cancer

Over the last decade, systematic characterization of gene expression or transcriptome profiling of cancer samples has been largely performed with either microarray technologies or Sanger sequencing methods [67]. Microarrays, which are based on synthesis of oligonucleotide probes and subsequent fluorescence through target hybridization, provide high throughput and low-cost measurement of mRNA abundance or gene expression [106]. They have been used to survey every major type of cancer [107, 108].

High-throughput hybridization-based microarrays have been widely used in cancer research as a tool to systematically profile gene expressions since their first introduction in 1995 [109]. As independent of the platform and the analysis methods [110, 111] used, a result of a microarray experiment is the most often a list of differentially expressed genes. This technology has greatly impacted our understanding of human diseases as well as clinical practice, such as tumour classification, treatment response and outcome prediction [112, 113]. Now, with the routine availability of high-throughput DNA sequencing, millions of sequencing fragments of lengths between thirty six and two hundred sixty two bp (base pair) and higher can be routinely profiled [114] using either SAGE like [115] or global sequencing strategies [116].

The systematic profiling of various cancer types using statistical methods has been among the first applications of microarray-based transcriptomic studies in the late 1990s [117-119] and has since remained at the forefront of applications targeted by new omics technologies. In order to directly answer questions such as detecting genes which are differential expressed in samples which exhibit a specific genomic alteration, the most direct method is to stratify the samples based on the genomic alteration and analyse the gene expression in relation to this stratification [120-122]. Another approach which does not require explicit stratification is to compute the correlation between each DNA copy number and each gene expression, or formulate the problem as a regression problem [123], in order to detect genomic loci whose amplification is strongly associated to variations in expression of some genes [124]. Comparative methods can also be employed on identified genes and their biological pathways as illustrated in the study of abiotic stress [125].

Major efforts have been put into analyzing microarray data. Many different methods are becoming readily available as comparison of several methods for Affymetrix microarray data analysis [126] used for disease profiling is studied [127]. One issue existing in most of these methods is how to interpret the results when hundreds of genes are found to be important. This includes the earlier approaches such as Fisher's exact test [128], Gene Set Enrichment Analysis (GSEA) [129], Parametric Analysis of Gene Set Enrichment (PAGE) [130]. There is huge literature available on normalisation of microarray data [131, 132], also comparison of different normalization strategies in the context of cDNA microarrays is also reported [133]. A review on statistical tests for differential expression in cDNA microarrays is also given [134]. The fold change can be found in some earlier literature on microarray data [135, 136]. Even though the performance of the different methods is in the same range, there can be significant differences. It is observed that there is no single best method as analysis depends on many factors including the data analysed, the number of samples and of features, and the experience of the programmer [137, 138].

The application of gene expression profiling in cancer has significantly aided the identification of genes involved in pathogenesis [139, 140] and progression [141, 142], has enabled the identification of prognostic expression signatures, and has provided a method for sub grouping heterogeneous cancer types [143]. Most studies have relied on gene expression microarrays to survey broadly the expression level of many genes such as ER and HER2 status usually measured by pathologists in the clinics can be recovered, allowing in principle the automatic classification of each tumour in one of the four class subtypes few genes [144].

This approach can be visualized as bottom up approach where computational biologists extract biological knowledge after analyzing data from huge microarray datasets [145] or next generation sequencing technologies data [146]. Such models are heuristic based which integrate observations while finding differentially expressed genes [147]. Data driven models are basically heuristics to make general little assumptions about the internal mechanisms of the system and are well suited in the situations of uncertainty [7, 148]. At present, an automatic ontological analysis approach, which helps us interpret such heuristic results, is the de facto standard for the secondary analysis of high throughput experiments.

1.3.9 Model Driven Systems Biology of Cancer

Systems biology aims at sketching the main traits of the cell, i.e. the major transitions, the signalling pathways or the interactions between key players involved in a biological process [103, 149]. Technologies like Fluorescence Resonance Energy Transfer (FRET) are used to observe nano-scale interactions inside living cells for sensitivity analysis and can improve nucleic acid detection [150]. The knowledge in networks has also increased tremendously over the past decades, due to the emergence of high throughput technologies and our capability of analysing them [151]. It is assumed long time back that cellular processes are based on complex networks of interacting genes and proteins [152]. More recently, cancer was referred to as systems biology or a network disease [91]. As a result, modelling cancer starts with the study of possible deregulations of normal cell cycle.

Cancer cells often fail to respond to external signals that would halt proliferation of normal cells, therefore in cancer cells, cell cycle checkpoint mechanisms that should stop the cycle in abnormal situations are altered. For all these reasons, the study of a normal cell cycle seems to be a good starting point to the study of cancer cell cycle [153]. The transitions from one phase to another are thus monitored by checkpoint controls. Since cancer is a disease of uncontrolled and excessive proliferation, some of these checkpoints malfunction in many different cancers. One of the first mathematical interpretation of cell cycle was published in 1962 [154]. Later in 1990s several groups developed complex models of cell cycle mechanisms [155]. Since then articles related to cell cycle modelling have exploded [156].

In 1974, [157] proposed the theory of a restriction point for animal cell proliferation. Many models have explored the restriction point from a theoretical point of view. Among them [158], studied the effect of varying concentrations of the genes that have an influence on the restriction point. Similarly [159], among others characterised the molecular features of the restriction point. [160] reproduced the experiment done by [161] that aimed at identifying the precise time of the restriction point in late G1 and to conclude [162] measured the importance of some cell cycle actors in the control of the positioning of the restriction point. From the clinical perspective, radiation oncology is prevalent and has become indispensible tool for cancer research [1, 163, 164].

One popular formalism used when modelling quantitative data is use of chemical kinetics [165] based on systems of nonlinear ordinary differential equations. The most appropriate type of networks for ordinary differential equation (ODE) modelling being reaction networks from biological point of view. Some works have been proposed on the translation of Boolean models to ODEs models [166] and that some methodological works are currently done on the automatic translation of reaction networks to influence networks. Some more recent computational analyses have been further performed on this model [162]. Theoretical works have explored some of these networks motifs and transposed the biochemical reaction network into an influence network [167] which requires agent based modelling.

Agent based modelling is a powerful simulation modelling technique that has seen a number of applications in the last few year, including cancer biology [168]. It allows one to encapsulate complex patterns of objects behaviour in the form of rules [169]. Agents interact in a competitive and repetitive fashion. Agent rules can incorporate physical and chemical laws as illustrated in the tutorial [170]. For example, in an agent based model of liver lobule response to intoxication by paracetamol, the rules governing cell behaviour were based on physical adhesive cellular properties and physical motion equations [171].

The long-term goal of mathematical modelling is to construct a virtual object that would mimic this object's behaviour in real conditions, understand, and predict the behaviour of perturbations. In recent times, mathematical modelling and simulations of biological processes has become an important tool. However, only a few models of solid avascular tumour growth [172] and multidimensional tumour growth models [173-175] have been proposed and a few models of tumour growth coupled with the process of angiogenesis [176, 177] were developed [178-180]. Recently, a few models incorporating chemotherapy drug treatments with/without angiogenesis have been proposed [181-183].

Thus, there is a great demand for multiscale modelling [174] that takes into account multiple spatial, temporal or structural scales and is of particular importance in studying cancer. There are numerous applications of multiscale modelling approach in cancer biology [184] leading towards physiological models.

1.3.10 Physiological Models

There are several independent physiological models reported [185-188] in the history of mathematics and medical science since first publication of "*De motu cordis*" in 1628 detailing cardiac yields in animals. These models are self-sufficient in their particular domains but for a full physiology of human beings, there is still huge research gap. With the advancement in computing power, the first human physiological model was proposed [189] which is freely available (http://faculty.pnc.edu/pwilkin/human.html) as a software package to advance the research in human physiology. The last five decades has witnessed major breakthroughs from an informatics point of view of the central dogma of molecular biology with the emergence of genomics leading to physiological functions [190]. Recent contributions to anatomy with statistical inferences based on classification methods like a principal component analysis [191] were drawn along with various integrative approaches for dynamic modelling and simulations of physiological time series has given promising insights [192, 193]. Also, a genomics view point can further led to the growth of new ideas and concepts guiding genome structural design or the realization of physiological states model [194].

The post genome era aided by next generation sequencing (NGS) technologies and current microarray database technology paves the path for functional genomics. This enables scientists to decipher novel genes ultimately leading towards physiological picture of the organism under observation. The success of physiological models lies in the validation of inferred genes with the wet lab experimental biological database [190]. There are several scientists and researchers throughout the world contributing towards human physiology but in independent manner. Therefore, there is an urgent need for collaboration and integration of useful biological data using contemporary bioinformatics to make holistic data warehouses or data grids where the integration can be done.

One of the best examples in this direction is available in the link (http://www.physiome.org/) for the IUPS Physiome Project, which still lacks support of different established independent groups in the physiological field. In recent times, multi-scale, multi-level models were proposed for integrating advancements in the field of biomedicine [195]. Also, systems biology approaches that lead to a physiological picture were proposed for research using top-down or bottom-up prototypes [196]. Due to the interdisciplinary nature of work, systems biology gives holistic solutions by unifying existing and evolving knowledge to interpret a full physiology. As such, systems biology has gained so much significance and has bridged the gap between experiments in vivo, in vitro, and in silico working in tandem for reproducible results [197]. Therefore, definite objectives for integrative physiology such as informational science are defined by computational multiscale modelling of physiological systems [198]. The complex nature of biological systems poses huge challenges for integrative physiology with the solution lying with computational framework integrating models and simulation results as described in PhySIM link (http://robotics.case.edu/PhySIM/index.html) [199]. Researchers forecast that translational medicine can become reality with the availability of descriptive physiological data [200]. These prophecies are need for proposing futuristic novel models to take on board information from various physiological states models that requires a global, collaborative, and integrative approach with omics and NGS technologies playing a dominant role.

1.4 OBJECTIVES OF THE PRESENT STUDY

Based on the motivation, problem statement and review of literature stated above, this thesis addresses these issues with four overall objectives:

1) To propose conceptual models to achieve a full physiology after reviewing the overall strategy ranging from functional genomics and its ontologies to functional understanding. The first objective is to provide some novel insights into how the biological data base-based, essentially descriptive, physiological models can be further functionally improved in terms of systems biology depending on the accessibility and integration of data.

2) There is a need for plant species-specific oriented repositories that will be continually developed and maintained. In order to illustrate how unambiguously representing a physiology of interest may be represented in an unambiguous and

computationally tractable way, the second objective is to implement a gene ontology data mining tool relevant for plant physiological models and provide a centralized plants physiology database as a new searching and investigating tool online.

3) To understand microarray database technology and the background of cancer, currently one of the most threatening diseases, the third objective is to extract functional genomics data associated with leukemia. To illustrate data driven systems biology of cancer based on computational statistical tools, gene expression profiling of leukemia using microarray data analysis is taken up as third objective.

4) To understand the knowledge based models for systems biology of cancer. The existing mathematical dynamical models lack of sufficient resolution to evaluate drug effect at the cellular, molecular and tissue level. Hence, the aim is to study the drug (EGFR inhibitor) using computational systems biology approach. To illustrate model driven systems biology of cancer based on dynamical mathematical models, multi-scale modelling of solid tumour growth and lung cancer treatment using systems biology pathway is taken up as the final objective.

1.5 ORGANIZATION OF THE THESIS

The present thesis has been organized into six chapters and the work included in each chapter has been presented in the following sequence:

Chapter 1, the current chapter, introduces and lays the foundation for research work presented in this thesis. The state-of-the-art survey of gene ontology data mining and systems biology of cancer related work has been presented and considerable effort was made to ensure that the presented material is supplemented by rich literature cross-references.

In Chapter 2, this work propose a gene ontology based models to achieve the goal of physiology after reviewing the overall strategy ranging from functional genomics and its ontologies to functional understanding. A collaborative research paradigm that integrates various elements to represent full physiology in particular domain is the focus of the work presented.

In Chapter 3, a gene ontology data mining tool is implemented after understanding the concept for a full physiology explained in the previous chapter. Plants Physiology Database (PPDB) that is based on mining a large amount of the gene ontology data currently available is presented. PPDB, a new searching and browsing tool is available at (http://www.iitr.ernet.in/ajayshiv/), so that plant physiology (biological process, cellular component, molecular function - BPCCMF) as a whole can be investigated.

In Chapter 4, a data driven systems biology of cancer approach is presented for gene expression profiling of leukaemia data set. This work presents step wise statistical approach based on heuristics for microarray data analysis. This study also describes the promising technology of microarray database bioinformatics (MDB). The gene ontology data mining or gene enrichment analysis results are also presented after finding the set of differential expressed genes. The normalisation results for all samples are shown in the Annexure B.

In Chapter 5, model driven systems biology of cancer approach for lung cancer treatment is presented. The multiscale (at cellular level, molecular level and tissue level) computational approach of tumour growth model is presented. The study in this chapter developed a mathematical model for tumour growth and angiogenesis that simulated a solid tumour's growth/progression with chemotherapy and anti-angiogenesis drugs using partial differential equations (PDE) modelling incorporating spatiotemporal processes. The downstream pathway of EGFR signalling is implemented using ordinary differential equations (ODE). The computer simulated results are shown for signalling pathways of TKI-EGFR and IGF1R regulation of various active cells, migrating cells, proliferative cells, apoptotic and quiescent cells that could be a united behaviour for the entire profile of tumour growth.

Chapter 6 concludes the work contained in the main body of the thesis and presents the suggestions for future work.

Chapter – 2

GENE ONTOLOGY BASED MODELS TO REALIZE PHYSIOLOGY

2.1 CHAPTER OVERVIEW

The design of an ontology that represents gene function is critical for addressing the challenge of integrating sequence data with the rapid increase in the amount of data from functional analyses of genes. Since genes are expressed in temporally and spatially characteristic patterns, their products often reside in specific cellular compartments and may be part of one or more multi-component complexes. Genes can have more than one product that are functionally distinct. An overall strategy clarifies how an ontology-based gene function may be implemented using genomic databases is herein dissected. Knowing that gene products possess one or more biochemical, physiological or structural function, the present strategy is suggested to lead towards physiological models. Thus, this work propose a conceptual model to achieve the goal of physiology after reviewing the overall strategy ranging from functional genomics and its ontologies to functional understanding.

The vital importance of the tremendous amount of biological knowledge contained in genomic databases has been investigated by analyzing an ontology-based gene function that leads to a physiological model. The overall progress accomplished in the functional understanding of genes through systems biology has arguably established bioinformatics and computational biology as a field of great interest. A system level understanding and the approach advocated in systems biology requires a change in our notion, from a reductionist approach to a holistic approach. While understanding the role of individual genes and proteins continues to be important, the focus has now shifted towards understanding a system’s structure, function and dynamics. morphological, physiological and molecular studies focus on what is really going on inside tissue cells are typically carried out on isolated cell populations due to known difficulties manifested by interference and interaction with surrounding cells, the present chapter is believed to somewhat contribute to overcoming the difficulties by considering gene ontology based models to realize physiology through the efficient management of biological data.

2.2 INTRODUCTION

A tremendous amount of biologically irrelevant data, out of which some usable information can be deduced, became available at the dawn of this century. This led to an emergent field called bioinformatics. It deals with the problem of handling and analyzing large datasets in a beneficial way for researchers. The bioinformatics world is inundated with individual genomic data that is being produced at a phenomenal rate. The flood of data introduces various challenges that bioinformatics scientists have to face. The various bioinformatics challenges can be described as organizing and sorting large-volume genomic data, that interprets and presents the functional impacts of genomic variations, integrating data to complex network models and translating these discoveries [201]. Besides, bioinformatics and systems biology can be presumably viewed as a key to many more unresolved issues.

The collaborated and integrated Human Genome Project (HGP), an international project aimed at the identification of the human genome sequence, its structure, and its regulation along with function of all sequenced genes and their products. Genes can have multiple products that are functionally distinct. To facilitate the functional analysis of encoded gene products, the international endeavour of the HGP has been expanded to sequence the genome of other life forms like the roundworm, mouse, plants, etc. Furthermore, representing gene functions is critical for addressing the major challenges of integrating sequence data with the abundant amount of data from analyzing gene functions. This requires a brief introduction on the basic terminology used in the work as discussed in next subsections.

2.2.1 Genomics

Genomics is comprised of two words: gene + omics, which means genome scale data as compared to single gene based data. Omics refers to large biological gene based databases. Many novel omics field came into existence after the popularity of genomics field like proteomics, transcriptomics, etc., although the basic backbone of this entire new field is sequenced genomic data. Genomic analysis investigates regular gene expression patterns in parallel with all genes comprehensively to understand functional genomics for the realization of a physiology. As such, genomics studies have great deal to offer in the field of health and drug discovery [202].

The foundation of genomics has been the next generation of sequencing technologies for inclusive biological information. This information can lead to deeper insights into the functional roles of genes under study. Genome sequences act as the base reservoir that is quickly emerging and needs continually improvements in sequencing technologies, quick ontology based annotation techniques and functional analysis of the revealed genes. Putting it all together can show the way to any genomic association study, which further helps in realizing the functional understanding [203].

2.2.2 Network Modelling

Gene network modelling is an important field that has its roots in genomics that gives a global outlook of genome data as compared to conventional genotype/phenotype research that focused on a single gene basis [204]. Network modelling analysis focuses on three phases starting with finding important nodes in the whole network, then the functional relationship with other nodes directly connected and afterwards with all other nodes indirectly connected through whole network. This led to a functional understanding between different networks of experimentally verified data that ultimately helps us to understand the physiology as a whole with different views. Since there are thousands of metabolic reactions working simultaneously and not independent of each other, graph based computational methods tackle the complexity of such networks and decodes several hidden biological properties [205]. Several gene based metabolic databases are available such as Kyoto Encyclopedia of Genes and Genomes (KEGG) [206, 207], MetaCyc [208, 209], and BioCyc [210], Wikipathways [211], pathway interaction database [212] accelerating biological discoveries.

The different types of biological networks in the field of systems biology are commonly represented as mixed graphs known as cell signalling pathways. The nodes in these networks are mostly proteins, but also can be metabolites, lipids, second messengers, or peptides. Interactions designate information flow and can be activation or inhibition. One of the first methods used to build those types of networks was yeast-2-hybrid (Y2H) screens. So experimentally, protein-protein interactions can be determined in high throughput using the yeast-2-hybrid screen system. However, evolving research is going on to connect the network of a cell to other things against connecting molecular components within cells as considered so far. For example, the Food and Drug Administration (FDA) approved drugs and their direct target proteins can be seen using network modelling. This is an example of a network that connects drugs to their molecular targets. These networks are bipartite graphs, containing two types of nodes. Nodes in biological networks can be connected through more abstract types of interactions. For example, genes can be connected based on the disease caused by mutations in those genes. In this particular example, each node corresponds to a distinct disorder, coloured based on the disorder class. A network representation that connects genes and proteins with more abstract concepts can be used for data integration. Anchoring many experiments with the genes identified by the experiments can be used to find dense regions in a bipartite graph of genes and experiments to find defined functional models and this was done by while analyzing the yeast network [213].

2.2.3 Biological Cataloguing using Gene Ontology

Gene Ontology (GO) is a buzzword these days as genomic datasets can be nerve-racking due to heaps of raw data produced from different laboratories across the world. The knowledge from these datasets can be gained from the abstract objects of gene ontology explaining the biological meaning of these objects having common relationships represented in the form of Directed Acyclic Graph (DAG) [33]. It acts as an incremental regular annotation tool sorting genes into three distinct classes as Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) within cell. On this pattern, the whole physiology is revealed as the physiological model consists of BP+CC+MF (collectively termed as BPCCMF) and each of these has its own ontology. Overall, GO annotation is a perfect model for complex genome biology to interpret physiology [27].

Although GO annotations may develop into a powerful tool in the future, they are currently limited and incomplete since they are dependent on database entries by individual investigators. This is particularly problematic since the translation of biological data into GO terms is highly dependent on a researchers view and scientific background on a given subject. In many cases, there is no perfectly fitting GO term available to describe the biological data, making a ‘best fit’ annotation necessary. Even more concerning is the very limited quality control of the data entries. In addition, just as initial GO assignments may initially be time consuming and difficult to make, the problem of keeping results and annotations up to date adds another layer of complexity to the problem. Therefore, investigating the sensitivity of GO annotations with a variety of database entries, which are relevant for the particular problem of research interest, are the primary concern. Moreover, gene ontology data mining is evolving so rapidly that biological discoveries now dependent on these genome databanks and has been transformed from biological sciences to informational sciences.

2.3 FROM GENOMICS TO PHYSIOLOGY

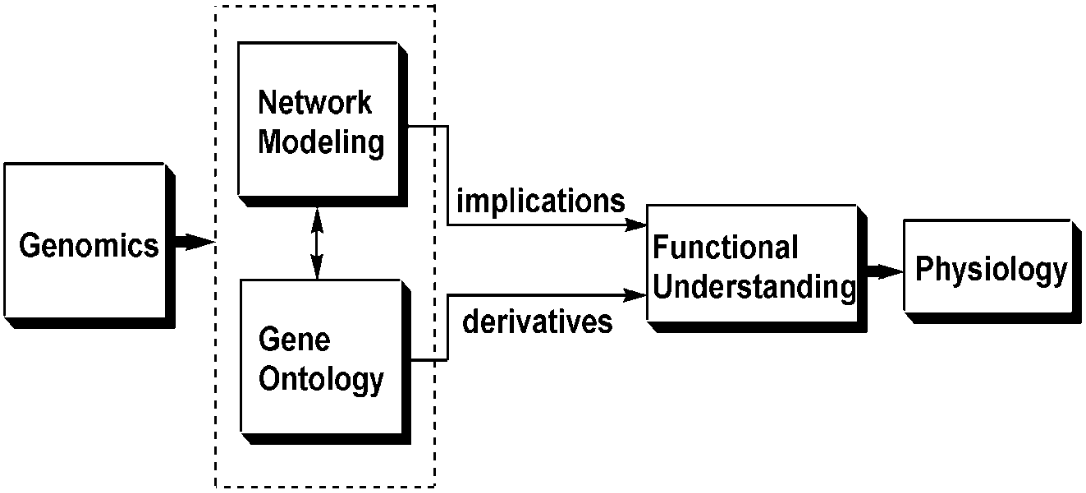


Figure 2.1 Dissected flow of information - from genomics to physiology

The combination of network models and GO with the input of functional genomics paves the path for a functional understanding, which ultimately leads to physiology models as shown in Figure 2.1. These physiology models are supposed to be experimentally tested and verified. Dynamic physiological models are represented as a series of mathematical equations or differential equations. This leads to chaotic behaviour of the system. As such, GO-based gene functions can lead to a better physiological model concealing mathematical complexity. However, the visualization of models is more convenient for comprehension rather than the quantification of a non-linear system behaviour using mathematical equations. If the schematic representation of a model is changed, it does not end up reflecting the same in a mathematical model. This can be a most time consuming and error prone operation. The proposed model is an enhanced integrative approach with different modules from genomics to physiology. The use of high computational techniques at various modules must ensure a programming interface for a viable flow of information between these modules. It is noteworthy that more work needs to be done in the area of descriptive physiological ontology that will boost the path to an integrative physiology ontology [199]. The research can arguably be intensified with functions assigned to identify genes along with organization and the control of genetic pathways in order to debug the physiology in detail [214]. The goals to be achieved by consistent computational modelling of physiological systems include deciphering biological information from false-negative and false-positive results and rationalizing coherent mechanism needed to grasp physiology [215].

Functional understanding is an implied and derived product of network modelling and GO. Ontology-based gene functions are needed to achieve a biological objective to which a specific gene or gene product contributes. The roles of genes can be well established after getting individual genes associated with other genes that are more accurately verified and annotated. To understand the aspect more clearly, GO tools can be used to collect information from different databases in order to see how biological processes interact with each other and to analyze their collective function by correlation methods relevant for the establishment of a functional gene network. This network always checks for GO updates to achieve satisfactory reliability in understanding individual gene functions that enable an understanding of the fully functional system [216]. Hence functional understanding is an automated process with the aid of a new form of text mining using GO and pathway ontologies. While GO considers gene function leading to physiology, pathway ontologies consider function as bio-chemical reactions and interactions leading to an ordered assembly of one or more molecular networks [217]. For example, any molecular function can be searched using the AmiGO browser for all possible combinations annotated to the particular GO term [218]. This makes GO able to optimally find entities with related functions and biological processes where traditional methods based on sequence similarity fail to match. Such functional understanding gives researchers an opportunity to know how sequences can be evolved in order to get the same output [219]. The more advanced functional understanding can be achieved using tools like the Gene Ontology for Functional Analysis (GOFFA) with several user friendly utilities [220]. In this way, it is possible to decipher GO data more effectively and dynamically leading to the generation of new hypotheses. Also, after proper functional understanding, functional characterizations of pathways can be developed on the sidelines of GO annotations by making functionality templates that further boost up the process of discovering new metabolic pathways [214]. Such a pathway representation was illustrated using the EcoCyc database [221, 222].

2.3.1 Two Way Associations - Network Analysis and Genome Annotation

Understanding ontologies is a way forward from biological knowledge contained in database, such as GenBank, to a physiological model conceivable through cellular components, biological processes, and molecular functions. In order to implement such a way forward, understanding networks through the usage of various pathway softwares combined with the knowledge of interactomics is highly recommended. All this knowledge comes in handy for researchers who understand ontologies and networks, which leads to the functional understanding of any organism as shown in Figure 2.2.

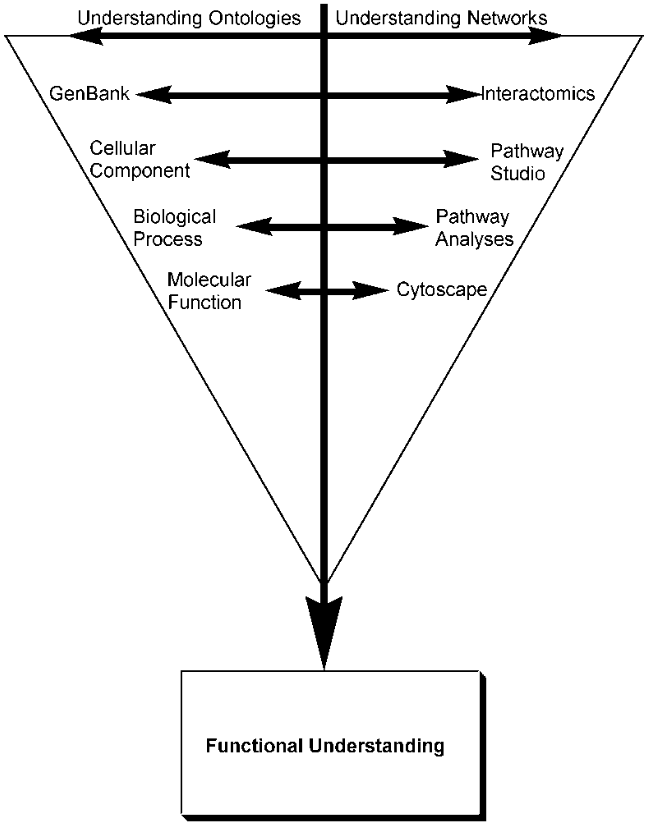


Figure 2.2 Modelling approach to understanding of ontology based gene function

This paves the path for deciphering a promising area of biological science known as microarray database (MDB) technology. MDB technology allows for the expression levels for hundreds or thousands genes. The differentially expressed genes can be understood from different pathway analysis tools applied to existing biological pathways and by using in-built databases and resources that are publically available [223]. The molecular function ontology component can be understood by using Cytoscape, an open source tool for integrating interaction networks with high-throughput expression data and other molecular states. Similarly, biological process ontology and cellular component ontology can be studied using two pathway software tools namely Pathway Studio (a proprietary product of Ariadne Genomics) [224] and Ingenuity Pathway Analysis (IPA), a proprietary product of Ingenuity Systems Inc. [225]. The Cytoscape software supports a plug-in extensibility to study various additional components of interest. For example, the

MiMI [226] plug-in annotates interactions from the Medical Subject Headings (MeSH) [227] and the Online Mendelian Inheritance in Man (OMIM) [228] as well as GO, and normalizes data from all of these different sources in order to merge all of the annotated and normalized data using alignment techniques. Recently, the most popular Rat Genome Database (RGD) [229] diagrams were made from the Pathway Studio and added to the Pathway Portal [230]. Another work that illustrated the enormous predictive capability aspect of metabonomics analyses for the determination of drug toxicity was the Ingenuity Pathway Analysis [231]. With the help of IPA, it is possible to perceive and comprehend biology at multiple levels by the integration of heterogeneous data. This gives an insight on molecular and chemical interactions, cellular phenotypes, and disease processes in the specific system. These three tools can use gene annotated databases from GenBank to interpret interactomics, from GO to interpret molecular biological functions, from Pathway Studio database to investigate cellular components and from the Ingenuity's Knowledge Base to extract information regarding various biological processes [232]. Network visualization and modelling tools are in fact a turning point for grasping data relationships [233]. By this way, scientists can analyze their or others experimental work, browse different pathway collections, do literature mining, and share their data for both results analysis and future reference with other scientists.

Due to the availability of pathway and genome databases, cell signalling pathways can be reconstructed using genome annotations and validated through supplementary information related to its physiology and microenvironment reactions. This requires a great effort to supplement both these independent fields of network modelling and annotated genome since certain missing links erupt in the network as all the sub pathways like catalysts might not be annotated in the genome database. There are several implementations like Pathologic module of Pathway Tools software set committed to querying and computing with BioCyc database [234] where whole signalling networks are reconstructed using annotated genomic databases [205]. The aim of all of this is to provide a clear picture of functional understanding.

2.3.2 Microarrays Heralded Functional Genomics Blueprint

As a central dogma revolves around molecular genetics [235], functional genomics generates an understanding of functionality at the genomics level by giving a know-how of gene structure by assigning functions to them and the evolution of biological systems. Here the main lead role can be played with MDB technology that can assign functions to genes. Therefore, microarray data can be treated as functional genomics data and provide a great leap in understanding the perturbations at the genome level. Many life scientists are of the opinion that differential gene expression obtained from microarray study helps in delineating a system level understanding of the biology of interest. This has allowed genes to be analysed by gathering established classifications of an organism. Hence, MDB is considered a breakthrough for functional genomics [236]. The biological interpretation of these microarrays can be performed by a general approach displayed in Figure 2.2.

Nowadays, the interrelated physiology and anatomy fields are activated by MDB to seek answers to life sciences old questions with the help of holistic practice. Integration of these fields with MDB technology are able to bring out evolutionary relationships with biological information inherent under the wraps of physiological models of any organism whether eukaryotes or prokaryotes. This integration can be achieved with existing and upcoming bioinformatics methods with a modular approach to understand the whole system at the genomic level, which is the backbone of functional genomics. As such, functional genomics can reveal physiology with information and technology principles working in tandem to assign a function to ten's of thousands of genome data concurrently under observation and know the working of complex biological systems. This is one of the greatest challenges, defining the relationship between functional genomics and physiological models and thus the integration of MDB comes into the limelight. Predictions in life sciences are not as easy with arithmetical precision as in conventional science due to fast changes within a cell, tissue, or organism under consideration. Therefore, principal bioinformatics methods and evolving tools are key to physiology, studying with universal format like Minimum Information about a Microarray Experiment (MIAME) [237] to annotate microarray data in its basic step. The annotation of microarray data in an unambiguous fashion can be easily managed by using GO [33] to have an insight of proper biological knowledge as data mined from microarray databases can be extracted and stored as a biological tractable term in the GO database, which helps the clarity of the system to be physiologically modelled.

The above explanation can be understood as a modelling approach towards physiology derived from Figure 2.1 using microarray technology as shown in Figure 2.3. This approach clearly shows that MDB along with the ability of bioinformatics to execute regular patterns of gene expression level by synchronizing diverse genome information leading to functional genomics in collaboration with GO. The next generation up-to-date bioinformatics processes evolve the functional understanding of the system as the knowledge stored in MDB heralds functional genomics to a more conceivable full physiology.

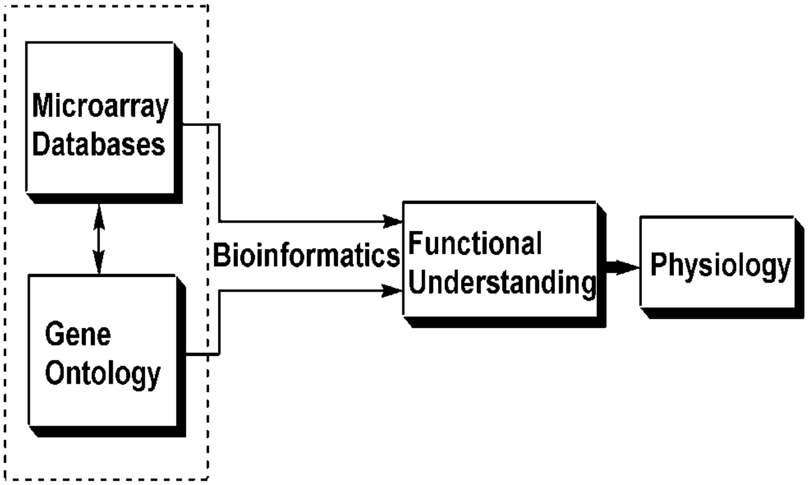


Figure 2.3 Modelling approach to physiology using microarray technology

This modelling framework can act as a guide for coupling the varied biological modules in order to understand the organism under observation as a single complex system for a meaningful physiology. Moreover, a meaningful physiology is tied up to automation in gene ontology module and this realization moves from a predictive to integrative to descriptive biology.

2.4 INTEGRATION OF BIOLOGICAL DATA

At present, all research indications speak in favour of the key challenge in integrative biology: providing physiological models that facilitate the development of novel drugs against diseases such as cancer, HIV, Alzheimer's disease, against which effective therapeutics currently does not exist. Even though such "full physiological models" are not always attainable due to inadequate biological data and/or their appropriate integration, functional genomics can be currently considered as a reliable functional basis upon which such models are expected to rely. Integrating systems biology models and biomedical ontologies is consequently needed in order to understand the background of many currently incurable diseases, primarily in terms of identifying new therapeutic targets as a necessary step for the target validation by relevant experimental techniques. Understand at the system level is an approach currently advocated in systems biology, yet it has also generated a significant debate in the literature. Thus, systems biology inspired collaborated and integrative methods are needed for modelling physiological processes. A novel idea is to integrate all the relevant biological data available to get the full physiological picture of the organism under observation as shown in Figure 2.4 (a) and Figure 2.4 (b).

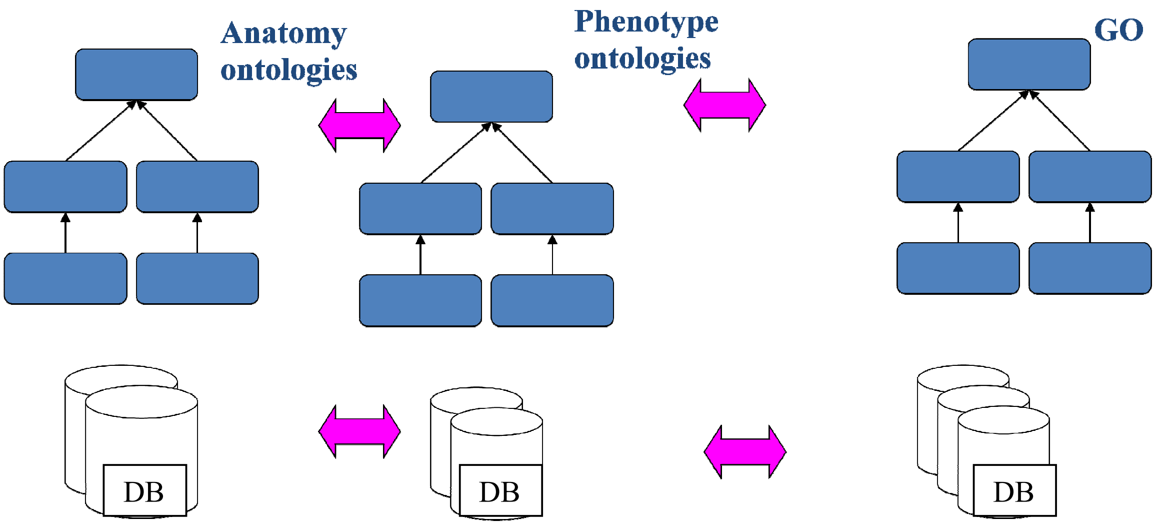


Figure 2.4 (a) Connecting the scattered islands

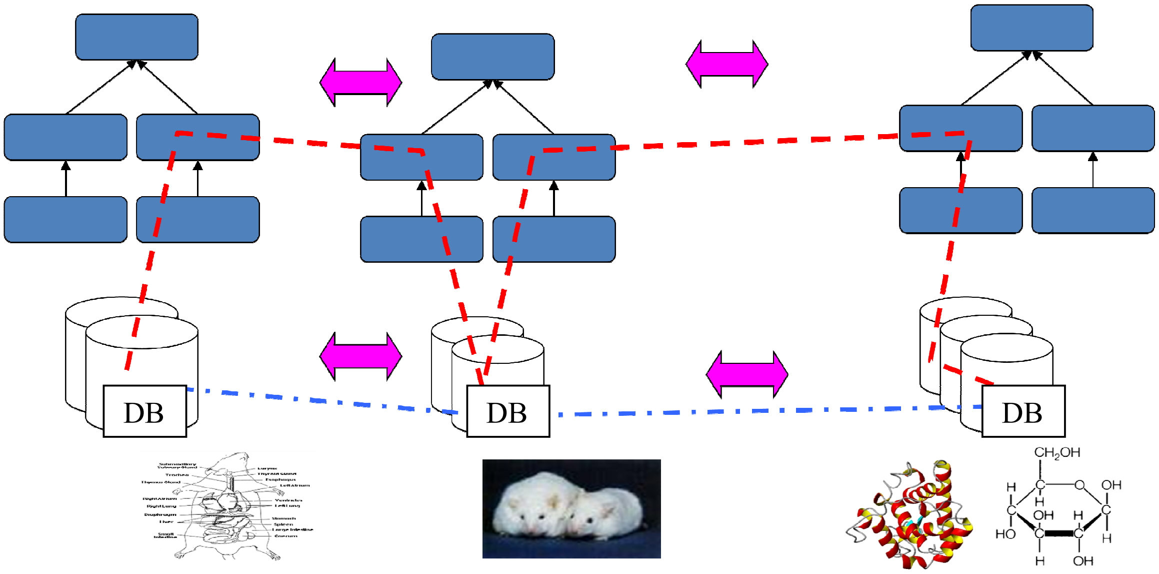


Figure 2.4 (b) Integrating and sharing for full physiology

The expected outcome of Figure 2.4 (a) and (b) is that data integration can lead to physiological models. DB in the Figure 2.4 (a) and (b) denotes the database ontologies produced by different scientists and researchers as their validated research output in their domains. As these DB ontologies are widely strewn, we can connect the scattered ontologies output provided by various laboratories throughout the world based on some standardization like GO. It is clear from the picture that integrating the data and sharing the data will lead to the ultimate goal of a full physiology. As such, a hypothetical full physiological model is supposed to have its full biological process (BP), full cellular component (CC), full molecular function (MF) and with its specific full ontologies respectively. Connecting individual ontologies from various data resources is a key step leading to a universal full physiological model. As such, this model is supposed to have its full BPCCMF with its specific full ontologies. After understanding the concept for a full physiology, the same has been implemented in the next chapter by using the power of contemporary bioinformatics. The aim of the proposed integrated model is to augment the prototype development processes for multiscale-multilevel physiological models in the future.

2.5 DISCUSSION AND CONCLUSION

Functional understanding is not only necessary but also a sufficient prerequisite for approaching a physiological model, since it is tied to functional genomics data recognized through ontology-based implications and derivatives. This standpoint was found critical for an explosive progress in systems biology over the last few years. There remains a giant information gap of how genes and physical traits are related and influence each other. While this is now the post genome era, their direct influence on specific phenotypes is still not well understood. A huge database warehouse is needed to integrate the relevant available data of genes and databases containing detailed information of physical traits. This will aid researches move seamlessly from genes to physical and physiological attributes to better understand their influence and effects.

Now days, a new breed of scientists called bioinformaticians or systems biologists are dealing with the deluge of data being produced using high throughput technologies like MDB. It is pertinent to mention here that these new breed of researches might not be perfect data miners but data mining should be one of the required skills. There is a huge demand of such researchers in this field since the majority of knowledgeable professionals in this field are from a computer science or information technology background that learns biological skills or biologists who later learn the art of computer technologies. Also at the same time, the new terminologies, concepts, and models are persistently coined in the emerging field with different meanings in this overlapping field. Existing scientific communities join the new field, exert various influences, and shape it in sometimes unexpected ways.

As the better understanding of many pathological conditions is the ultimate goal of the full physiological models is systems biology. These systems approaches to biology emphasize the structure and dynamic behaviour of biological systems and the interactions that occur within them. Systems biology depends on the accessibility and integration of data across different domains and levels of granularity. Biomedical ontologies, which are ultimately expected to facilitate this integration of data, are frequently used to annotate bio-simulation models in systems biology.

High throughput computational methods and informational resources have become an integral part of descriptive gene ontology based research that will integrate with validated experimental works [238]. However, data integration for a fuller physiological and biological picture includes a collaboration of gene ontology data mining, physiological and phenotype data that has several levels of complexities. The lower level of complexity lies in integrating similar databases with different languages and protocols that can be user queried and with unified responses. The challenge is how to present and translate multiple programming languages and protocols. The middle level of complexity lies in the correlation of integrated data to derive useful information from numerous pools of combined data. These correlations derived information will help in testing and generating innovative hypothesis about biological processes. The upper level of complexity lies in using the information from these two levels to make a foundation for dynamical models to simulate and analyse the results [95].

The development of such a gene ontology based solution will lead to genome wide associated studies with a capability for functional genomics as a whole. From this descriptive ontology based physiological model, further tailored prototypes can be developed that will ultimately lead to a full physiological picture of any organism under observation. The envision and success of such a complex full physiological tailored system to be functional, intelligible and trustable depend on the validated gene ontology databanks. Such a systems biology solution needs the continuous involvement of the modeller and best high performance computing automated methods since these gene ontology databanks are rapidly evolving.

2.6 AUTHOR’S RESEARCH CONTRIBUTION

Sharma, A.S., Gupta, H.O., Mitrasinovic, P.M. (2012). “From Ontology-Based Gene Function to Physiological Model” Current Bioinformatics, 7(4), 436-446.

Chapter – 3
GENE ONTOLOGY DATA MINING TOOL FOR INVESTIGATION OF PLANTS PHYSIOLOGY

3.1 CHAPTER OVERVIEW

Representing the way forward from functional genomics and its ontology to a functional understanding and physiological model in a computationally tractable fashion is one of the ongoing challenges faced by computational biology. To tackle this, we propose an application to contemporary database management for the development of PPDB, a searching and browsing tool for the Plants Physiology Database that is based on mining a large amount of the gene ontology data that is currently available. The working principles and search options associated with the PPDB are publicly available and freely accessible on-line (http://www.iitr.ernet.in/ajayshiv/) through a user-friendly environment generated by Drupal-6.24. Since genes are expressed in temporally and spatially characteristic patterns and that their functionally distinct products often reside in specific cellular compartments and may be part of one or more multi-component complexes, this work is intended to be relevant for investigating the functional relationships of gene products at a system level with the goal of a full physiology.

PPDB is a new searching and browsing tool for the Plants Physiology Database, so that plant physiology (biological process, cellular component, molecular function - BPCCMF) as a whole can be investigated. The Plants Physiological model was built using computing power as an abstraction of the plants gene ontology data with common vocabulary, systemization of knowledge, standardization and easy to use functionality that acts as an educational resource for plant biologists and bioinformaticians. It contains the latest full compendium of curated expression data entries for plants and is freely offered to researchers and companies worldwide via open access. Due to a substantial flexibility in the structure and organization of the Plants Physiology Database, it is capable of being conveniently upgraded in the future with new data entries from various sources and more efficient approaches to data mining.

3.2 INTRODUCTION

The completion of the Human Genome Project in 2003 [239], advances in the field of sequence analysis, and the recent completion of the pilot phase of the 1000 Genomes project [240] paved the way to creating models that can decipher the functions of living systems. Functional genomics employs various techniques such as microarrays, proteomics and transcriptomics to describe the complete picture of any gene function and gene interactions [241]. It focuses on the central dogma of molecular biology involving DNAs, RNAs and proteins as well as post translational effects in order to provide viable insights into the functions and behaviour of genes. With the advent of functional genomics, an unprecedented amount of data has been generated which is not arranged in an ordered fashion. For this reason, scientists are first engaged in creating databases containing physiologically meaningful information extracted from a huge amount of data. Many groups emerged and created databases such as Protein Data Bank, Swiss-Prot, GenBank, FlyBase when the sequencing of DNA started. They focused within their particular scientific group of people without integrating the knowledge lying in these databases [242]. For this research gap to be filled, representatives from major groups allied with mouse, drosophila and yeast databases founded the Gene Ontology (GO) Consortium in 1998 to collaborate on ways of assimilating the information contain in these databases [243].

The concept of gene ontology was introduced for annotating the data in a more relevant and tractable form, so that usable information from the databases can be extracted. Even though the more general definitions of GO were previously proposed, an aspect of biological relevance is a tool for the unification of biology [23]. It is in fact a large public database providing a set of controlled gene products vocabulary based on their role within a cell. It also holds data from diverse data resources to generate gene annotation data [24]. This facility is being extensively exploited by researchers from various backgrounds. It can be used to express viral gene functions and describe the three major components that are the heart of physiology, namely, biological processes (BP), cellular component (CC) and molecular function (MF) [collectively termed as (BPCCMF)]; however, it does not directly describe these three major components underlying physiology in a way that is unique to any particular disease. For instance, tumour genesis is the physiology of abnormal functional deviations in eukaryote kingdom and cannot be considered as a legitimate GO term [244]. Thus, the challenge is to see how GO terms may be employed to understand important pathophysiological conditions such as cancer. A network model of tumour genesis was in fact found to depict the GO cellular processes like proliferation, apoptosis, differentiation, mitogenesis, and immune function that are intimately involved in the occurrence of tumour genesis [245]. The potential solution for such challenges also lies in unification of key GO terms related to pathophysiological state of affairs to grasp the meaningful information. Based on this GO terms unification, the lymphomas pattern construction in the herpes virus infected tissues was stated and well explained [246, 247]. There are numerous virus genes with similar BPCCMF that act as host genes. The patterns found in the oncogenes of host derived transcription factors [248] and virokines genes training the alteration of host immunity [249, 250] are relevant findings. Such breakthroughs have paved the way for the development of 700 novel terms for plant pathogens, most of which are associated with the interactions between the species [251] to model host pathogen systems using GO.

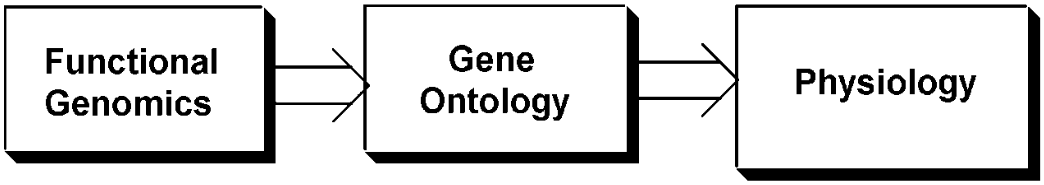


Figure 3.1 From functional genomics to physiology via gene ontology.

Physiology and anatomy are closely related scientific domains that work in tandem. Both must be integrated to understand the entire physiology mechanism of any system. The integration of mathematical models is needed in order to understand how the system works as a whole. It can also be done in a modular fashion in order to understand physiology of a specific module that is part of the overall system. A more conceivable descriptive physiology is rooted in functional genomics and gene ontology as shown in Figure 3.1. One of the best examples is that of the Physiome Project dedicated to understanding animal physiology by integrating knowledge from various resources and network models [252].

Due to the explosion of genomic data, physiology is becoming more quantified, relying on an extensive utilization of computer power [216, 253]. Accordingly, the present work demonstrates how contemporary database management can be combined with a data mining approach in order to construct an annotated gene ontology database relevant for plant physiological models. The model needs to feature the relationships between gene products on one side, and BPCCMF on the other side, highlighting the fact that gene products, similar many other functional entities, contain domains that are responsible for diverse molecular functions and participate in alternative interactions with other bioentities.

3.2.1 From Gene Ontology to Functional Understanding

Physiology is the science of how functions take place in living systems. This requires a thorough understanding of how organisms, organ systems, organs, cells and biomolecules carry out their chemical and physical functions. Gene products (proteins) are essential for the function of the cellular environment and are considered the workforce of living systems at a molecular level. To approach a full physiological model, a huge amount of biological knowledge contained in various databases needs to be sorted out by differentiating different types of data subjected to a double integration: i) vertically from atomic and molecular levels, over cell and organ levels, all the way to the level of a whole organism and (ii) horizontally comprising gene, anatomy and phenotype data. Connecting individual ontologies from various data resources is a key step leading to a universal full physiological model. As such, this model is supposed to have a full BPCCMF with specific full ontologies.

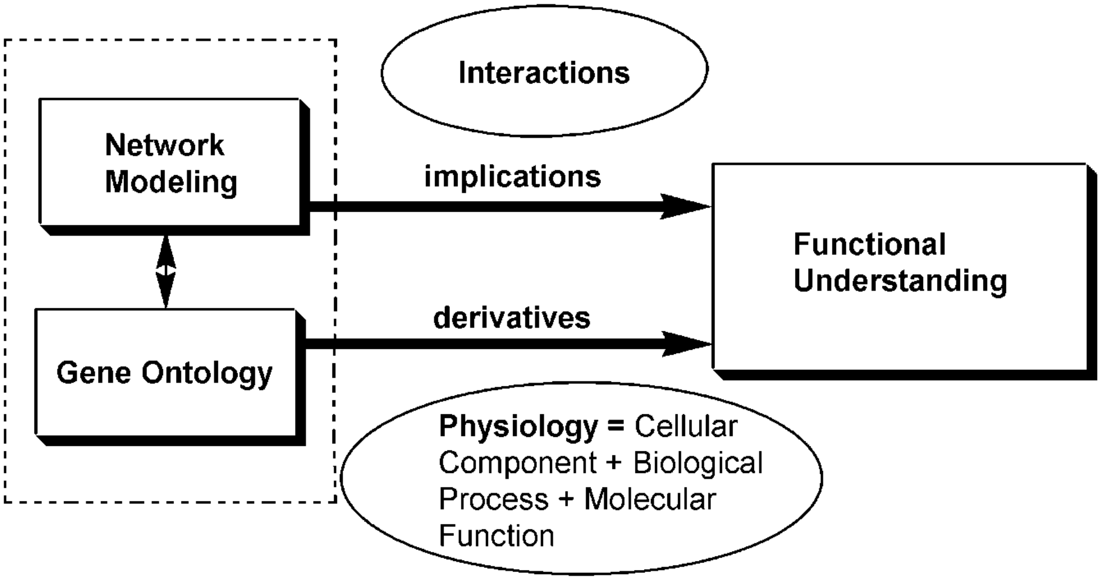


Figure 3.2 From gene ontology to functional understanding.

Annotation deals with a gene product associated to a particular BPCCMF underlying physiology determined by specific evidence or references from various database resources. GO has categorized entities into thousands of genetic attributes, managed in a hierarchical order through mutual relations. It essentially provides insights into BPCCMF by defining three distinct types of ontology. BP ontology is a specific biological objective that is contributed by a gene product. CC ontology is the position in the cell compartment where gene products are in a surface-dynamic condition. It describes locations by defining the levels of sub cellular structures and macromolecular complexes. Molecular function ontology refers to the biochemical activity of a gene product [29]. This function is a single activity reflected through binding with other bioentities to maintain the stability of the complexes and contribute to the conversion of one entity to another [254]. These three gene ontology terms describing physiology as a whole in conjunction with network modelling, derive a functional understanding of the entire system as shown in Figure 3.2. Functional understanding is the main challenge, since all functions at different levels can be related to each other like biological processes [255]. Therefore, the interactions between genes can globally be observed by means of a network of functionally-related genes in terms of GO annotation [217, 256].

3.3 DEVELOPMENT OF PPDB USING GO ANNOTATION

The primary aim of the PPDB is to explore plant physiology (BPCCMF) only in terms of gene ontologies and their relationships, giving ready information access to understand plant specific genomics [30]. This could be considered as the use case for the GO database and AmiGO [218], the official web browser and search engine (http://amigo.geneontology.org) of GOC provided by Berkeley Bioinformatics Open-source Projects. The PPDB prototype acquired significant traits for the plant research community by providing: 1) a controlled vocabulary related to plants, 2) searching and browsing for the GO terms in the Plants Physiology Database, and 3) viewing the associated ancestors and children terms with their detailed descriptions.

3.3.1 Motivation and Origin of PPDB

The pioneering works of Gene Ontology Consortium (GOC) resource (http://geneontology.org/) transformed the manner of thinking of genomics biology. GO started its operation in 1998 as a collaborative attempt for knowledge integration from FlyBase (Drosophila), the Saccharomyces Genome Database (SGD) and the Mouse Genome Informatics (MGI) project [29]. Currently, more than fifteen bio-curator groups are working for manual and automatic annotations of GO. GO is the benchmark genomics product ontology based informational database warehouse functionally cross-linked with different species and annotated knowledge databases. This is extremely useful informational resource for classes of species with restricted experimentally validated biological knowledge where electronic/computer inferred annotations occasionally provides and fill the biological information gap [30]. All gene ontological data will be released on a periodical basis, freely offered and can be downloaded in varied formats from the GO webpage (http://www.geneontology.org/GO.downloads.database.shtml). This will motivate researchers of particular scientific communities to use this gene ontological database warehouse to build specific tailor-made databases to fulfil their needs from the standard GO ontology repository. Similarly, the Plant Ontology Consortium (POC) resource (http://www.plantontology.org/) initiated their work by integrating three species-specific ontologies for flowering plant researchers to perform comparative analyses. The plant ontology browser and software were also developed by the GOC [257]. Recently, the POC is considering shifting the base to open source Drupal 6 (https://drupal.org/) content management system (CMS) as illustrated in wiki page (http://wiki.plantontology.org/index.php/Plant_Ontology_Web_Site_Update:_2013).

While several related database tools have been published with distinguished capabilities and limitations, there is no committed tool available that renders comprehensive genomic information about plant physiology exclusively like PPDB to the best of author's knowledge. Moreover, user needs are not satisfied by the existing investigating tools due to the absence of a collaborative and controlled vocabulary on plant physiology. Hence, a plant physiology database was developed as a new investigating tool to give ready information access. This will save valuable time and effort for the relevant researchers. Further, users can download the latest updated database and its database schema on plant physiology. The technical knowhow or manual is also provided for life scientists to build their own local PPDB mirror, which is missing in most of investigating tools published to the best of author's knowledge. Besides this, the GO structure is depicted by DAG (Directed Acyclic Graph) where terms are connected by edges (two transitive relations) within a hierarchy using is-a and part-of relationship. Due to the cross-references across other databases in GO, the GO terms are species independent. However, a quantity of low-level nodes/terms in DAG could possibly be related to a particular species (for example plants) although DAG nodes/terms integrity at high-level are not species dependent relative. These factors gave the insight for compiling more than 200 plants gene ontology data entries currently available in GO promoting plants (species-specific) database.

PPDB was developed from scratch using data mining in a GO database with keyword ‘plants’ and ‘plant’ with Drupal 6 to make an investigation tool for understanding plant physiology as a whole. The GO datasets used in PPDB are downloaded from the GOC resource website. A subset of GO is termed as GO Slim (http://www.geneontology.org/GO.slims.shtml) which is a user created database fitting to one’s own requirement. PPDB can be taken as ‘slim’ related to plant ontology curated from database release_name as 2012-03-24 and release_type as assocdb. It catalogs extracted GO terms specifically for plant species with fine-grained plant ontology. The focus was to develop a simplified investigation tool for giving a physiological understanding of plant biology by coupling the Drupal CMS and Gene Ontology database. PPDB will act as a continuing resource tool distinctively for the scientific society studying plant biology that will ease the discovery of biological processes.

3.3.2 Computer Hardware and Software

The operating system and other application software were launched on the Dell Precision T7400 Tower Workstation machine with multi-core Intel Xeon processors and an advanced memory, as well as the RAID (Redundant Array of Independent Disks) options in order to power through the most complex applications aimed at maximizing the performance of the PPDB. All the software support for building PPDB was provided through open source software and utilities that are available free of charge. Two different kinds of the software services employed were responsible for preparing the database system and for preparing the PPDB website, respectively. The basic software backbone of the database system is Linux (http://www.ubuntu.com/), Apache (http://www.apache.org/), MySQL (http://www.mysql.com/) and PHP (http://www.php.net/) [collectively known as LAMP system]. ‘Ubuntu 11.10 Oneiric Ocelot’ Linux provided the operating system, while Apache 2.2.20 acted as the web server. MySQL 5.1.58 was exploited to provide the relational database management system, while PHP 5.3.6 (The Hypertext Preprocessor) was in charge of the scripting language support for rendering dynamic web pages.

PPDB web site was created using an open source CMS platform namely Drupal-6.24. There are numerous motivating biological database internet frameworks based upon this CMS because of its gaining popularity inside the computational biologist community.

The strong and protractible approach through its numerous modules with powerful community supported themes provided a look and feel of the web site. PPDB is regulated by a green-clean theme and can be downloaded freely from the Drupal website (https://drupal.org/project/green-clean). This offers a versatile approach to style, management, and organizing content maintained in a dynamic fashion rather than static pages. PPDB can be integrated with different Drupal frontends for ontologies (like Tripal [68], DBSF [258], EMBRACE [69], RNA-Seq Atlas [70], CASIMIR [71], PepX [72]; these are only few examples within the growing list of Drupal usage) to expand into comparative genomics oriented database in the near future. It may also use Drupal's inbuilt searching method or integrate Google search for data mining, a filtering database if required, and catalogue of the plant-specific genomic data accordingly as it grows with community support.

3.3.3 Data mining and Database Design

Ontological databases are exponentially increasing with scientific discoveries through the dilation of storage power in computers. It is a herculean task for the plant community to extract and study the desired information from a physiological viewpoint in a minimum time from large databases. Data mining holds the key for the knowledge discovery process in huge databases by extracting or mining knowledge from colossal quantities of data [38]. The vital aspect of data mining is data retrieval in order to find knowledge in a database warehouse for additional utilization and to present the newly obtained knowledge in a user friendly manner.

Data mining for PPDB focuses on data retrieval process using exact keywords like 'plant' or 'plants' to advance the feature of information access. All the relevant data stem from the GO database that works as a repository. An in-depth knowledge of the Relational Database Management System (RDBMS) functionality was required for the successful search. Searching in a huge database was consequently performed by executing Structured Query Language (SQL) queries required for flexibility in data mining. The stepwise procedure for identifying annotated plant-specific genomic data with supervised data mining tool is as follows:

Step 1: Export the latest full GO database on MySQL database. [Raw Data Collection]

Step 2: Extract relevant tables using SQL queries. [Feature Extraction]

Step 3: Use of the keywords 'plant' and 'plants' for further refinement. [Feature Selection]

Step 4: Classify the search results based on the different ontologies BPCCMF. [Feature Classification]

Step 5: Store the obtained results on PPDB schema. [Training Dataset]

Step 6: Design and code the interface for the end user. [Testing and Evaluation]

Based on this procedure, each individual plant or plants entry was searched in all the 44 tables exported having more than two hundred ninety million rows of records. It was found that 35 tables contained genes or proteins related to plants. Having performed an extensive search to find out relevant information, more than 200 entries were associated with the particular keywords that are scattered at different source tables. All the search results were afterwards discriminated using their ontologies, such as biological process, cellular component and molecular function. This classified and annotated plant genomic data with 149 BP terms, 36 CC terms, 21 MF terms were finally stored in the Plants Physiology Database. It allows appending of scripting languages and CMS for producing web pages with this annotated genomic data.

For better understanding of the functionality, the database schema and the ontology database can be downloaded by clicking on the downloads link on website. The heart of PPDB schema is eight relational tables encompassing the blueprint of the database as shown in Figure 3.3. The database tables can be described by the key legend as follows. The key, located next to id INT(11), is the Primary Key of the table. The F Signature represents the Foreign Key to another table. A gray diamond key is related to a column that cannot have a NULL value, while a white diamond key denotes a column which can have a NULL value. The relationship between any two tables A and B, denoted by the interconnecting lines, means a Foreign Key relationship between A and B. The colour of the links does not indicate anything important, except distinction among various relationships.

3.3.4 Availability

The PPDB is available in the public domain through a user friendly environment accessible at www.iitr.ernet.in/ajayshiv/ or www.iitr.ac.in/ajayshiv. The on-line tool is interoperable and is tested using the Firefox, Chrome and Internet Explorer browsers.

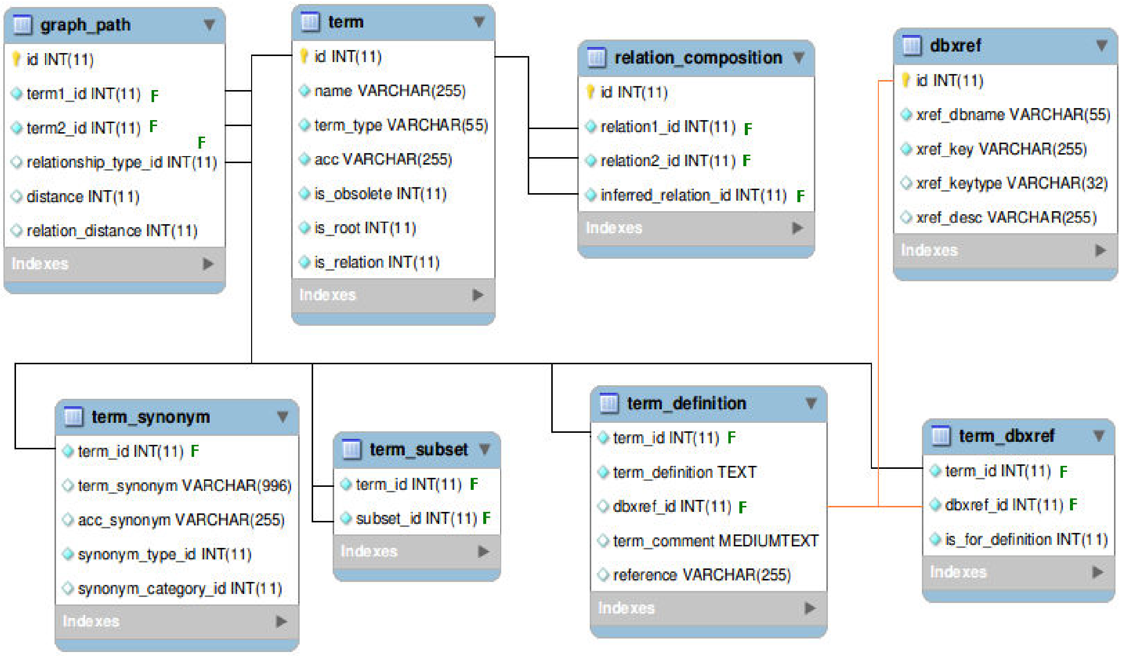


Figure 3.3 Blueprint of PPDB

Bearing in mind both the growing *in silico* approach towards biological information and the lack of collaborative and controlled vocabulary information on plant physiology, the Plants Physiology Database was constructed. The PPDB as browsing tool is explained in detail by taking a search example illustrating gene ontology data mining tool in Annexure A.

3.4 MAKING A LOCAL PPDB MIRROR

The whole data repository can be obtained from the latest PPDB database that is available under the downloads link at the PPDB web site. The detailed stepwise description or manual of installing the database with examples of SQL queries is available as well. One is supposed to have full super user privileges on the UNIX variant machine in order to be able to perform the following steps:

i) open the terminal from dash and type *terminal* and press *enter* or either press *ctrl+alt+T,*

ii) give the command *apt-get install php5 mysql-server apache2* (this command will install the latest version 5 of PHP and the MySQL server with the Apache http server),

iii) give the command *gunzip ppdb.gz* in the directory where the downloaded PPDB database is placed,

iv) give the command *mysql –u root –p*

which gives the mysql prompt to give further commands as shown in step v) below.

v) *mysql> CREATE DATABASE pp;*

mysql> GRANT ALL ON pp. TO ‘ ‘@’localhost’;*

mysql> quit

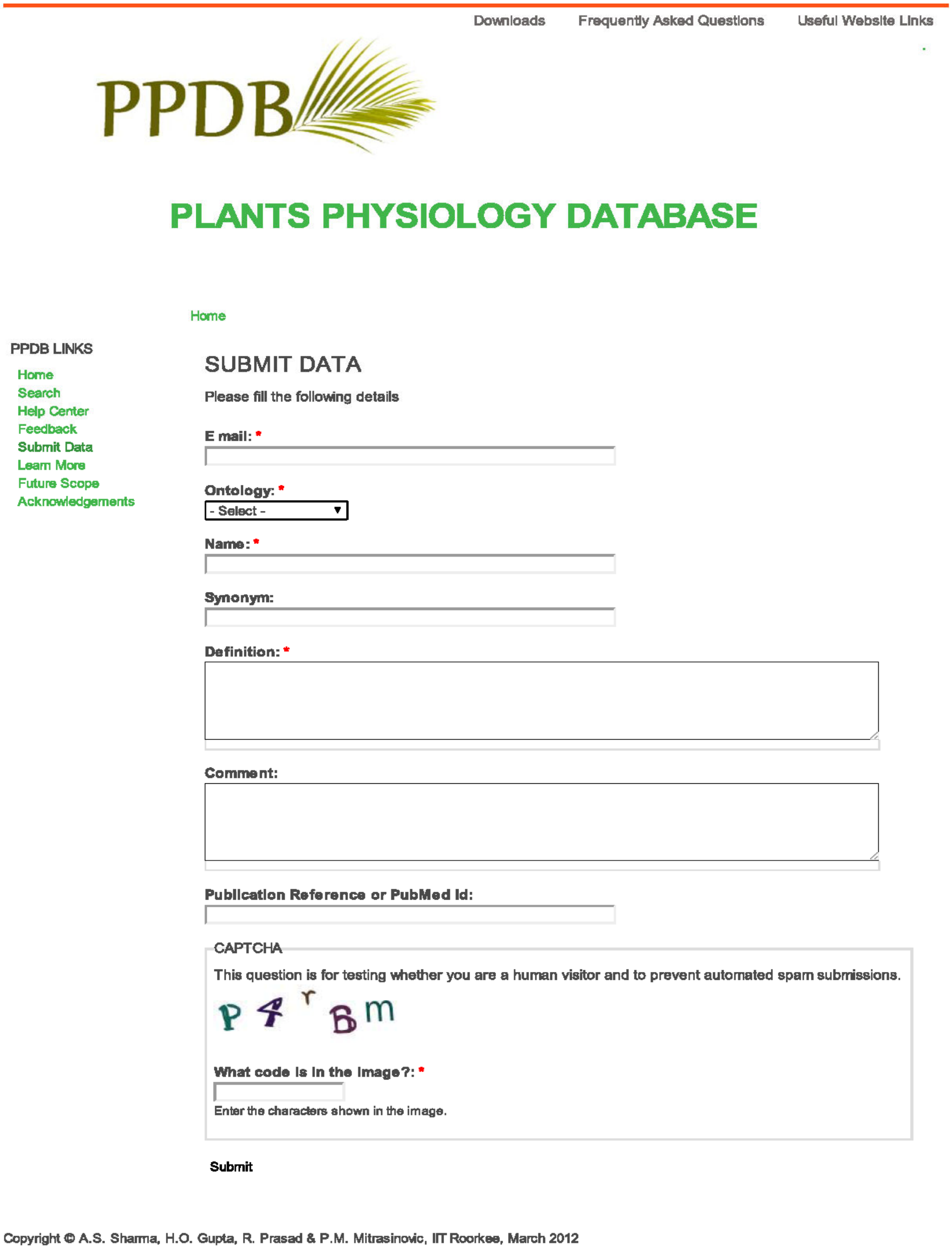
mysql pp < ppdb.

This protocol makes a local mirror server with the full PPDB database. The major benefit is that a multidisciplinary training of a new generation of scientists and engineers may be feasible.

3.5 CHALLENGES, COMMUNITY AND SUPPORT

The PPDB in its preliminary stage can grow with small incremental changes, similar to vocabulary works (like conventional branches of biology) that are always incremental. The biological data acquisition process grows over time with community support. To promote usage of a database by larger plant scientific community, a quick demonstration of the website is available on front page of the website or by clicking on the ‘Help Center’ section on the left panel of the website. This section also provides support a page that offers customized technical support to meet needs of bench scientists and researchers. The continuing work of developing a physiological database underlies great challenges in maintaining and renewing the biological ontology database.

The PPDB interface to the ontological database permits browsing and searching, as well as submitting gene annotations via a ‘Submit Data’ webpage shown in Figure 3.4. The project envisions active involvement from the scientific community to contribute to the growth of this ontological database and with help of collaborations with existing database resources and promising plant scientists. As annotated submissions to the PPDB grow, expert moderators are needed (from members of biological branches and plants research community) to ensure that curated plant gene ontologies fulfil the requirements of the entire user community. There are enormous benefits of community support, for example, the revisions of GONUTS [259] by non-staff members have recently surpassed the EcoliWiki [260] due to the large community base [218, 261]. The community groups interested in collaborating with PPDB can contribute plant genomic databases, open source software resources and other relevant suggestions for by email to the author.



Downloads Frequently Asked Questions Useful Website Links

PPDB

PLANTS PHYSIOLOGY DATABASE

Home

PPDB LINKS

- Home
- Search
- Help Center
- Feedback
- Submit Data
- Learn More
- Future Scope
- Acknowledgements

SUBMIT DATA

Please fill the following details

E mail: *

Ontology: *

- Select - ▼

Name: *

Synonym:

Definition: *

Comment:

Publication Reference or PubMed Id:

CAPTCHA

This question is for testing whether you are a human visitor and to prevent automated spam submissions.

What code is in the image?: *

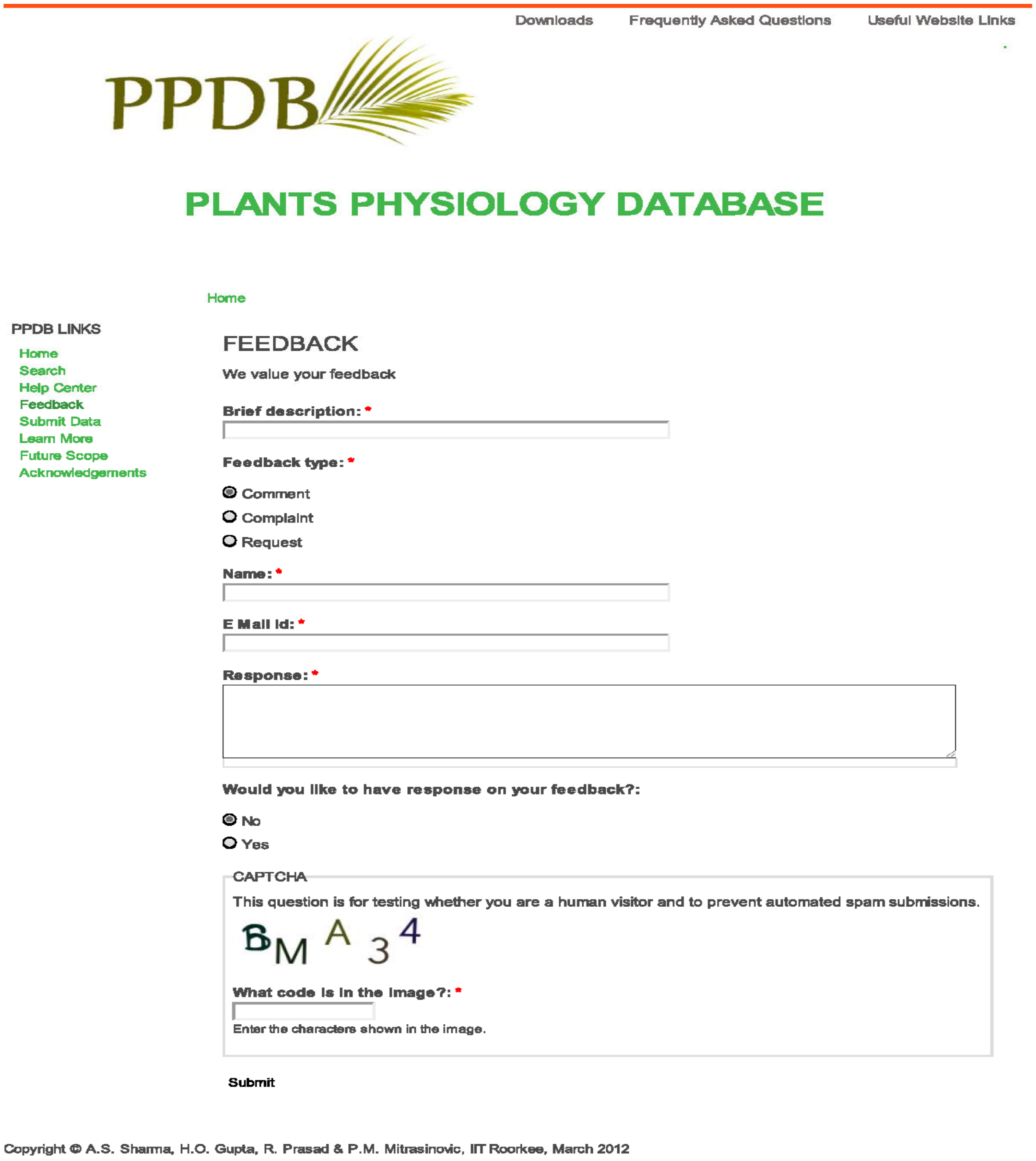
Enter the characters shown in the image.

Submit

Copyright © A.S. Sharma, H.O. Gupta, R. Prasad & P.M. Mitrasinovic, IIT Roorkee, March 2012

Figure 3.4 Submitting the gene annotations via 'Submit Data' webpage

PPDB also provides a feedback mechanism shown in Figure 3.5 to address the issues and problems that will be discovered by potential users. The database will be updated bi-yearly taking into account community feedback and ensuring the management of version controlled releases of PPDB. The foremost challenge is the precise computation on submitted genomic annotations since it is often unsystematic, ambiguous, presenting them in computer tractable form. PPDB can be extended to work with communities and journals to have a 'collaborative and controlled PPDB catalogue' to allow easier data access giving insights into plant biology that will facilitate scientific progression.



PLANTS PHYSIOLOGY DATABASE

Home

PPDB LINKS

- Home
- Search
- Help Center
- Feedback
- Submit Data
- Learn More
- Future Scope
- Acknowledgements

FEEDBACK

We value your feedback

Brief description: *

Feedback type: *

● Comment
○ Complaint
○ Request

Name: *

E Mail Id: *

Response: *

Would you like to have response on your feedback?:

● No
○ Yes

CAPTCHA

This question is for testing whether you are a human visitor and to prevent automated spam submissions.

What code is in the image?: *

Enter the characters shown in the image.

Submit

Copyright © A.S. Sharma, H.O. Gupta, R. Prasad & P.M. Mitrasinovic, IIT Roorkee, March 2012

Figure 3.5 Display showing various options for providing feedback

3.6 DISCUSSION AND CONCLUSION

The methodology for building PPDB is dynamic in nature, whereas manual curation is a static approach. Manual curation is best suited if updating is not done frequently. The volume of data for bioinformatics is increasing on a daily basis, so it is hard to maintain huge databases using manual curation. Once the mining of data related to plants is done, the queries once written will not be changed whenever the database is changed at the source website. The database only has to be updated. By relying on several plus points of dynamic methods of data mining, this computationally extensive approach was used for developing PPDB.

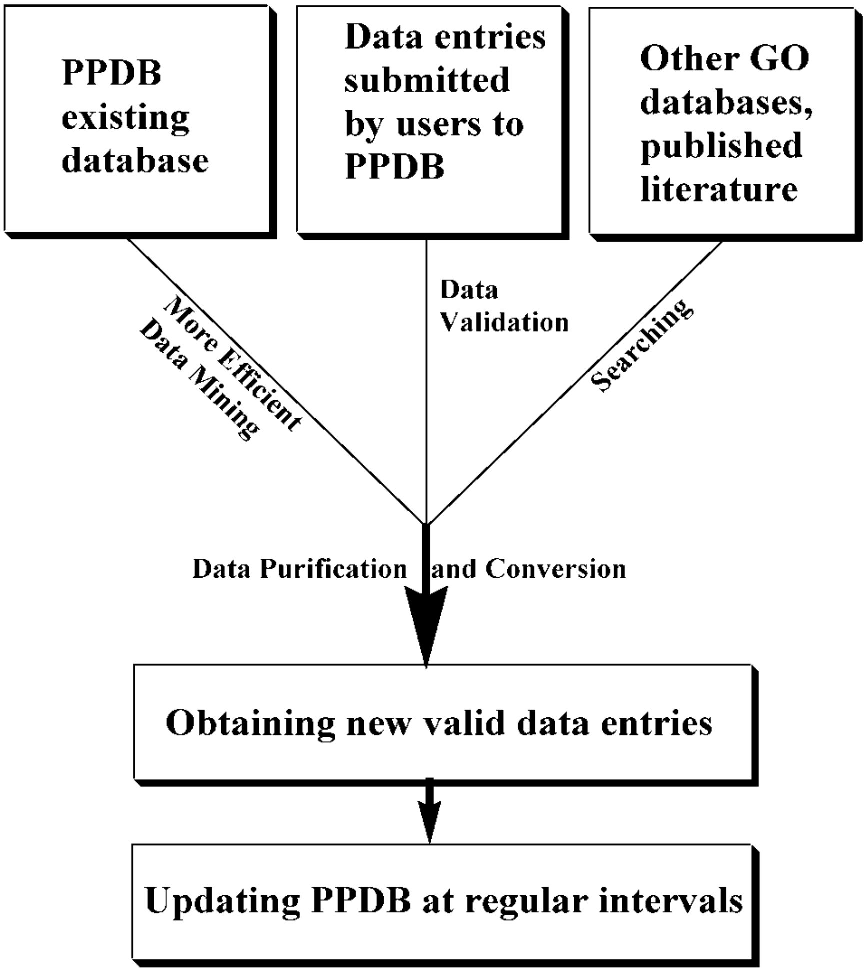

PPDB existing database
Data entries submitted by users to PPDB
Other GO databases, published literature
More Efficient Data Mining
Data Validation
Searching
Data Purification and Conversion
Obtaining new valid data entries
Updating PPDB at regular intervals

Figure 3.6 Outline of the future PPDB developments

The present development is the first phase of the overall PPDB project. The central repository of PPDB can support non RDBMS data resources, such as flat data files, OBO and XML files. PPDB may be the starting point for the development of several other database tools extending the support of the system for the plants physiology database. An outreach facility may also be included for bioinformaticians and researchers who want to share their ideas on plant gene ontology data. The term enrichment analysis can be done in near future by using the Account Login option, as provided in the Help Center. The future plans are best described by Figure 3.6. The PPDB website allows user to submit data entries by filling an on-line form under the Submit Data section of the website. However, before updating the PPDB at regular intervals, data validity needs to be ensured. Similarly, some more efficient data mining techniques for the existing PPDB database are going to be provided (for example, users may customize Drupal's internal search engine with advanced modes or integrate Google search for data mining), while keeping an eye on the other GO databases and published literature with information on plant physiology incorporating the manual curation. The data purification and suitable conversion will be needed as part of the quality control of data in order to obtain valid new data entries for the PPDB. Updating the PPDB at regular intervals will take care of the "tsunami" of data available worldwide. Thus, Figure 3.6 is an outline of the future developments in the framework of phase 2 of the PPDB effort.

3.7 AUTHOR'S RESEARCH CONTRIBUTION

Sharma, A.S., Gupta, H.O., Prasad, R. (2014) "PPDB - A tool for investigation of plants physiology based on gene ontology" Accepted for publication in Springer Interdisciplinary Sciences: Computational Life Sciences.

Chapter – 4

MICROARRAY DATA ANALYSIS FOR LEUKEMIA

4.1 CHAPTER OVERVIEW

In order to illustrate a data driven system on the biology of cancer based on computational statistical tools, a gen55-60,69-72e expression profiling of leukaemia was performed with a focus on the central dogma of microarray data analysis. First, this study describes the promising technology of microarray database bioinformatics (MDB) which takes the centre stage after completion of the human genome project (HGP). This technology in collaboration with computational systems biology research serves as key to functional genomics via gene ontology data mining/enrichment analysis to find regularities in large genome datasets produced at a phenomenal rate. Based on these approaches, this work uses a refined statistical approach based on heuristics. The medical diagnosis of two subtypes of cancer ALL (acute lymphoblastic leukemia) and AML (acute myeloid leukemia) is very complex owing to their clinical resemblance. As such, heuristics models are well suited for identifying which genes are differentially expressed in two different types of cancer patients. The results of the microarray data analysis provides invaluable information that can pave the way for innovative opportunities for early diagnosis of malignancies and building explicit disease sketches.

4.2 INTRODUCTION

Molecular biology revolution started with ground breaking work of James Watson and Francis Crick who solved the structure of DNA (deoxyribonucleic acid) with a two paired chain of chemical bases known as the double helix in a spiral form. This structure is composed of four different nucleotides A (adenine), T (thymine), C (Cytosine), G (Guanine) alias bases. Base properties help in the creation of chemical bonds A-T and C-G by complementary characteristic according to the Watson-Crick rule in two strands of DNA, which gave new insights in the field of molecular biology [262]. In light of such advancements, researchers are toiling to reveal the series of actions in genetic information such as how does DNA make RNA (ribonucleic acid) formulate proteins thus defining their interrelationship and role within cells [235]. The research has progressed in this direction with the initial draft of the human genome sequencing project to understand biological information corresponding to Homo sapiens [263]. The completion of the HGP in 2003 [264] and similar sequencing projects for other organisms like yeast, human fly, roundworm and the common mouse [265] produced an avalanche of data. As of 2013, 2568 organisms have been completely sequenced in different laboratories across the world, which is evident from tens of thousands of gene expression studies published [266]. This explosion of data from different resources can be coined as biological big data. All this leads to information mining driven by high throughput computing power thus empowering the thinking of life scientists, computational biologists and researchers ultimately breeding an emerging discipline called bioinformatics.

Bioinformatics is presumed to be a solution to the biological goals and mission defined by the National Institutes of Health, USA (NIH) presenting an opportunity and ambition with great challenges [267]. While there is a long way to go, advancements in research capabilities with the evolving bioinformatics field are leading to a visualization of biological landscapes from different perspectives. Bioinformatics is an interplay exploiting basic sciences such as mathematics, physics, chemistry, computer science and biological sciences such as molecular biology, structural biology, pathway biology to present a comprehensive picture. To understand complex biological systems, many science fields come under the umbrella of bioinformatics with numerous applications. Moreover, defining a new field of science is always a difficult task since young disciplines have fuzzy borders.

Along similar lines, in order for the field of data driven systems biology to mature, novel statistical and computational analysis methods are needed to deal with the growing amount of high-throughput data from genomics and genetics experiments. Systems biology is a pragmatic approach which consists of four steps [268]: first, large scale data are collected to describe all the components of the system for instance at the DNA level, RNA level and so forth. Second, the components of the system are systematically perturbed by genetic means, drugs or controlled environments that are monitored, if possible, on a global scale with all genes assayed. Third, a model is built and iteratively refined so that its predictions fit experimental observations; finally, specific perturbations are designed and performed to test models and distinguish between competing hypothesis [89]. It is a shift from a traditionally ‘reductionist’ to a more holistic, integrative, system-based approach to understanding the dynamics and organizational principles of living systems [269].

4.3 MICROARRAY DATABASE BIOINFORMATICS (MDB) TECHNOLOGY

This is a massive technology used to estimate a synchronized gene level expression database of genes placed in certain order using high performance computing (HPC) power. This technology allows a global gene expression space to be built on probes with any organism that has a sequenced genome. With this state-of-art in the field technology, RNA abundance can be monitored in parallel fashion [270] which facilitates the generation of hypothesis about an entire genome under study using global gene expression space. Moreover, the expressive data can be interpreted by consequently mining the microarray database taking a whole genome into account [271] since in the past only a single gene by gene basis explanation was put into practice. As such, this technology is a fascinating tool to unravel the secrets of life in contrast to traditionally available polymerase chain reaction (PCR) technologies, western blot and northern blot methods. It helps in clarifying why gene expression levels fluctuate under various conditions/stages like when a cell is in the developing stage. An important application evolved from which is the study of patterns in normal and diseased cells. Overall, this technology has witnessed rapid growth in the exploration of gene expressions and the varied domain of genomics. As such, microarray databases heralded a functional genomics blueprint.

4.3.1 MDB Warehouses

Microarray database technology yields manifold data in the form of image data and gene expression matrices before and after the microarray analysis. Since the microarray data is massive and costly to produce, certain companies and research groups like Affymetrix, Alphagene, Biodot, Broad Institute, Genometrix, Qiagen, OncorMed [236] make their data available to the world through online repositories. These public vaults are known as microarray database warehouses. The wealth of shared microarray data is of great importance to the scientific community, which can do their analysis and develop innovative computer algorithms to further recondition the data. Also, the data from diverse MDB warehouses can also be integrated to demystify biological perplexity which is also an upcoming field. Microarray data can be submitted to Gene Expression Omnibus (GEO) [272] on the National Centre for Biotechnology Information (NCBI) database server or ArrayExpress [273] on the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI) database server as soon as it is published [274].

4.4 DESIGN OF MICROARRAYS

There are many types of microarray designs available these days depending on the substrate used [274]. Broadly, these can be classified as single-colour microarrays and two-colour microarrays. Single-colour microarrays are known by name oligonucleotide microarrays whereas two-coloured microarrays are known by name complimentary DNA (cDNA) microarrays alias two-coloured microarrays. Single-colour microarrays use an in-situ method for probe making with perfect match oligo and mismatch oligo to accurately calibrate with reference to sequence to check for any cross hybridization of genetic data as depicted in Figure 4.1.

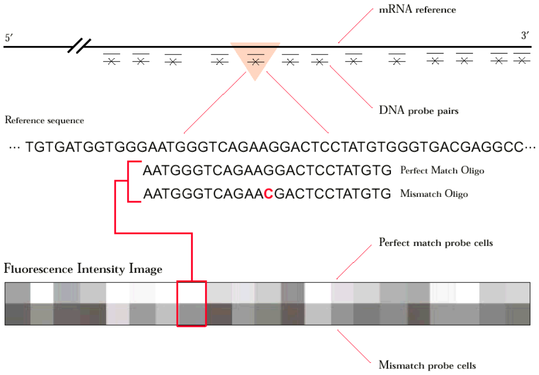

mRNA reference
5′
3′
DNA probe pairs
Reference sequence
··· TGTGATGGTGGGAATGGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC···
AATGGGTCAGAAGGACTCCTATGTG Perfect Match Oligo
AATGGGTCAGAACGACTCCTATGTG Mismatch Oligo
Perfect match probe cells
Fluorescence Intensity Image
Mismatch probe cells

Figure 4.1 Single-colour microarray design

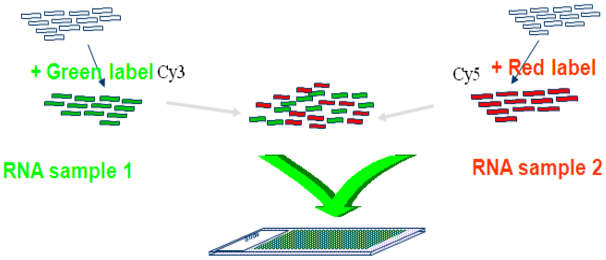


Figure 4.2 Two-colour microarray design

Two-colour microarrays use a spotting method for probe making with cDNA obtained from two samples labeled with fluorescent Cy3/Cy5 (Cyanine3/5 - green/red) dyes generally used for comparing two conditions such as normal versus diseased RNA samples or before and after treatment data. These labelled samples are further hybridized as shown in Figure 4.2.

4.4.1 Working of MDB

First, the basic microarray technological terms are explained as:

i) ***Feature:*** A microarray element.

- ***ii)*** ***Array(Probe):*** A feature that adheres on a solid surface in an ordered fashion at known x,y coordinates having known nucleotide sequence.
- ***iii)*** ***Labelled Sample(Target):*** A dataset of unknown nucleotide sequences to be hybridized with prepared array.

The working of MDB technology is depicted in Figure 4.3. Initially, probes are prepared on a microarray substrate with known nucleotide sequences ultimately adding a target sample according to a complementary characteristic. The RNA or target is picked from reference sequence adding Cy3/Cy5 to it.

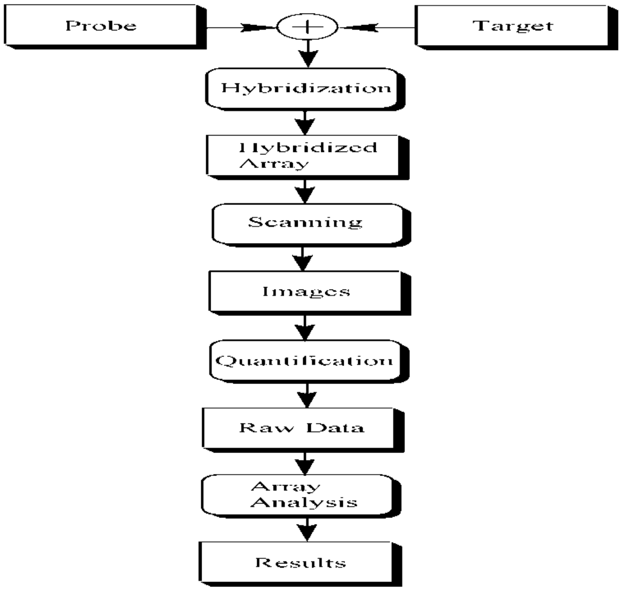


Figure 4.3 Working of Microarray Database technology

Both the probe and the target are hybridized to each other according to hybridization principle of A combining with T and C combining with G counterpart. There is a washing process for removing un-hybridized samples. After this, the hybridized array remains there to study transcriptional and enzyme activities of genes under observation in immobilized DNA. Since the sample was already labelled before hybridization, the fluorescence capability of Cy3/Cy5 allows seeing the spots needed to be measured when excited by a laser during the scanning process by using an imaging device. The images thus produced can be quantified by subtracting the background intensities. Then, the colour intensities produce raw data using relative gene expression level in the form of a gene expression data matrix. This data matrix is represented in the aX(b+c) format where 'a' corresponds to the number of genes expressed, 'b' characterizes the conditions expressed for each gene and 'c' symbolizes the other features describing a particular gene.

This quantified raw data matrix is further statistically scrutinized in the microarray analysis process to get the desired output results for which the experiment is performed or hypothesis is drawn.

4.4.2 Systems Biology Workflow to Analyze MDB

Analyzing MDB in a systems biology workflow depends on several initial steps in order to prepare high-throughput experiments so the results are accessible for biological analysis and modelling. While these steps are not typically used to define systems biology, they are deemed necessary for enabling a systems biology approach [275]. Once a biological and/or clinical question is posed, the workflow happens in the following order:

First, the experimental design is defined to efficiently answer the problem, then high throughput experiments are done according to Figure 4.3. After that, a scanner makes an analysis of the microarray, sequencing slides or phenotyping screening, and creates images that are processed using relevant algorithms that quantify the raw signal. This is followed by background correction that corrects the systematic sources of variability to improve the signal-to-noise ratio. The quality of data is reviewed for both the image analysis and background correction steps. At this stage, the information from the background correction is still rough. Useful biological information relevant for biologists needs to be extracted from the data. Once this information is extracted, the data can be used in a transversal analysis for clinical biostatistics, classification or systems biology approaches. Finally, the results need to be validated, interpreted, which and produce leads for new experiments.

A bioinformatics workflow and computational systems biology approach are cynical processes involving the acquisition of data and pre-processing, modelling and analysis. The integration and sharing of knowledge sustains the capabilities of this cycle to predict and explain the behaviour of biological systems. Therefore, for a successful workflow, there needs to be a robust enabling processes that annotates, manages and computes data [89].

4.5 DATA FOR MICROARRAY ANALYSIS

The raw data microarray cel files/arrays used in this work are obtained from the Broad Cancer Institute, USA and are available in the open domain in supplementary files at weblink:

(http://www.broadinstitute.org/cgi- in/cancer/publications/pub_paper.cgi?mode=view&paper_id=63) as shown below in Table 4.1. The Human Genome *Array(Probe)* HGu95 and HGu95Av2 are obtained from Affymetrix website [276] after making account on the website.

Table 4.1 Raw Microarray data cel files

No	Array
1	CL2001011101AA
2	CL2001011102AA
3	CL2001011104AA
4	CL2001011105AA
5	CL2001011108AA
6	CL2001011109AA
7	CL2001011110AA
8	CL2001011111AA
9	CL2001011112AA
10	CL2001011113AA
11	CL2001011114AA
12	CL2001011116AA
13	CL2001011118AA
14	CL2001011119AA
15	CL2001011120AA
16	CL2001011121AA
17	CL2001011122AA
18	CL2001011123AA
19	CL2001011124AA
20	CL2001011134AA
21	CL2001011150AA
22	CL2001011151AA
23	CL2001011153AA
24	CL2001011154AA
25	CL2001011126AA
26	CL2001011127AA
27	CL2001011128AA
28	CL2001011129AA
29	CL2001011130AA
30	CL2001011131AA
31	CL2001011132AA
32	CL2001011133AA
33	CL2001011137AA
34	CL2001011138AA
35	CL2001011139AA
36	CL2001011140AA
37	CL2001011142AA
38	CL2001011143AA
39	CL2001011144AA
40	CL2001011146AA
41	CL2001011149AA
42	CL2001011152AA

Each of these 42 cel files mentioned in Table 4.1 contains one .CEL file per chip, which contains PM (Perfect Match) and MM (Mismatch) values for each probe in the chip and the Presence/Absence calls for one per probe set. They can be interpreted as a statistical test of the spot foreground intensity in the experimental sample respect to the background intensity distribution. Also, separate PM/MM values are converted into a single expression matrix containing one column per chip with absolute intensity values and one row per probe set in the given raw data.

The biological example from this research work [108] showed that there is a genetic translocation that occurs in the leukemia dataset on the assumption that the disease is different due to certain mutations. In our work, we consider the microarray database of Human Genome *Array(Probe)* HGu95 shown in Fig. 5.1 and HGu95Av2 obtained from Affymetrix [276] and *Labelled Sample(Target)* of leukemia dataset samples obtained from the Broad Institute [277]. The low-level analysis (making background correction of microarray data, normalization of the data) is performed for quality control to extract specific features. The results are summarized with an aim to quantify and compare large scale gene expression data (with the assumption there are only a few significantly affected genes). The analysis of microarray data is performed using dChip [278] and R [279] software to summarize the results in addition to a calculation of the gene-level expression entities and outlier inspection for obtaining the pre-processed output. Rcom software package is required to run R in the company of dChip.

4.5.1 Affymetrix Array Data and Background Correction

The Affymetrix HGu95 probe is a standard array format with a feature size of 20μm, with oligonucleotide probe length of 25-mer with approximately 16 probe pairs or sequences in Human Genome u95 set as per the critical specifications mentioned in its datasheet [276]. In the hybridization process, even after washing there is some disproportion left on the substrate that acts as a noise. It is important to record the correct intensity strength in gene expression level studies. As such, a background correction becomes necessary for making the intensities analogous. The background correction is done by Affymetrix Microarray Suite 5.0 (MAS 5.0) proprietary method. Background correction produces five major files namely .exp (having experimental information), .dat (having image obtained after scanning), .cel (quantified matrix values), .cdf (chip description file), .chp (having levels of gene expression) for a particular probe (for example HGu95 probe). The core output file is a .cel binary file containing Perfect Match

(PM) Oligo and Mismatch (MM) Oligo as illustrated in Figure 4.1 with intensity values for each probe to be measured for the gene expression level. This is computed using presence and absence calls, which are a perfectly quality control measure thus producing microarray data gene expression matrix.

4.6 RESULTS AND DISCUSSION OF MICROARRAY ANALYSES

The total genes to be analyzed were 12600 and the group maa was created containing 42 leukemia microarray datasets, which were treated as 42 samples and a single sample group. This is done since more samples in a single group augment the prospect of picking quality samples to improve gene expression calculation and to discover patterns from an analysis. In addition, the more target oligos are present in the sample, the better the model fit and outlier detection. In the preliminary step, we analyzed the gene annotation terms with supplied gene information and a sample information file along with gene IDs in the Entrez Gene (http://www.ncbi.nlm.nih.gov/gene), 2354 redundant probe sets with 10272 unique genes comprising of 1200 gene ontology terms, 1200 protein domain terms, 872 pathway terms and 372 chromosome terms were found. Further steps involved in the microarray analysis alongside results and discussion are given below:

4.6.1 Microarray Data Normalization

Background correction removes background noise for the reliability of the experimental performance whereas microarray probes are made comparable and scaled for multiplicative factor. There is also a huge possibility that scanned images having inequality on the overall brightness or errors that crept in due to technical issues like batch effects, machinery and other miscellaneous factors. In this case, microarray data normalization becomes a prequalifying criteria to start an analysis of microarray data to make these arrays analogous for designing good experiments. The focal point of normalization is the renovation of microarray data by eliminating non-pertinent aberrations. Since this step is crucial for a better analysis in the research work, normalized results for all the 42 samples are shown in Annexure B.

Normalization, which is referred to as a low-level analysis, corrects the systematic sources of variability to improve the signal-to-noise ratio in order to make more accurate biological and/or clinical interpretations. Historically, normalization begins in the area of messenger RNA (mRNA) expression microarrays. Lowess normalization proposed bicolour microarrays. Normalization is still an active research area for current high-throughput technologies [89].

Inherent sources of variability directly affect signal measurements. In a certain manner, blocking in experimental designs already incorporates the effects that need to be corrected. Normally, correcting the batch effect is s part of normalization process. Yet, blocking cannot account for every variability source. In fact, every experiment is singular and indicates a specific variability that needs correction. For example, spatial artefacts are often seen in microarray experiments and there are many spatial normalization methods for correcting them for gene expression [280], comparative genomics hybridization (CGH) [281] and DNA methylation [282] microarrays. The CGH and method MicroArray NORmalisation (MANOR) [283] improves the signal-to-noise ratio on array and should be habitually used during an analysis of microarray data.

Researchers have also suggested using the ITerative and Alternative normalisation and Copy number calling for affymetrix Snp arrays (ITALICS) [284] based on multiple regression to correct the effect of the GC-content that affects Affymetrix GeneChip SNP arrays. For the next generation sequencing (NGS) data, systems biologists are given free access to the available software FREEC [285] and suggested methods for correcting the effect of GC-content on read counts in NGS data [286].

While the effect of both spatial bias and GC-content can be readily discerned, the shape and magnitude of the bias varies from one study to another. Therefore, normalization has to be adaptive and specified for each experiment. It is very important to identify and investigate all parameters that can bias the signal and these need to be discussed with platform providers and the operator in charge of the platform, since they are familiar with the protocol steps affecting signal measurement. The data normalization is a vital step that needs to be carefully considered since it can affect reliability, accuracy and validity of downstream analysis [287].

In this work, we make use of a simple method invariant set normalization method [110, 278] to normalize microarray data by picking one array in the set with a median overall future intensity to ensure the same weight of RNA across all the samples. The rationale is to fine-tune the overall chip intensity of the arrays to a comparable level. In this method, we designate one of the microarrays as a baseline microarray after computing measures of central tendency. The remaining microarrays are normalized against this baseline microarray after assigning ranks to similar genes and plotting a median curve from first to last, ranking invariant probes by fine-tuning the entire intensity values by fitting the normalized microarray data using smoothing curve. The scatter plot of perfect match and mismatch data before normalization and after normalization along with M-A plot before normalization and after normalization results are shown in Annexure B.

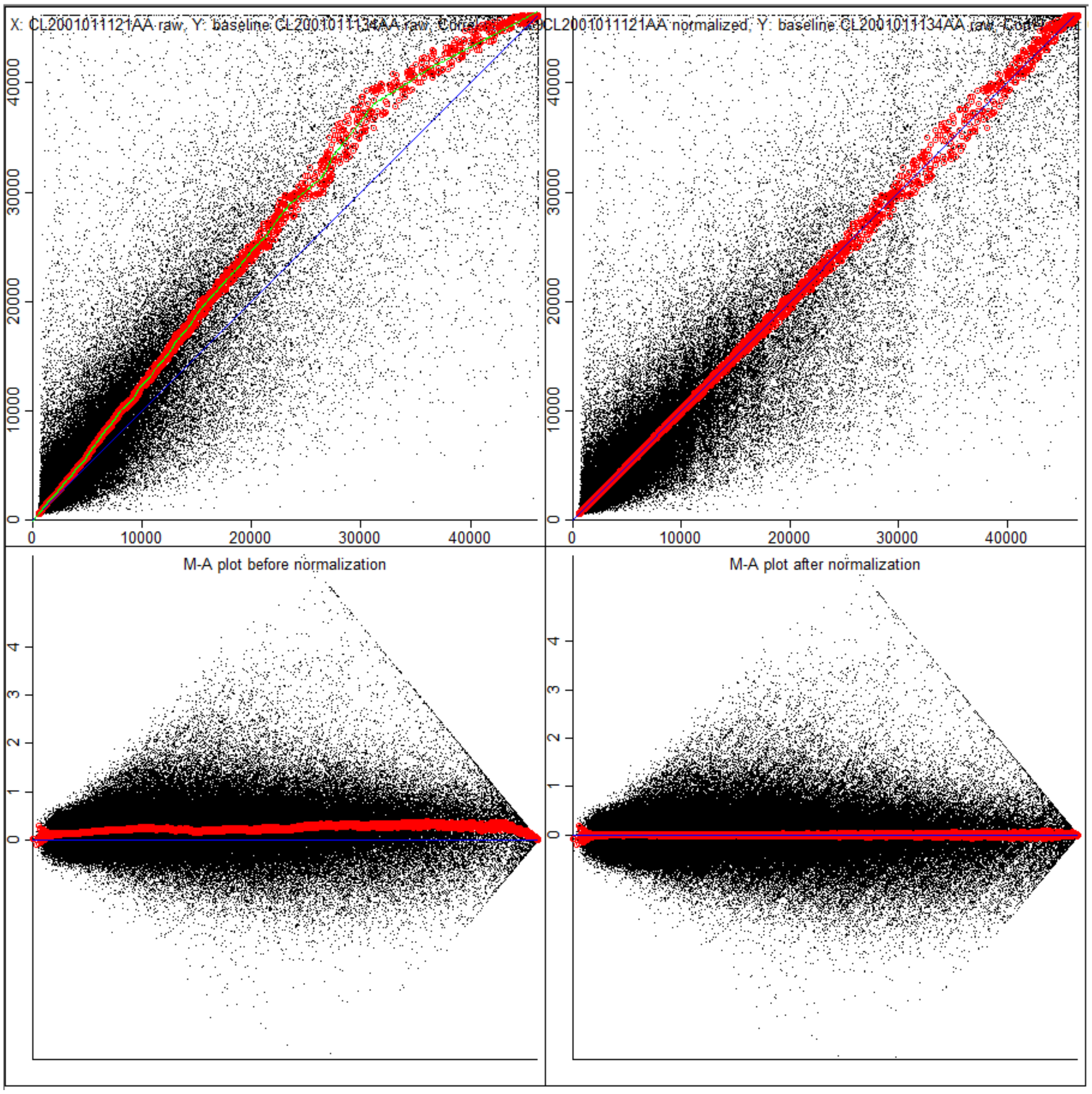


Figure 4.4 Normalization result of CL2001011121AA

Figure 4.4 portrayed the case of a single normalization result showing the output before normalization and after normalization of the microarray CL2001011121AA with a baseline microarray CL2001011134AA. The normalized results were produced for all 42 samples in similar manner and are shown in Annexure B. The running median curve through rank invariant probes becomes the new y=x curve as shown and all values are adjusted accordingly. The data obtained from arrays come in the form of fluorescent Red

(Cy5 or R) and Green (Cy3 or G) dye intensities and is fitted at the slope of the line around one. Another better representation for gene expression space is done by a forty-five degree scale rotation using M-A plots. In the case of gene expression plots, we take logs since increase and decrease are symmetric under log. In biological terms, $\log_2$ means a two-fold change. M-A plots or MvA plots are also shown in this figure, which converts hybridized intensity values to log ratios by computing mean and then transforming with log function. Most of the genes in this plot are located at 0 because log(1)=0.

4.6.2 Model Based Expression Index (MBEI) Method

MBEI is a robust weighted average method that considers the average probe intensity within one probe set where varying levels of gene expression are down weighted [110]. We found expression/signal values using 'Model-based expression' modelling method and 'Mismatch probe (PM/MM difference)' as a background subtraction with 5th percentile region (applied PM only method). The results (maa_array_summary file) of this method analysis after calculation of the measures of central tendency is shown in Table 4.2.

Table 4.2 Microarray output analysis maa_array_summary

No	Array	Raw Median Intensity	P call percentage	Percentage Array outlier	Percentage Single outlier
1	CL2001011101AA	1519	48.2	0.76	0.069
2	CL2001011102AA	1202	38.3	2.645	0.402
3	CL2001011104AA	1795	49.5	0.594	0.147
4	CL2001011105AA	1106	36.9	2.519	0.371
5	CL2001011108AA	1512	38.7	3.065	0.175
6	CL2001011109AA	1592	46.6	0.626	0.133
7	CL2001011110AA	1447	46.3	0.309	0.06
8	CL2001011111AA	1435	47.5	0.238	0.081
9	CL2001011112AA	1258	49.8	0.95	0.166
10	CL2001011113AA	1133	43.8	0.958	0.105
11	CL2001011114AA	1817	45.7	1.022	0.099
12	CL2001011116AA	1624	46	1.204	0.298
13	CL2001011118AA	1533	34.8	0.784	0.249
14	CL2001011119AA	1342	42.4	3.849	0.211
15	CL2001011120AA	2305	37.3	1.014	0.235
16	CL2001011121AA	1234	43.5	0.99	0.138
17	CL2001011122AA	1460	47.4	1.275	0.465
18	CL2001011123AA	1654	44.7	0.8	0.578
19	CL2001011124AA	3127	36.8	10.233	0.991
20	CL2001011134AA	1465	45.2	1.449	0.462
21	CL2001011150AA	1898	45.6	0.855	0.255

22	CL2001011151AA	1592	47.1	1.243	0.248
23	CL2001011153AA	1071	44.4	0.744	0.16
24	CL2001011154AA	1235	44.4	1.236	0.391
25	CL2001011126AA	1295	46.6	0.594	0.195
26	CL2001011127AA	1181	44.5	2.067	0.287
27	CL2001011128AA	1691	46.5	0.293	0.204
28	CL2001011129AA	1184	39.4	0.38	0.115
29	CL2001011130AA	1174	37.4	1.996	0.283
30	CL2001011131AA	945	40.3	0.879	0.318
31	CL2001011132AA	1248	41.4	0.594	0.182
32	CL2001011133AA	1160	42.4	0.578	0.201
33	CL2001011137AA	812	36.1	4.776	0.61
34	CL2001011138AA	1953	46.9	0.824	0.188
35	CL2001011139AA	1980	46.8	1.378	0.105
36	CL2001011140AA	898	37.8	1.727	0.347
37	CL2001011142AA	1541	43	0.317	0.149
38	CL2001011143AA	1663	40.7	0.475	0.179
39	CL2001011144AA	1267	38.4	1.972	0.191
40	CL2001011146AA	1852	40.6	0.515	0.251
41	CL2001011149AA	1684	45.6	0.76	0.151
42	CL2001011152AA	1777	43.1	0.863	0.1

We use MBEI because it diminishes inconsistency for low gene expression approximations and eradicates the cross-hybridizing probes and the consistently mismatch/negative probes.

4.6.3 Outlier Detection

The output summary file obtained in the previous step was the result of the model fitting process which shows that CL2001011121AA has 0.99 % array outliers and 0.138 % of single outliers are for whole microarray dataset/probeset. After calculating MBEI values as above, we can view cel image files, which are updated and the outliers are overlayed in the image at the same time. The percentage of probe sets are termed as array outlier in one array, the percentage of probe pairs are termed as a single outlier in one array. The array outliers displayed in white colour for elevated standard error and single outliers were displayed in purple for discounted dimensions on behalf of the whole probeset as shown in Figure 4.5.

On similar lines, we can view cel images for all the samples but this subject of microarray image analysis is very beautifully explained in the latest published book [288].

Since array outliers and single outliers are marked after calculating gene expression values, the MBEI method down weighted high standard errors represented by array outliers might be produced while pooling replicates or defective fold changes. The image spikes represented by single outliers were replaced during the model fitting process to get rid of poor results. The smaller quantity of outliers indicates good and reliable samples whereas samples with a large number of outliers need to be checked. They should be removed and the steps mentioned in 4.6.2 and 4.6.3 should be redone.

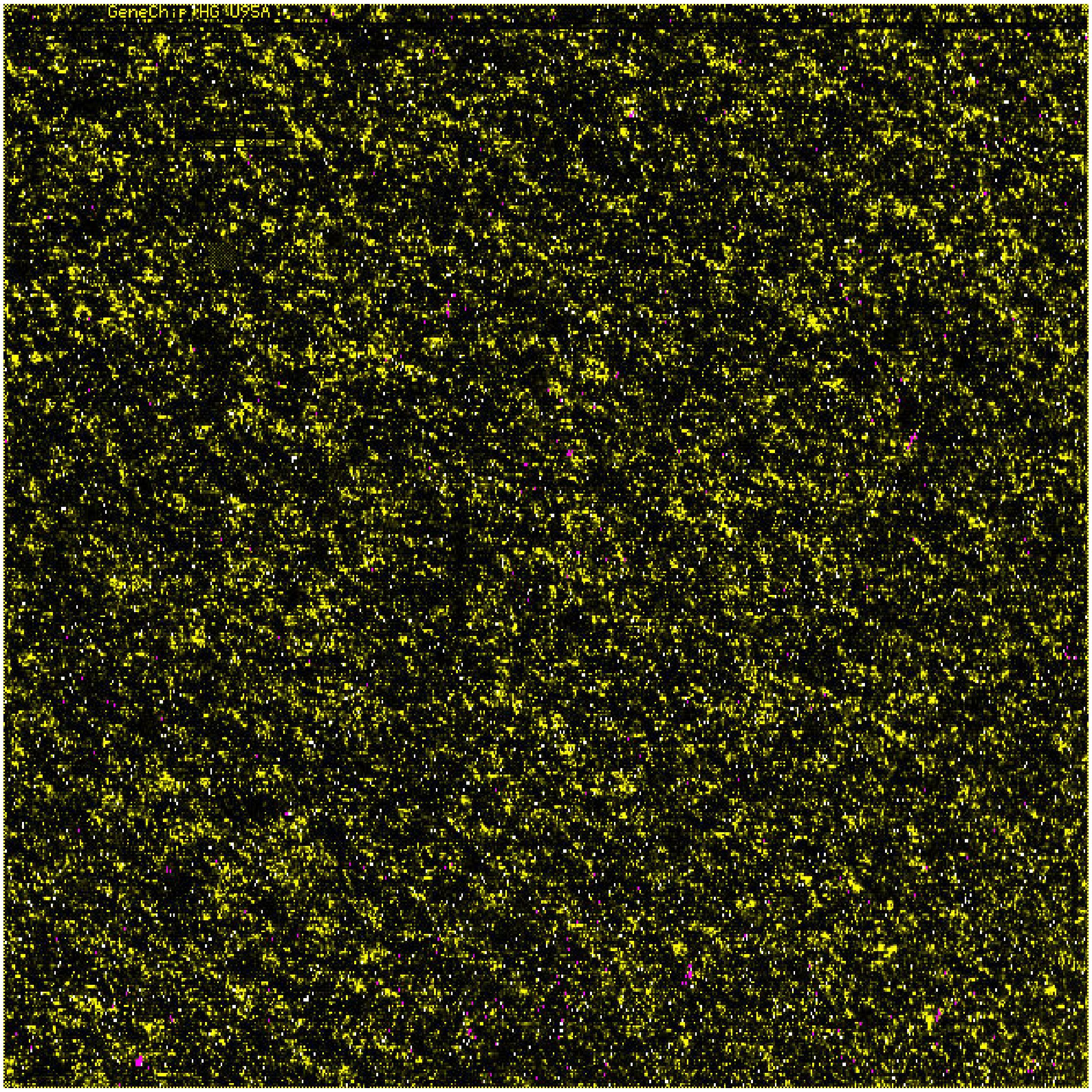


Figure 4.5 CL2001011121AA cel image range covers 1st percentile (black) to 95th percentile (yellow). Outliers are shown in white and purple

4.6.4 Filter Genes to Find Interesting Genes

In this step, we take the one group as baseline and the other as experiment group. Then filtering criteria is made based on absolute differences using the lower bound of fold change. The baseline group consists of ALL samples and Experiment group consists of

MLL samples with variations across samples taken as 0.50 < Standard deviation / Mean < 1000.00 and the P call percent in the array used must be greater than or equal to 20%. The results of the investigation will be more dependable by initially excluding genes with little or no variation across the samples or genes that are missing in the bulk of the samples. On comparing the ALL and MLL samples, 604 differential expressed genes were found which satisfied the comparison filtering criteria (using lower bound fold change, absolute differences) with a false discovery rate (number) of fifty permutations. Figure 4.6 shows

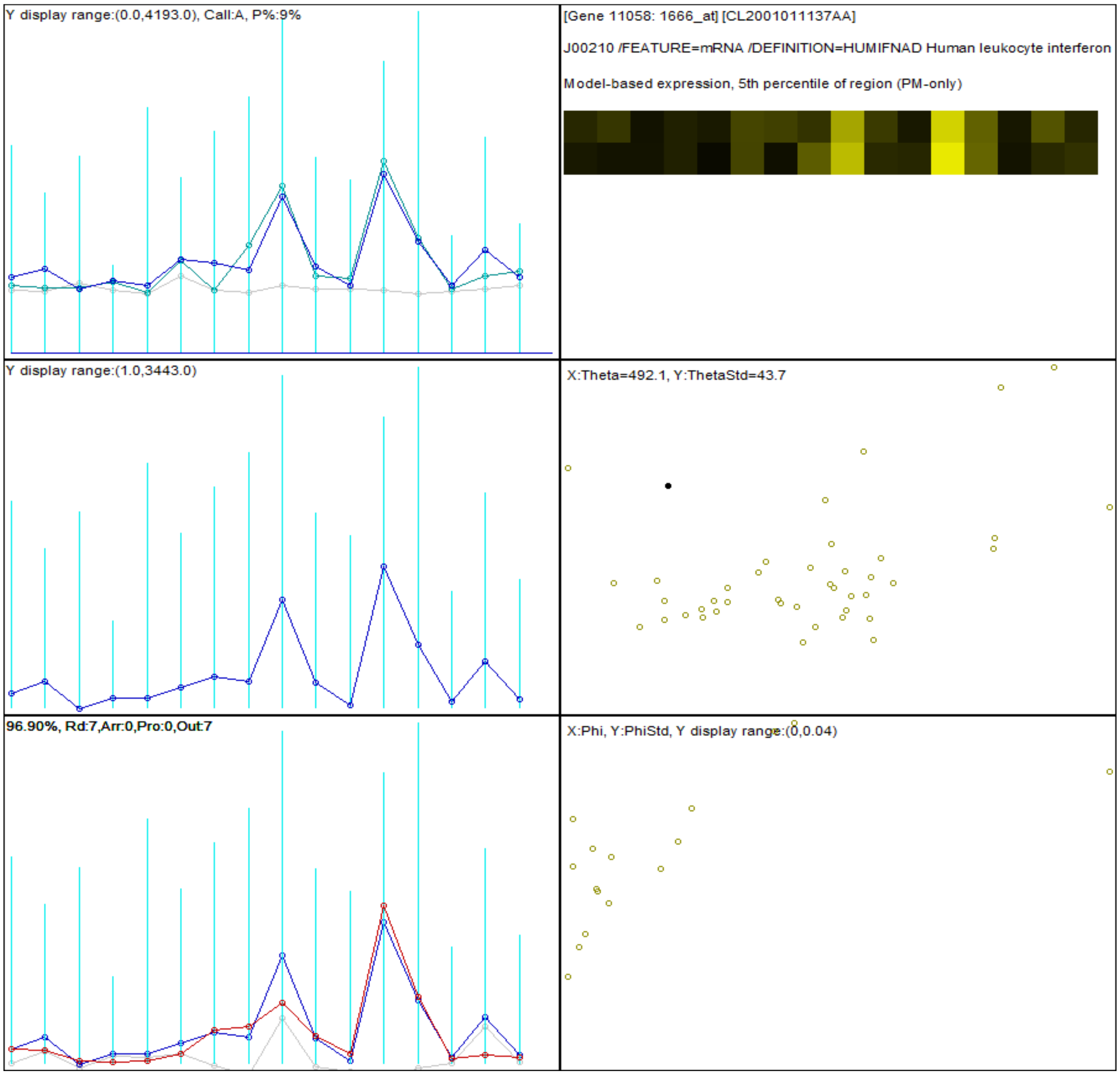


Figure 4.6 Probeset Panel for human leukocyte interferon gene

the result of the human leukocyte interferon gene found in the probeset panel with six quadrants. The first quadrant shows 16 probes with PM/MM/BG values in CL2001011121AA array sorted in order of increasing MBEI values, the second shows the heatmap, the third shows perfect match values only because of the PM method used, the fourth shows plotting of MBEI values against standard errors with colour dot as outlier, the fifth is for checking residuals which can be informative at times and the sixth shows the sensitivity of probes plotted against their standard errors. For this particular gene in question, it took seven rounds of iterations with 96.90% of variation having 0 array outliers, 0 probe outliers and 7 single outliers. Similarly, we can see the probeset panels for all the revealed differential expressed genes for further analysis. These significant genes act as pre-processed output for high-level analysis of genomic data.

4.6.5 Hierarchical Clustering of Samples

The basic hierarchical clustering was performed on the genes found in the previous step using Euclidean distance for calculating the distance to find the nearest neighbour and merge nodes. Biological replicates rely heavily on the same cell line grown-up in diverse tableware whereas in a systems biology analysis there should be more replicate samples treated under one group as we have done. By using filtered genes and replication, the process can corroborate the quality of the microarray analysis and circumvent the absent call variant genes in the microarray. The result of the microarray analysis after hierarchical clustering of samples is shown in the heatmap file available at www.iitr.ac.in/ajayshiv/cluster.html since the file size is too big to be incorporated here. It can be used to interpret highly expressed genes in a certain assembly of samples and the closeness of clusters. It is noted that genes with analogous patterns across gene expression space fit in a related functional group.

4.6.6 Correlation Matrix Based on Genetic Association Study

Figure 4.7 shows the correlation matrix based on genetic association study of clustered genes with a comparable expression mould with a given gene. In this correlation matrix based on a correlation coefficient, red indicates highly similar correlated, white points out no similarity at all and blue depicts a negative correlation.

4.6.7 Gene Ontology Data Mining / Gene Annotation Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a statistical test that can identify sets of genes belonging to a particular biological category, which plays an important role in distinguishing between two classes of gene expression data. The test is particularly sensitive since small changes that are coordinated across the set can be detected. The test helps reveal the biological mechanisms responsible for the difference between the two classes because the test set has an a priori biological theme. The prior biological knowledge exists for the feature annotations and are available in different databases. The genes were classified in this step depending on their functional category and as a result anchor genes were mapped to chromosomes. The result of mapping of all chromosomes is shown in the heatmap file available at www.iitr.ac.in/ajayshiv/chromo.html since the file size is too big to be incorporated here.

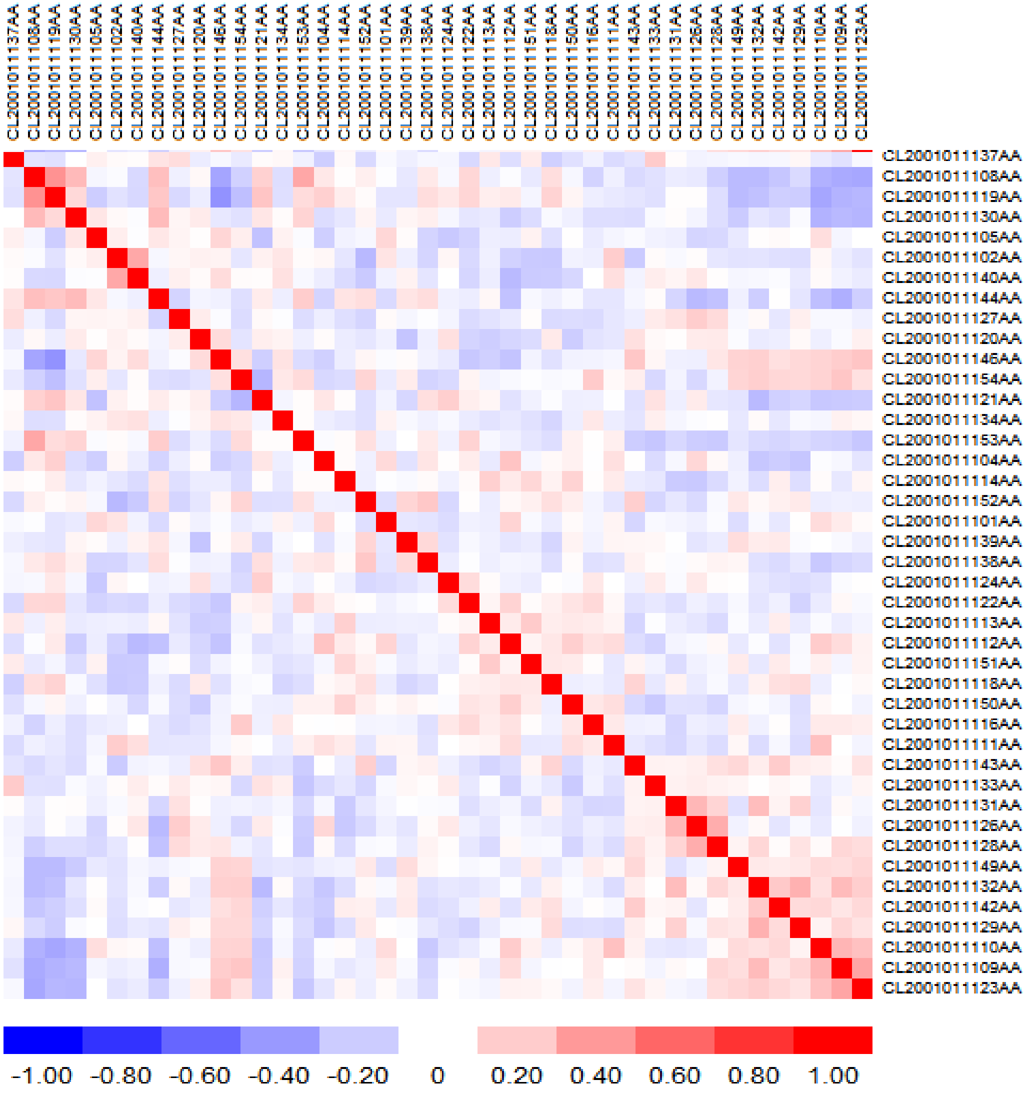


Figure 4.7 Correlation matrix

A gene annotation enrichment analysis was made on clustered genes that have the following classification parameters: "'C1' identifies the number of genes in a cluster or list with this annotation term, 'C2' indicates the number of annotated genes in this cluster or list, 'C3' recognizes the number of all genes on array that have this annotation term, 'C4' is used to categorize all annotated genes on an array. The ‘P-value’ is used to establish the binomial approximated p-value for hypergeometric distribution.” From an analysis, there were 2335 probe sets for the same genes in all other samples and were discarded from the annotation enrichment analysis. The results after calculation based on gene ontology data mining reported 80 significant genes with 358 estimated false positives and 357554 cluster-term pairs evaluated at p-value threshold 0.001000 as shown in Table 4.3.

Table 4.3 Results based on Gene Ontology Data Mining

C1	C2	C3	C4	P-value	Term Name
4	50	48	5898	0.000749	actin binding
5	436	10	5898	0.000983	adherens junction
29	48	2198	5898	0.000948	binding
50	93	2198	5898	0.000879	binding
7	7	2198	5898	0.000998	binding
86	1195	283	5898	0.000170	biosynthesis
5	12	350	5898	0.000409	cell cycle
18	133	350	5898	0.000924	cell cycle
5	139	30	5898	0.000779	cell cycle arrest
42	53	3310	5898	0.000374	cell growth and/or maintenance
72	98	3310	5898	0.000294	cell growth and/or maintenance
4	16	167	5898	0.000890	cell organization and biogenesis
10	256	61	5898	0.000392	chaperone
165	595	1305	5898	0.000771	cytoplasm
4	18	110	5898	0.000300	cytoplasm organization and biogenesis
4	33	76	5898	0.000837	cytoskeletal protein binding
5	50	76	5898	0.000465	cytoskeletal protein binding
4	18	60	5898	0.000029	cytoskeleton organization and biogenesis
64	1195	179	5898	0.000014	cytosol
19	1195	36	5898	0.000204	cytosolic large ribosomal subunit (sensu Eukarya)
43	2276	64	5898	0.000488	cytosolic ribosome (sensu Eukarya)
18	1671	27	5898	0.000952	cytosolic small ribosomal subunit (sensu Eukarya)
6	63	93	5898	0.000484	defense/immunity protein
8	484	24	5898	0.000963	di-, tri-valent inorganic cation transport
9	252	58	5898	0.000995	endopeptidase inhibitor
12	110	222	5898	0.000928	enzyme regulator
19	1195	38	5898	0.000394	eukaryotic 43S pre-initiation complex
18	1671	27	5898	0.000952	eukaryotic 48S initiation complex
8	507	23	5898	0.000989	Exocytosis
6	410	16	5898	0.000999	feeding behavior
7	265	33	5898	0.000825	G2/M transition of mitotic cell cycle
4	26	97	5898	0.000819	GTP binding
4	31	73	5898	0.000565	GTPase
4	88	29	5898	0.000981	guanyl-nucleotide exchange factor
12	37	704	5898	0.000863	hydrolase
5	8	704	5898	0.000992	hydrolase
4	4	884	5898	0.000505	integral plasma membrane protein
153	250	3023	5898	0.000980	intracellular
20	1195	40	5898	0.000281	large ribosomal subunit
10	391	45	5898	0.000999	lipid biosynthesis
6	146	45	5898	0.000963	lipid biosynthesis
12	98	236	5898	0.000558	lipid metabolism
82	1715	185	5898	0.000163	macromolecule biosynthesis
11	17	1501	5898	0.000737	membrane
181	407	2151	5898	0.000549	metabolism

4	265	9	5898	0.000793	mitochondrion organization and biogenesis
4	12	172	5898	0.000297	mitotic cell cycle
5	5	1039	5898	0.000170	nucleic acid binding
142	629	1039	5898	0.000905	nucleic acid binding
5	7	784	5898	0.000689	nucleus
4	18	96	5898	0.000179	organelle organization and biogenesis
4	6	472	5898	0.000539	organogenesis
11	449	46	5898	0.000966	oxidoreductase, acting on CH-OH group of donors
7	29	291	5898	0.000424	physiological processes
4	88	29	5898	0.000981	plasma glycoprotein
7	9	1211	5898	0.000373	plasma membrane
5	152	29	5898	0.001000	plasma protein
9	252	58	5898	0.000995	protease inhibitor
82	1715	185	5898	0.000163	protein biosynthesis
200	1195	798	5898	0.000927	protein metabolism
27	914	89	5898	0.000965	protein tyrosine kinase
6	12	608	5898	0.000638	receptor
5	20	215	5898	0.000630	receptor signaling protein
14	139	211	5898	0.000496	regulation of cell cycle
7	107	82	5898	0.000780	reproduction
4	10	289	5898	0.000953	response to pest/pathogen/parasite
5	12	403	5898	0.000784	response to stress
52	1634	117	5898	0.000823	ribonucleoprotein complex
52	2276	83	5898	0.000660	ribosome
64	969	259	5898	0.000996	RNA binding
5	20	220	5898	0.000699	RNA polymerase II transcription factor
7	107	82	5898	0.000780	sexual reproduction
7	100	88	5898	0.000786	small GTPase regulatory/interacting protein
18	1671	27	5898	0.000952	small ribosomal subunit
6	99	67	5898	0.000978	specific RNA polymerase II transcription factor
4	110	23	5898	0.000961	steroid biosynthesis
45	2276	69	5898	0.000669	structural constituent of ribosome
51	800	236	5898	0.000935	structural molecule
17	435	95	5898	0.000865	transcriptional activator
4	5	584	5898	0.000443	transferase

The results after calculation based on protein domain data annotation reported 8 significant protein domains, 98 estimated false positive and 97945 cluster-term pairs evaluated at p-value threshold 0.001000 as shown in Table 4.4.

Table 4.4 Results based on Protein Domain annotation

C1	C2	C3	C4	P-value	Term Name
4	22	127	6567	0.000774	Cytochrome c heme-binding site
9	398	37	6567	0.000510	Dbl domain (dbl/cdc24 rhoGEF family)
4	80	5	6567	0.000001	Dynein heavy chain
4	11	267	6567	0.000716	Immunoglobulin/major histocompatibility complex
9	372	43	6567	0.000906	Protein kinase C, phorbol ester/diacylglycerol binding
5	110	44	6567	0.000923	Steroid hormone receptor
5	18	253	6567	0.000477	Zn-finger, C2H2 type

The results after calculation based on pathway annotation reported 2 significant pathways, 16 estimated false positive and 15889 cluster-term pairs evaluated at p-value threshold 0.001000 as shown in Table 4.5.

Table 4.5 Results based on Pathway annotation

C1	C2	C3	C4	P-value	Term Name
45	553	82	1722	0.000414	Cytoplasmic Ribosomal Proteins
9	129	33	1722	0.000908	Nuclear Receptors

The results after calculation based on chromosome annotation reported 31 significant, 113 estimated false positive and 112809 cluster-term pairs assessed at p-value threshold 0.001000 as shown in Table 4.6.

Table 4.6 Results based on Chromosome annotation

C1	C2	C3	C4	P-value	Term Name
13	132	304	8477	0.000977	10
5	22	304	8477	0.000936	10
5	25	241	8477	0.000613	10q
10	102	241	8477	0.000679	10q
4	7	494	8477	0.000350	11
9	59	322	8477	0.000370	11q
5	55	81	8477	0.000186	12q24
5	19	345	8477	0.000804	16
6	28	345	8477	0.000789	16
4	21	160	8477	0.000587	16q
4	24	160	8477	0.000997	16q
10	927	27	8477	0.000957	16q24
4	24	146	8477	0.000710	17p
5	100	62	8477	0.000886	17q11
4	99	37	8477	0.000982	17q25
19	283	236	8477	0.000438	19p
4	10	289	8477	0.000240	19q
4	9	469	8477	0.000942	1p
8	502	33	8477	0.000918	1p13
4	97	36	8477	0.000823	1q42
6	19	527	8477	0.000773	2
4	8	527	8477	0.000853	2
11	94	307	8477	0.000608	2q
5	22	307	8477	0.000978	2q
5	22	307	8477	0.000978	2q
121	1645	467	8477	0.000979	3
4	10	241	8477	0.000120	3p
6	1342	7	8477	0.000998	3q12
4	40	93	8477	0.000966	4p
5	350	16	8477	0.000594	5q11
28	401	308	8477	0.000901	6p
23	474	185	8477	0.000395	6p21

4.7 DISCUSSION AND CONCLUSION

The list of significant genes or differentially expressed genes in provisions of probabilities was found. This helps to find the functional relationships between genes in MDB warehouses by linking annotations of GO. Moreover, clusters with filter genes are able to distinguish ALL and MLL in an unsupervised clustering. Systems biology application of MDB enhances further research in diagnostics, prognostics, disease markers, target validation and targeted therapies. The recent case with a precautionary double mastectomy on finding the BRCA1 gene with only 87 percent probable chance of acquiring the disease shows the promising nature of this field. We endow with upcoming field of microarray bioinformatics that will ignite the passion of life scientists and budding researchers to work in this field. To the best of author's knowledge, no comparative experimental test has been carried out on systems biology processes that clearly distinguish between truly differentially expressed genes and false positives. This can be a future work with great potential. The analysis on the leukemia microarray data is to be done for another input data set – the latest Affymetrix-deposited HG files: HG-U133 gene info2.xls, HG-U133 gene info2 Gene Ontology.xls and HG-U133 gene info2 Protein Domain.xls (or other most recent files, if deposited in the meantime) in the near future. In addition, high throughput technologies produce a huge amount of data that requires reliable annotations (also termed metadata) in order to provide significant biological and/or clinical interpretations and fit to be used in systems biology approaches. Therefore, for any experiment, the quality and availability of gene annotations is essential for biological discoveries. To conclude, the modern era of functional genomics propelled by MDB technology paves the way to elucidate physiological model of any disease. Therefore, bioinformatics applications and systems biology will be able to facilitate in decoding the life sciences concealed knowledge to beat diseases using a personalized medicine decision support system.

4.8 AUTHOR'S RESEARCH CONTRIBUTION

Sharma, A.S., Gupta, H.O., Prasad, R., Mitrasinovic, P.M. (2013) "Microarray Database Bioinformatics usher Functional Genomics to unveil Biological Knowledge underlying Physiology" International Journal of Advanced Research in Computer Science and Software Engineering, 3(10), 38-45.

Chapter – 5

MULTISCALE MODELLING OF SOLID TUMOUR GROWTH FOR LUNG CANCER TREATMENT

5.1 CHAPTER OVERVIEW

This study developed a mathematical model for tumour growth and angiogenesis that simulated a solid tumour's growth/progression with chemotherapy and anti-angiogenesis drugs using partial differential equation (PDE) modelling. The PDE compartmental model incorporated spatiotemporal processes including cellular and tissue-mediated diffusion, cellular transport and migration, cell proliferation, angiogenesis, apoptosis, vessel maturation and formation to model tumour progression and transition from avascular to vascular growth. The angiogenesis process coupled with the solid tumour growth model on a reaction–diffusion kinetics framework portrayed the spatiotemporal development of the generalised functions of a tumour's micro-environment viz., nutrients and growth factors that regulate the tumour's growth during angiogenesis. Most cancers involve an endothelial growth factor receptor/extracellular signal-regulated kinases (EGFR/ERK) signalling pathway, which are related to the cell-division cycle promoting tumour cells. Treatment is studied from tyrosine kinase inhibitors (TKI) in EGFR signalling, which are distributed through the blood vessels of a tumour's microvasculature. This showed a huge potential for in-vitro experiments due to the availability of clinical and expression data information, which helps in learning about the responses to treatment. Using ordinary differential equations to model the systems pathway of downstream pathway of EGFR signalling (SOS→RAS→RAF→MEK→ ERK→PI3K→AKT), we performed computational simulations to determine the facilitation of glucose, oxygen, tumour angiogenesis factor (TAF), drug (TKI), tumour growth factor alpha (analogue of EGFR) and angiogenesis inhibitor. The simulation results showed signalling pathways of TKI-EGFR and IGF1R regulation of various active cells, migrating cells, proliferative cells, apoptotic and quiescent cells could be a united behaviour for the entire profile of tumour growth. The results established the dual role behaviour played by angiogenesis as TKI-EGFR and VEGF inhibitors are furnished to diminish tumour incursion. In addition, the neovasculature can transport nutrients to neoplasm cells to continue cell metabolism, thus enhancing the rate of cell endurance.

Hence, simulation results suggest that the co-expression of EGFR and IGF1R activates a higher number of ERK receptors compared to down and over-expressions. There is a good agreement between the simulations, an experimental wild type mouse model and clinical data.

5.2 INTRODUCTION

In recent times, mathematical modelling and simulations of biological processes has become an important tool. However, only a few models of solid avascular tumour growth [172] and multidimensional tumour growth models [173-175] have been proposed and a few models of tumour growth coupled with the process of angiogenesis were developed [178-180]. Recently, a few models incorporating chemotherapy drug treatments with/without angiogenesis have been proposed [181-183].

In multiscale modelling of biological systems, biological species, such as an organism, a cell, and an organelle or sub cellular structure are often explored at different microscopic levels. Studying biological elements at different levels (molecular level, cellular level, tissue level, etc.) helps to understand the whole system as a single entity. However, the system takes into account spatial modelling (a single cell through whole body) and temporal modelling (nanosecond through entire lifespan) that the entire biological system will be thoroughly understood by recognizing how the processes at different levels work together. Systems biology inspires the development of a model using a non-reductionist approach, starting with the simplest basic model [289]. This is because biological systems are synergistic and integrative behaviour happens in a non-predictable fashion. Most studies have focused only on a single scale such as genes, proteins, cells, tissues, organs, organ systems, and the entire body. Currently, the emphasis is on developing novel and efficient algorithms, methods, tools and scientific approaches to seamlessly integrate the path from microscale to macroscale.

Malignant tumours usually develop out of a few cells that have vanished or no longer respond to normal regulatory mechanisms, probably due to mutations and/or an altered gene expression [290, 291]. As a tumour progresses, the genetic instability causes sustained alterations in the biological mechanisms such as cell invasion, angiogenesis, and metastasis, resulting in complex tumour formations. Moreover, the tissue that surrounds tumours, including avascular tumours, diffuses a supply of nutrients to the tumours. Anterior to the development of a blood supply, these tumours are unable to establish an adequate supply of nutrients that maintains the tumour's cell mass despite a continuous supply of nutrients on the surface of the tumour. The limited availability of vital nutrients (glucose and oxygen) will appear within the cell mass as noticeable gradations. Tumour cells react to this by self-induced changes in their physiological and metabolic processes along with altered genes and protein expressions [292, 293]. Furthermore, it would be advantageous to have better insight into heterogeneous microenvironments on the growth regulation and malignant development of avascular tumours [294].

Multi-cellular tumour spheroids frequently used with in vitro models of avascular tumour growth for studying fluctuations in physiological status along with micro-environmental factors that naturally occur in tumours [295, 296]. Spheroid tumour cells can be cultivated under controlled conditions with specific parameters including tumour volume, the number of cells, viable and fraction of necrotic cells, and saturation size by regulating the nutrient supply [297, 298]. Spheroids receive their supply of nutrition by diffusion from extracellular bodies or the surface of the cell [292, 296]. Therefore, when the mass increases, the nutrient-deprived inner regions of the cell are changed. Spheroids acquire several key attributes of avascular tumours, including proliferation arrest, metabolic changes, gene alterations and protein expression, necrosis and a resistance to drugs. Furthermore, spheroid growth curves demonstrate the same movements as nodular tumours in live bodies, including quasi-exponential growth and saturation in size [299]. A descriptive model that clarifies the controlled growth and viability in spheroids proposes that during the initial development, growth viability factors reach every cell in the spheroid. During this initial development, a mass is made up of growing cells. During growth, the core of the spheroid experiences a decrease in the concentration of growth factors leading to a fall in the threshold value and becomes inactive. However, the spheroid remains active because of the outer proliferating cells, yet there is a continuing decrease of the central concentration of viability factors. After the concentration of viability factors are under the threshold value, necrotic cell death happens and the spheroids attain the centre of the necrosis regions. Controlled experiments indicate that simple molecules such as oxygen and glucose are the viability factors in spheroids [297, 298]. In addition, limited data indicates small protein factors inhibit growth [300]. Currently, factors influencing a tumour's growth or inhibit a tumour's viability system, including avascular tumours in live bodies, have not been substantially identified.

Earlier models of tumour growth that involved cellular dynamics are basic empirical mathematical expressions [299, 301], rate equations of cell populations [295, 302-305], cellular automata models of interacting cells [306, 307], cellular geometry [308], cellular automata and continuous chemical and blood flow. A few models employed a fusion of cellular spaces for representing cell behaviour and mathematical equations on the flow of plasma and chemicals for tumour growth modelling [309, 310]. Recently, three dimensional tumour growth coupled with angiogenesis and chemotherapy has been modelled [311]. In this article, we propose a multi-scale, multi-cellular model along with anti-angiogenesis and endothelial growth factor receptor (EGFR) inhibitor treatment. At the cellular scale, the model incorporates the growth of tumours, migration, tumour multiplication, quiescence and apoptosis. For the extracellular level, the model integrates diffusion, convection, consumption, secretion, production and nutrition uptake such as glucose, oxygen, drug, growth factors viz., VEGF inhibitors and tumour angiogenesis factors. At the molecular level, drug and growth factors interact through a systems pathway analysis. Data from earlier multi-cellular spheroids experiments governed the simulation constraints [292, 298, 299].

The EGFR family initially documented (170 kDa protein on the membrane of A431 epidermoid cells) forms a genus of the transmembrane receptor tyrosine kinase superfamily members (HER1/ ERBB1, HER2 / ERBB2, HER3 / ERBB3, HER4 / ERBB4) found through probing complimentary DNA molecules synthesized from EGFR [312-314]. These receptors are often found in various epithelial, mesenchymal, and neural origin tissues. In fact, several types of cancers such as breast, brain, throat, colon, kidney pancreas, and ovary overexpress EGFR [315, 316]. At least 60% of non-small cell lung cancers (NSCLCs) display an overexpression of EGFR, yet no overexpression is noticed in small cell lung cancer [317]. The reason for EGFR overexpression may be due to several epigenetic mechanisms, gene amplification, and oncogenic viruses [315]. It has been shown that EGFR expression is associated with a poor prognosis [318]. EGFR ligands could also be influential with lung tumorogenesis. NSCLCs express EGF, tumour growth factor-alpha and amphiregulin, and stimulate EGFR and its posterior signalling pathways [319]. With the majority of patients, specific reactions from a small portion of NSCLC that responds to these agents are determined by acquired mutations/alterations in EGFR tyrosine kinase domain with a reaction to selective inhibitors such as gefitinib or erlotinib [320-322]. It is very common that EGFR mutations involve NSCLC with women, Asians, individuals with lung cancer and non-smokers who have tumours [323, 324]. EGFR mutations are hardly ever seen in small cell-lung-cancer, squamous cell carcinomas of the lung, or other avascular malignancies. Therefore, the stimulation of somatic EGFR mutations properly belongs to a genus of NSCLC. The primary drug set (EGFR-TKI) inhibiting posterior receptor signalling such as gefitinib and erlotinib selectively target the intracellular EGFR domain. The acquired mutation of EGFR is the most significant predictors of a cells reaction to TKIs [316, 317]. Moreover, EGFR-TKIs for NSCLC with EGFR alterations notably increased the actual response time and chance of amelioration-free existence as evident from positive random trials in contrast to standard platinum therapy. Eight-five percent of lung cancers are NSCLC, which have a high mortality rate after the lung resection, therefore, there is a need to study how EGFR (HER1/ERBB1) and IGF1R signalling pathways works in cancer treatment. Hence, we developed a model of EGFR inhibitors in conjunction with IGF1R signalling pathway.

5.3 MATHEMATICAL MODELLING OF TUMOUR GROWTH

Several factors play a major role in tumour growth, such as nutrient supply, metabolite waste clearance and tissue density have been extensively studied in earlier tumour growth models [325-327]. In our model, we treat nutrient concentrations and growth promoters as typical leading factors of vascular and avascular tumours. Previous models have examined tissue pressure, cell growth and the movement toward areas of lower pressure [328-330]. It is assumed that oxygen and glucose control the metabolic activities of tumour cells such as consumption and secretion rates of oxygen and glucose and anti-angiogenesis (or VEGF inhibitor) and chemotherapy drug uptake rate. Here, it is again assumed that tumour angiogenesis factor-induced vessels are unable to inhabit the necrotic core of tumours due to a cell's density and pressure. Hence, as TAF concentrations, interstitial pressure and tissue density increase the centre of a tumour, the angiogenic vasculature becomes more compact and convoluted. The size and age of the vessels develops iteratively after budding from pre-existing vessels, therefore, tumour vessels differ spatiotemporally in their capacity to maintain basic functions, such as nutrition and waste elimination. Vascular growth supplies the necessary nutrition to otherwise nutrient deficient tumours, which enables the growth of tumours and ultimately to metastasis. Here in our model, we assumed both the tumour growth and angiogenesis is integrated together.

It is commonly known that the intensified growth of tumour cells increases the need of critical nutrients required for synthesising DNA to RNA making proteins, which supply carbon for creating the assimilation force of a tumour cell. On the other hand, it is difficult for hydrophilic aliments such as micronutrients, amino acids, fatty acids, glucose and vitamins to transit through the cytomembrane [331]. In our model, we consider nutrient concentrations of oxygen (required by cells for aerobic metabolism) and glucose as generic elements. The oxygen accessibility to cells directly influences tumour cell processes (growth metabolism, angiogenesis and metastasis) [332, 333]. The circulatory system is comprised of several types of blood vessels. The three kinds of blood vessels viz., capillaries, veins and arteries transport oxygen that is bound to haemoglobin to the entire body. Arteries transport oxygen rich blood to the capillaries away from the heart, which quickly moves into the tumour tissue. Healthy tissue also eliminates waste by moving it from the cell back to the haemoglobin, which is carried to the heart and lungs. Therefore, the spatiotemporal development for oxygen levels (a) is presumed to happen by diffusion, proliferation of red-blood-cells, the elimination through tumour cells, and convection as follows:

$$\frac{\partial a}{\partial t} = D_a \nabla^2 a - \nabla(\vec{u}.a) + \rho_a(k_v,(pv-p))\delta_{\sum V} - \lambda_a(N_i)\delta V_T \tag{1}$$

where

$$\rho_a(k_v,(pv-p)) = \rho_{a0} R_i W \tag{2}$$

$$W = \begin{Bmatrix} \dfrac{pv-p}{pv}, \text{ if } pv-p>0 \\ 0, \text{ otherwise} \end{Bmatrix} \tag{3}$$

and R_i is the radius of the vessel taken as $R_i = \dfrac{Age_i}{Age_i + k_{AR2}} k_{AR11}$ (4)

$$N_i = \frac{n}{n+1}\exp(-5(1-a)^4) \text{ and} \tag{5}$$

$$\frac{dV_i}{dt} = \frac{N_i}{N_i+1} s_{active} = -s_{quiescent} \tag{6}$$

In the above Equation1, the first term is the diffusion term, the second being convection term, third term is O_2 released by the blood cells and the last term being O_2 uptake by the cells. Vessels are pruned when $Age_i = 0$. All the rate constants are taken from the earlier experimental data given in Table 5.1.

Table 5.1 Parameters and constants used in the model

Symbol	Value	Unit	Description	Reference
D_a	8x10^{-14}	m^2/s	Oxygen diffusion coefficient	[325]
ρ_a	6.8x10^{-4}	mol/(m^{3}s)	Oxygen supply rate	[325]
λ_a	0 6.8x10^{-4}	ml/(m^{3}s)	Oxygen consumption rate	[294]
D_b	6.7x10^{-7}	cm^2/s (5.2-7.2)	glucose diffusion coefficient	[168]
ρ_b	3x10^{-5}	mol/(m^{3}s)	glucose permeability rate	[168]
λ_b	0.28	m mol/hr	glucose uptake by cells	[168]
λ_{b1}	0.2	m	glucose decay rate	[168]
D_c	1.2x10^{-13}	m^2/s	TAF diffusion coefficient	[311]
ρ_c	2x10^{-9}	mol/(m^{3}s)	TAF secretion rate	[311]
λ_c	0	ml/cm^{3}s	TAF consumption rate	[311]
D_d	1.5x10^{-14}	m^2/s	Drug diffusion coefficient	[311]
ρ_d	2.5x10^{-7}	ml/(cm^{3}s)	Drug consumption rate	[311]
λ_d	1x10^{-8}	ml/(cm^{3}s)	Drug decay rate	[311]
D_e	2.9x10^{-7}	cm^2/s	VEGF diffusion rate	[334]
ρ_e	0.6	m mol/hr	VEGF secretion rate	[334]
λ_e	0.1x10^{-4}	cm/s	VEGF permeability	[334]
λ_{e1}	0.2	[range 0.1-0.4]	VEGF decay rate	[334]
k_{active}	1.0		Rate of CVE addition by active cells	[311]
$k_{quiescent}$	0.1		Rate of CVE addition by quiescent cells	[311]
k_{AR1}	1.0		Vessel radius constant	[311]
k_{AR2}	500.0		Vessel radius constant	[311]
K	4.5x10^{-15}	cm^2/mmHg^{2}sec	Interstitium’s hydraulic conductivity	[311]
Lp	2.8x10^{-9}	m/mm Hg-sec	Microvascular wall’s hydraulic conductivity	[335]
P	1.49x10^{-9}	m/sec	Vascular permeability coefficient	[311]
σ_T	0.82		Average osmotic reflection coefficient	[335]
σ_D	0.1		Average osmotic reflection coefficient	[311]
p_V	0.3546	mmHg	Osmotic pressure of plasma	[311]
ρ_i	0.2667	mmHg	Osmotic pressure of interstitial fluid	[311]
ρ_V	30	mmHg	Capillary/vascular pressure	[336]
p_0	60	mmHg	Tumor pressure	[337]
d_V	1.0	mol/m^3	Interstitial drug concentration	[311]
a_0	8.4	mol/m^3	Standard nutrient O_2 concentration	Estimated
b_0	10.5	mol/m^3	Standard waste (Co_2) concentration	Estimated
c_0	4.361024	kg/m^3	Standard TAF concentration	[338]
d_0	2.13	mol/m^3	Standard drug concentration	Estimated
e_0	1.0	mol/m^3	Standard VEGF concentration	Estimated

In the similar fashion, the concentration of glucose (*b*) is governed by the following equation:

$$\frac{\partial b}{\partial t} = D_b \nabla^2 b - \nabla(\vec{u}.b) + \rho_b(N_i)\delta\mu_T - \lambda_b(k_v,(pv-p))\delta\sum V - \lambda_{b1}(N_i)b^2\delta\sum T \qquad (7)$$

where $\rho_b(N_i) = \rho_{b0}N_i$ and $\lambda_b(k_v,(pv-p)) = \lambda_{b0}R_iW$ (8)

In Equation 7, the first term is diffusion, the second being convection, third term is glucose source, fourth term is glucose uptake by the cells and the last term is natural decay of glucose. Tumour cells require substantial levels of glucose and oxygen and collect waste during the tumour's growth process due to poor waste elimination dynamics. Henceforth, it reduces biosynthesis, cell activity or results in cell death.

Tumour cells consume nutrients faster than normal cells, which further causes a lack of oxygen in avascular tumours. Due to the lack of oxygen, tumour cells secrete a tumour angiogenesis factor (TAF) that stimulates new vessels from vasculature cells in oxygen depleted regions [339]. Generally, the drugs can be classified into different types such as cytotoxic, anti-angiogenic inhibitors or growth factors. Anti-angiogenic drugs reduce the rate of ATF secretion, which restrains the development of neovasculature cells.

$$\frac{\partial c}{\partial t} = D_c \nabla^2 c - \nabla(\vec{u}.c) + \rho_c(a)\delta\sum\Omega_T - \lambda_c(k_v)\delta\sum V \qquad (9)$$

where $\lambda_c(k_v) = \lambda_{c0}R_i$ and $\rho_c(a) = \rho_{c0}(1-a)$; $\lambda_c(k_v) = \lambda_{c0}R_i$ (10)

In Equation 9 above, the first term is diffusion, the second a convection term, third term is TAF released by tumour cells, the fourth term is the removal of TNF through blood vessels in which the rate of secretion, $\rho_c(a)$ is directly proportional to the oxygen level, *Dc* is the diffusion coefficient of TAF in tumour tissue, λ_c is supply rate of TNF through new vessels as defined by [335, 340]. The TKI inhibitor of EGFR signalling [341] is governed by the following equation

$$\frac{\partial d}{\partial t} = D_d \nabla^2 d - \nabla(\vec{u}.d) + \rho_d(k_v,(pv-p),d_v(t))\delta\sum V - \lambda_{d1}(k_v,(N_i)d^2\delta\sum T - \lambda_{d2}d \qquad (11)$$

where $\lambda_d(k_v,(N_i)) = \lambda_{d0}R_iW$ and

$$\rho_d(k_v,(pv-p,d_v(t))) = \frac{L_pS(k_v)}{V}(p_v - p_i - \sigma_T(\pi_v - \pi_i)(1-\sigma_D)d_v(t) + P\frac{S(k_v)}{V}(d_v(t)-d_i)\frac{P_{ev}}{e^{P_{ev}}-1} \qquad (12)$$

In Equation 11 above, the first term is diffusion, the second is convection, third term is drugs released by cells, the fourth term is uptake by cells and the last term is drug decay. In Equation12, " L_p is the hydraulic conductivity of tumour microvasculature wall, $S(k_v)/V$ is the surface area per unit volume of drug transport in the tumour, p_v is the vascular pressure, p_i is interstitial pressure, π_v is the osmotic pressure of the drug, π_i is the osmotic pressure of the interstitium, σ_T is the average osmotic reflection coefficient of the drug, P_{ev} is the Peclet number, defined as the ratio of convection to diffusion magnitude across the capillary wall" [311, 342]. Previous studies showed the control of tumour growth in wild type mammals by infusion of angiogenesis inhibitors. Therefore, we incorporated vascular endothelial growth factor (VEGF) inhibitors, which play a role in anti-angiogenesis process [334] that is given by

$$\frac{\partial e}{\partial t} = D_e \nabla^2 e - \nabla(\vec{u}.e) + \rho_e(k_v,(pv-p),e_v(t))\delta\sum V - \lambda_{e1}(k_v,(N_i)e^2\delta\sum T - \lambda_{e2}e \quad (13)$$

where $\lambda_e(k_v,(N_i)) = \lambda_{e0}R_iW$ and $\rho_e(k_v,(pv-p) = \rho_{e0}R_iW$ (14)

In the above Equation13, the first term is diffusion term, the second being convection term, third term drug released by cells, the fourth term is uptake by cells and the last term being drug decay. We define the state of the cell such as apoptosis, migration, proliferation and active as follows. Let D_m and G_m are the drug and glucose concentrations of the cells

$$A_T = \begin{Bmatrix} D_m, G_m < d_T, necrosis \\ A_T < D_m, G_m < d_T, reversible\, queiscent \end{Bmatrix} \quad (15)$$

where A_T is the active threshold and d_T is dead threshold respectively;

For cell migration, we estimate the mean value of phospholipase C gamma1 μ(PLCγ), is a protein in the body, which plays a role in cancer metastasis and subsequently blocking it stopped cancer from spreading. The migration potential of the cell is taken as the rate of change of the PLCγ is given by

$$M_p = \frac{d[PCL\gamma]}{dt} \quad (16)$$

$$\begin{Bmatrix} M_p > \mu[PCL\gamma], cell\ migrates \\ M_p < \mu[PCL\gamma], cell\ proliferates \end{Bmatrix} \quad (17)$$

The probability of choosing the "attractive location" in migration position is taken by

$$P_i = {}^{\phi G_i}\!\Big/_{F_i + (1-\phi)\varepsilon_i} \tag{18}$$

where G_i is the concentration of glucose at location i, F_i is the concentration of fibronectin and ε_i is the error term, which is Gaussian distribution with mean 0 and variance 1 and $\phi \in (0,1)$ representing the search precision [343] which is considered as 0.7. The probability of migration of tip endothelial cell is defined as

$$P_i = \left(\alpha \frac{w_v}{w_v + V_f} \nabla V_f + \beta \nabla F \right) l_k,\ k = 1, \dots 4 \tag{19}$$

where the first term is chemotaxis in response to VEGF [334] and second term in response to fibronectin. α, w_v and β are the positive rate constants, V_f and F are the concentration of VEGF and fibronectin respectively, l_k is the directional vector along k th direction (left, right, top and bottom).

Tissue regeneration and vessel branching process in two and three dimensional grid is given as

$$B_i^{\theta} = \frac{V_l * (A_{Vf}^{l})}{(1 + w_l * A_{Dm}^{l})} \tag{20}$$

$$\text{where } V_l = \begin{Bmatrix} 1/cell^2 \in [7,26] \\ 1/cell \in [5,6] \\ 1/cell \in [3,4] \\ 1/cell \in [1,2] \end{Bmatrix} \tag{21}$$

(A_{Vf}^{l}) and (A_{Dm}^{l}) is the concentration of VEGF and drug respectively at l^{th} candidate site and its neighbors and w_1 is small value taken as 0.01.

5.3.1 ODE Model of EGFR and IGF1R Signalling Pathways

In this section, we formulate the system of ordinary differential equations (Equations 22 - 39) of all the species (SOS-RAS-RAF-MEK-ERK-PI3K-AKT) involved in the EGFR signalling pathways as shown in Figure 5.1.

$$\frac{dC_1}{dt} = -k_1 * C_1 \tag{22}$$

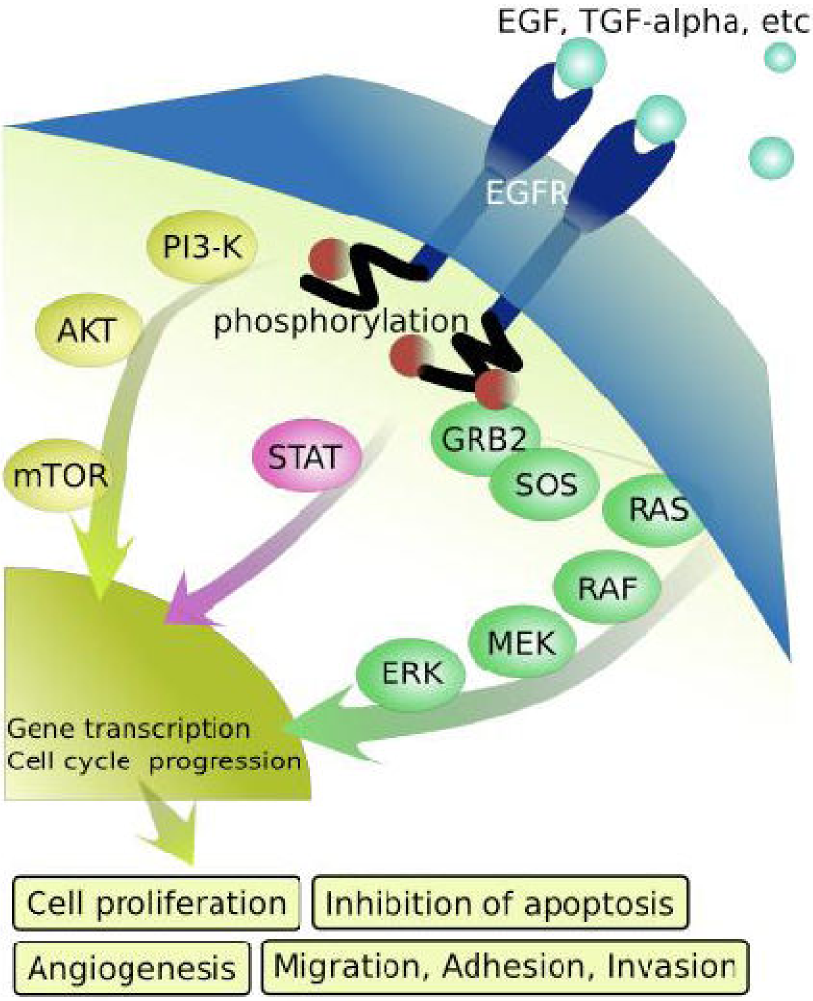


Figure 5.1 Schematic diagram of EGFR Signaling Pathway

$$\frac{dC_2}{dt} = -k_2 * C_2 \tag{23}$$

$$\frac{dC_3}{dt} = k_3 * C_1 * \frac{C_4}{(k_E + C_4)} + C_2 * k_4 * \frac{C_4}{(k_I + C_4)} - C_{13} * k_5 * \frac{C_3}{(k_{p90} + C_3)} \tag{24}$$

$$\frac{dC_4}{dt} = k_3 * C_1 * \frac{C_4}{(k_E + C_4)} + C_2 * k_4 * \frac{C_4}{(k_I + C_4)} + C_{13} * k_5 * \frac{C_3}{(k_{p90} + C_3)} \tag{25}$$

$$\frac{dC_5}{dt} = C_3 * k_6 * \frac{C_6}{(k_{M\_Ras} + C_6)} - C_6 * k_{rgab} * \frac{C_5}{(k_{M\_RasGab} + C_5)} \tag{26}$$

$$\frac{dC_6}{dt} = C_3 * k_7 * \frac{C_5}{(k_{M\_Ras\_Rasgb} + C_5)} - C_3 * k_{Ras\_sos} * \frac{C_6}{(k_{M\_Rassos} + C_6)} \tag{27}$$

$$\frac{dC_7}{dt} = RafPP * k_7 * \frac{C_8}{(k_{M\_Rafpp} + C_8)} - C_5 * k_{Raf\_Ras} * \frac{C_7}{(k_{M\_Raf\_Ras} + C_7)} + C_{17} * k_{Raf\_Act} * \frac{C_8}{(k_{M\_RafAct} + C_8)} \tag{28}$$

$$\frac{dC_8}{dt} = C_5 * k_8 * \frac{C_7}{(k_{M\_Raf\_Ras} + C_7)} - Rafpp2 * k_{Raf\_Rafpp} * \frac{C_8}{(k_{M\_Raf\_Rafpp} + C_8)} - C_{17} * k_{RafAct} * \frac{C_8}{(k_{M\_RafAct} + C_8)} \tag{29}$$

$$\frac{dC_9}{dt} = pp2A * k_9 * \frac{C_{10}}{(k_{M\_MEK\_PP2A} + C_{10})} - C_8 * k_{Mek\_Raf} * \frac{C_9}{(k_{M\_Mek\_Raf} + C_9)} \tag{30}$$

$$\frac{dC_{10}}{dt} = C_8 * k_{10} * \frac{C_9}{(k_{M\_Mek\_Raf} + C_9)} - PP2A * k_9 * \frac{C_9}{(k_{M\_Mek\_Raf} + C_9)} - pp2A * k_{Erkpp2A} * \frac{C_{10}}{(k_{M\_Erk\_PP2A} + C_{10})} \tag{31}$$

$$\frac{dC_{11}}{dt} = C_{10} * k_{11} * \frac{C_{12}}{(k_{M\_Erk\_Mek} + C_{12})} - pp2A * k_{Erkpp2A} * \frac{C_{11}}{(k_{M\_Erk\_pp2A} + C_{11})} \tag{32}$$

$$\frac{dC_{12}}{dt} = pp2A * k_{12} * \frac{C_{11}}{(k_{M\_Erk\_pp2A} + C_{11})} - C_{10} * k_{Erk\_Mek} * \frac{C_{12}}{(k_{M\_Erk\_Mek} + C_{12})} \tag{33}$$

$$\frac{dC_{13}}{dt} = C_{11} * k_{13} * \frac{C_{14}}{(k_{M\_p90Rsk\_Erk} + C_{14})} - k_{p90RSK} * C_{13} \tag{34}$$

$$\frac{dC_{14}}{dt} = k_{14} * C_{13} - C_{11} * k_{p90RSK\_ERK} * \frac{C_{14}}{(k_{M\_p90PSK\_ERK} + C_{14})} \tag{35}$$

$$\frac{dC_{15}}{dt} = C_2 * k_{15} * \frac{C_{16}}{(k_{PIK3\_IGF1R} + C_{16})} + * k_{PIK3\_EGFR} * C_1 * \frac{C_{16}}{(k_{M\_PIK3\_EGFR} + C_{16})} + C_5 * k_{PIK3\_Ras} * \frac{C_{16}}{(k_{M\_PIK3\_Ras} + C_{16})}$$
$$- k_{f\_PI3\_kactive} * C_{15} \tag{36}$$

$$\frac{dC_{16}}{dt} = -C_1 * k_{15} * \frac{C_{16}}{(k_{M\_PIK3\_IGF1R} + C_{16})} - k_{PIK3\_EGFR} * C_1 * \frac{C_{16}}{(k_{M\_PIK3\_EGFR} + C_{16})} - C_5 * k_{PIK3\_Ras} * \frac{C_{16}}{(k_{M\_PIK3\_Ras} + C_{16})}$$
$$+ k_{PI3\_kactive} * C_{15} \tag{37}$$

$$\frac{dC_{17}}{dt} = C_{15} * k_{16} * \frac{C_{18}}{(k_{M\_AKT\_PIK3} + C_{18})} - k_{Akt} * C_{17} \tag{38}$$

$$\frac{dC_{18}}{dt} = -C_{15} * k_{16} * \frac{C_{18}}{(k_{M\_AKT\_PIK3} + C_{18})} + k_{Akt} * C_{17} \tag{39}$$

The initial concentrations of all the species are taken from previous experiments [344] given in Table 5.2 .

Table 5.2 Initial conditions of the state variables

Species	**Initial value**
C1	IGFR_active_0=600.0
C2	EGFR_active_0=8000.0
C3	SOS_0=0.0
C4	DSOS_0=120000.0
C5	Ras_active_0=0.0
C6	Akt_0=600000.0
C7	Ras_0=120000.0
C8	Raf_0=120000.0
C9	Raf_active_0=0.0
C10	Mek_0=600000.0
C11	Mek_active_0=0.0
C12	Erk_active_0=0.0
C13	Erk_0=600000.0
C14	P90Rsk_active_0=0.0
C15	P90Rsk_0=120000.0
C16	PIK3_active_0=0.0
C17	PIK3_0=120000.0
C18	Akt_active_0=0.0

The chemical kinetics and the rate constants are taken from earlier studies and some are estimated as given in Table 5.3. The constant enzyme values RafPP_0 = 120000; PP2A_0 = 120000; RasGapActive_0 = 120000 are taken from previous research [344].

Table 5.3 Rate constants

Constants	**Reference**
$K_{1}=0.02$;	[344]
$K_{2}=0.02$;	(Estimated)
$K_{3}=694.731$;	[345]
$K_{4}=500$;	(Estimated)
$K_{E}=6086070$;	[345]
$K_{I}=100000$;	(Estimated)
$K_{5}=161197$;	[345]
$K_{6}=1509.36$;	[345]
$K_{7}=0.126329$;	[345]
$K_{8}=0.884096$;	[345]
$K_{9}=2.83243$;	[345]
$K_{10}=185.759$;	[345]
$K_{11}=9.85367$;	[345]
$K_{12}=8.8912$;	[345]
$K_{13}=0.0213697$;	[345]
$K_{14}=0.005$;	[344]
$K_{15}=10.6737$;	[345]
$K_{16}=0.0566279$;	[345]
$k_{f\_PI3K\_active}=0.005$;	[345]
$k_{M\_p90Rsk\_Erk}=763523$;	[345]
$k_{Ras\_SOS}=32.344$;	[345]
$k_{M\_Ras\_SOS}=35954.3$;	[345]
$k_{M\_ERK\_MEK}=1007340$;	[345]
$k_{M\_DSOS\_p90Rsk}=896896$;	[345]
$k_{M\_PIK3\_IGF1R}=184912$;	[345]
$k_{PIK3\_EGFR}=10.6737$;	(Estimated)
$k_{M\_PIK3\_EGFR}=184912$;	(Estimated)
$k_{M\_Akt\_PIK3}=653951$;	[345]
$k_{Akt}=0.005$;	[344]
$k_{M\_Erk\_PP2A}=3496490$;	[345]
$k_{PIK3\_Ras}=0.0771067$;	[345]
$k_{M\_PIK3\_Ras}=272056$;	[345]
$k_{M\_Raf\_Ras}=62464.6$;	[345]
$k_{M\_Mek\_Raf}=4768350$;	[345]
$k_{Raf\_Act}=15.1212$;	[345]
$k_{M\_Raf\_Act}=119355$;	[345]
$k_{M\_Ras\_RasGab}=1432410$;	[345]
$k_{M\_MEK\_PP2A}=518753$;	[345]
$k_{M\_Raf\_RafPP}=1061.71$;	[345]

5.4 RESULTS

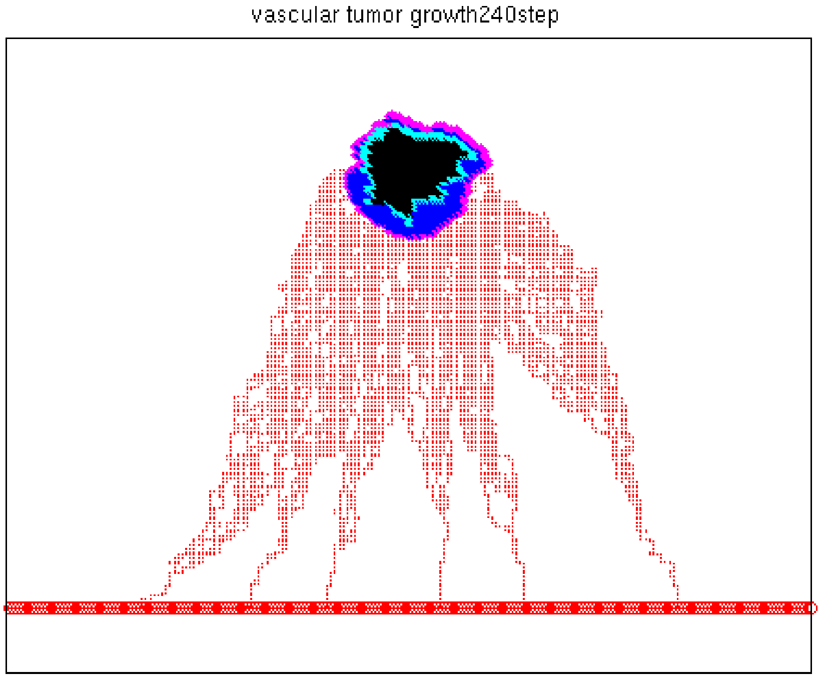


Figure 5.2(a)

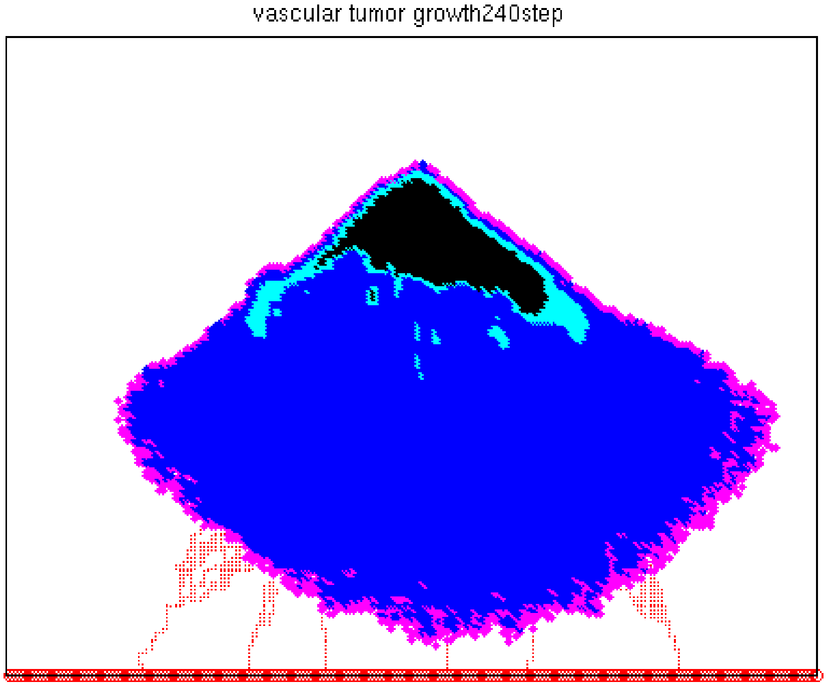


Figure 5.2(b)

Figure 5.2(a) Vascular tumour growth under drug treatment up to 240 hours of simulation. Various cells are colour coded: necrotic cells (black), active cells (blue), quiescent cells (cyan) and proliferative cells (pink); **Figure 5.2(b)** Under no treatment

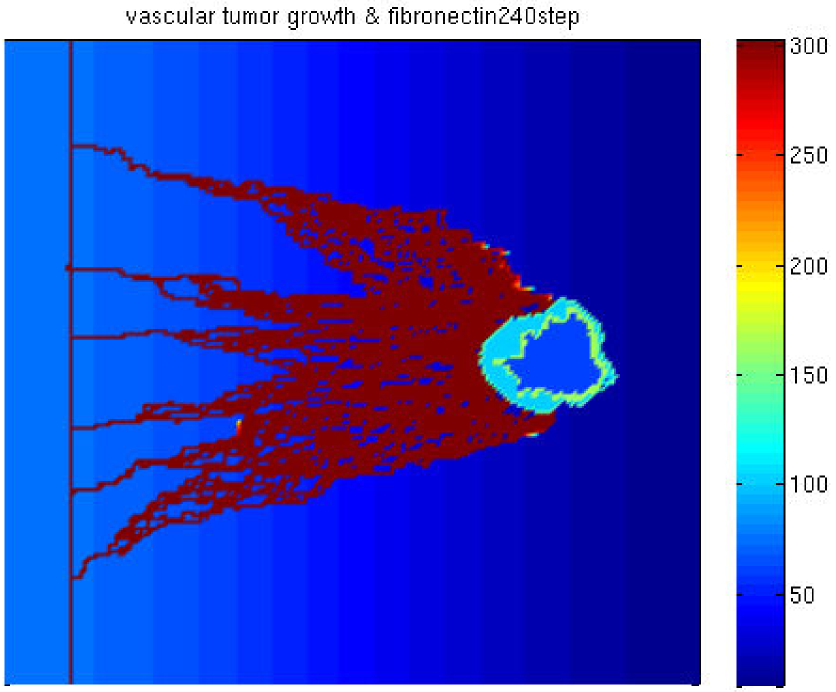


Figure 5.3(a)

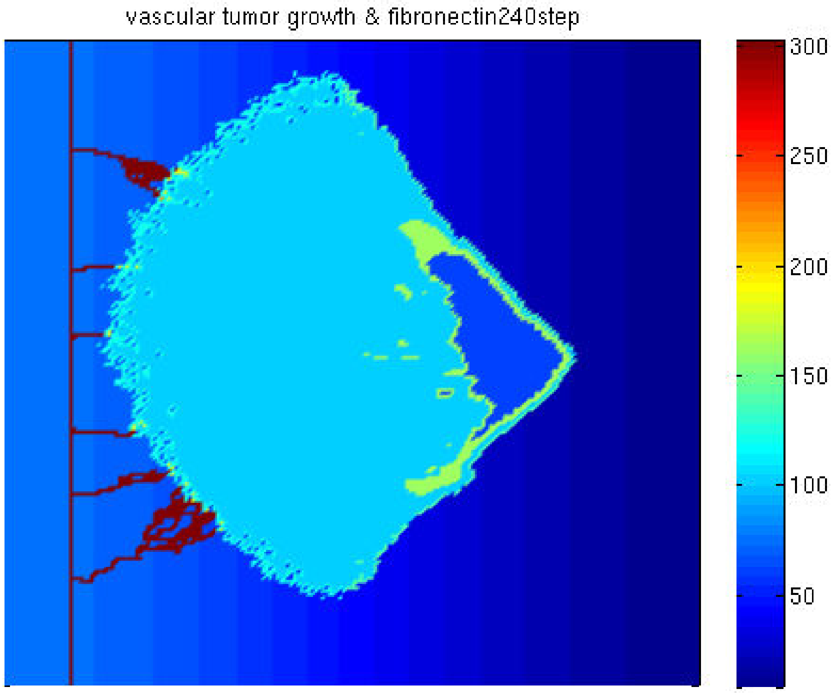


Figure 5.3(b)

Figure 5.3(a) The distribution of fibronectin under treatment in various cell regions of vascular tumour growth as colour coded up to 240 hours of simulation; **Figure 5.3(b)** No treatment

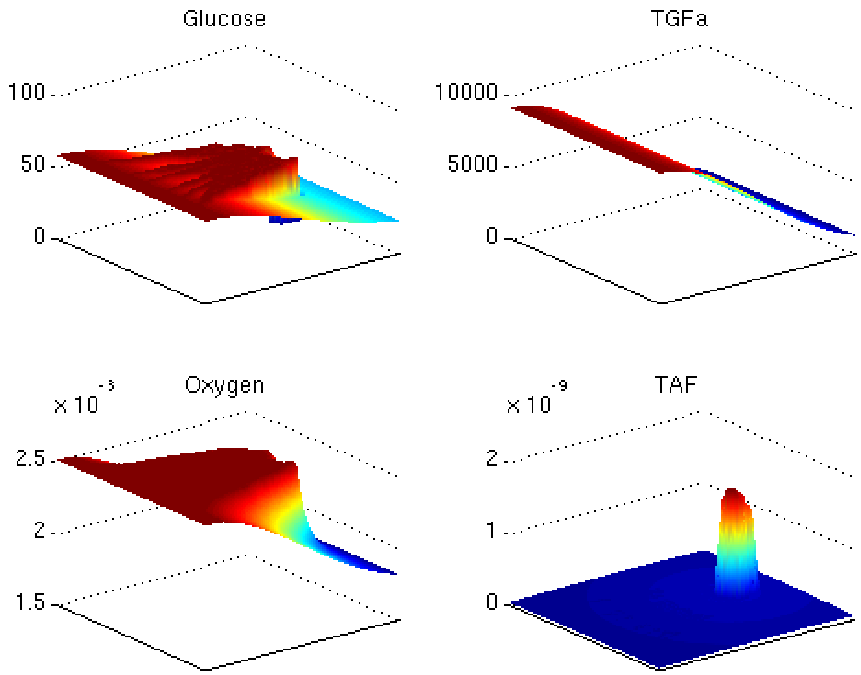


Figure 5.4(a)

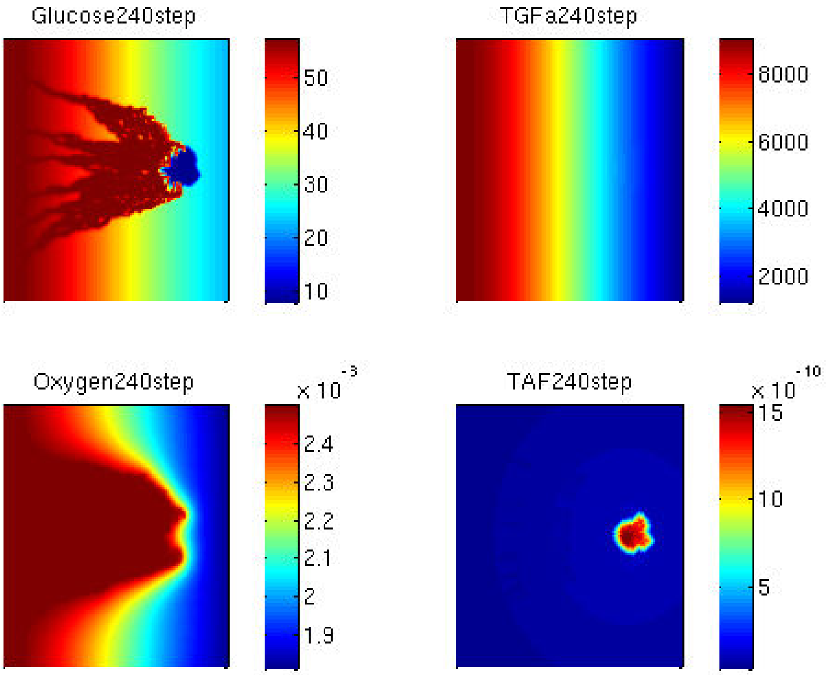


Figure 5.4(b)

Figure 5.4(a) The 3D plots of glucose, oxygen, TGF-alpha and TAF concentrations under drug treatment up to 240 hours of simulation; **Figure 5.4(b)** Concentration profiles of glucose, oxygen, TNFalpha and TAF among various cell regions are colour coded up to 240 hours of simulation;

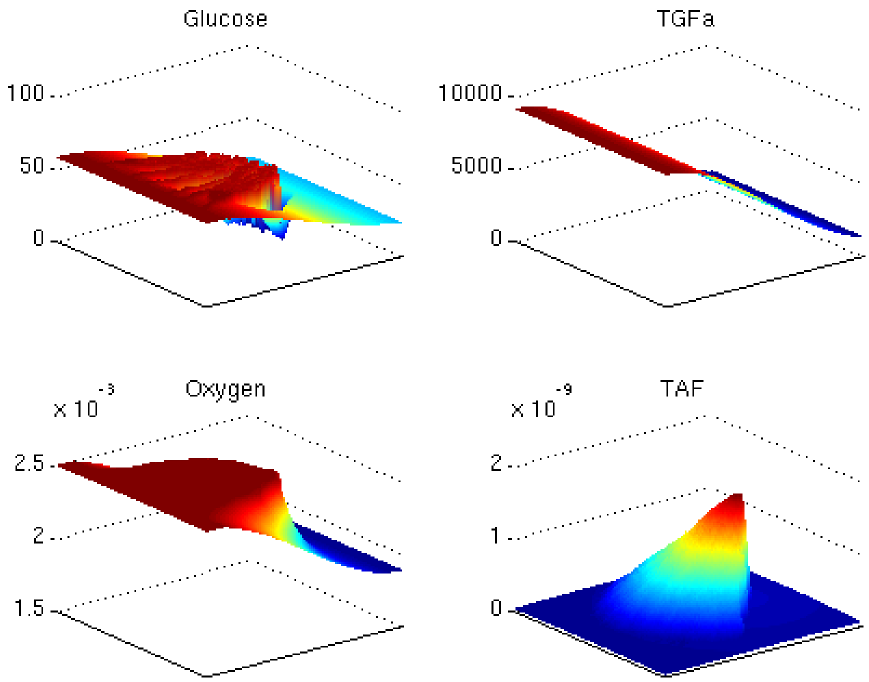


Figure 5.4(c)

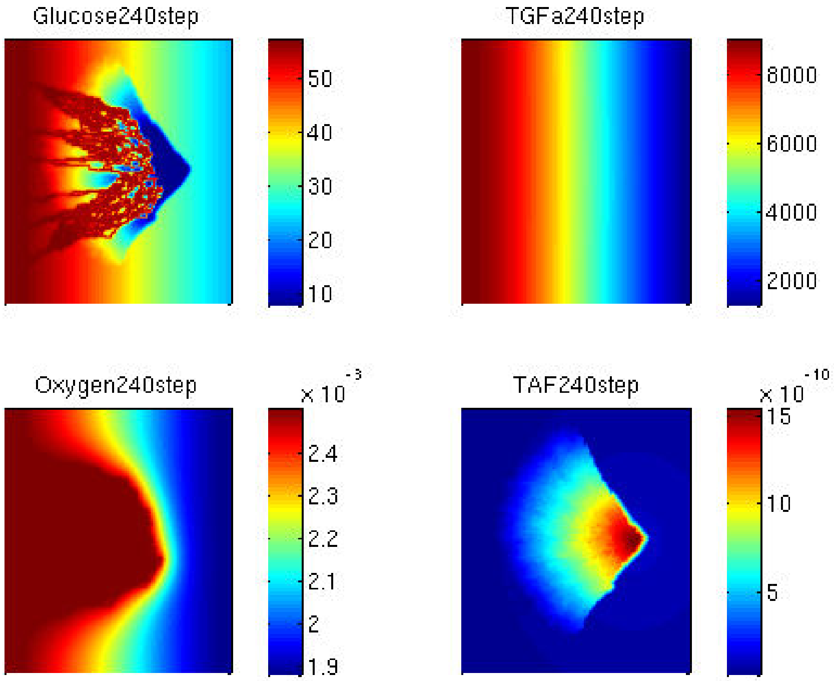


Figure 5.4(d)

Figure 5.4(c) The 3D plots of glucose, oxygen, TGF-alpha and TAF concentrations under no treatment up to 240 hours of simulation; **Figure 5.4(d)** Concentration profiles of glucose, oxygen, TNF-alpha and TAF under no treatment among various cell regions are colour coded up to 240 hours of simulation

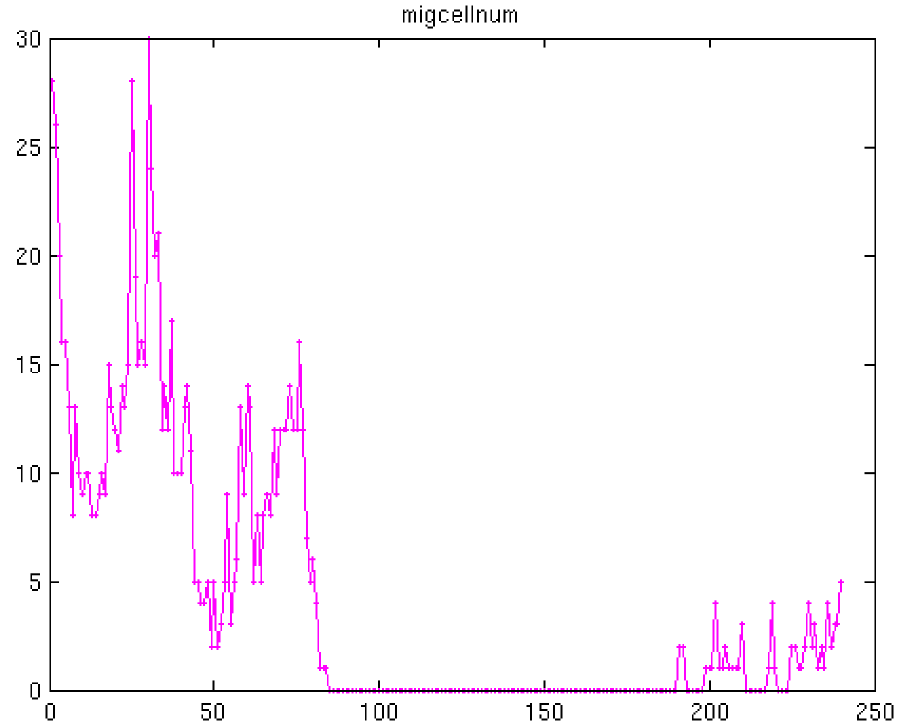


Figure 5.5(a)

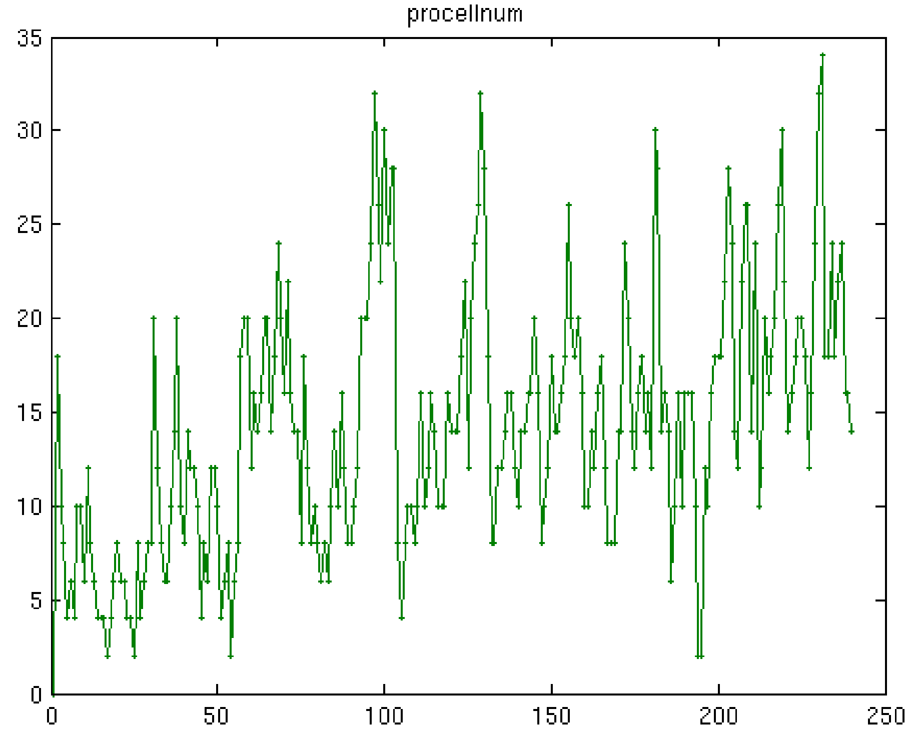


Figure 5.5(b)

Figure 5.5(a) Profile of migrating cells vs. time; **Figure 5.5(b)** proliferative cells vs time under drug treatment up to 240 hours of simulation

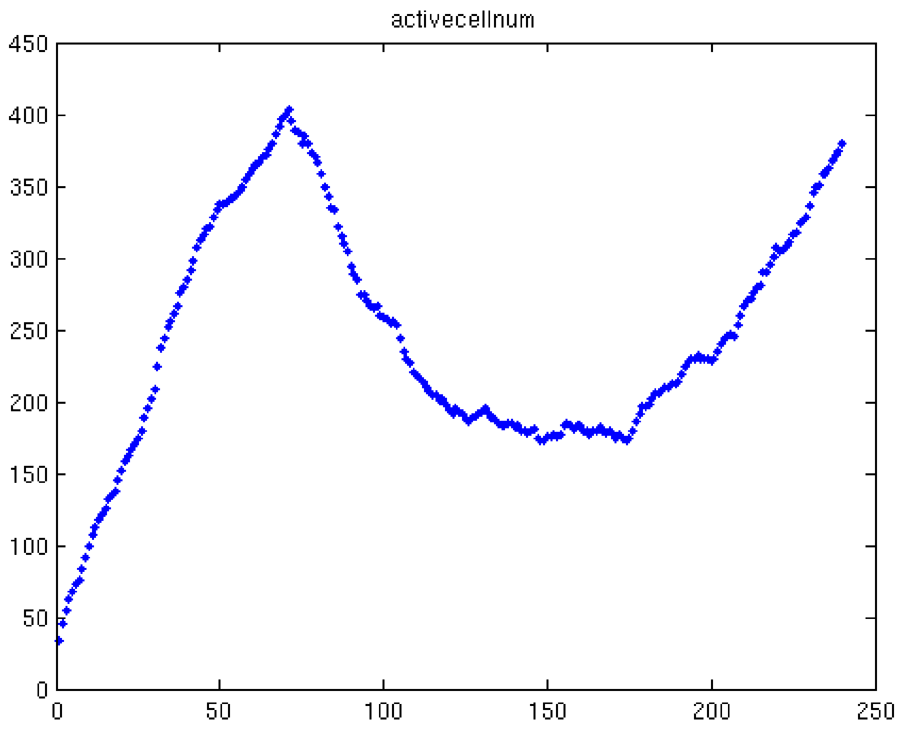


Figure 5.6(a)

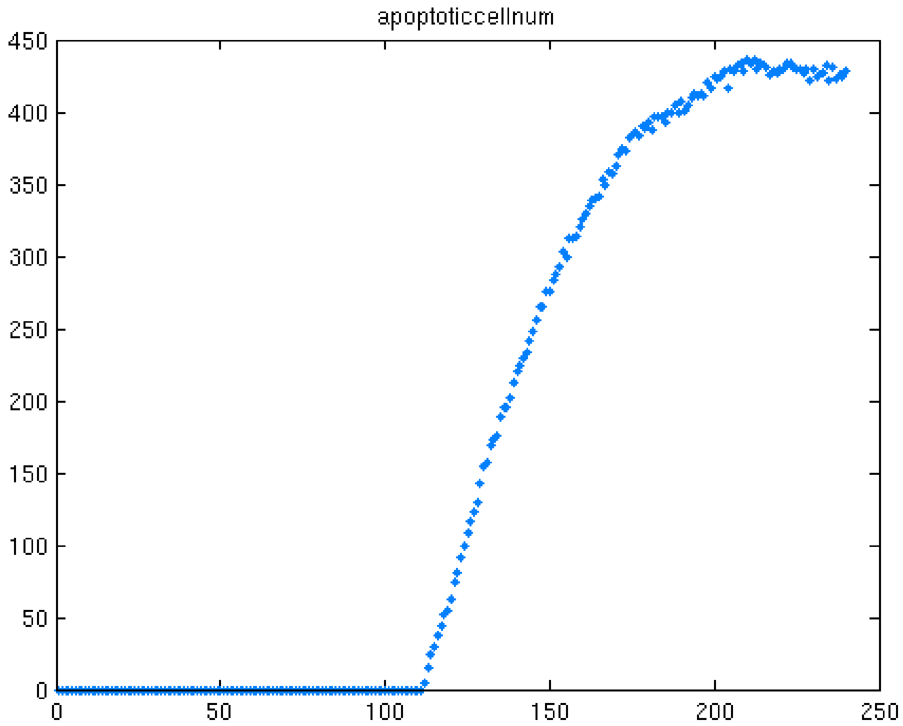


Figure 5.6(b)

Figure 5.6(a) Profiles of active cells vs. time; **Figure 5.6(b)** apoptotic cells vs. time under drug treatment up to 240 hours of simulation

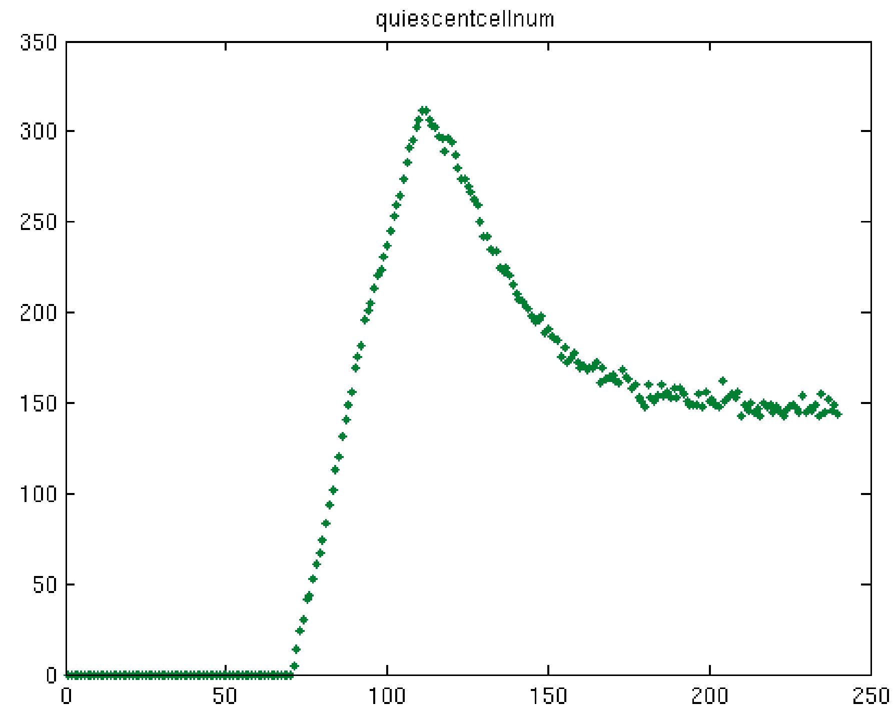


Figure 5.7(a)

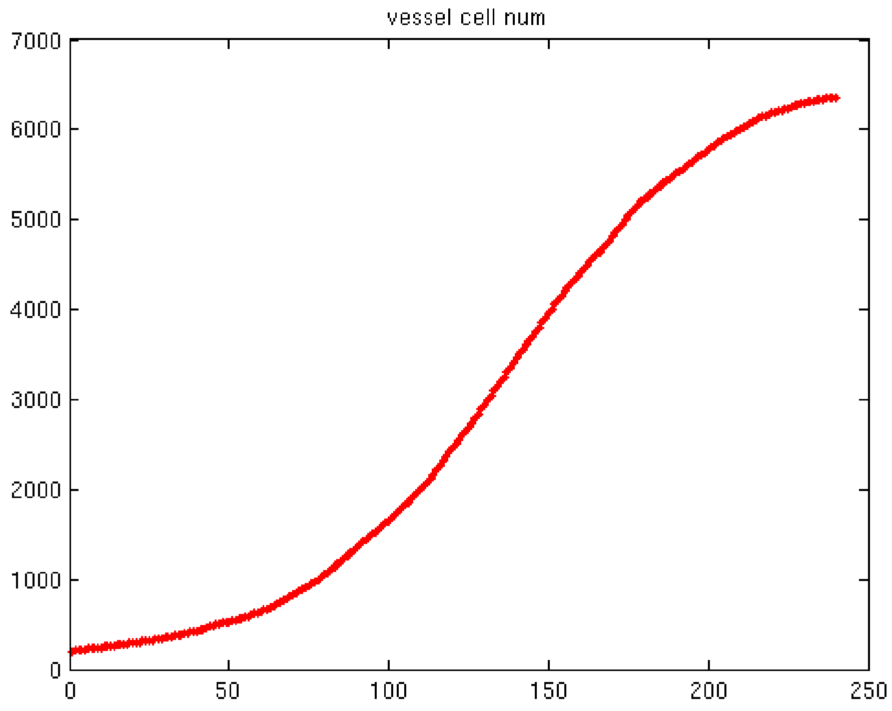


Figure 5.7(b)

Figure 5.7(a) Profiles of quiescent cells vs. time; **Figure 5.7(b)** vessel cells vs. time under drug treatment up to 240 hours of simulation

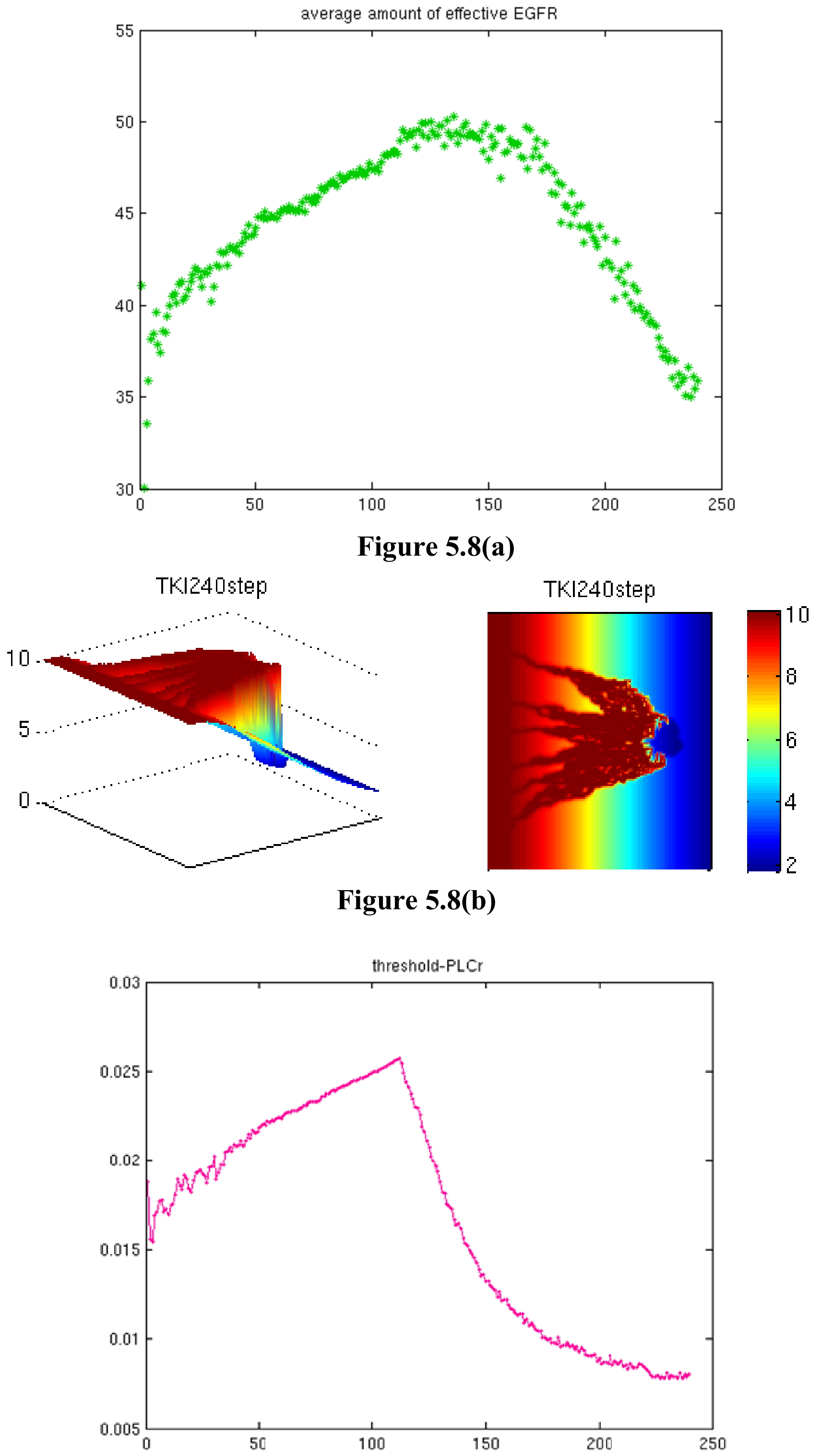


Figure 5.8(a)

Figure 5.8(b)

Figure 5.8(c)

Figure 5.8(a) Effective amount of EGFR vs. time; **Figure 5.8(b)** Drug (TKI) distribution among various cell regions (colour coded); **Figure 5.8(c)** Threshold of PLCγ vs. time up to 240 hours of simulation

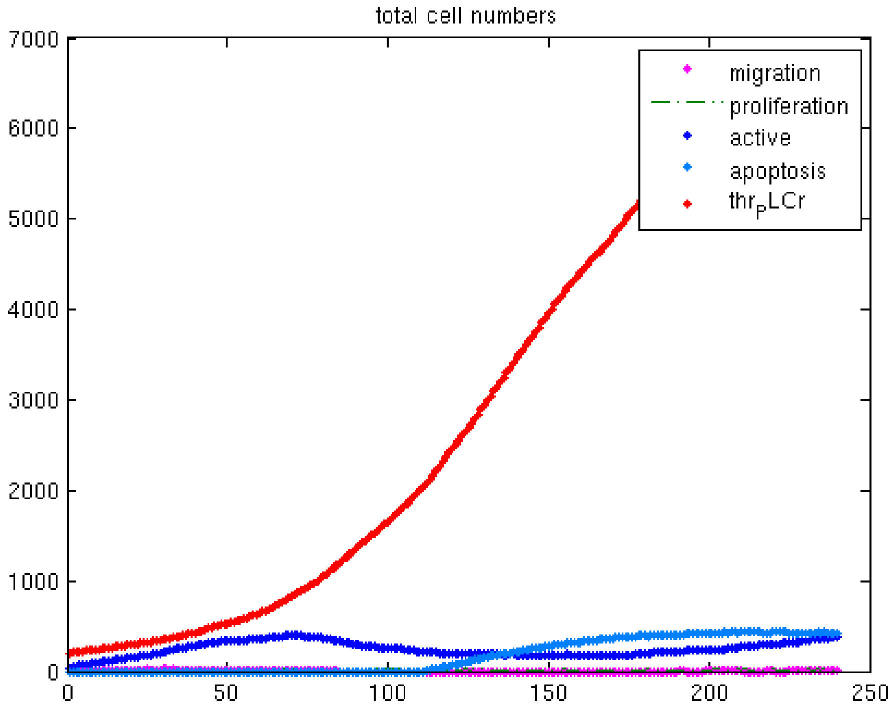


Figure 5.9(a)

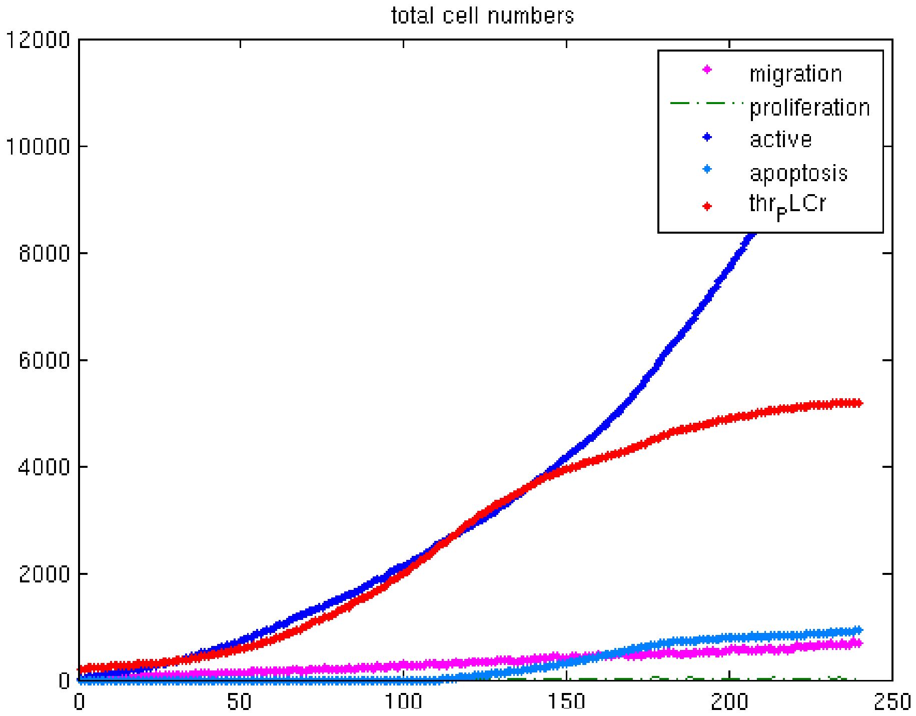


Figure 5.9(b)

Figure 5.9(a) Prolife of various cell types with drug (TKI); **Figure 5.9(b)** Profiles of various cell types without drug up to 240 hours simulation

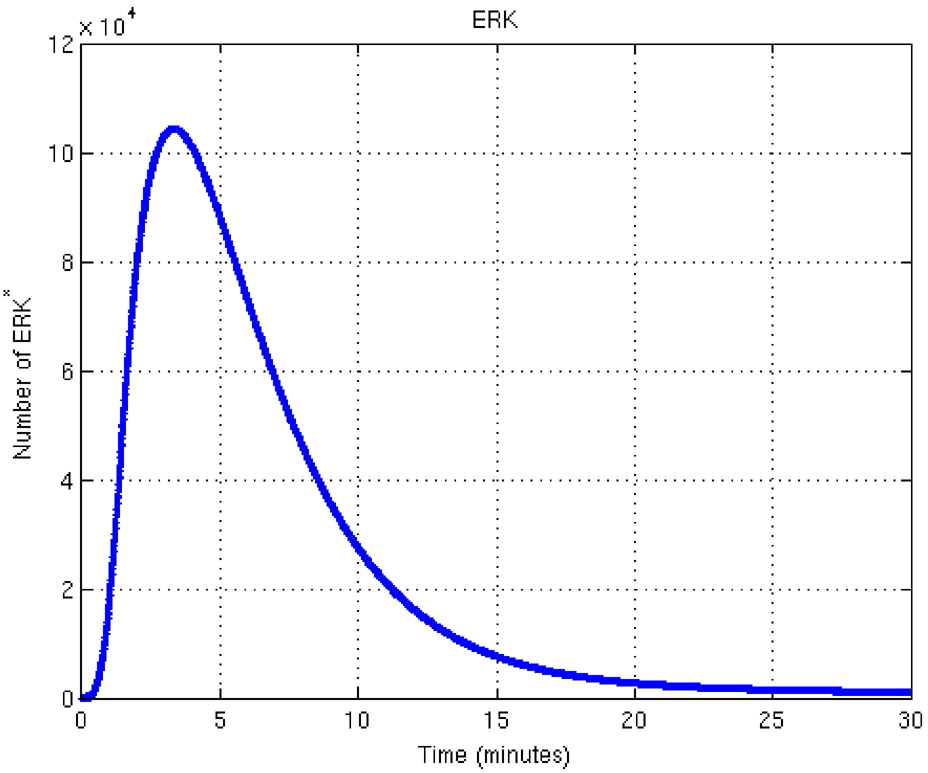


Figure 5.10 Number of active ERK receptors vs. time

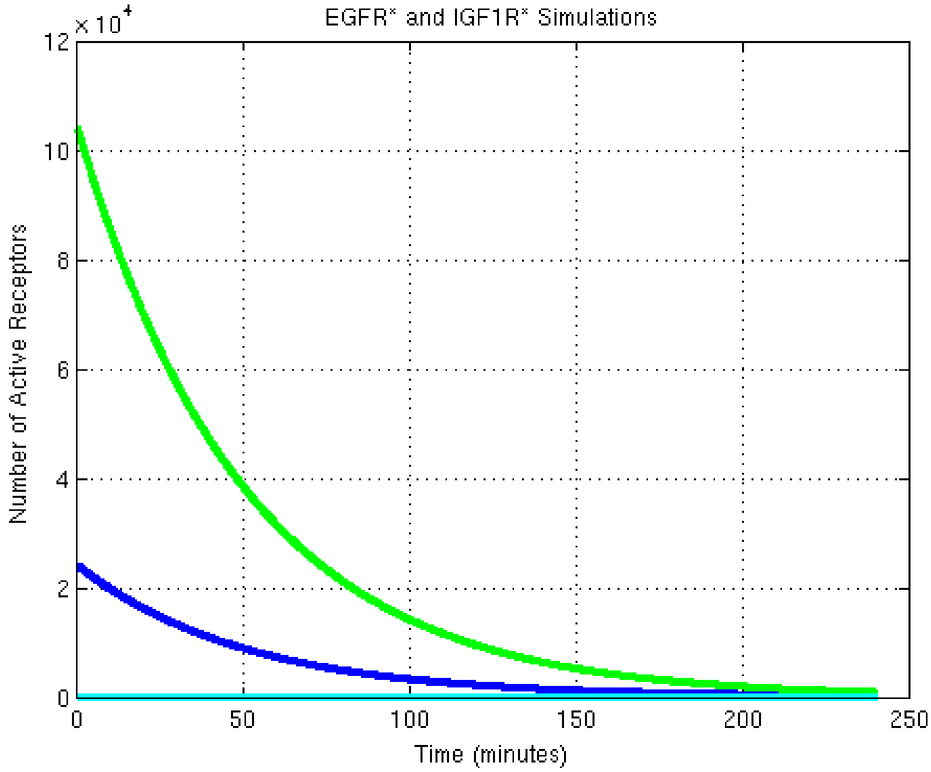


Figure 5.11 Number of active EGFR and IGF1R receptors vs. time

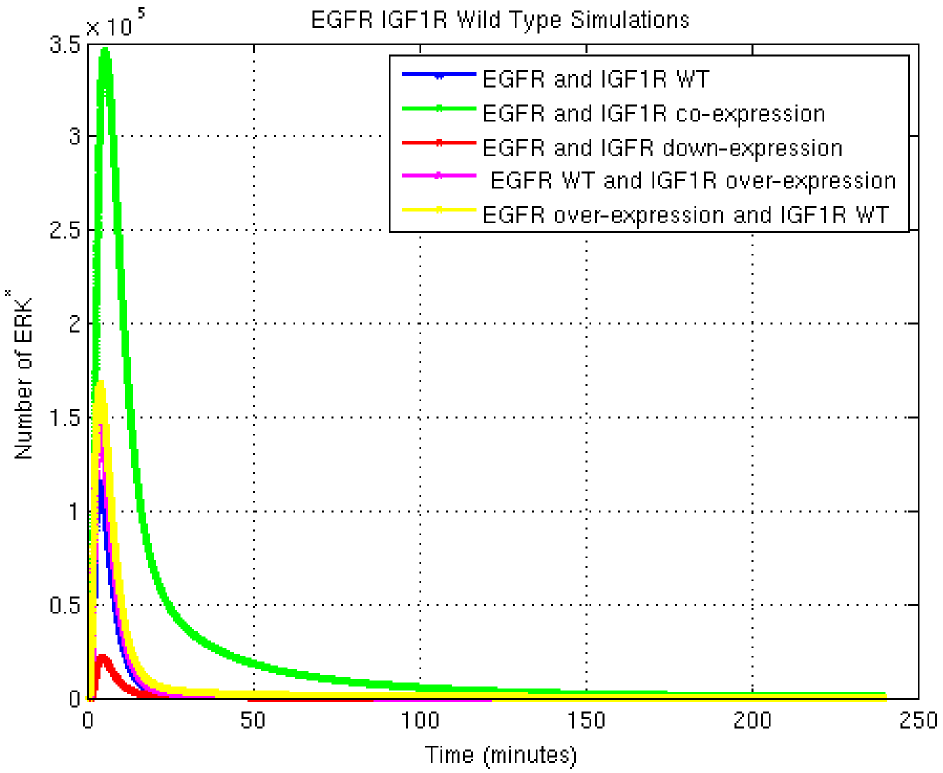


Figure 5.12 Number of ERK receptors activated by EGFR and IGF1R simulations in wild type

The simulation of in vitro tumour growth along with cell distributions in various zones have been done at each given time point. We initially took the nutrient concentrations such as oxygen, glucose, tumour angiogenesis factors, drug and vascular endothelial growth factors to be homogeneous a0, b0, c0, d0 and e0, respectively. The boundary conditions were set for each nutrient, drug and growth factors for numerical computation purposes were made. Equations were normalized and Neumann boundary conditions were applied with zero flux around the boundary and solved with PDESolver in Matlab2010a. At initial time point t0, the number of cells in two-dimensional and three-dimensional grids was placed accordingly. Tumour growth was modelled with respect to time utilizing the biophysical quantities such as oxygen, glucose, TAF, drug, and the level of VEGF were the calculated data point. Drugs were introduced to the blood stream and the relative level was measured at each grid point chronologically. Tumour growth was simulated for 240 hours, which gave enough time for a transition from an avascular to a vascular growth. Certain assumptions of model were made a) angiogenesis is induced by tumour cells that produce new vessels from previous vessels quiescent state; b) cell cycle of 24 hours; c) cells reversibly enter quiescence when cell nutrient intake is less than a particular threshold value; d) cells irreversibly become necrotic when cell nutrient intake will be zero; and e) tumour pressure [337] and capillary pressure [336].

Figure 5.2(a) shows tumour volume and morphology changes during the avascular and vascular tumour growth under treatment and without treatment Figure 5.2(b). Once it transforms to a vascular stage, the tumour was characterized by an expansion of different cell types that are colour coded: active cells (blue), the necrotic core (black), quiescent cells (cyan) and proliferative cells (pink). The distribution of the fibronectin under drug treatment in different cell regions are colour coded as active cell zone (blue), quiescent zone (demarcated with yellow) and proliferative zone (cyan) as shown in Figure 5.3(a) and without treatment Figure 5.3(b). During the transformation from the avascular state to vascular state in tumour growth process, the cells quickly uptake oxygen and release a huge amount of waste during protein or DNA synthesis or cell proliferation. Figure 5.4(a) shows the 3D plot of the distribution of metabolic state profiles under drug treatment such as oxygen, glucose, TAF and TGF-alpha (analogue of EGFR); Figure 5.4(b) represents concentrations of various zones under drug treatment that are colour coded, active cells (blue), quiescent (yellow), proliferative (cyan). Figure 5.4(c) shows the 3D plot of the distribution of metabolic state profiles under no drug treatment such as oxygen, glucose, TAF and TGF-alpha; Figure 5.4(d) represents concentrations of various zones under no drug treatment, which are colour coded as active cells (blue), quiescent (yellow), proliferative (cyan). Below the threshold value, the oxygen level drops, the endothelial cells (EC) become quiescent and EC move to high VEGF, low drug concentration. This condition impairs the tumour and the cell cycle prolongs, resulting in the transition from avascular to vascular and causes cell death at the core tumour region. The number of tumour cells does not change; yet, the percentage of dead cells to inactive cells augments over time. The angiogenesis process is initiated and the supply the nutrients for further renewal of cell proliferation to support the growth of tumour. The simulations show that the vessels appeared to be confided to the outer edges of the tumour, resulting in more oxygen near the edge of the tumour and less or no oxygen at the tumour core, glucose accretion occurred throughout the tumour mass. However, the most concentrated level of oxygen was found just inside the outer edge of tumour, possibly from the proximity of active cells and the inability to adequately eliminate waste. Tumour angiogenesis factors tend to focus more within the tumour's core, especially in low oxygen and glucose content regions were categorized by high waste content. Active cells were found near the edges of tumours were there are higher levels of oxygen and sufficient waste elimination. This was seen in a large number of tumours and not within dead regions of the tumour as shown in Figure 5.4(b). Tumour actuated angiogenesis was studied for the local interstitial pressure using a sophisticated pattern based model for endothelial cells which induced a proliferation of TAF gradients. Thus, the model calculates tumour interstitial pressure and the blood vessel circulation's intricate relationship gauges the growth factors accessibility and spatio-temporal deviation in nutrients.

Figure 5.5(a) shows the profile of the distribution of migrating cells with time variation. One can observe that the number of migrating cells gradually decreases and becomes almost zero after 85 hours until 190 hours and proliferative cells shows the oscillatory behaviour as shown in Figure 5.5(b), first increase after 85 hours of simulation and then falls drastically and further increases. This is due to the action of drug and anti-angiogenesis or VEGF inhibitor treatment. Figure 5.6(a) represents the profiles of the active cells showing increase up to 85 hours and gradually decreases until 150 hours and becomes stable and again raises after 180 hours whereas the apoptotic cells are almost zero up to 105 hours and later gradually increases up to 200 hours of simulation and becomes stable as shown in Figure 5.6(b). The reason is that once the drug and anti-angiogenesis inhibits the number of active cells and thereafter falls gradually and the apoptotic cells start to increase when there are no metabolites supplied to the cells. The quiescent cell population is zero until 70 hours and gradually increases until 105 hours and then falls down as shown in Figure 5.7(a). This is because of a reduction of cell metabolites after 105 hours of treatment. The plot between vessel numbers with time variation shows the behaviour of a sigmoid curve that initially increases and gets saturation after certain period as shown in Figure 5.7(b). Thus, the model indicates that endothelial cells move to high VEGF gradient. In contrast to low-pressure tumours owing to a robust outwards interstitial fluid convection, high-pressure tumours with lower drug concentrations and low crowded vessels are prone to mature as a dendritic cell formation. Furthermore, there will be a reduction in oxygen influx as well as residuals, favouring necrosis or else cell quiescence in the core tumour region. As a result, one could anticipate a more rapid proliferation of vascularized cells at the edge of the tumour following a dendritic structure within. In this study, the tumour morphology with respect to tyrosine kinase inhibitor in

EGFR/ERK, we tested a tumour model both with and without the drug. The simulation results in Figure 5.8(a) shows that the effective average amount of EGFR increases until 120 hours and then gradually falls, which is an inverted parabola curve. Figure 5.8(b) indicates the drug profile (TKI, EGFR inhibitor) distribution among various zones of tumour regions that are colour coded up to 240 hours of simulation. Figure 8(c) shows the profile of the threshold value of PLCγ, showing TKI molecules binding by EGFR lower the effective EGFR and accounts for the low PLCγ value. PLCγ value initially increases up to 105 hours and then gradually decreases, which indicates a fewer number of cell migration. We performed our simulations both under the conditions of drug concentration and with no drug. Profiles of different cell types such as proliferative cells, active cells, migration cells, quiescent cells, apoptotic cells and endothelial cells with drug treatment is shown in Figure 5.9(a) and a profile of all the cell types without drug treatment is shown in Figure 5.9(b). We observed that the proliferation rate of active cells become reduced, made a more solid tumour morphology and produced significant cell apoptosis with TKI induced EGFR inhibitor treatment. When no treatment is given, the results showed diminutive tumour growth control yield and a subsequent increase of cells in the necrotic region as well as active region.

We framed the ODEs for the downstream signalling of EGFR signalling pathway. We solved the system of ODEs using the Runge-Kutta method (ODE15solver) in MATLAB2010a. Figure 5.10 shows the profile of active ERK receptors activated by the TKI inhibitor in the EGFR signalling pathway. We ran the simulation for 30 minutes and found that the average number of active ERK receptors initially increases and then falls after 3 minutes, showing the behaviour of an inverted parabola as shown in Figure 5.10. Since there are studies of non-small cell lung cancer and brain tumours like glioma and glioblastoma multiforme showing only TKI inhibitor in EGFR signalling that gives resistance to the tumour cells, we implemented the inhibition of EGFR in conjunction with IGF1R signalling pathways. The simulations of EGFR and IGF1R up to 240 minutes in Figure 5.11 shows the plots of the wild type expression (blue), down-expression (cyan) and over-expression (green) with the number of active EGFR and IGF1R receptors with respect to time. We performed the wild type simulations of EGFR and IGF1R receptors and found the number of active ERK receptors are higher for co-expression of EGFR along with IGF1R compared to the other conditions such as EGFR and IGF1R wild type expression, EGFR and IGF1R down-expression, EGFR wild type and over-expression of

IGF1R, over-expression of EGFR and IGF1R wild type as shown in Figure 5.12. The simulation results are consistent with an earlier noteworthy immunohistochemistry study showing a higher co-expression of IGF1R plus EGFR acts as a prognostic basis representing NSCLC initial stage cases.

In order to investigate the tumour growth model for robustness, we performed a sensitivity analysis to assess if any specific parameters have affected the outcome. For this, several factors such as diffusion rate, supply and consumption rate of nutrients (oxygen, glucose), TAF, drug and VEGF values are altered by ±1%, ±5%, and ±10% in the parameters table used in the model. We measured the number of active cells, examined the change of system response to the base level activity, and found that the kinetic rates of oxygen, glucose, TAF, VEGF, drug and vascular pressure are sensitive variables.

5.5 DISCUSSION

Understanding how an entire systems work together is often done by studying biological systems at the micro to macro level by using multi-scale modelling, which can incorporate a wide range of time spans, from nanoseconds to years and from a single gene to an entire genome. This model produces an integrative understanding of different aspects of biological systems. The proposed multi-scale tumour growth model integrated with angiogenesis process uses a system of partial differential equations based on reaction–diffusion kinetics, which illustrates the development of nutrient delivery such as glucose, oxygen, drugs, tumour angiogenesis factors and vascular endothelial growth factors. These factors control and influence the growth of new vessels from existing vessels in the case of tumours. Laboratory models have shown that cells migrate towards blood vessels and develop over time. Tumour cells are attracted to the higher nutrient levels near blood cells, which may account for this movement. At the same time, blood vessels also move in the direction of the tumour's core and form a complex network of vessels. This may be due to the high VEGF-gradient that is in proximity to the tumour, which draws endothelial cells thereby intensifying vessel branching [346]. The activity of tumour cells indicates the toxicity of a given treatment. Higher levels of drugs can also bring about tumour cell transitions (active, inactive or dead). The body level pharmacokinetics models were used to determine the intravascular drug concentration as proposed [347, 348]. To simplify the model, drug levels remained the same or on par with constant intravenous drips. Nevertheless, the plasma drug levels verses time is easily simulated with population kinetics modelling and the drug dose was affected via extracellular area, vessel radius and the force disparity in blood capillaries [349]. The dose of drug age as a result was more likely to disseminate in a near to low force territory and not capable of diffusing within the interior tumour-compartment. The presence of a drug in a tumour's interstitial region can be affected by the varying pressure levels found in capillaries and outside of cells so that drugs find the path of least resistance by diffusing into low-pressure areas away from the centre of the tumour cell. The normal growth rate of a vessel was assumed to be 0.6mm/day in normal tissue [350], which will be increased for vessels surrounding the tumour region. Despite earlier studies that imitated angiogenesis by tip cell division [334, 351, 352], we considered the influence of the nutrient availability to the cells for vessel maturation. Interstice-tumour-pressure was found to be much higher than the pressure of the surrounding tissue and the core of the tumour had the highest level of pressure [340]. Few studies [337, 353] have suggested that the high tumour pressure is due to the compact nature of tumour cells rapid mitosis, which pushes normal tissue trapping interstice fluids and increases pressure at the core of the tumour.

The multi-scale modelling of the tumour at different levels (cellular, molecular and tissue) coupled with the angiogenesis process is evaluated with the TKI inhibition of EGFR signalling pathway and VEGF inhibitor. At the tissue level of the model, the tumours inter cell space-pressure accounts for pressure from tumour mitosis and vascular perfusion. It can also determine the vascular and tumour pressure during tumour proliferation. The proposed model is not restricted to avascular tumours and can be applicable to all kinds of tumours. We built a system of partial differential equations to model the metabolites and growth factors such as oxygen, glucose, drug, TAF distributions and VEGF inhibitors at intratumoral levels by incorporating tumour pressure-induced interstitial fluid convection. Anti-angiogenic inhibitors target pre-angiogenic vessels and may cause apoptosis of the endothelial cells.

The regulation of the EGFR family receptor tyrosine kinases depends on precise interactions between interconnected parts and systems. The alteration of a single protein or sequence can severely disrupt its regulation. Yet, mutations are the entry point for therapies of targeted inhibition of dysregulated EGFR signalling. In fact, the transfer of DNA during the normal regulation of EGFR family signalling is one the most studied of the receptor tyrosine kinase family, especially at the atomic-level with new studies continually providing new insights and novel directions. A recent study discovered the asymmetric homodimer assembly is critical for kinase activation [354]. Other studies examined the systems that trigger mutations and anti-mutation structures specifically in the area of kinase, have given deep insights regarding the mechanisms of EGFR family activation and resistance to small molecule inhibitors. Our results indicate that treatment by EGFR inhibitors slows the advancement of the tumour. The trussing of TKI molecules towards EGFR lowers the EGFR in force, which effects the low threshold value of PLCγ and diminution in the number of cell migration. Consequently, it reduces tumour invasion, which is consistent with other simulation studies [341]. However, due to the twin facets of angiogenesis, the survival rate of the tumour cell does not always diminish. Lately created vessels deliver a considerable quantity of EGFR inhibition molecules to obstruct the EGFR signalling passageway guiding tumour growth inhibition in the initial phase resulting in diminished cell endurance. In tandem, the new-found vessels transport a large amount of oxygen and glucose to tumour growth cells that boosted cell endurance at advanced phases. The effects of the dual roles of angiogenesis disclose that cancer progression can be arrested or decreased by using EGFR inhibitor treatment and insulin-like growth factor 1 (IGF1) receptors at the same time. Our simulations demonstrate that our proposed model shows good agreement with experimental data on the knock out mouse model and great potential for translating it to clinical work.

The computational simulations performed in this work demonstrate the potential of integrating a drug evaluation coupled with angiogenesis in the proposed model. Multiple cell types of tumour progression, including active cells, proliferative cells, quiescent cells and apoptotic cells along with drug and growth factors were computer-generated. Moreover, we integrated several tissue specific biophysical parameters such as diffusion and secretion rates of metabolites, drug, growth factors, tissue density, vascular pressure and intratumoral, etc., depicting a three dimensional tumour growth blueprint. Due to the huge potential of the modular approach, the model can be extended in nature and incorporate additional functions or properties related to drugs, tissue perfusion and comprehensive signalling pathways. The model can be tailored for patient specific therapeutics with some other additional inputs and standardization. The model proposed in this study predicts and its potential applicability will be a valuable prognostic platform for drug discovery and therapeutic planning. Future work can further refine the model by incorporating several other factors for insights using in silico three-dimensional tumour growth models moving from 'bench side' inventions to a patients 'bedside'. However, simulations for high dimensional in silico approaches necessitate high throughput computing power besides tedious computational time and cost.

5.6 CONCLUSION

This study presents a new solid tumour growth multi-scale-modelling at the cellular, molecular and tissue levels with an angiogenesis module and TKI in EGFR signalling pathway integrated with anti-angiogenesis or IGF1R inhibitor treatment. A system of PDEs framework were used to alter the level of various nutrients supplied to the cells such as glucose, oxygen, drug, TAF and VEGF in the tumour microenvironment at the cellular level. VEGF angiogenesis inhibitors were shown by blood vessel formations and maturation owing to the tumour cells because of the infused nutrients forming new blood vessels at the tissue level. A system of ODEs was used to simulate the posterior mechanics of EGFR/ERK signalling integrated with IGF1R signalling pathway at the molecular level. Our simulation results give a demonstration of joint actions of active cells, migrating cells, proliferative cells and quiescent cells regulated by the active ERK receptors including EGFR and IGF1R receptors depicting sketch of tumour evolution. The dual effects of angiogenesis were also found on EGFR inhibition treatment to decrease the tumour invasion, as well as supply nutrients to tumour compartments to promote cell endurance at later time stages. Thus, our simulations suggest that EGFR and IGFR1 co-expression activates more ERK receptors when compared to down-expression or over-expression of EGFR and IGF1R. There is a good agreement between the simulations, an experimental mouse model and clinical data.

5.7 AUTHOR’S RESEARCH CONTRIBUTION

Sharma, A.S., Rallabandi, V.P.S., Gupta, H.O., Prasad, R. (2014) “Multiscale Modeling of Solid Tumor Growth and Treatment: Lung Cancer as Case Study” Submitted to Elsevier Mathematical Biosciences.

Chapter – 6

CONCLUSION AND FUTURE SCOPE OF WORK

6.1 CONCLUSIONS

Despite more biological data and faster computers, the primary driving force is intellectual activity that will spur new ideas, approaches, allow the further development of the field and its future establishment as the crown of biological sciences. This work tried to meet all four objectives in a single piece of work and the results indicated that we have succeeded in taking constructive steps toward each goal. As such, these works hope to advance the understanding of contemporary bioinformatics and cancer systems biology relevant to research in life sciences.

The design of an ontology that represents gene function is critical for addressing the challenge of integrating sequence data, which has seen a rapid increase in the amount of data from functional analyses of genes. The overall strategy presented in this thesis clarifies how an ontology-based gene function may be implemented using genomic databases. The vital importance of the tremendous amount of biological knowledge contained in genomic databases has been investigated by analyzing an ontology-based gene function that leads to a physiological model. The proposed model contributed to overcome the difficulties by considering gene ontology based models to realize physiology through the efficient management of biological data. Thus, a huge database warehouse is needed to integrate the relevant available data of genes and databases containing detailed information of physical traits. This will aid researches move seamlessly from genes to physical and physiological attributes to better understand their influence and effects. The development of such a gene ontology based solution will lead to genome wide associated studies with the capability of functional genomics as a whole. From this descriptive ontology based physiological model, further tailored prototypes can be developed that will ultimately lead to a full physiological picture of any organism under observation. The success of such a promising, complex, and full physiological tailored system that is functional, intelligible and reliable depend on the validated gene ontology databanks. Such systems biology solution requires the continuous involvement of a modeller and the best high performance computing automated methods that can use rapidly evolving gene ontology databanks.

After understanding the concept of a full physiology model, the same has been implemented using the power of contemporary bioinformatics. The Plants Physiological model (PPDB) was built after mining data from gene ontology (GO) as an abstraction of the plants gene ontology data with a common vocabulary, systemization of knowledge, standardization and easy to use functionality that acts as an educational resource for plant biologists and bioinformaticians. This classified and annotated plant genomic data with 149 BP terms, 36 CC terms, 21 MF terms, and 37 annotated terms were finally stored in the Plants Physiology Database. Hence, a plant physiology database was developed as a new investigating tool to give ready information access and the plant physiology (biological process, cellular component, molecular function - BPCCMF) as a whole can be investigated. This will save valuable time and effort for relevant researchers. Further, users can download the latest updated database and its database schema on plant physiology. The technical knowhow or manual is also provided for life scientists to build their own local PPDB mirror, which is missing in most of the investigating tools published to the best of the author's knowledge. The working principles and search options associated with the PPDB are publicly available and freely accessible on-line (http://www.iitr.ernet.in/ajayshiv/) through a user-friendly environment. It contains the latest full compendium of curated expression data entries for plants and is freely offered to researchers and companies worldwide via open access.

Through this thesis, we hope to enlighten the reader on the computational systems biology approaches applied to cancer research. These approaches offer promising insights to defeat cancer. The heuristics models are well studied and suited for identifying which genes are differentially expressed in two different types of cancer patients. The results of the microarray data analysis provides invaluable information that can pave the way for innovative opportunities for an early diagnosis of malignancies and building explicit disease sketches. The data driven systems biology of cancer found the list of significant genes or differentially expressed genes in provisions of probabilities for leukemia dataset. This helps to find the functional relationships between genes in MDB warehouses by linking annotations from GO. Moreover, clusters with filter genes were able to distinguish ALL and MLL in unsupervised clustering. In addition, the quality and availability of gene annotations is essential for biological discoveries. The modern era of functional genomics propelled by MDB technology paves the way to elucidate physiological model of any disease. Therefore, bioinformatics applications and systems biology will be able to facilitate in decoding the life sciences concealed knowledge to beat diseases using a personalized medicine decision support system.

The knowledge based cancer systems biology study presents a new solid tumour growth multi-scale-modelling at the cellular, molecular and tissue levels with an angiogenesis module and TKI in EGFR signalling pathway integrated with anti-angiogenesis or IGF1R inhibitor treatment. A system of PDEs framework were used to alter the level of various nutrients supplied to the cells such as glucose, oxygen, drug, TAF and VEGF in the tumour microenvironment at the cellular level. VEGF angiogenesis inhibitors were shown by blood vessel formations and maturation owing to the tumour cells because of the infused nutrients forming new blood vessels at the tissue level. A system of ODEs was used to simulate the posterior mechanics of EGFR/ERK signalling integrated with IGF1R signalling pathway at the molecular level. Our simulation results give a demonstration of joint actions of active cells, migrating cells, proliferative cells and quiescent cells regulated by the active ERK receptors including EGFR and IGF1R receptors depicting sketch of tumour evolution. The dual effects of angiogenesis were also found on EGFR inhibition treatment to decrease the tumour invasion, as well as supply nutrients to tumour compartments to promote cell endurance at later time stages. Thus, our simulations suggest that EGFR and IGFR1 co-expression activates more ERK receptors when compared to down-expression or over-expression of EGFR and IGF1R. There is a good agreement between the simulations, an experimental mouse model, and clinical data.

6.2 A WISH FOR DEVELOPING COUNTRIES

Systems biology of cancer aims at designing tailored treatment for each patient. One can also ask if the cost of this personalized medicine is affordable. Progress based on new technologies is often expensive. To start with, it will be based on the availability of targeted therapeutic molecules for all investigation points. This means developing hundreds of new drugs.

On one hand, pharmaceutical companies focus on blockbuster drugs applicable to most people. This is a model of the average patient and the opposite of personalized medicine. With systems biology and personalized medicine comes preventive medicine. This prevention should bring important gains in terms of reducing the price of costly treatments and of course patient suffering. Making personalized medicine a reality is therefore primarily a matter of political will and organization [89].

6.3 FUTURE SCOPE

The aim of the proposed integrated model is to augment the prototype development processes for multiscale-multilevel physiological models in the future. Moreover, a model has to be regarded as a transient object that assists biologists. It is in the nature of scientific models to become obsolete. Thus, a model has to be superseded by more refined and more up-to-date models.

Due to a substantial flexibility in the structure and organization of the Plants Physiology Database, it is capable of being conveniently upgraded in the future with new data entries from various sources and approaches that are more efficient for data mining. The same implementation can be extended for other organisms, pathological conditions, etc., after mining gene ontological data in near future.

To the best of author's knowledge, no comparative experimental test has been carried out on systems biology processes that clearly distinguish between truly differentially expressed genes and false positives. This can be a future work with great potential. The analysis on the leukemia microarray data can be done for another input data set, the latest Affymetrix deposited HG-U133 files (or other most recent files, if deposited in the meantime) in the near future.

The work related to the structure of dynamical modelling can be extended for 3D simulation integrating more factors like blood flow, intratumoral pressure. In addition, high throughput technologies produce a huge amount of data that requires reliable annotations (also termed metadata) in order to provide significant biological and/or clinical interpretations and fit to be used in systems biology approaches.

Contemporary bioinformatics along with systems biology applications facilitate in decoding the life sciences concealed knowledge to win over diseases using personalized medicine decision support system in the near future. As such, four areas with novel techniques are required to realize the potential of personalized medicine: 1) Handling high throughput genomic data; 2) deducing the functional genomics and the effect of genomic variation; 3) integrating and relating complex genetic interactions with phenotypes using systems biology approaches; and 4) translating these findings into medical practice. Processing individual genomic data will be a big issue. High throughput genomic data profiling can replicate current leukemia classification based on present-day techniques in genetic marking, morphology, immunophenotype. Specificity and sensitivity are high for subcategories with specific therapeutic consequences and turnaround time will also be short. It can complement prevalent diagnostic standards and may even define superior classification systems in the near future.

Due to rapidly evolving field, the issues related to literature mining which is incarnation of all prior knowledge having huge potential. This comes into picture while doing thorough review of literature related to this research work. For instance, nearly seventeen million citations of published articles in the MEDLINE database alone exist. It is found that it is complex task to search relevant data in biomedical domain. The centralized data resource with proper cataloging and annotation for related literature is also one of the needs of the hour for gaining a quicker overview. This can be taken into account as one of the future work.

REFERENCES

[1] S. Agrawal, A. K. Maurya, K. Shrivastava, S. Kumar, M. Pant, and S. K. Mishra, "Training the trainees in radiation oncology with telemedicine as a tool in a developing country: a two-year audit," *International journal of telemedicine and applications,* vol. 2011, p. 1, 2011.

[2] J. Warrington, *Metaphysics* vol. 1: Dent, 1956.

[3] R. Neches, R. E. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator*, et al.*, "Enabling technology for knowledge sharing," *AI magazine,* vol. 12, p. 36, 1991.

[4] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge acquisition,* vol. 5, pp. 199-220, 1993.

[5] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, "What are ontologies, and why do we need them?," *IEEE Intelligent systems,* vol. 14, pp. 20-26, 1999.

[6] N. F. Noy and M. Klein, "Ontology evolution: Not the same as schema evolution," *Knowledge and information systems,* vol. 6, pp. 428-440, 2004.

[7] A. K. Nagar and D. Sokhi, "On wavelet-based adaptive approach for gene comparison," *International journal of intelligent systems technologies and applications,* vol. 5, pp. 104-114, 2008.

[8] A. Farquhar, R. Fikes, and J. Rice, "The ontolingua server: A tool for collaborative ontology construction," *International journal of human-computer studies,* vol. 46, pp. 707-727, 1997.

[9] S. Staab and R. Studer, *Handbook on ontologies*: Springer, 2010.

[10] Y. A. Tijerino, D. W. Embley, D. W. Lonsdale, Y. Ding, and G. Nagy, "Towards ontology generation from tables," *World Wide Web,* vol. 8, pp. 261-285, 2005.

[11] H. S. Pinto and J. P. Martins, "Ontologies: How can they be built?," *Knowledge and Information Systems,* vol. 6, pp. 441-464, 2004.

[12] O. Corcho, "Ontology based document annotation: trends and open research problems," *International Journal of Metadata, Semantics and Ontologies,* vol. 1, pp. 47-57, 2006.

[13] R. Poli, M. Healy, and A. Kameas, *Theory and applications of ontology: Computer applications*: Springer, 2010.

[14] M. Cristani and R. Cuel, "A survey on ontology creation methodologies," *International Journal on Semantic Web and Information Systems (IJSWIS),* vol. 1, pp. 49-69, 2005.

[15] G. Schreiber, B. Wielinga, and W. Jansweijer, "The KACTUS view on the ‘O’word," in *IJCAI workshop on basic ontological issues in knowledge sharing*, 1995, pp. 159-168.

[16] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider, "Sweetening ontologies with DOLCE," in *Knowledge engineering and knowledge management: Ontologies and the semantic Web*, ed: Springer, 2002, pp. 166-181.

[17] M. Gruninger and M. S. Fox, "The logic of enterprise modelling," in *Modelling and Methodologies for Enterprise Integration*, ed: Springer, 1996, pp. 140-157.

[18] D. Fensel, C. Bussler, Y. Ding, V. Kartseva, M. Klein, M. Korotkiy*, et al.*, "Semantic web application areas," in *NLDB Workshop*, 2002.

[19] O. Corcho, M. Fernández-López, A. Gómez-Pérez, and A. López-Cima, "Building legal ontologies with METHONTOLOGY and WebODE," in *Law and the semantic web*, ed: Springer, 2005, pp. 142-157.

[20] T.-L. Wong, W. Lam, and E. Chen, "Automatic domain ontology generation from web sites," *Journal of Integrated Design and Process Science,* vol. 9, pp. 29-38, 2005.

[21] J. B. Bard and S. Y. Rhee, "Ontologies in biology: design, applications and future challenges," *Nature Reviews Genetics,* vol. 5, pp. 213-222, 2004.

[22] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters*, et al.*, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature biotechnology,* vol. 25, pp. 1251-1255, 2007.

[23] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry*, et al.*, "Gene Ontology: tool for the unification of biology," *Nature genetics,* vol. 25, pp. 25-29, 2000.

[24] J. Z. Wang, Z. Du, P. S. Yu, and C.-F. Chen, "An Efficient Online Tool to Search Top-N Genes with Similar Biological Functions in Gene Ontology Database," in *Bioinformatics and Biomedicine, 2007. BIBM 2007. IEEE International Conference on*, 2007, pp. 406-411.

[25] B. Madan and P. Mishra, "Overexpression, purification and characterization of organic solvent stable lipase from Bacillus licheniformis RSP-09," *Journal of molecular microbiology and biotechnology,* vol. 17, pp. 118-123, 2009.

[26] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide*, et al.*, "Big data: The future of biocuration," *Nature,* vol. 455, pp. 47-50, 2008.

[27] G. O. Consortium, "The Gene Ontology (GO) database and informatics resource," *Nucleic acids research,* vol. 32, pp. D258-D261, 2004.

[28] M. Popescu and D. Xu, *Data mining in biomedicine using ontologies*: Artech House, 2009.

[29] M. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger*, et al.*, "The Gene Ontology (GO) database and informatics resource," *Nucleic acids research,* vol. 32, pp. D258-61, 2004.

[30] P. Mutowo-Meullenet, R. P. Huntley, E. C. Dimmer, Y. Alam-Faruque, T. Sawford, M. J. Martin*, et al.*, "Use of Gene Ontology Annotation to understand the peroxisome proteome in humans," *Database: the journal of biological databases and curation,* vol. 2013, 2013.

[31] G. O. Consortium, "The Gene Ontology in 2010: extensions and refinements," *Nucleic acids research,* vol. 38, pp. D331-D335, 2010.

[32] P. Du, G. Feng, J. Flatow, J. Song, M. Holko, W. A. Kibbe*, et al.*, "From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations," *Bioinformatics,* vol. 25, pp. i63-i68, 2009.

[33] G. O. Consortium, "Gene Ontology annotations and resources," *Nucleic acids research,* vol. 41, pp. D530-D535, 2013.

[34] S. Myhre, H. Tveit, T. Mollestad, and A. Lægreid, "Additional gene ontology structure for improved biological reasoning," *Bioinformatics,* vol. 22, pp. 2020-2027, 2006.

[35] C. Creighton and S. Hanash, "Mining gene expression databases for association rules," *Bioinformatics,* vol. 19, pp. 79-86, 2003.

[36] J. Han, H. Cheng, D. Xin, and X. Yan, "Frequent pattern mining: current status and future directions," *Data Mining and Knowledge Discovery,* vol. 15, pp. 55-86, 2007.

[37] R. Alves, D. S. Rodriguez-Baena, and J. S. Aguilar-Ruiz, "Gene association analysis: a survey of frequent pattern mining from gene expression data," *Briefings in bioinformatics,* vol. 11, pp. 210-224, 2010.

[38] S. Chakrabarti, E. Cox, E. Frank, R. H. Güting, J. Han, X. Jiang*, et al.*, *Data Mining: Know It All: Know It All*: Morgan Kaufmann, 2008.

[39] J. Han and A. Fu, "Mining multiple-level association rules in large databases," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 11, pp. 798-805,

1999.

[40] U. M. Fayyad, "Data mining and knowledge discovery: Making sense out of data," *IEEE Intelligent Systems,* vol. 11, pp. 20-25, 1996.

[41] S. Gupta, V. Bhatnagar, and S. K. Wasan, "Architecture for knowledge discovery and knowledge management," *Knowledge and information systems,* vol. 7, pp. 310-336, 2005.

[42] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 26, pp. 97-107, 2014.

[43] J. Rung and A. Brazma, "Reuse of public genome-wide gene expression data," *Nature Reviews Genetics,* vol. 14, pp. 89-99, 2012.

[44] S. Gupta, S. Wasan, and V. Bhatnagar, "On mining of data," *IETE J RES,* vol. 47, pp. 5-17, 2001.

[45] C.-A. Wu, W.-Y. Lin, C.-L. Jiang, and C.-C. Wu, "Toward intelligent data warehouse mining: An ontology-integrated approach for multi-dimensional association mining," *Expert Systems with Applications,* vol. 38, pp. 11011-11023, 2011.

[46] M. Hilario, P. Nguyen, H. Do, A. Woznica, and A. Kalousis, "Ontology-based meta-mining of knowledge discovery workflows," in *Meta-Learning in Computational Intelligence*, ed: Springer, 2011, pp. 273-315.

[47] M. Záková, P. Kremen, F. Zelezny, and N. Lavrac, "Automating knowledge discovery workflow composition through ontology-based planning," *Automation Science and Engineering, IEEE Transactions on,* vol. 8, pp. 253-264, 2011.

[48] P. Brezany, I. Janciak, and A. M. Tjoa, "Ontology-based construction of grid data mining workflows," 2007.

[49] M. Cannataro and D. Talia, "Semantics and knowledge grids: building the next-generation grid," *Intelligent Systems, IEEE,* vol. 19, pp. 56-63, 2004.

[50] T. T. Quan, S. C. Hui, A. C. M. Fong, and T. H. Cao, "Automatic generation of ontology for scholarly semantic web," in *The Semantic Web–ISWC 2004*, ed: Springer, 2004, pp. 726-740.

[51] U. Priss, "Formal concept analysis in information science," *ARIST,* vol. 40, pp. 521-543, 2006.

[52] A. Borgida and F. Giunchiglia, "Importing from functional knowledge bases-a preview," 2007.

[53] R. Studer, V. R. Benjamins, and D. Fensel, "Knowledge engineering: principles and methods," *Data & knowledge engineering,* vol. 25, pp. 161-197, 1998.

[54] B. Swartout, R. Patil, K. Knight, and T. Russ, "Toward distributed use of large-scale ontologies," in *Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems*, 1996, pp. 138-148.

[55] A. Maedche, B. Motik, L. Stojanovic, R. Studer, and R. Volz, "Ontologies for enterprise knowledge management," *IEEE Intelligent Systems,* vol. 18, pp. 26-33, 2003.

[56] C. Marinica and F. Guillet, "Knowledge-based interactive postmining of association rules using ontologies," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 22, pp. 784-797, 2010.

[57] D. Won and D. McLeod, "Ontology-driven rule generalization and categorization for market data," in *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, 2007, pp. 917-923.

[58] J. T. Wang, M. J. Zaki, H. T. Toivonen, and D. Shasha, *Introduction to data mining in bioinformatics*: Springer, 2005.

[59] S. Falcon and R. Gentleman, "Using GOstats to test gene lists for GO term association," *Bioinformatics,* vol. 23, pp. 257-258, 2007.

[60] G. Dennis Jr, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane*, et al.*, "DAVID: database for annotation, visualization, and integrated discovery," *Genome biol,* vol. 4, p. P3, 2003.

[61] S. Gupta, M. Mishra, N. Sen, R. Parihar, G. R. Dwivedi, F. Khan*, et al.*, "DbMDR: A Relational Database for Multidrug Resistance Genes as Potential Drug Targets," *Chemical biology & drug design,* vol. 78, pp. 734-738, 2011.

[62] M. M. Javidi, M. Sohrabi, and M. K. Rafsanjani, "Intrusion detection in database systems," in *Communication and Networking*, ed: Springer, 2010, pp. 93-101.

[63] M. K. Rafsanjani and Z. A. Varzaneh, "Intrusion Detection By Data Mining Algorithms: A Review," *Journal of New Results in Science,* vol. 2, pp. 76-91, 2013.

[64] V. Goyal, S. K. Gupta, and S. Saxena, "Query rewriting for detection of privacy violation through inferencing," in *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services*, 2006, p. 28.

[65] S. Haider, B. Ballester, D. Smedley, J. Zhang, P. Rice, and A. Kasprzyk, "BioMart Central Portal—unified access to biological data," *Nucleic acids research,* vol. 37, pp. W23-W27, 2009.

[66] M. Manuja and D. Garg, "Semantic web mining of un-structured data: Challenges and opportunities," *International Journal of Engineering (IJE),* vol. 5, p. 268, 2011.

[67] G. Dellaire, J. N. Berman, and R. J. Arceci, *Cancer Genomics: From Bench to Personalized Medicine*: Academic Press, 2013.

[68] S. P. Ficklin, L.-A. Sanderson, C.-H. Cheng, M. E. Staton, T. Lee, I.-H. Cho*, et al.*, "Tripal: a construction toolkit for online genome databases," *Database: the journal of biological databases and curation,* vol. 2011, 2011.

[69] S. Pettifer, D. Thorne, P. McDermott, T. Attwood, J. Baran, J. C. Bryne*, et al.*, "An active registry for bioinformatics web services," *Bioinformatics,* vol. 25, pp. 2090-2091, 2009.

[70] M. Krupp, J. U. Marquardt, U. Sahin, P. R. Galle, J. Castle, and A. Teufel, "RNA-Seq Atlas—a reference database for gene expression profiling in normal tissue by next-generation sequencing," *Bioinformatics,* vol. 28, pp. 1184-1185, 2012.

[71] D. Smedley, P. Schofield, C.-K. Chen, V. Aidinis, C. Ainali, J. Bard*, et al.*, "Finding and sharing: new approaches to registries of databases and services for the biomedical sciences," *Database: the journal of biological databases and curation,* vol. 2010, 2010.

[72] P. Vanhee, J. Reumers, F. Stricher, L. Baeten, L. Serrano, J. Schymkowitz*, et al.*, "PepX: a structural database of non-redundant protein–peptide complexes," *Nucleic acids research,* vol. 38, pp. D545-D551, 2010.

[73] M. Bhargava and A. Sharma, "DNA barcoding in plants: Evolution and applications of< i> in silico</i> approaches and resources," *Molecular phylogenetics and evolution,* vol. 67, pp. 631-641, 2013.

[74] B. Beavis, D. Gessler, S. Rhee, D. Rokhsar, D. Main, L. Mueller*, et al.*, "Plant Biology Databases: A Needs Assessment November 16, 2005," 2005.

[75] S. Y. Rhee, J. Dickerson, and D. Xu, "Bioinformatics and its applications in plant biology," *Annu. Rev. Plant Biol.,* vol. 57, pp. 335-360, 2006.

[76] Z. Lacroix, "Designing Efficient User-Friendly Biological Data Management Systems," *OMICS A Journal of Integrative Biology,* vol. 7, pp. 113-115, 2003.

[77] S. Mukherjea, "Information retrieval and knowledge discovery utilising a biomedical Semantic Web," *Briefings in bioinformatics,* vol. 6, pp. 252-262, 2005.

[78] H. Tawfik, O. Anya, and A. K. Nagar, "Understanding clinical work practices for cross-boundary decision support in e-health," *Information Technology in Biomedicine, IEEE Transactions on,* vol. 16, pp. 530-541, 2012.

[79] V. Wadhwa and D. Garg, "Multi-Objective Single Facility Location Problem: a Review," *International Review on Modelling & Simulations,* vol. 4, 2011.

[80] S. Maji and D. Garg, "Progress in Gene Prediction: Principles and Challenges," *Current Bioinformatics,* vol. 8, pp. 226-243, 2013.

[81] A. K. Kahlon, S. Roy, and A. Sharma, "Molecular docking studies to map the binding site of squalene synthase inhibitors on dehydrosqualene synthase of Staphylococcus aureus," *Journal of Biomolecular Structure and Dynamics,* vol. 28, pp. 201-210, 2010.

[82] S. Roy, N. Maheshwari, R. Chauhan, N. K. Sen, and A. Sharma, "Structure prediction and functional characterization of secondary metabolite proteins of Ocimum," *Bioinformation,* vol. 6, p. 315, 2011.

[83] R. Sareen, U. Bornscheuer, and P. Mishra, "A microtiter plate assay for the determination of the synthetic activity of protease," *Analytical biochemistry,* vol. 333, pp. 193-195, 2004.

[84] A. Saxena, K. P. Tripathi, S. Roy, F. Khan, and A. Sharma, "Pharmacovigilance: effects of herbal components on human drugs interactions involving cytochrome P450," *Bioinformation,* vol. 3, 2008.

[85] S. P. de Visser, J. S. Valentine, and W. Nam, "A biomimetic ferric hydroperoxo porphyrin intermediate," *Angewandte Chemie International Edition,* vol. 49, pp. 2099-2101, 2010.

[86] S. Shaik and S. P. De Visser, "Computational approaches to cytochrome P450 function," in *Cytochrome P450*, ed: Springer, 2005, pp. 45-85.

[87] S. d. Visser, "Density functional theory (DFT) and combined quantum mechanical/molecular mechanics (QM/MM) studies on the oxygen activation step in nitric oxide synthase enzymes," *Biochemical Society Transactions,* vol. 37, p. 373, 2009.

[88] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *cell,* vol. 100, pp. 57-70, 2000.

[89] E. Barillot, L. Calzone, P. Hupe, J.-P. Vert, and A. Zinovyev, *Computational systems biology of cancer*: CRC Press, 2012.

[90] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell,* vol. 144, pp. 646-674, 2011.

[91] J. J. Hornberg, F. J. Bruggeman, H. V. Westerhoff, and J. Lankelma, "Cancer: a systems biology disease," *Biosystems,* vol. 83, pp. 81-90, 2006.

[92] E. Wang, *Cancer systems biology*: CRC Press, 2010.

[93] H. M. Lodhi and S. H. Muggleton, *Elements of computational systems biology* vol. 8: John Wiley & Sons, 2010.

[94] H. Kitano, *Foundations of systems biology*: MIT press Cambridge, 2001.

[95] E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, H. Lehrach, and R. Herwig, *Systems biology*: John Wiley & Sons, 2013.

[96] A. K. Konopka, *Systems biology: principles, methods, and concepts*: CRC Press, 2006.

[97] F. Boogerd, F. J. Bruggeman, J.-H. S. Hofmeyr, and H. V. Westerhoff, *Systems biology: philosophical foundations*: Elsevier, 2007.

[98] L. Alberghina and H. V. Westerhoff, *Systems Biology: Definitions and Perspectives. Topics in Current Genetics*: Springer, 2005.

[99] H. Kitano, "Computational systems biology," *Nature,* vol. 420, pp. 206-210, 2002.

[100] K. Najarian, *Systems biology and bioinformatics: a computational approach*: CRC Press, Inc., 2009.

[101] D. J. Wilkinson, *Stochastic modelling for systems biology*: CRC press, 2011.

[102] M. Ullah and O. Wolkenhauer, *Stochastic approaches for systems biology*: Springer, 2011.

[103] M. Patel, *The role of model integration in complex systems modelling: An example from cancer biology*: Springer, 2010.

[104] J. DiStefano III, *Dynamic Systems Biology Modeling and Simulation*: Academic Press, 2014.

[105] C. Sonnenschein and A. M. Soto, "Why systems biology and cancer?," in *Seminars in cancer biology*, 2011, pp. 147-149.

[106] G. Hardiman, *Microarray Innovations: Technology and Experimentation*: CRC Press, 2010.

[107] S. Knudsen, *Guide to analysis of DNA microarray data*: John Wiley & Sons,

2005.

[108] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden*, et al.*, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature genetics,* vol. 30, pp. 41-47, 2002.

[109] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science,* vol. 270, pp. 467-470, 1995.

[110] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proceedings of the National Academy of Sciences,* vol. 98, pp. 31-36, 2001.

[111] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application," *Genome Biol,* vol. 2, pp. 1-11, 2001.

[112] P. F. Macgregor, "Gene expression in cancer: the application of microarrays," *Expert review of molecular diagnostics,* vol. 3, pp. 185-200, 2003.

[113] S. B. Amin, P. K. Shah, A. Yan, S. Adamia, S. Minvielle, H. Avet-Loiseau*, et al.*, "The dChip survival analysis module for microarray data," *BMC bioinformatics,* vol. 12, p. 72, 2011.

[114] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics,* vol. 10, pp. 57-63, 2009.

[115] A. S. Morrissy, R. D. Morin, A. Delaney, T. Zeng, H. McDonald, S. Jones*, et al.*, "Next-generation tag sequencing for cancer gene expression profiling," *Genome research,* vol. 19, pp. 1825-1835, 2009.

[116] M. Garber, M. G. Grabherr, M. Guttman, and C. Trapnell, "Computational methods for transcriptome annotation and quantification using RNA-seq," *Nature methods,* vol. 8, pp. 469-477, 2011.

[117] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov*, et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science,* vol. 286, pp. 531-537, 1999.

[118] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald*, et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature,* vol. 403, pp. 503-511, 2000.

[119] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees*, et al.*, "Molecular portraits of human breast tumours," *Nature,* vol. 406, pp. 747-

752, 2000.

[120] S. Bungaro, M. C. Dell'Orto, A. Zangrando, D. Basso, T. Gorletta, L. Lo Nigro*, et al.*, "Integration of genomic and gene expression data of childhood ALL without known aberrations identifies subgroups with specific genetic hallmarks," *Genes, Chromosomes and Cancer,* vol. 48, pp. 22-38, 2009.

[121] J. R. Pollack, T. Sørlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning*, et al.*, "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors," *Proceedings of the National Academy of Sciences,* vol. 99, pp. 12963-12968, 2002.

[122] M. Morley, C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman*, et al.*, "Genetic analysis of genome-wide variation in human gene expression," *Nature,* vol. 430, pp. 743-747, 2004.

[123] P. P. Shah, A. P. Singh, M. Singh, N. Mathur, M. C. Pant, B. N. Mishra*, et al.*, "Interaction of cytochrome P4501A1 genotypes with other risk factors and susceptibility to lung cancer," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* vol. 639, pp. 1-10, 2008.

[124] J. Peng, J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J. R. Pollack*, et al.*, "Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer," *The annals of applied statistics,* vol. 4, p. 53, 2010.

[125] S. S. Dhawan and A. Sharma, "Analysis of differentially expressed genes in abiotic stress response and their role in signal transduction pathways," *Protoplasma,* vol. 251, pp. 81-91, 2014.

[126] K. Shedden, W. Chen, R. Kuick, D. Ghosh, J. Macdonald, K. R. Cho*, et al.*, "Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data," *BMC bioinformatics,* vol. 6, p. 26, 2005.

[127] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf*, et al.*, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics,* vol. 4, pp. 249-264, 2003.

[128] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita*, et al.*, "A systems biology approach for pathway level analysis," *Genome research,* vol. 17, pp. 1537-1545, 2007.

[129] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A.

Gillette*, et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, pp. 15545-15550, 2005.

[130] S.-Y. Kim and D. J. Volsky, "PAGE: parametric analysis of gene set enrichment," *BMC bioinformatics,* vol. 6, p. 144, 2005.

[131] G. K. Smyth and T. Speed, "Normalization of cDNA microarray data," *Methods,* vol. 31, pp. 265-273, 2003.

[132] J. Quackenbush, "Microarray data normalization and transformation," *Nature genetics,* vol. 32, pp. 496-501, 2002.

[133] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed, "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics,* vol. 19, pp. 185-193, 2003.

[134] X. Cui and G. A. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biol,* vol. 4, p. 210, 2003.

[135] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes," *Proceedings of the National Academy of Sciences,* vol. 93, pp. 10614-10619, 1996.

[136] G. Gibson, "Microarrays in ecology and evolution: a preview," *Molecular ecology,* vol. 11, pp. 17-24, 2002.

[137] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American statistical association,* vol. 97, pp. 77-87, 2002.

[138] A.-C. Haury, P. Gestraud, and J.-P. Vert, "The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures," *PloS one,* vol. 6, p. e28210, 2011.

[139] M. L. Whitfield, L. K. George, G. D. Grant, and C. M. Perou, "Common markers of proliferation," *Nature Reviews Cancer,* vol. 6, pp. 99-106, 2006.

[140] H. Goodarzi, O. Elemento, and S. Tavazoie, "Revealing global regulatory perturbations across human cancers," *Molecular cell,* vol. 36, pp. 900-911, 2009.

[141] C. Sotiriou and M. J. Piccart, "Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?," *Nature Reviews Cancer,* vol. 7, pp. 545-553, 2007.

[142] J. S. Reis-Filho and L. Pusztai, "Gene expression profiling in breast cancer: classification, prognostication, and prediction," *The Lancet,* vol. 378, pp. 1812-1823, 2011.

[143] C. G. A. Network, "Comprehensive molecular portraits of human breast tumours," *Nature,* vol. 490, pp. 61-70, 2012.

[144] C. Desmedt, B. Haibe-Kains, P. Wirapati, M. Buyse, D. Larsimont, G. Bontempi*, et al.*, "Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes," *Clinical Cancer Research,* vol. 14, pp. 5158-5165, 2008.

[145] T. L. Mei-Ling, "Analysis of Microarray Gene Expression Data," ed: Springer US, 2004.

[146] W. Wu, *MicroRNA and Cancer*: Springer, 2011.

[147] H. Gohlmann and W. Talloen, *Gene expression studies using Affymetrix microarrays*: CRC Press, 2009.

[148] T. Speed, *Statistical analysis of gene expression microarray data*: CRC Press, 2004.

[149] A. Chowdhury, A. Konar, P. Rakshit, and A. K. Nagar, "A Multi-Objective Evolutionary Approach to Evaluate the Designing Perspective of Protein-Protein Interaction Network," *International Journal of Computer Information Systems and Industrial Management Applications,* vol. 6, pp. 445 - 465, 2014.

[150] R. Gupta, P. Mishra, and A. Mittal, "Enhancing nucleic acid detection sensitivity of propidium iodide by a three nanometer interaction inside cells and in solutions," *Journal of nanoscience and nanotechnology,* vol. 9, pp. 2607-2615, 2009.

[151] M. Dehmer, F. Emmert-Streib, A. Graber, and A. Salvador, *Applied statistics for network biology: methods in systems biology*: John Wiley & Sons, 2011.

[152] C. H. Waddington, "The strategy ofthe genes," *London: Allen,* vol. 86, 1957.

[153] C. J. Sherr, "Cancer cell cycles," *Science,* vol. 274, pp. 1672-1677, 1996.

[154] A. Koch and M. Schaechter, "A model for statistics of the cell division process," *Journal of general microbiology,* vol. 29, pp. 435-454, 1962.

[155] J. J. Tyson, B. Novak, G. M. Odell, K. Chen, and C. Dennis Thron, "Chemical kinetic theory: understanding cell-cycle regulation," *Trends in biochemical sciences,* vol. 21, pp. 89-96, 1996.

[156] A. Csikász-Nagy, "Computational systems biology of the cell cycle," *Briefings in*

bioinformatics, vol. 10, pp. 424-434, 2009.

[157] A. B. Pardee, "A restriction point for control of normal animal cell proliferation," *Proceedings of the National Academy of Sciences,* vol. 71, pp. 1286-1290, 1974.

[158] B. Aguda and Y. Tang, "The kinetic origins of the restriction point in the mammalian cell cycle," *Cell proliferation,* vol. 32, pp. 321-335, 1999.

[159] Z. Qu, W. R. MacLellan, and J. N. Weiss, "Dynamics of the cell cycle: checkpoints, sizers, and timers," *Biophysical journal,* vol. 85, pp. 3600-3611, 2003.

[160] B. Novak and J. J. Tyson, "A model for restriction point control of the mammalian cell cycle," *Journal of theoretical biology,* vol. 230, pp. 563-579, 2004.

[161] A. Zetterberg, O. Larsson, and K. G. Wiman, "What is the restriction point?," *Current opinion in cell biology,* vol. 7, pp. 835-842, 1995.

[162] R. Conradie, F. J. Bruggeman, A. Ciliberto, A. Csikász-Nagy, B. Novák, H. V. Westerhoff*, et al.*, "Restriction point control of the mammalian cell cycle via the cyclin E/Cdk2: p27 complex," *FEBS journal,* vol. 277, pp. 357-367, 2010.

[163] M. Pant, X.-Y. Liao, Q. Lu, S. Molloi, E. Elmore, and J. Redpath, "Mechanisms of suppression of neoplastic transformation in vitro by low doses of low LET radiation," *Carcinogenesis,* vol. 24, pp. 1961-1965, 2003.

[164] M. Rastogi, M. Srivastava, K. S. Chufal, M. Pant, K. Srivastava, and M. B. Bhatt, "Mitomycin and fluorouracil in combination with concomitant radiotherapy: A potentially curable approach for locally advanced head and neck squamous cell carcinoma," *Japanese journal of clinical oncology,* vol. 35, pp. 572-579, 2005.

[165] C. M. Guldberg and P. Waage, "Studies concerning affinity," *CM Forhandlinger: Videnskabs-Selskabet i Christiana,* vol. 35, p. 1864, 1864.

[166] D. M. Wittmann, J. Krumsiek, J. Saez-Rodriguez, D. A. Lauffenburger, S. Klamt, and F. J. Theis, "Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling," *BMC systems biology,* vol. 3, p. 98, 2009.

[167] A. Fauré, A. Naldi, C. Chaouiya, and D. Thieffry, "Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle," *Bioinformatics,* vol. 22, pp. e124-e131, 2006.

[168] L. Zhang, Z. Wang, J. A. Sagotsky, and T. S. Deisboeck, "Multiscale agent-based cancer modeling," *Journal of mathematical biology,* vol. 58, pp. 545-559, 2009.

[169] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 99, pp. 7280-7287, 2002.

[170] C. M. Macal and M. J. North, "Tutorial on agent-based modelling and simulation," *Journal of Simulation,* vol. 4, pp. 151-162, 2010.

[171] H. G. Holzhütter, D. Drasdo, T. Preusser, J. Lippert, and A. M. Henney, "The virtual liver: a multidisciplinary, multilevel challenge for systems biology," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine,* vol. 4, pp. 221-235, 2012.

[172] J. A. Sherratt and M. A. Chaplain, "A new mathematical model for avascular tumour growth," *Journal of mathematical biology,* vol. 43, pp. 291-312, 2001.

[173] S. Dormann and A. Deutsch, "Modeling of self-organized avascular tumor growth with a hybrid cellular automaton," *In silico biology,* vol. 2, pp. 393-406, 2002.

[174] B. Ribba, O. Saut, T. Colin, D. Bresch, E. Grenier, and J.-P. Boissel, "A multiscale mathematical model of avascular tumor growth to investigate the therapeutic benefit of anti-invasive agents," *Journal of theoretical biology,* vol. 243, pp. 532-541, 2006.

[175] V. Andasari, R. T. Roper, M. H. Swat, and M. A. Chaplain, "Integrating intracellular dynamics using CompuCell3D and Bionetsolver: applications to multiscale modelling of cancer cell growth and invasion," *PloS one,* vol. 7, p. e33726, 2012.

[176] R. Langer, H. Conn, J. Vacanti, C. Haudenschild, and J. Folkman, "Control of tumor growth in animals by infusion of an angiogenesis inhibitor," *Proceedings of the National Academy of Sciences,* vol. 77, pp. 4331-4335, 1980.

[177] R. S. Kerbel, "Tumor angiogenesis: past, present and the near future," *Carcinogenesis,* vol. 21, pp. 505-515, 2000.

[178] A. Shirinifard, J. S. Gens, B. L. Zaitlen, N. J. Popławski, M. Swat, and J. A. Glazier, "3D multi-cell simulation of tumor growth and angiogenesis," *PLoS One,* vol. 4, p. e7190, 2009.

[179] H. Perfahl, H. M. Byrne, T. Chen, V. Estrella, T. Alarcón, A. Lapin*, et al.*, "Multiscale modelling of vascular tumour growth in 3D: the roles of domain size and boundary conditions," *PloS one,* vol. 6, p. e14790, 2011.

[180] M. M. Olsen and H. T. Siegelmann, "Multiscale agent-based model of tumor

angiogenesis," *Procedia Computer Science,* vol. 18, pp. 1026-1035, 2013.

[181] X. Zhou, "An integrated multiscale mechanistic model for cancer drug therapy," *ISRN Biomathematics,* vol. 2012, 2012.

[182] J. Wang, L. Zhang, C. Jing, G. Ye, H. Wu, H. Miao*, et al.*, "Multi-scale agent-based modeling on melanoma and its related angiogenesis analysis," *Theoretical Biology and Medical Modelling,* vol. 10, p. 41, 2013.

[183] S. Kapoor, V. S. Rallabandi, C. Sakode, R. Padhi, and P. K. Roy, "A patient-specific therapeutic approach for tumour cell population extinction and drug toxicity reduction using control systems-based dose-profile design," *Theoretical Biology and Medical Modelling,* vol. 10, p. 68, 2013.

[184] T. S. Deisboeck, Z. Wang, P. Macklin, and V. Cristini, "Multiscale cancer modeling," *Annual review of biomedical engineering,* vol. 13, 2011.

[185] A. Krogh, "The supply of oxygen to the tissues and the regulation of the capillary circulation," *The Journal of physiology,* vol. 52, pp. 457-474, 1919.

[186] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of physiology,* vol. 117, p. 500, 1952.

[187] H. T. Milhorn Jr, R. Benton, R. Ross, and A. C. Guyton, "A mathematical model of the human respiratory control system," *Biophysical Journal,* vol. 5, pp. 27-46, 1965.

[188] A. C. Guyton, T. G. Coleman, and H. J. Granger, "Circulation: overall regulation," *Annual review of physiology,* vol. 34, pp. 13-44, 1972.

[189] T. G. Coleman and J. E. Randall, "A Comprehensive Physiological Model," *The Physiologist,* vol. 26, 1983.

[190] J. Hocquette, I. Cassar-Malek, A. Scalbert, and F. Guillou, "Contribution of genomics to the understanding of physiological functions," *J Physiol Pharmacol,* vol. 60, pp. 5-16, 2009.

[191] F. M. Weber, D. U. Keller, S. Bauer, G. Seemann, C. Lorenz, and O. Dossel, "Predicting tissue conductivity influences on body surface potentials—an efficient approach based on principal component analysis," *Biomedical Engineering, IEEE Transactions on,* vol. 58, pp. 265-273, 2011.

[192] A. Loewe, M. Wilhelms, F. Fischer, E. Scholz, and O. Dössel, "Impact of hERG Mutations on Simulated Human Atrial Action Potentials," *Biomedical*

Engineering/Biomedizinische Technik, 2013.

[193] M. Pfeifer, G. Lenis, and O. Dössel, "A General Approach for Dynamic Modeling of Physiological Time Series," *Biomedical Engineering/Biomedizinische Technik,* 2013.

[194] J. A. Shapiro, "Revisiting the central dogma in the 21st century," *Annals of the New York Academy of Sciences,* vol. 1178, pp. 6-28, 2009.

[195] P. Kohl and M. Viceconti, "The virtual physiological human: computer simulation for integrative biomedicine II," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* vol. 368, pp. 2837-2839, 2010.

[196] P. Kohl, E. Crampin, T. Quinn, and D. Noble, "Systems biology: an approach," *Clinical Pharmacology & Therapeutics,* vol. 88, pp. 25-33, 2010.

[197] P. Kohl and D. Noble, "Systems biology and the virtual physiological human," *Molecular Systems Biology,* vol. 5, 2009.

[198] D. A. Beard, J. B. Bassingthwaighte, and A. S. Greene, "Computational modeling of physiological systems," *Physiological genomics,* vol. 23, pp. 1-3, 2005.

[199] E. Z. Erson and M. C. Çavuşoğlu, "Design of a framework for modeling, integration and simulation of physiological models," *Computer methods and programs in biomedicine,* vol. 107, pp. 524-537, 2012.

[200] R. L. Hester, R. Iliescu, R. Summers, and T. G. Coleman, "Systems biology and integrative physiological modelling," *The Journal of physiology,* vol. 589, pp. 1053-1060, 2011.

[201] G. H. Fernald, E. Capriotti, R. Daneshjou, K. J. Karczewski, and R. B. Altman, "Bioinformatics challenges for personalized medicine," *Bioinformatics,* vol. 27, pp. 1741-1748, 2011.

[202] F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer, "A vision for the future of genomics research," *Nature,* vol. 422, pp. 835-847, 2003.

[203] J. Zhou, *Microbial functional genomics*: John Wiley & Sons, 2004.

[204] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics,* vol. 5, pp. 101-113, 2004.

[205] P.-Y. Bourguignon, J. van Helden, C. Ouzounis, and V. Schächter, "Computational analysis of metabolic networks," in *Modern Genome Annotation*, ed: Springer, 2008, pp. 329-351.

[206] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG

resource for deciphering the genome," *Nucleic acids research,* vol. 32, pp. D277-D280, 2004.

[207] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic acids research,* p. gkr988, 2011.

[208] R. Caspi, H. Foerster, C. A. Fulcher, R. Hopkinson, J. Ingraham, P. Kaipa*, et al.*, "MetaCyc: a multiorganism database of metabolic pathways and enzymes," *Nucleic acids research,* vol. 34, pp. D511-D516, 2006.

[209] R. Caspi, T. Altman, J. M. Dale, K. Dreher, C. A. Fulcher, F. Gilham*, et al.*, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic acids research,* vol. 38, pp. D473-D479, 2010.

[210] M. Latendresse, S. Paley, and P. D. Karp, "Browsing metabolic and regulatory networks with BioCyc," in *Bacterial Molecular Networks*, ed: Springer, 2012, pp. 197-216.

[211] T. Kelder, M. P. van Iersel, K. Hanspers, M. Kutmon, B. R. Conklin, C. T. Evelo*, et al.*, "WikiPathways: building research communities on biological pathways," *Nucleic acids research,* vol. 40, pp. D1301-D1307, 2012.

[212] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay*, et al.*, "PID: the pathway interaction database," *Nucleic acids research,* vol. 37, pp. D674-D679, 2009.

[213] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir, "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, pp. 2981-2986, 2004.

[214] A. Cakmak, M. Kirac, M. R. Reynolds, Z. M. Ozsoyoglu, and G. Ozsoyoglu, "Gene ontology-based annotation analysis and categorization of metabolic pathways," in *Scientific and Statistical Database Management, 2007. SSBDM'07. 19th International Conference on*, 2007, pp. 33-33.

[215] J. A. Dow and S. A. Davies, "Integrative physiology and functional genomics of epithelial function in a genetic model organism," *Physiological Reviews,* vol. 83, pp. 687-729, 2003.

[216] S. Meier and C. Gehring, "A guide to the integrated application of on-line data mining tools for the inference of gene functions at the systems level,"

Biotechnology journal, vol. 3, pp. 1375-1387, 2008.

[217] P. D. Thomas, H. Mi, and S. Lewis, "Ontology annotation: mapping genomic regions to biological function," *Current opinion in chemical biology,* vol. 11, pp. 4-11, 2007.

[218] S. Carbon, A. Ireland, C. J. Mungall, S. Shu, B. Marshall, and S. Lewis, "AmiGO: online access to ontology and annotation data," *Bioinformatics,* vol. 25, pp. 288-289, 2009.

[219] M. G. Giglio, C. W. Collmer, J. Lomax, and A. Ireland, "Applying the Gene Ontology in microbial annotation," *Trends in microbiology,* vol. 17, pp. 262-268, 2009.

[220] H. Sun, H. Fang, T. Chen, R. Perkins, and W. Tong, "GOFFA: gene ontology for functional analysis–a FDA gene ontology tool for analysis of genomic and proteomic data," *BMC bioinformatics,* vol. 7, p. S23, 2006.

[221] I. M. Keseler, C. Bonavides-Martínez, J. Collado-Vides, S. Gama-Castro, R. P. Gunsalus, D. A. Johnson*, et al.*, "EcoCyc: a comprehensive view of Escherichia coli biology," *Nucleic acids research,* vol. 37, pp. D464-D470, 2009.

[222] I. M. Keseler, J. Collado-Vides, A. Santos-Zavaleta, M. Peralta-Gil, S. Gama-Castro, L. Muñiz-Rascado*, et al.*, "EcoCyc: a comprehensive database of Escherichia coli biology," *Nucleic acids research,* vol. 39, pp. D583-D590, 2011.

[223] J. Hedegaard, C. Arce, S. Bicciato, A. Bonnet, B. Buitenhuis, M. Collado-Romero*, et al.*, "Methods for interpreting lists of affected genes obtained in a DNA microarray experiment," in *BMC proceedings*, 2009, p. S5.

[224] A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo, "Pathway studio—the analysis and navigation of molecular networks," *Bioinformatics,* vol. 19, pp. 2155-2157, 2003.

[225] Á. Jiménez-Marín, M. Collado-Romero, M. Ramirez-Boo, C. Arce, and J. J. Garrido, "Biological pathway analysis by ArrayUnlock and ingenuity pathway analysis," in *BMC proceedings*, 2009, p. S6.

[226] J. Gao, A. S. Ade, V. G. Tarcea, T. E. Weymouth, B. R. Mirel, and H. Jagadish, "Integrating and annotating the interactome using the MiMI plugin for cytoscape," *Bioinformatics,* vol. 25, pp. 137-138, 2009.

[227] S. J. Nelson, W. D. Johnston, and B. L. Humphreys, "Relationships in medical subject headings (MeSH)," in *Relationships in the organization of knowledge*, ed: Springer, 2001, pp. 171-184.

[228] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic acids research,* vol. 33, pp. D514-D517, 2005.

[229] S. N. Twigger, M. Shimoyama, S. Bromberg, and R. Team, "Rat Genome Database," *Update,* pp. D658-D662, 2007.

[230] V. Petri, M. Shimoyama, G. T. Hayman, J. R. Smith, M. Tutaj, J. de Pons*, et al.*, "The rat genome database pathway portal," *Database,* vol. 2011, p. bar010, 2011.

[231] H. Lv, L. Liu, Y. Zhang, T. Song, J. Lu, and X. Chen, "Ingenuity pathways analysis of urine metabonomics phenotypes toxicity of gentamicin in multiple organs," *Molecular Biosystems,* vol. 6, pp. 2056-2067, 2010.

[232] A. Bonnet, S. Lagarrigue, L. Liaubet, C. Robert-Granié, M. SanCristobal, and G. Tosser-Klopp, "Pathway results from the chicken data set using GOTM, Pathway Studio and Ingenuity softwares," in *BMC proceedings*, 2009, p. S11.

[233] A. Droit, G. G. Poirier, and J. M. Hunter, "Experimental and bioinformatic approaches for interrogating protein–protein interactions to determine protein function," *Journal of molecular endocrinology,* vol. 34, pp. 263-280, 2005.

[234] M. Krummenacker, S. Paley, L. Mueller, T. Yan, and P. D. Karp, "Querying and computing with BioCyc databases," *Bioinformatics,* vol. 21, pp. 3454-3455, 2005.

[235] F. Crick, "Central dogma of molecular biology," *Nature,* vol. 227, pp. 561-563, 1970.

[236] M. Schena, R. A. Heller, T. P. Theriault, K. Konrad, E. Lachenmeier, and R. W. Davis, "Microarrays: biotechnology's discovery platform for functional genomics," *Trends in biotechnology,* vol. 16, pp. 301-306, 1998.

[237] A. Brazma, "Minimum information about a microarray experiment (MIAME)–successes, failures, challenges," *The Scientific World Journal,* vol. 9, pp. 420-423, 2009.

[238] U. Yu, S. H. Lee, Y. J. Kim, and S. Kim, "Bioinformatics in the post-genome era," *Journal of biochemistry and molecular biology,* vol. 37, pp. 75-82, 2004.

[239] K. Banihashemi, "Iranian human genome project: Overview of a research process among Iranian ethnicities," *Indian journal of human genetics,* vol. 15, p. 88, 2009.

[240] S. Katzman, J. A. Capra, D. Haussler, and K. S. Pollard, "Ongoing GC-biased

evolution is widespread in the human genome and enriched near recombination hot spots," *Genome biology and evolution,* vol. 3, p. 614, 2011.

[241] A. S. Sharma, H. O. Gupta, R. Prasad, and P. M. Mitrasinovic, "Microarray Database Bioinformatics usher Functional Genomics to unveil Biological Knowledge underlying Physiology," *International Journal of Advanced Research in Computer Science and Software Engineering,* vol. 3, pp. 38-45, 2013.

[242] S. E. Lewis, "Gene Ontology: looking backwards and forwards," *Genome Biol,* vol. 6, p. 103, 2005.

[243] J. Blake, J. Corradi, J. Eppig, D. Hill, J. Richardson, and M. Ringwald, "Creating the gene ontology resource: design and implementation," 2001.

[244] F. M. McCarthy, T. J. Mahony, M. S. Parcells, and S. C. Burgess, "Understanding animal viruses using the Gene Ontology," *Trends in microbiology,* vol. 17, pp. 328-335, 2009.

[245] W. Jiang, X. Li, S. Rao, L. Wang, L. Du, C. Li*, et al.*, "Constructing disease-specific gene networks using pair-wise relevance metric: application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements," *BMC systems biology,* vol. 2, p. 72, 2008.

[246] S. Kumar, J. J. Buza, and S. C. Burgess, "Genotype-dependent tumor regression in Marek's disease mediated at the level of tumor immunity," *Cancer Microenvironment,* vol. 2, pp. 23-31, 2009.

[247] J. J. Buza and S. C. Burgess, "Different signaling pathways expressed by chicken naive CD4+ T cells, CD4+ lymphocytes activated with staphylococcal enterotoxin B, and those malignantly transformed by Marek's disease virus," *Journal of proteome research,* vol. 7, pp. 2380-2387, 2008.

[248] G. Klein, "Perspectives in studies of human tumor viruses," *Frontiers in bioscience: a journal and virtual library,* vol. 7, pp. d268-74, 2002.

[249] M. Klouche, G. Carruba, L. Castagnetta, and S. Rose-John, "Virokines in the pathogenesis of cancer: focus on human herpesvirus 8," *Annals of the New York Academy of Sciences,* vol. 1028, pp. 329-339, 2004.

[250] S. A. Smith and G. J. Kotwal, "Virokines: novel immunomodulatory agents," *Expert opinion on biological therapy,* vol. 1, pp. 343-357, 2001.

[251] T. Torto-Alalibo, C. Collmer, and M. Gwinn-Giglio, "The Plant-Associated Microbe Gene Ontology (PAMGO) Consortium: community development of new

Gene Ontology terms describing biological processes involved in microbe-host interactions," *BMC microbiology,* vol. 9, p. S1, 2009.

[252] J. B. Bassingthwaighte, "Strategies for the physiome project," *Annals of Biomedical Engineering,* vol. 28, pp. 1043-1058, 2000.

[253] E. J. Crampin, M. Halstead, P. Hunter, P. Nielsen, D. Noble, N. Smith*, et al.*, "Computational physiology and the physiome project," *Experimental Physiology,* vol. 89, pp. 1-26, 2004.

[254] A. D. Diehl, J. A. Lee, R. H. Scheuermann, and J. A. Blake, "Ontology development for biological systems: immunology," *Bioinformatics,* vol. 23, pp. 913-915, 2007.

[255] A. S. Sharma, H. O. Gupta, and P. M. Mitrasinovic, "From Ontology-Based Gene Function to Physiological Model," *Current Bioinformatics,* vol. 7, pp. 436-446, 2012.

[256] L. Tari, C. Baral, and P. Dasgupta, "Understanding the Global Properties of Functionally-Related Gene Networks Using the Gene Ontology," in *Pacific Symposium on Biocomputing*, 2005, pp. 209-220.

[257] P. Jaiswal, S. Avraham, K. Ilic, E. A. Kellogg, S. McCouch, A. Pujar*, et al.*, "Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages," *Comparative and functional genomics,* vol. 6, pp. 388-397, 2005.

[258] A. Papanicolaou and D. G. Heckel, "The GMOD Drupal bioinformatic server framework," *Bioinformatics,* vol. 26, pp. 3119-3124, 2010.

[259] D. P. Renfro, B. K. McIntosh, A. Venkatraman, D. A. Siegele, and J. C. Hu, "GONUTS: the gene ontology normal usage tracking system," *Nucleic acids research,* vol. 40, pp. D1262-D1269, 2012.

[260] B. K. McIntosh, D. P. Renfro, G. S. Knapp, C. R. Lairikyengbam, N. M. Liles, L. Niu*, et al.*, "EcoliWiki: a wiki-based community resource for Escherichia coli," *Nucleic acids research,* vol. 40, pp. D1270-D1277, 2012.

[261] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, and R. Apweiler, "QuickGO: a web-based tool for Gene Ontology searching," *Bioinformatics,* vol. 25, pp. 3045-3046, 2009.

[262] J. D. Watson and F. H. Crick, "Molecular structure of nucleic acids," *Nature,* vol. 171, pp. 737-738, 1953.

[263] D. J. Kevles and L. E. Hood, *The code of codes: Scientific and social issues in the human genome project*: Harvard University Press, 1993.

[264] F. S. Collins, M. Morgan, and A. Patrinos, "The Human Genome Project: lessons from large-scale biology," *Science,* vol. 300, pp. 286-290, 2003.

[265] C. Holdrege and J. Wirz, "Life beyond genes: Reflections on the human genome project," *Context,* vol. 5, pp. 14-19, 2001.

[266] T. R. o. t. U. o. California. (2014, 14 JULY 2014). *Genomes OnLine Database*. Available: http://www.genomesonline.org/index

[267] U. National Institutes of Health. (2014, 14 JULY 2014). *Mission*. Available: http://www.nih.gov/about/mission.htm

[268] T. Ideker, T. Galitski, and L. Hood, "A new approach to decoding life: systems biology," *Annual review of genomics and human genetics,* vol. 2, pp. 343-372, 2001.

[269] W. Dubitzky and F. Azuaje, *Artificial intelligence methods and tools for systems biology* vol. 5: Springer, 2004.

[270] N. Moseyko, T. Zhu, H.-S. Chang, X. Wang, and L. J. Feldman, "Transcription profiling of the early gravitropic response in Arabidopsis using high-density oligonucleotide probe microarrays," *Plant Physiology,* vol. 130, pp. 720-728, 2002.

[271] W.-J. Cao, H.-L. Wu, B.-S. He, Y.-S. Zhang, and Z.-Y. Zhang, "Analysis of long non-coding RNA expression profiles in gastric cancer," *World journal of gastroenterology: WJG,* vol. 19, p. 3658, 2013.

[272] T. Barrett and R. Edgar, "[19] Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis," *Methods in enzymology,* vol. 411, pp. 352-369, 2006.

[273] H. Parkinson, U. Sarkans, N. Kolesnikov, N. Abeygunawardena, T. Burdett, M. Dylag*, et al.*, "ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments," *Nucleic acids research,* vol. 39, pp. D1002-D1004, 2011.

[274] S. Russell, L. A. Meadows, and R. R. Russell, *Microarray technology in practice*: Academic Press, 2008.

[275] S. Ghosh, Y. Matsuoka, Y. Asai, K.-Y. Hsin, and H. Kitano, "Software for systems biology: from tools to integrated platforms," *Nature Reviews Genetics,* vol. 12, pp. 821-832, 2011.

[276] I. Affymetrix. (2014, 14 JULY 2014). *Affymetrix*. Available: http://www.affymetrix.com/estore/

[277] B. I. G. D. A. Center. (2014, 14 JULY 2014). *Cancer Program Data Sets*. Available: http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi

[278] G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger, *The analysis of gene expression data: an overview of Methods and Software*: Springer, 2003.

[279] R. G. a. R. Ihaka. (2014, 14 JULY 2014). Available: http://www.r-project.org/

[280] C. Workman, L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, H. B. Nielser*, et al.*, "A new non-linear normalization method for reducing variability in DNA microarray experiments," *Genome biol,* vol. 3, pp. 1-16, 2002.

[281] P. Neuvial, P. Hupé, I. Brito, S. Liva, É. Manié, C. Brennetot*, et al.*, "Spatial normalization of array-CGH data," *BMC bioinformatics,* vol. 7, p. 264, 2006.

[282] M. H. M. Saied, "Next-Generation Sequencing Analysis of DNA Methylation in Acute Myeloid Leukaemia," University of London London United Kingdom, 2012.

[283] A. Koren, I. Tirosh, and N. Barkai, "Autocorrelation analysis reveals widespread spatial biases in microarray experiments," *BMC genomics,* vol. 8, p. 164, 2007.

[284] G. Rigaill, P. Hupé, A. Almeida, P. La Rosa, J.-P. Meyniel, C. Decraene*, et al.*, "ITALICS: an algorithm for normalization and DNA copy number calling for Affymetrix SNP arrays," *Bioinformatics,* vol. 24, pp. 768-774, 2008.

[285] V. Boeva, A. Zinovyev, K. Bleakley, J.-P. Vert, I. Janoueix-Lerosey, O. Delattre*, et al.*, "Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization," *Bioinformatics,* vol. 27, pp. 268-269, 2011.

[286] J. M. Rizzo and M. J. Buck, "Key principles and clinical applications of “next-generation” DNA sequencing," *Cancer Prevention Research,* vol. 5, pp. 887-900, 2012.

[287] P. Stafford and M. Brun, "Three methods for optimization of cross-laboratory and cross-platform microarray expression data," *Nucleic acids research,* vol. 35, p. e72, 2007.

[288] K. Fraser, Z. Wang, and X. Liu, *Microarray Image Analysis: An Algorithmic Approach*: CRC Press, 2010.

[289] M. Savageau, "Reconstructionist molecular biology," *The New Biologist,* vol. 3, pp. 190-197, 1991.

[290] H. Osada and T. Takahashi, "Genetic alterations of multiple tumor suppressors and oncogenes in the carcinogenesis and progression of lung cancer," *Oncogene,* vol. 21, pp. 7421-7434, 2002.

[291] J. E. Klaunig and L. M. Kamendulis, "The role of oxidative stress in carcinogenesis," *Annu. Rev. Pharmacol. Toxicol.,* vol. 44, pp. 239-267, 2004.

[292] W. Mueller-Klieser, "Tumor biology and experimental therapeutics," *Critical reviews in oncology/hematology,* vol. 36, pp. 123-139, 2000.

[293] T. A. Brown, "Studying genomes," 2002.

[294] P. Vaupel, H. Fortmeyer, S. Runkel, and F. Kallinowski, "Blood flow, oxygen consumption, and tissue oxygenation of human breast cancer xenografts in nude rats," *Cancer research,* vol. 47, pp. 3496-3503, 1987.

[295] J. Landry, J. Freyer, and R. Sutherland, "A model for the growth of multicellular spheroids," *Cell Proliferation,* vol. 15, pp. 585-594, 1982.

[296] L. A. KUNZ-SCHUGHART, M. Kreutz, and R. Knuechel, "Multicellular spheroids: a three-dimensional in vitro culture system to study tumour biology," *International journal of experimental pathology,* vol. 79, pp. 1-23, 1998.

[297] J. P. Freyer and R. M. Sutherland, "Regulation of growth saturation and development of necrosis in EMT6/Ro multicellular spheroids by the glucose and oxygen supply," *Cancer research,* vol. 46, pp. 3504-3512, 1986.

[298] J. P. Freyer and R. M. Sutherland, "Proliferative and clonogenic heterogeneity of cells from EMT6/Ro multicellular spheroids induced by the glucose and oxygen supply," *Cancer Research,* vol. 46, pp. 3513-3520, 1986.

[299] M. Marušić, Ž. Bajzer, J. Freyer, and S. Vuk-Pavlović, "Analysis of growth of multicellular tumour spheroids by mathematical models," *Cell proliferation,* vol. 27, pp. 73-94, 1994.

[300] J. P. Freyer, "Role of necrosis in regulating the growth saturation of multicellular spheroids," *Cancer research,* vol. 48, pp. 2432-2439, 1988.

[301] M. Marušić, S. Vuk-Pavlovic, and J. P. Freyer, "Tumor growth< i> in vivo</i> and as multicellular spheroids compared by mathematical models," *Bulletin of mathematical biology,* vol. 56, pp. 617-631, 1994.

[302] K. Groebe and W. Mueller-Klieser, "Distributions of oxygen, nutrient, and metabolic waste concentrations in multicellular spheroids and their dependence on spheroid parameters," *European biophysics journal,* vol. 19, pp. 169-181, 1991.

[303] K. Groebe and W. Mueller-Klieser, "On the relation between size of necrosis and diameter of tumor spheroids," *International Journal of Radiation Oncology**

Biology Physics,* vol. 34, pp. 395-401, 1996.

[304] C. Chen, H. Byrne, and J. King, "The influence of growth-induced stress from the surrounding medium on the development of multicell spheroids," *Journal of Mathematical Biology,* vol. 43, pp. 191-220, 2001.

[305] T. L. Jackson and H. M. Byrne, "A mathematical model to study the effects of drug resistance and vasculature on the response of solid tumors to chemotherapy," *Mathematical biosciences,* vol. 164, pp. 17-38, 2000.

[306] K. Borkenstein, S. Levegrün, and P. Peschke, "Modeling and computer simulations of tumor growth and tumor response to radiotherapy," *Radiation research,* vol. 162, pp. 71-83, 2004.

[307] Y. Mansury, M. Kimura, J. Lobo, and T. S. Deisboeck, "Emerging patterns in tumor systems: simulating the dynamics of multicellular clusters with an agent-based spatial agglomeration model," *Journal of Theoretical Biology,* vol. 219, pp. 343-370, 2002.

[308] E. L. Stott, N. F. Britton, J. A. Glazier, and M. Zajac, "Stochastic simulation of benign avascular tumour growth using the Potts model," *Mathematical and Computer Modelling,* vol. 30, pp. 183-198, 1999.

[309] T. Alarcon, H. Byrne, and P. Maini, "Towards whole-organ modelling of tumour growth," *Progress in biophysics and molecular biology,* vol. 85, pp. 451-472, 2004.

[310] T. Alarcón, H. M. Byrne, and P. K. Maini, "A multiple scale model for tumor growth," *Multiscale Modeling & Simulation,* vol. 3, pp. 440-475, 2005.

[311] L. Tang, A. L. van de Ven, D. Guo, V. Andasari, V. Cristini, K. C. Li*, et al.*, "Computational Modeling of 3D Tumor Growth and Angiogenesis for Chemotherapy Evaluation," *PloS one,* vol. 9, p. e83962, 2014.

[312] N. E. Hynes and H. A. Lane, "ERBB receptors and cancer: the complexity of targeted inhibitors," *Nature Reviews Cancer,* vol. 5, pp. 341-354, 2005.

[313] T. Mitsudomi, S. Morita, Y. Yatabe, S. Negoro, I. Okamoto, J. Tsurutani*, et al.*, "Gefitinib versus cisplatin plus docetaxel in patients with non-small-cell lung cancer harbouring mutations of the epidermal growth factor receptor (WJTOG3405): an open label, randomised phase 3 trial," *The lancet oncology,* vol. 11, pp. 121-128, 2010.

[314] A. Gschwind, O. M. Fischer, and A. Ullrich, "The discovery of receptor tyrosine kinases: targets for cancer therapy," *Nature Reviews Cancer,* vol. 4, pp. 361-370,

2004.

[315] F. Hirsch, M. Varella-Garcia, and F. Cappuzzo, "Predictive value of EGFR and HER2 overexpression in advanced non-small-cell lung cancer," *Oncogene,* vol. 28, pp. S32-S37, 2009.

[316] N. Normanno, A. De Luca, C. Bianco, L. Strizzi, M. Mancino, M. R. Maiello*, et al.*, "Epidermal growth factor receptor (EGFR) signaling in cancer," *Gene,* vol. 366, pp. 2-16, 2006.

[317] A. Gazdar, "Activating and resistance mutations of EGFR in non-small-cell lung cancer: role in clinical response to EGFR tyrosine kinase inhibitors," *Oncogene,* vol. 28, p. S24, 2009.

[318] A.-P. Meert, B. Martin, P. Delmotte, T. Berghmans, J.-J. Lafitte, C. Mascaux*, et al.*, "The role of EGF-R expression on patient survival in lung cancer: a systematic review with meta-analysis," *European Respiratory Journal,* vol. 20, pp. 975-981, 2002.

[319] S. V. Sharma, D. W. Bell, J. Settleman, and D. A. Haber, "Epidermal growth factor receptor mutations in lung cancer," *Nature Reviews Cancer,* vol. 7, pp. 169-181, 2007.

[320] T. J. Lynch, D. W. Bell, R. Sordella, S. Gurubhagavatula, R. A. Okimoto, B. W. Brannigan*, et al.*, "Activating mutations in the epidermal growth factor receptor underlying responsiveness of non–small-cell lung cancer to gefitinib," *New England Journal of Medicine,* vol. 350, pp. 2129-2139, 2004.

[321] J. G. Paez, P. A. Jänne, J. C. Lee, S. Tracy, H. Greulich, S. Gabriel*, et al.*, "EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy," *Science,* vol. 304, pp. 1497-1500, 2004.

[322] W. Pao, V. Miller, M. Zakowski, J. Doherty, K. Politi, I. Sarkaria*, et al.*, "EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, pp. 13306-13311, 2004.

[323] L. V. Sequist, D. W. Bell, T. J. Lynch, and D. A. Haber, "Molecular predictors of response to epidermal growth factor receptor antagonists in non–small-cell lung cancer," *Journal of Clinical Oncology,* vol. 25, pp. 587-595, 2007.

[324] G. J. Riely, W. Pao, D. Pham, A. R. Li, N. Rizvi, E. S. Venkatraman*, et al.*, "Clinical course of patients with non–small cell lung cancer and epidermal

growth factor receptor exon 19 and exon 21 mutations treated with gefitinib or erlotinib," *Clinical Cancer Research,* vol. 12, pp. 839-844, 2006.

[325] C.-H. Wang and J. Li, "Three-dimensional simulation of IgG delivery to tumors," *Chemical Engineering Science,* vol. 53, pp. 3579-3600, 1998.

[326] H. B. Frieboes, J. S. Lowengrub, S. Wise, X. Zheng, P. Macklin, E. L. Bearer*, et al.*, "Computer simulation of glioma growth and morphology," *Neuroimage,* vol. 37, pp. S59-S70, 2007.

[327] S. Sanga, J. P. Sinek, H. B. Frieboes, M. Ferrari, J. P. Fruehauf, and V. Cristini, "Mathematical modeling of cancer progression and response to chemotherapy," 2006.

[328] G. R. DiResta, S. S. Nathan, M. W. Manoso, J. Casas-Ganem, C. Wyatt, T. Kubo*, et al.*, "Cell proliferation of cultured human cancer cells are affected by the elevated tumor pressures that exist in vivo," *Annals of biomedical engineering,* vol. 33, pp. 1270-1280, 2005.

[329] S. S. Nathan, G. R. DiResta, J. E. Casas-Ganem, B. H. Hoang, R. Sowers, R. Yang*, et al.*, "Elevated physiologic tumor pressure promotes proliferation and chemosensitivity in human osteosarcoma," *Clinical cancer research,* vol. 11, pp. 2389-2397, 2005.

[330] M. Hofmann, M. Guschel, A. Bernd, J. Bereiter-Hahn, R. Kaufmann, C. Tandi*, et al.*, "Lowering of tumor interstitial fluid pressure reduces tumor cell proliferation in a xenograft tumor model," *Neoplasia (New York, NY),* vol. 8, p. 89, 2006.

[331] V. Ganapathy, M. Thangaraju, and P. D. Prasad, "Nutrient transporters in cancer: relevance to Warburg hypothesis and beyond," *Pharmacology & therapeutics,* vol. 121, pp. 29-40, 2009.

[332] J. A. Bertout, S. A. Patel, and M. C. Simon, "The impact of O2 availability on human cancer," *Nature Reviews Cancer,* vol. 8, pp. 967-975, 2008.

[333] Y. Chen, R. Cairns, I. Papandreou, A. Koong, and N. C. Denko, "Oxygen consumption can regulate the growth of tumors, a new perspective on the Warburg effect," *PLoS One,* vol. 4, p. e7033, 2009.

[334] A. R. Anderson and M. Chaplain, "Continuous and discrete mathematical models of tumor-induced angiogenesis," *Bulletin of mathematical biology,* vol. 60, pp. 857-899, 1998.

[335] L. T. Baxter and R. K. Jain, "Transport of fluid and macromolecules in tumors. I. Role of interstitial pressure and convection," *Microvascular research,* vol. 37, pp.

77-104, 1989.

[336] Y. Boucher and R. K. Jain, "Microvascular pressure is the principal driving force for interstitial hypertension in solid tumors: implications for vascular collapse," *Cancer Research,* vol. 52, pp. 5110-5114, 1992.

[337] C.-H. Heldin, K. Rubin, K. Pietras, and A. Östman, "High interstitial fluid pressure—an obstacle in cancer therapy," *Nature Reviews Cancer,* vol. 4, pp. 806-813, 2004.

[338] J. C. Luo, S. Yamaguchi, A. Shinkai, K. Shitara, and M. Shibuya, "Significant expression of vascular endothelial growth factor/vascular permeability factor in mouse ascites tumors," *Cancer research,* vol. 58, pp. 2652-2660, 1998.

[339] J. Folkman, "The role of angiogenesis in tumor growth," in *Seminars in cancer biology*, 1992, pp. 65-71.

[340] L. T. Baxter and R. K. Jain, "Transport of fluid and macromolecules in tumors. II. Role of heterogeneous perfusion and lymphatics," *Microvascular research,* vol. 40, pp. 246-263, 1990.

[341] X. Sun, L. Zhang, H. Tan, J. Bao, C. Strouthos, and X. Zhou, "Multi-scale agent-based brain cancer modeling and prediction of TKI treatment response: Incorporating EGFR signaling pathway and angiogenesis," *BMC bioinformatics,* vol. 13, p. 218, 2012.

[342] J. W. Baish, T. Stylianopoulos, R. M. Lanning, W. S. Kamoun, D. Fukumura, L. L. Munn*, et al.*, "Scaling rules for diffusive drug delivery in tumor and normal tissues," *Proceedings of the National Academy of Sciences,* vol. 108, pp. 1799-1803, 2011.

[343] Y. Mansury and T. S. Deisboeck, "The impact of "search precision" in an agent-based tumor model," *Journal of theoretical biology,* vol. 224, pp. 325-337, 2003.

[344] R. J. Orton, M. E. Adriaens, A. Gormand, O. E. Sturm, W. Kolch, and D. R. Gilbert, "Computational modelling of cancerous mutations in the EGFR/ERK signalling pathway," *BMC systems biology,* vol. 3, p. 100, 2009.

[345] K. S. Brown, C. C. Hill, G. A. Calero, C. R. Myers, K. H. Lee, J. P. Sethna*, et al.*, "The statistical mechanics of complex signaling networks: nerve growth factor signaling," *Physical biology,* vol. 1, p. 184, 2004.

[346] P. Carmeliet and R. K. Jain, "Principles and mechanisms of vessel normalization for cancer and other angiogenic diseases," *Nature reviews Drug discovery,* vol. 10, pp. 417-427, 2011.

[347] L. Tang, J. Su, D.-S. Huang, D. Y. Lee, K. C. Li, and X. Zhou, "An Integrated Multiscale Mechanistic Model for Cancer Drug Therapy," *ISRN Biomathematics,* vol. 2012, p. 12, 2012.

[348] J. Dingemanse and S. Appel-Dingemanse, "Integrated pharmacokinetics and pharmacodynamics in drug development," *Clinical pharmacokinetics,* vol. 46, pp. 713-737, 2007.

[349] A. I. Minchinton and I. F. Tannock, "Drug penetration in solid tumours," *Nature Reviews Cancer,* vol. 6, pp. 583-592, 2006.

[350] H. Brem and J. Folkman, "Inhibition of tumor angiogenesis mediated by cartilage," *The Journal of experimental medicine,* vol. 141, pp. 427-439, 1975.

[351] M. A. Chaplain, S. R. McDougall, and A. Anderson, "Mathematical modeling of tumor-induced angiogenesis," *Annu. Rev. Biomed. Eng.,* vol. 8, pp. 233-257, 2006.

[352] F. Milde, M. Bergdorf, and P. Koumoutsakos, "A hybrid model for three-dimensional simulations of sprouting angiogenesis," *Biophysical journal,* vol. 95, pp. 3146-3160, 2008.

[353] T. Tanaka, N. Yamanaka, T. Oriyama, K. Furukawa, and E. Okamoto, "Factors regulating tumor pressure in hepatocellular carcinoma and implications for tumor spread," *Hepatology,* vol. 26, pp. 283-287, 1997.

[354] Z. Zhang, A. L. Stiegler, T. J. Boggon, S. Kobayashi, and B. Halmos, "EGFR-mutated lung cancer: a paradigm of molecular oncology," *Oncotarget,* vol. 1, p. 497, 2010.

[355] D. Chamovitz, *What a plant knows: a field guide to the senses*: Macmillan, 2012.

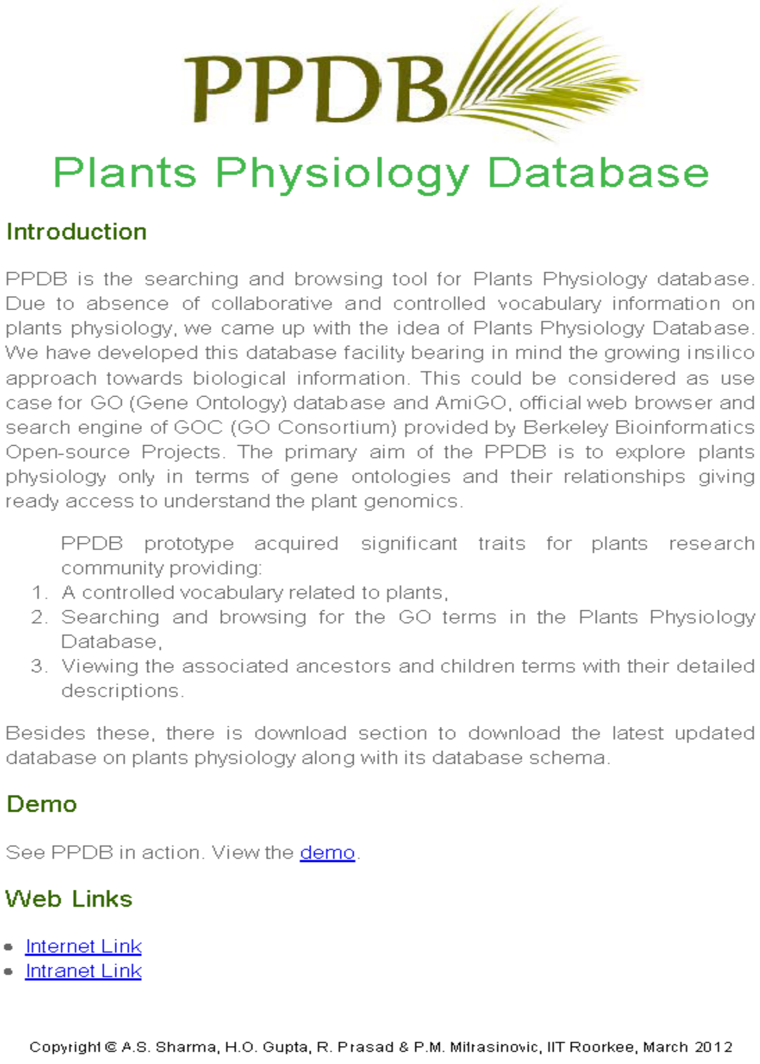
[356] M. Schwaiger, A. Schönauer, A. F. Rendeiro, C. Pribitzer, A. Schauer, A. F. Gilles*, et al.*, "Evolutionary conservation of the eumetazoan gene regulatory landscape," *Genome research,* vol. 24, pp. 639-650, 2014.

[357] Y. Moran, D. Fredman, D. Praher, X. Z. Li, L. M. Wee, F. Rentzsch*, et al.*, "Cnidarian microRNAs frequently regulate targets by cleavage," *Genome research,* vol. 24, pp. 651-663, 2014.

Annexure – A

PLANTS PHYSIOLOGY DATABASE AS BROWSING TOOL

The PPDB is available in the public domain through a user friendly environment accessible at www.iitr.ernet.in/ajayshiv/ or www.iitr.ac.in/ajayshiv as shown in Figure A.1.



PPDB

Plants Physiology Database

Introduction

PPDB is the searching and browsing tool for Plants Physiology database. Due to absence of collaborative and controlled vocabulary information on plants physiology, we came up with the idea of Plants Physiology Database. We have developed this database facility bearing in mind the growing insilico approach towards biological information. This could be considered as use case for GO (Gene Ontology) database and AmiGO, official web browser and search engine of GOC (GO Consortium) provided by Berkeley Bioinformatics Open-source Projects. The primary aim of the PPDB is to explore plants physiology only in terms of gene ontologies and their relationships giving ready access to understand the plant genomics.

PPDB prototype acquired significant traits for plants research community providing:

1. A controlled vocabulary related to plants,
2. Searching and browsing for the GO terms in the Plants Physiology Database,
3. Viewing the associated ancestors and children terms with their detailed descriptions.

Besides these, there is download section to download the latest updated database on plants physiology along with its database schema.

Demo

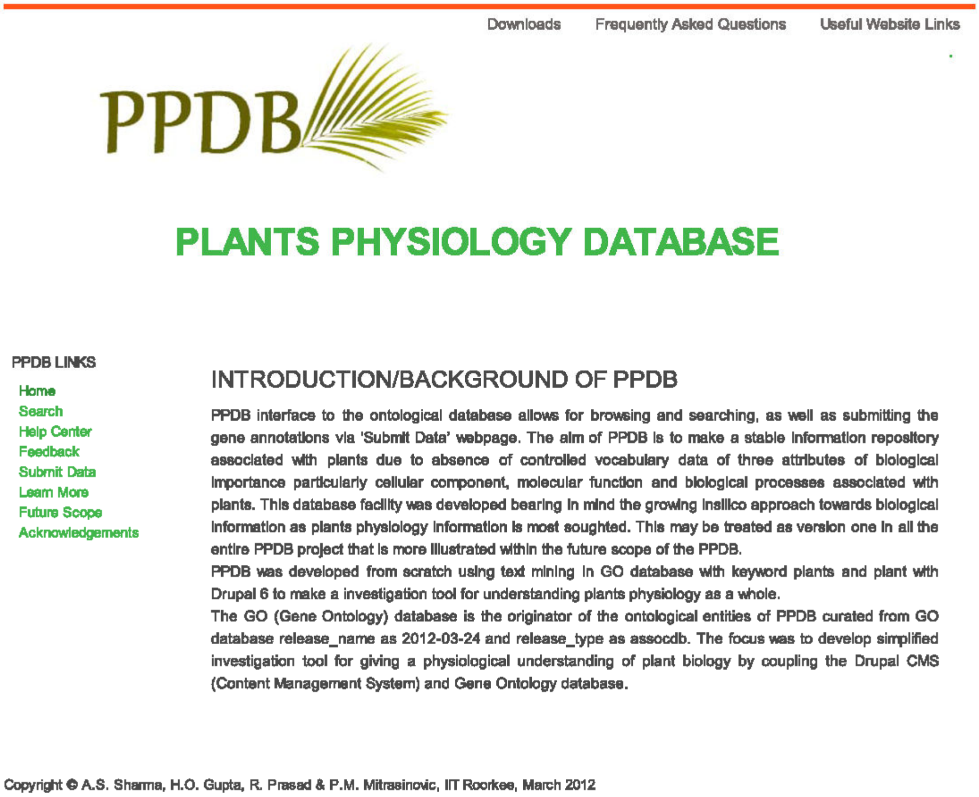
See PPDB in action. View the demo.

Web Links

- Internet Link
- Intranet Link

Figure A.1 PPDB Front Webpage

After clicking the Internet Link under Web Links as shown in Figure 1, a display of the webpage is shown in Figure A.2.



PLANTS PHYSIOLOGY DATABASE

PPDB LINKS

- Home
- Search
- Help Center
- Feedback
- Submit Data
- Learn More
- Future Scope
- Acknowledgements

INTRODUCTION/BACKGROUND OF PPDB

PPDB interface to the ontological database allows for browsing and searching, as well as submitting the gene annotations via 'Submit Data' webpage. The aim of PPDB is to make a stable information repository associated with plants due to absence of controlled vocabulary data of three attributes of biological importance particularly cellular component, molecular function and biological processes associated with plants. This database facility was developed bearing in mind the growing insilico approach towards biological information as plants physiology information is most soughted. This may be treated as version one in all the entire PPDB project that is more illustrated within the future scope of the PPDB.

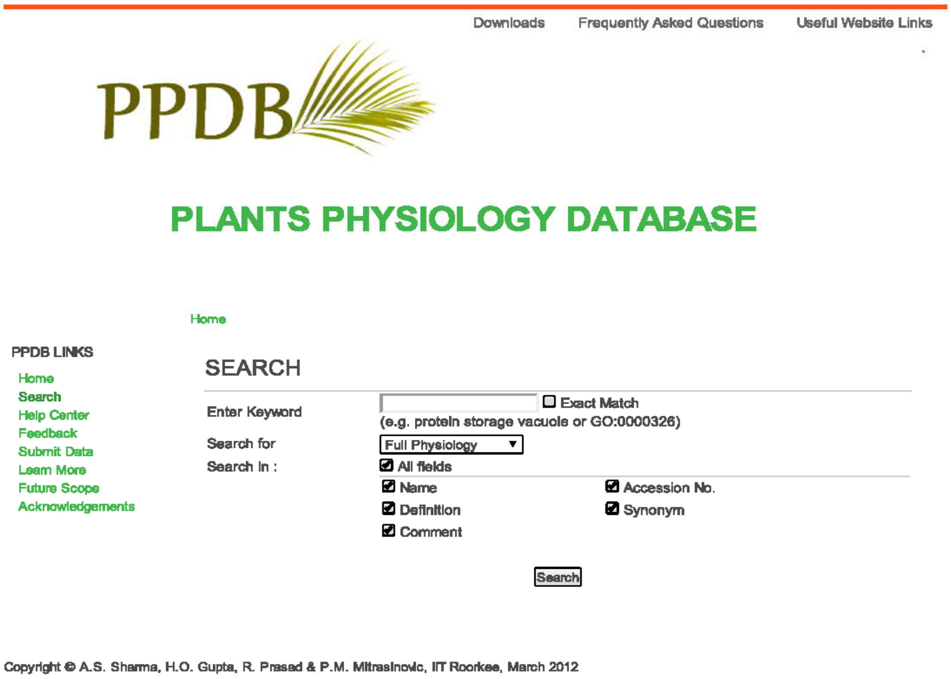
PPDB was developed from scratch using text mining in GO database with keyword plants and plant with Drupal 6 to make a investigation tool for understanding plants physiology as a whole.

The GO (Gene Ontology) database is the originator of the ontological entities of PPDB curated from GO database release_name as 2012-03-24 and release_type as assocdb. The focus was to develop simplified investigation tool for giving a physiological understanding of plant biology by coupling the Drupal CMS (Content Management System) and Gene Ontology database.

Copyright © A.S. Sharma, H.O. Gupta, R. Prasad & P.M. Mitrasinovic, IIT Roorkee, March 2012

Figure A.2 Webpage after clicking the Internet Link under Web Links

To indicate some of the functional features of PPDB, a search example for searching 'protein storage vacuole or GO:0000326' is described below. The initial display of PPDB that is relevant for retrieving data from the database is given in Figure A.3. Besides entering keywords or a specific GO accession number, the "SEARCH FOR" option offers the possibility of specifying either "Full Physiology" (BPCCMF), (biological process, cellular component, molecular function) three constitutive terms of physiological GO or any single one. In the instructive exercise, the search was performed by choosing "Enter Keywords = GO:0000326" and "Search for = Full Physiology", while the "Search in" option was associated with the "All fields" option checked. The display showing the search results is given in Figure A.4 with a broader view of showing major fields like "Accession No.", "Name", "Ontology/Type" and "EMBL-EBI Link".



Downloads | Frequently Asked Questions | Useful Website Links

PPDB

PLANTS PHYSIOLOGY DATABASE

Home

PPDB LINKS

- Home
- Search
- Help Center
- Feedback
- Submit Data
- Learn More
- Future Scope
- Acknowledgements

SEARCH

Enter Keyword ☐ Exact Match
(e.g. protein storage vacuole or GO:0000326)

Search for: Full Physiology ▼

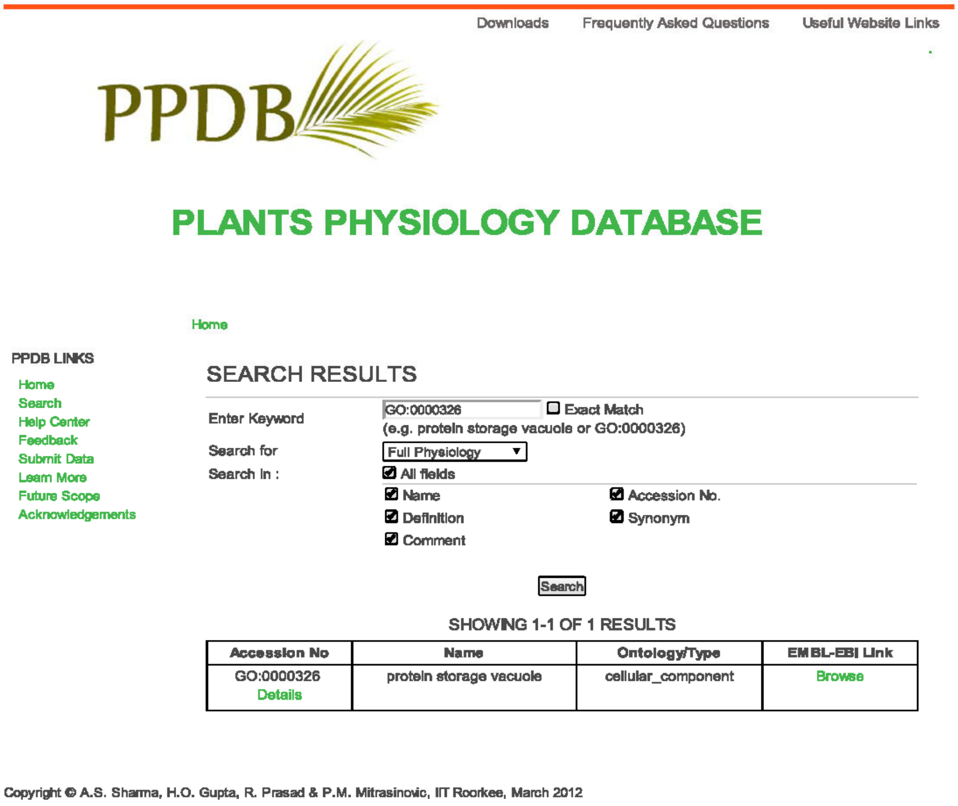
Search in : ☑ All fields

☑ Name ☑ Accession No.
☑ Definition ☑ Synonym
☑ Comment

Search

Copyright © A.S. Sharma, H.O. Gupta, R. Prasad & P.M. Mitrasinovic, IIT Roorkee, March 2012

Figure A.3 Display showing various search options of PPDB



Downloads | Frequently Asked Questions | Useful Website Links

PPDB

PLANTS PHYSIOLOGY DATABASE

Home

PPDB LINKS

- Home
- Search
- Help Center
- Feedback
- Submit Data
- Learn More
- Future Scope
- Acknowledgements

SEARCH RESULTS

Enter Keyword: GO:0000326 ☐ Exact Match
(e.g. protein storage vacuole or GO:0000326)

Search for: Full Physiology ▼

Search in : ☑ All fields

☑ Name ☑ Accession No.
☑ Definition ☑ Synonym
☑ Comment

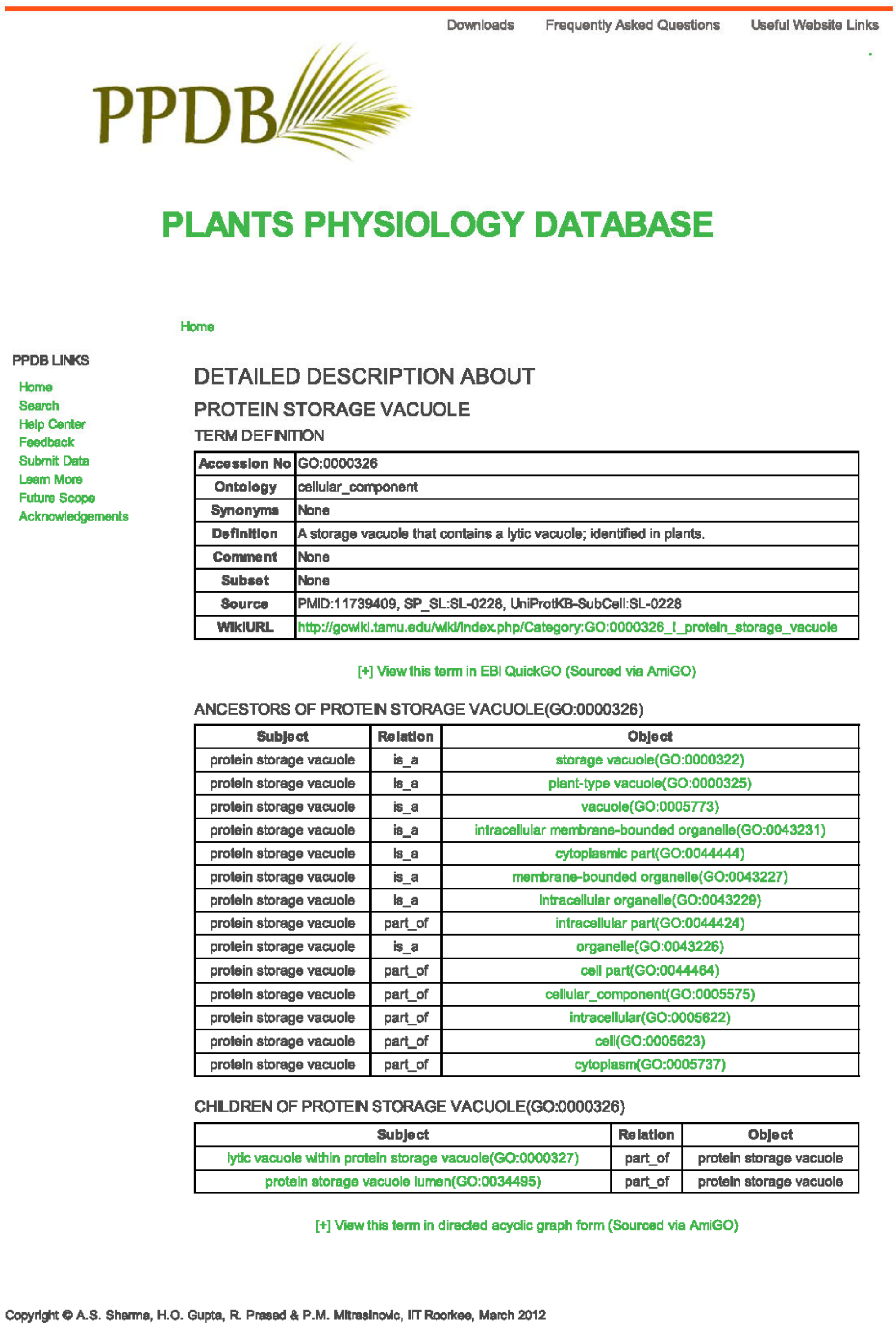
Search

SHOWING 1-1 OF 1 RESULTS

Accession No	Name	Ontology/Type	EMBL-EBI Link
GO:0000326 Details	protein storage vacuole	cellular_component	Browse

Copyright © A.S. Sharma, H.O. Gupta, R. Prasad & P.M. Mitrasinovic, IIT Roorkee, March 2012

Figure A.4 The search results obtained by specifying Enter Keywords = GO:0000326



PLANTS PHYSIOLOGY DATABASE

Home

PPDB LINKS

- Home
- Search
- Help Center
- Feedback
- Submit Data
- Learn More
- Future Scope
- Acknowledgements

DETAILED DESCRIPTION ABOUT

PROTEIN STORAGE VACUOLE

TERM DEFINITION

Accession No	GO:0000326
Ontology	cellular_component
Synonyms	None
Definition	A storage vacuole that contains a lytic vacuole; identified in plants.
Comment	None
Subset	None
Source	PMID:11739409, SP_SL:SL-0228, UniProtKB-SubCell:SL-0228
WikiURL	http://gowiki.tamu.edu/wiki/index.php/Category:GO:0000326_!_protein_storage_vacuole

[+] View this term in EBI QuickGO (Sourced via AmiGO)

ANCESTORS OF PROTEIN STORAGE VACUOLE(GO:0000326)

Subject	Relation	Object
protein storage vacuole	is_a	storage vacuole(GO:0000322)
protein storage vacuole	is_a	plant-type vacuole(GO:0000325)
protein storage vacuole	is_a	vacuole(GO:0005773)
protein storage vacuole	is_a	intracellular membrane-bounded organelle(GO:0043231)
protein storage vacuole	is_a	cytoplasmic part(GO:0044444)
protein storage vacuole	is_a	membrane-bounded organelle(GO:0043227)
protein storage vacuole	is_a	intracellular organelle(GO:0043229)
protein storage vacuole	part_of	intracellular part(GO:0044424)
protein storage vacuole	is_a	organelle(GO:0043226)
protein storage vacuole	part_of	cell part(GO:0044464)
protein storage vacuole	part_of	cellular_component(GO:0005575)
protein storage vacuole	part_of	intracellular(GO:0005622)
protein storage vacuole	part_of	cell(GO:0005623)
protein storage vacuole	part_of	cytoplasm(GO:0005737)

CHILDREN OF PROTEIN STORAGE VACUOLE(GO:0000326)

Subject	Relation	Object
lytic vacuole within protein storage vacuole(GO:0000327)	part_of	protein storage vacuole
protein storage vacuole lumen(GO:0034495)	part_of	protein storage vacuole

[+] View this term in directed acyclic graph form (Sourced via AmiGO)

Copyright © A.S. Sharma, H.O. Gupta, R. Prasad & P.M. Mitrasinovic, IIT Roorkee, March 2012

Figure A.5 Detailed description of GO:0000326 obtained by clicking on "Details" in the Accession No. column

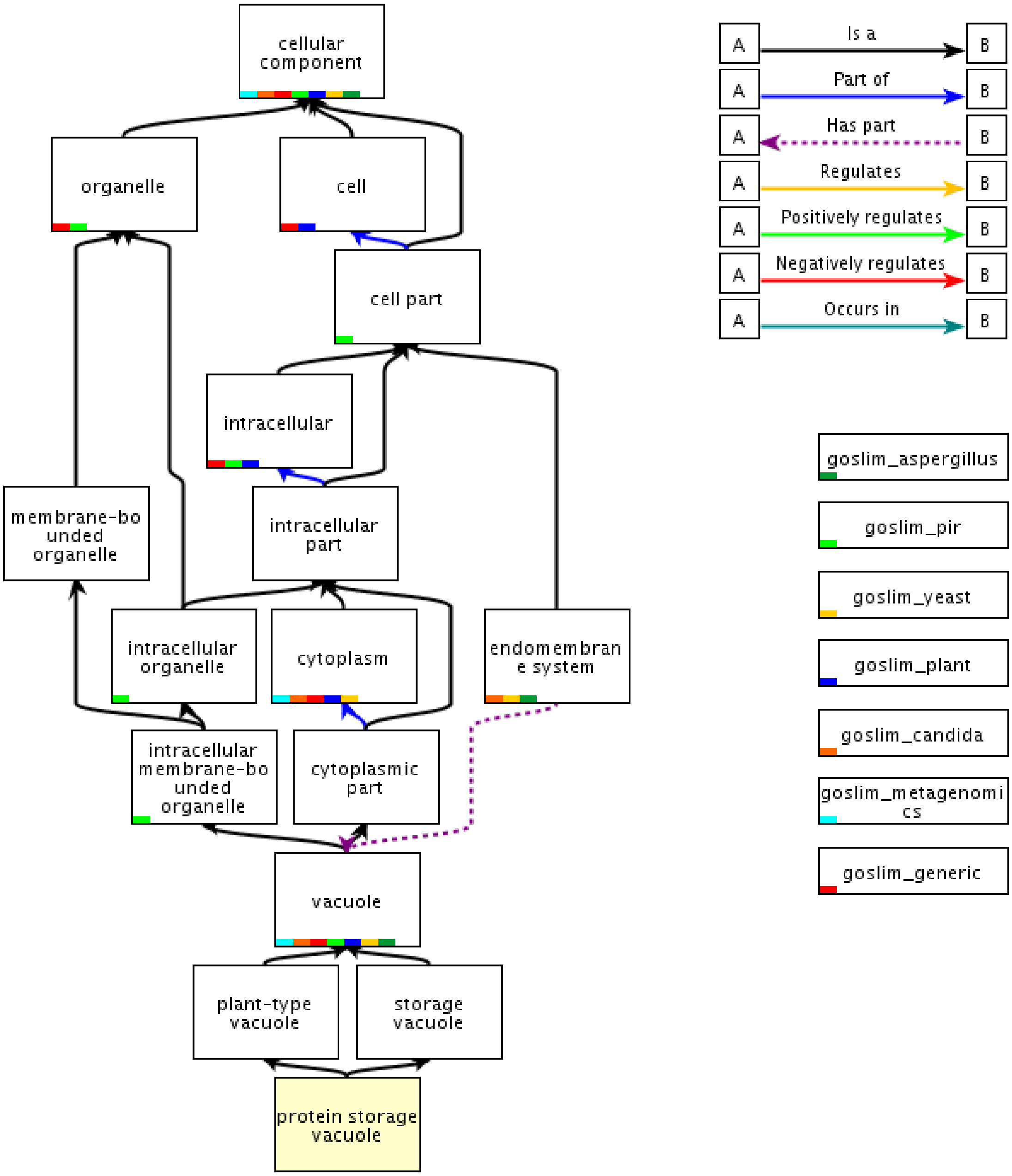
A thorough description of GO:0000326 can be accessed by clicking on "Details" given in the Accession No. column. The detailed description of term name represents Accession No, Ontology, Synonyms, Definition, Comment, Subset, Source, WikiURL as shown in Figure A.5.

Every term has a unique identifier and a name defined by the Gene Ontology Consortium [23, 243]. This was done in order to unify the information with other available databases. The Accession No. is prefixed with the GO to form a unique zero-padded seven digit identifier. Definition and Comment deal with added information about the term.

The Subset fields attribute shows up the GO slim term if it exists for the searched term. Source shows a reference from where the term is taken, e.g. a PubMed ID or some other reference. WikiURL indicates usage comments for the term on the Gene Ontology Normal Usage Tracking System (GONUTS) [259] having editable wiki for every GO entity for registered users, while the EMBL-EBI Link links to the European Molecular Biology Laboratory - European Bioinformatics Institute database in order to provide a wide background of information for researchers [23, 29, 243].

The description is followed by two tables containing the objects that are the ancestors and children of GO:0000326 in terms of their functional and ontological interrelationships, respectively. Besides offering the possibility to further browse and analyze each of the objects directly from the tables, it is important to note two additional links namely, "[+] View this term in EBI QuickGO (Sourced via AmiGO) [261]", "[+] View this term in directed acyclic graph form (Sourced via AmiGO)" being above and below the tables, respectively show the PNG (Portable Network Graphics) image [218].

By clicking on the upper link, a scheme illustrating functional relationships of the 14 ancestor Gene Ontologies of GO:0000326 pops up Figure A.6. Expanding this link provides a further link to a very useful resource 'EBI's QuickGO Reference Manual' [261] for purposes of information. From a physiological point of view, this type of scheme is quite handy for analyzing the locations and interactions of sub cellular structures and macromolecular complexes being in their active states, taking into account that the sub cellular structures are mainly gene products that have different biological and chemical functions. In other words, the expression of genes in temporally and spatially characteristic patterns is associated with the structurally distinct products that reside in specific cellular compartments and may be part of one or more multi-component complexes. By selecting the bottom link, another scheme featuring only the ontological relationships of the 14 ancestor GO terms shows up, see Figure. A.7.



QuickGO - http://www.ebi.ac.uk/QuickGO

Figure A.6 The functional relationships of the 14 ancestor Gene Ontologies of GO:0000326, elucidating the definitions, interactions and locations of sub cellular structures and macromolecular complexes being in their active states

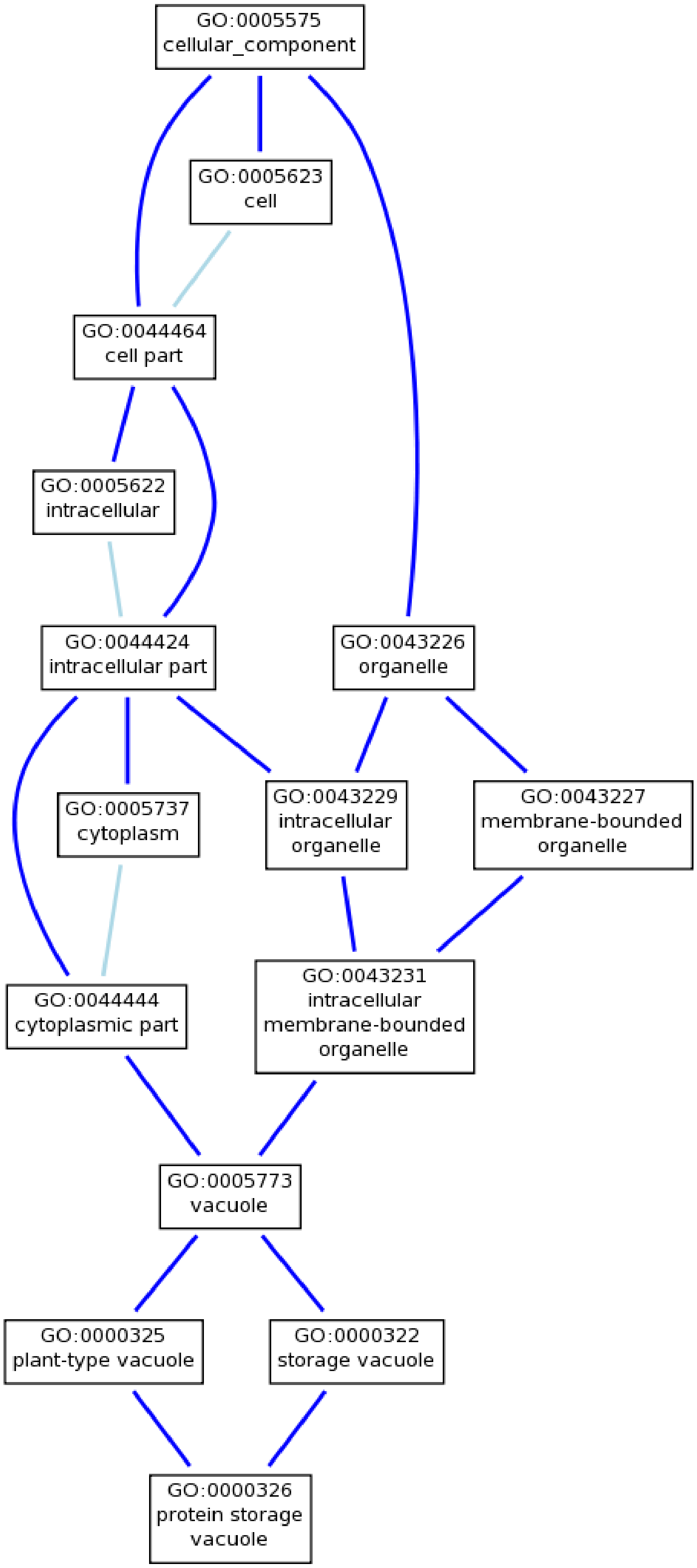


Figure A.7 Ontological relationships of the 14 ancestor GO terms of GO:0000326 in Directed Acyclic Graph form

This short example illustrates that the plant physiology of interest can be unambiguously represented using the PPDB in a computationally tractable way. By selecting adequate search options depending upon specific physiology, the PPDB is able to trace very intricate GO relationships in the sense of the cause-consequence considerations, thus substantially facilitating collection of novel insights into physiology of plants at a system level.

Annexure – B

NORMALIZED RESULTS

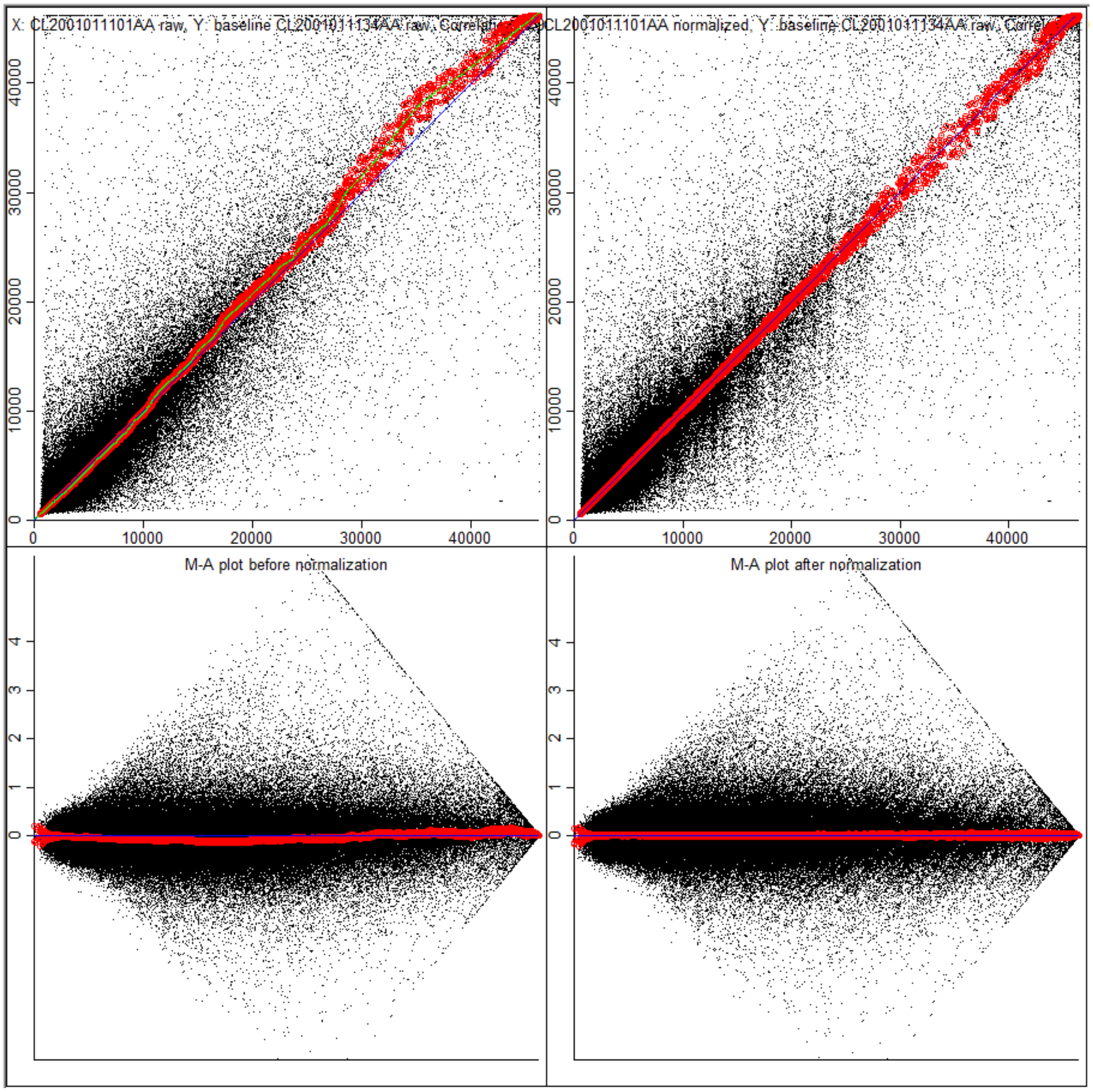


Figure B.1 Normalization result of CL2001011101AA

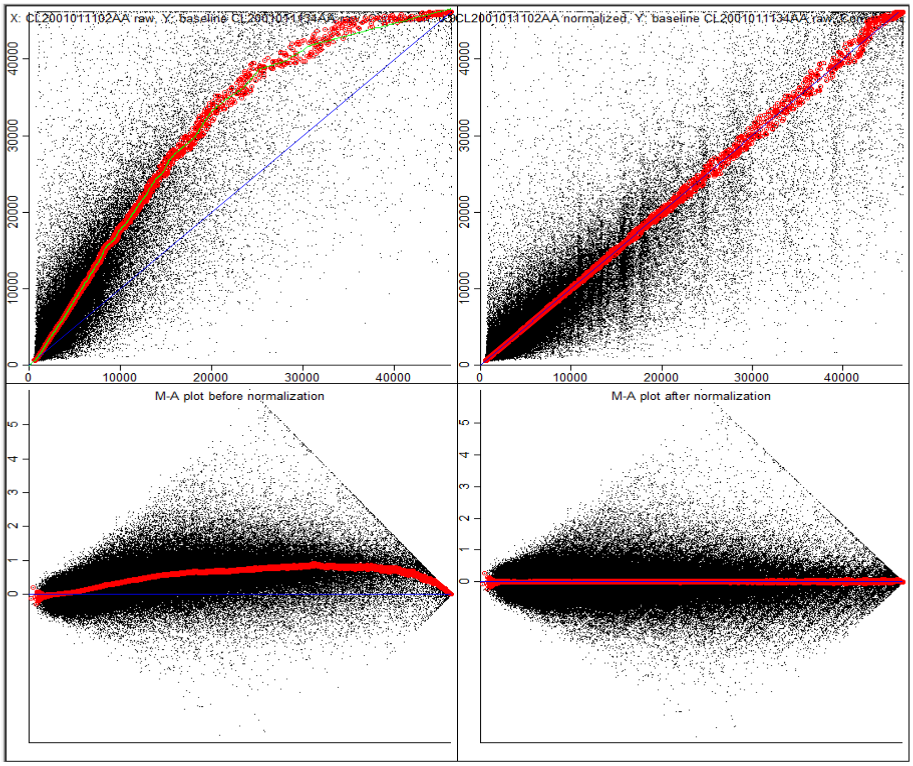


Figure B.2 Normalization result of CL2001011102AA

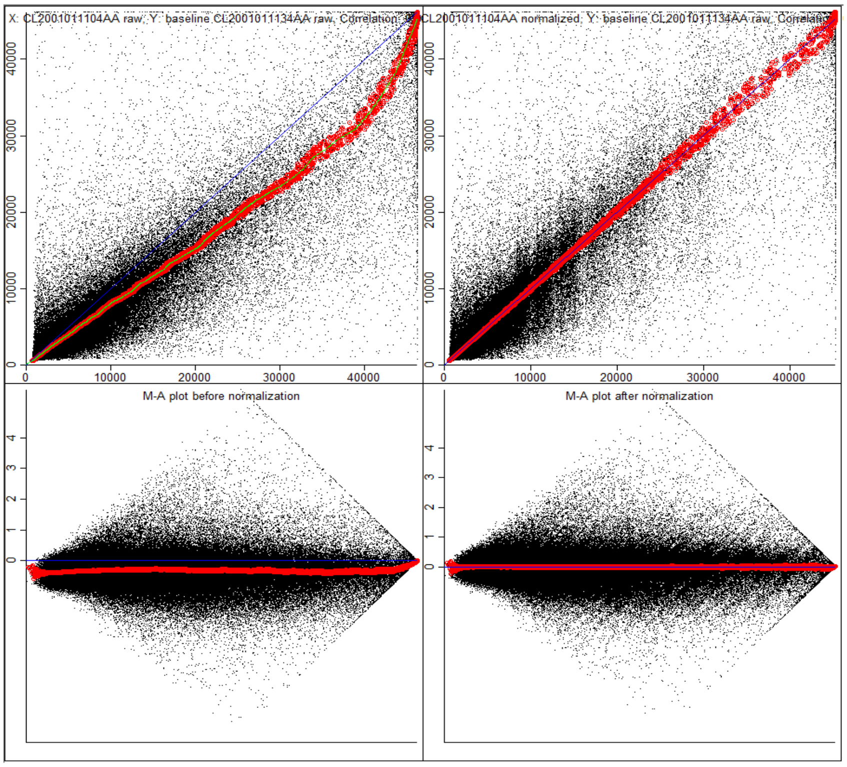


Figure B.3 Normalization result of CL2001011104AA

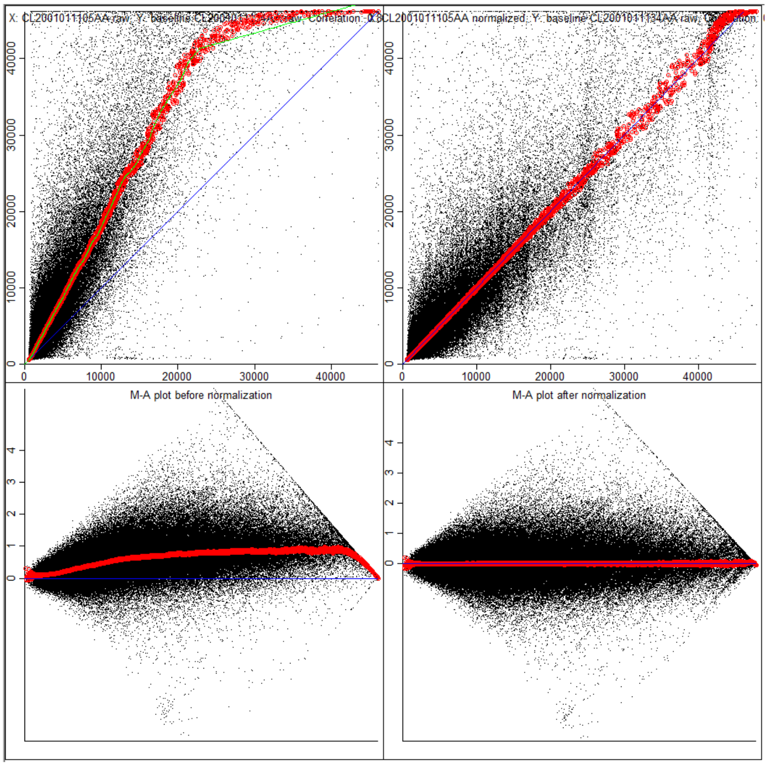


Figure B.4 Normalization result of CL2001011105AA

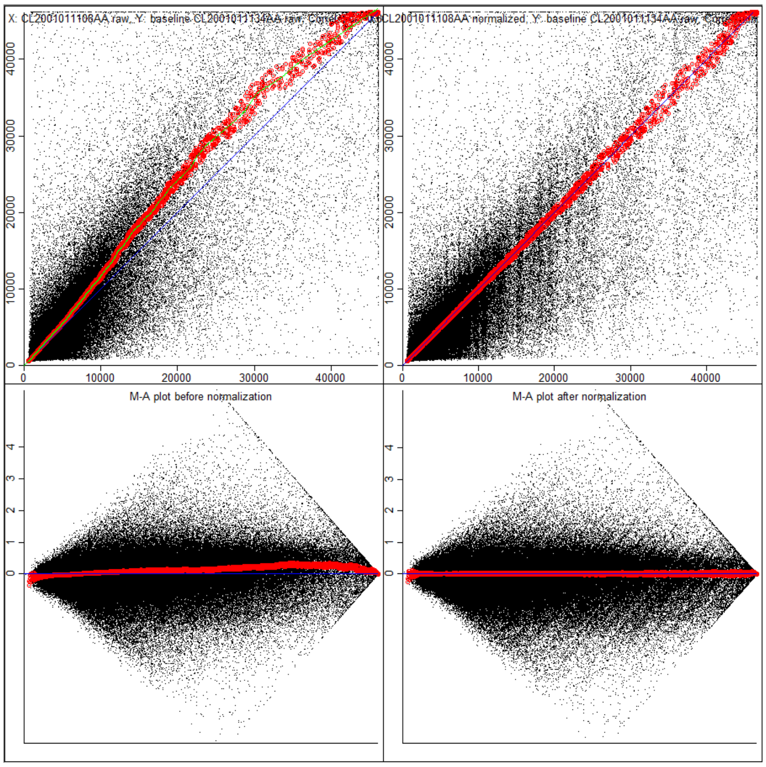


Figure B.5 Normalization result of CL2001011106AA

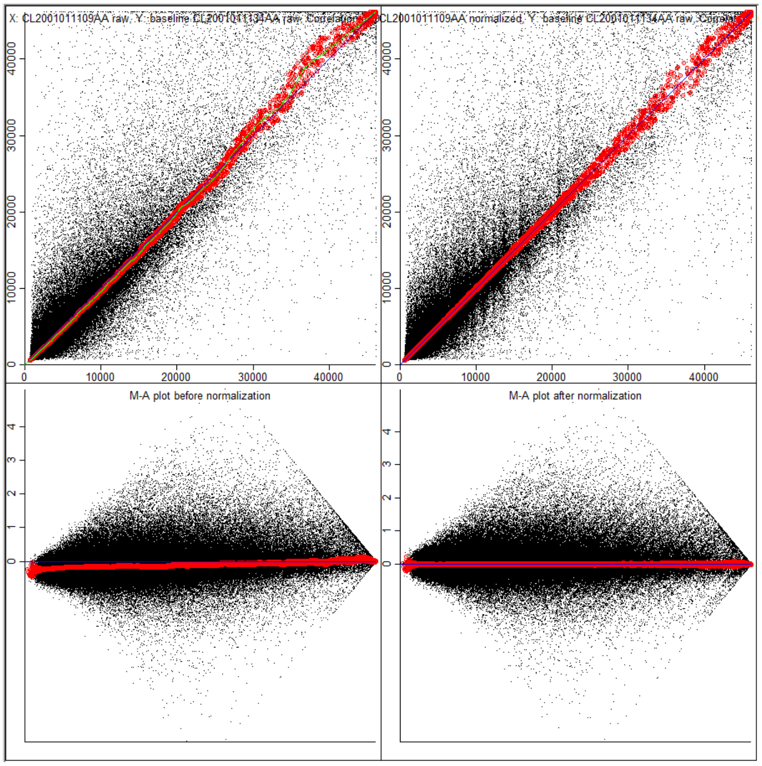


Figure B.6 Normalization result of CL2001011109AA

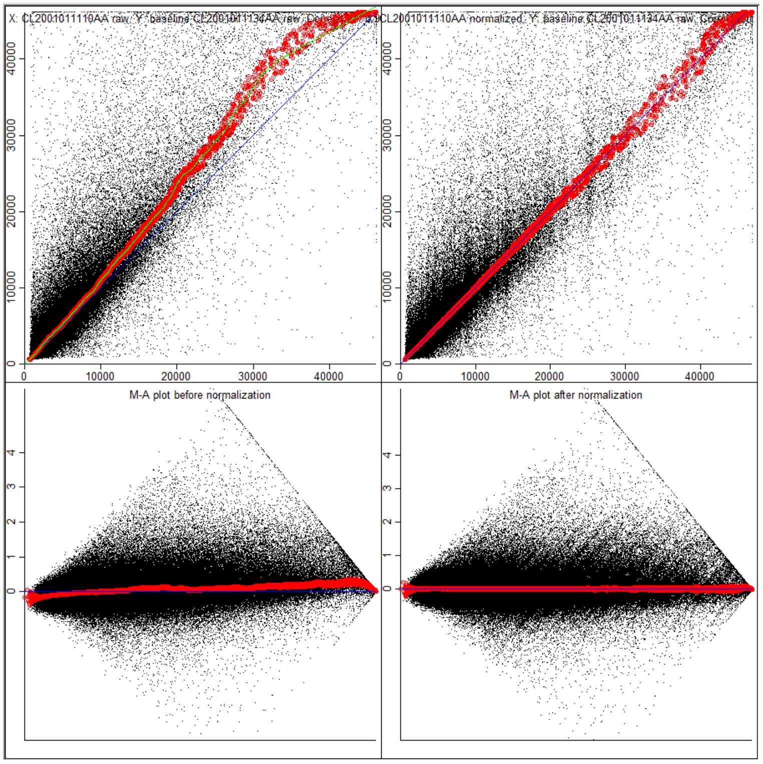


Figure B.7 Normalization result of CL2001011110AA

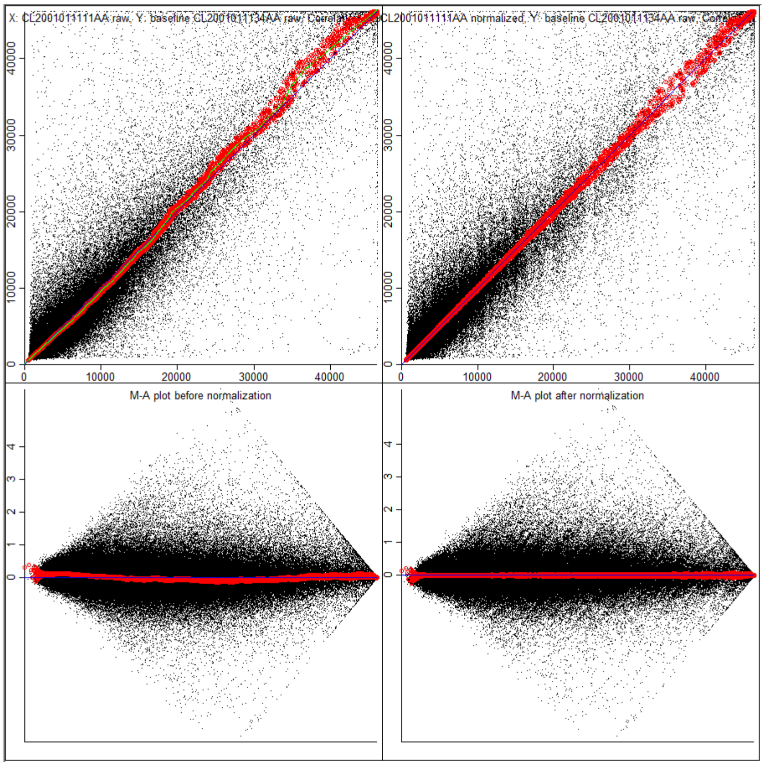


Figure B.8 Normalization result of CL2001011111AA

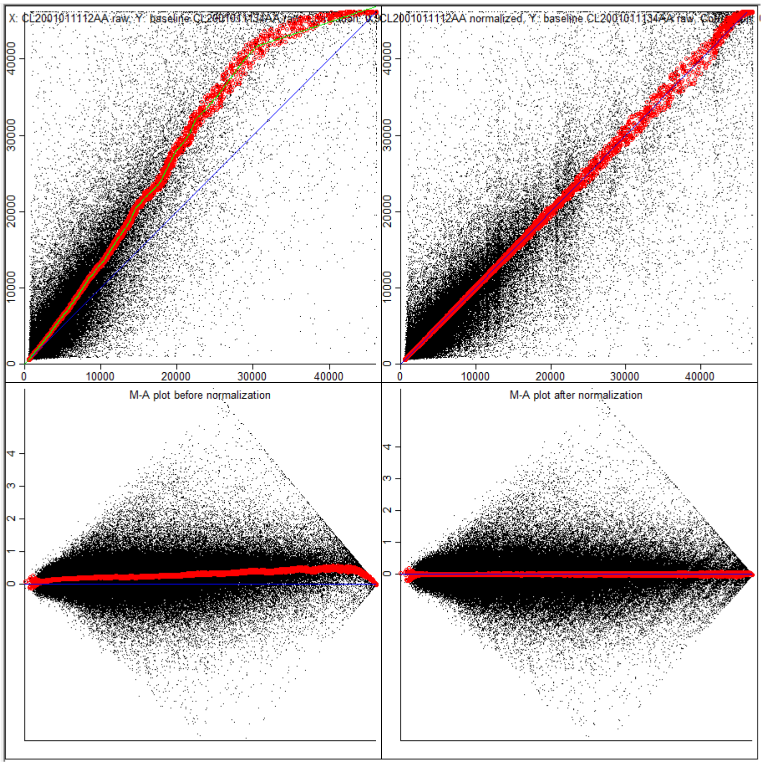


Figure B.9 Normalization result of CL2001011112AA

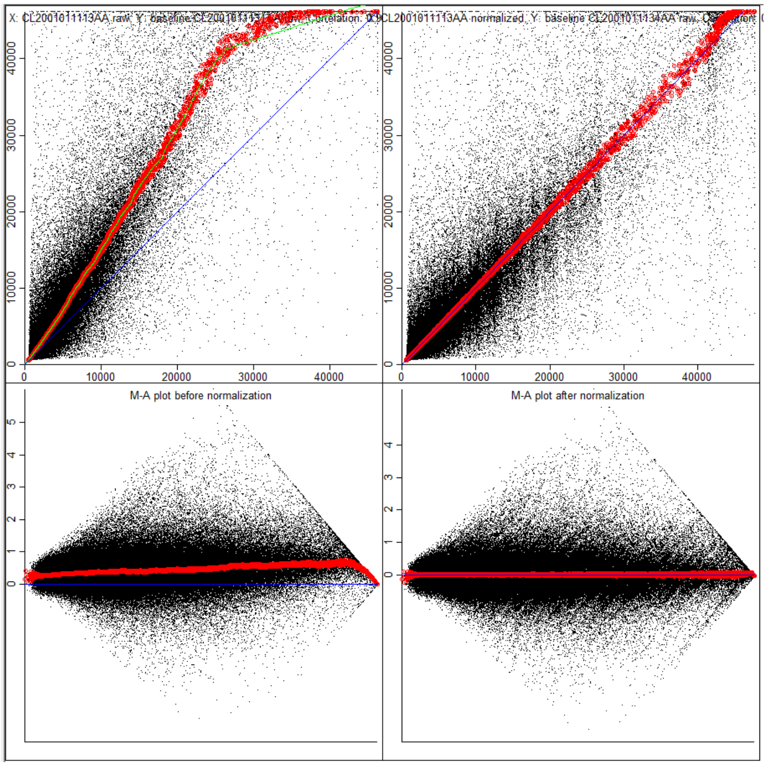


Figure B.10 Normalization result of CL2001011113AA

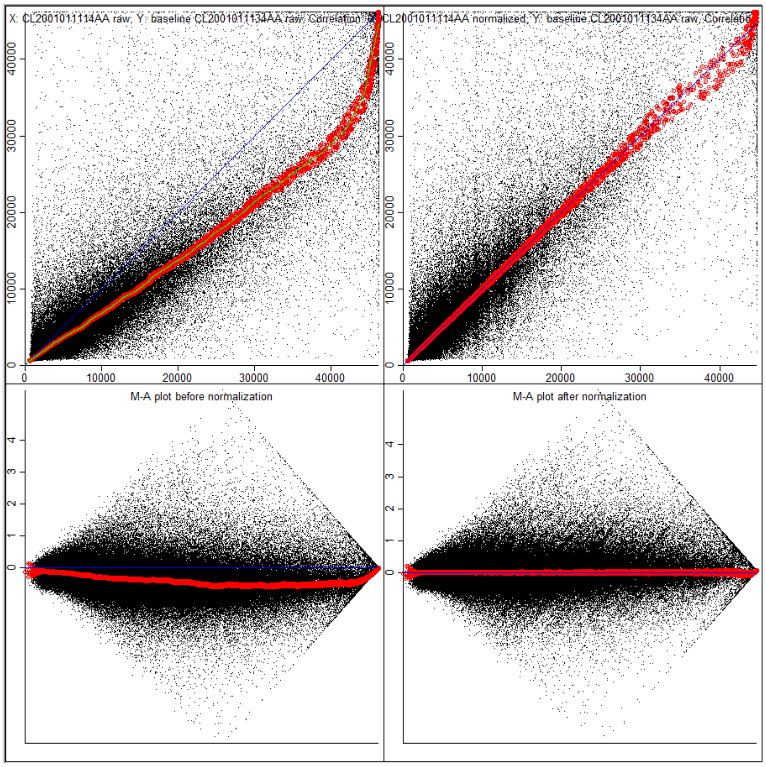


Figure B.11 Normalization result of CL2001011114AA

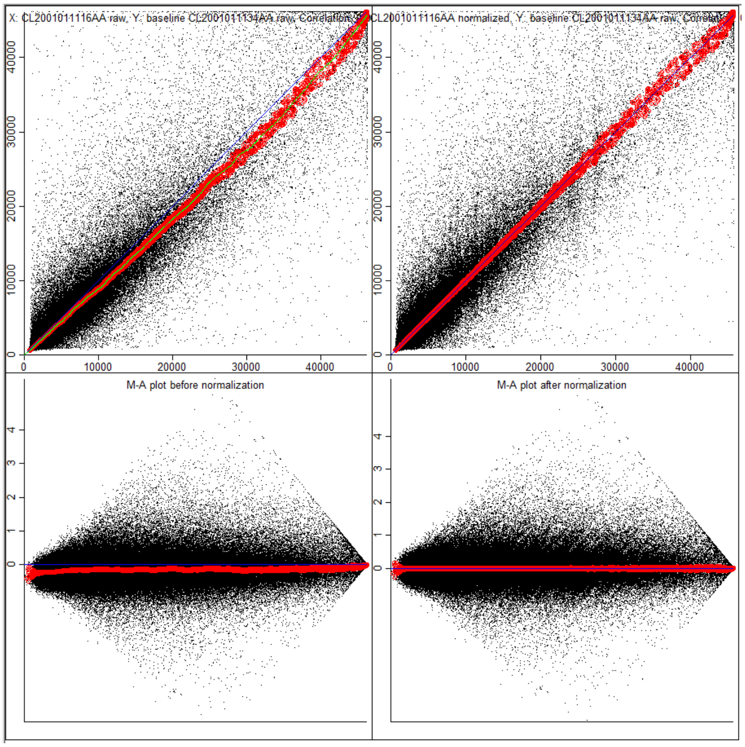


Figure B.12 Normalization result of CL2001011116AA

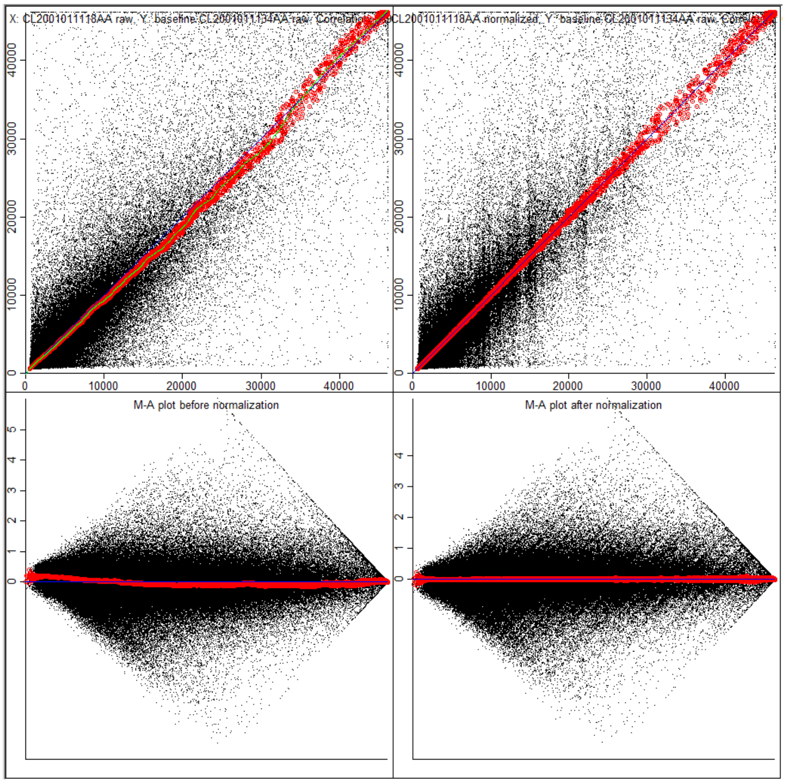


Figure B.13 Normalization result of CL2001011118AA

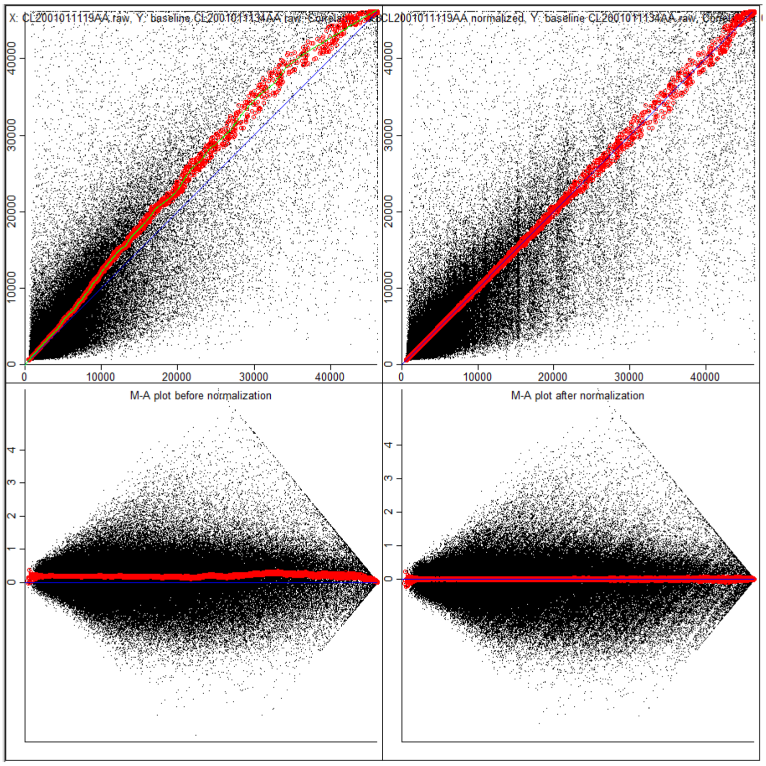


Figure B.14 Normalization result of CL2001011119AA

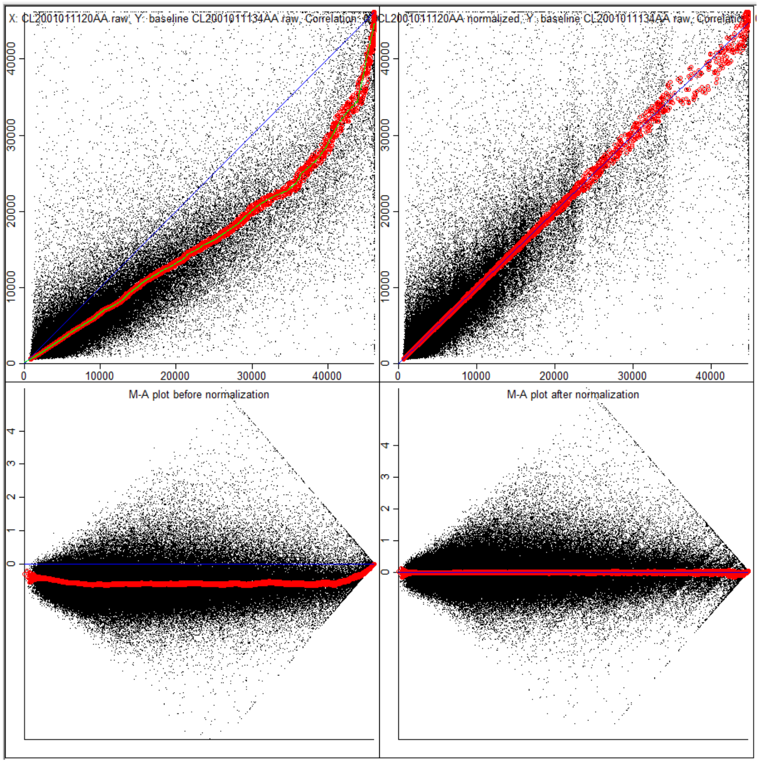


Figure B.15 Normalization result of CL2001011120AA

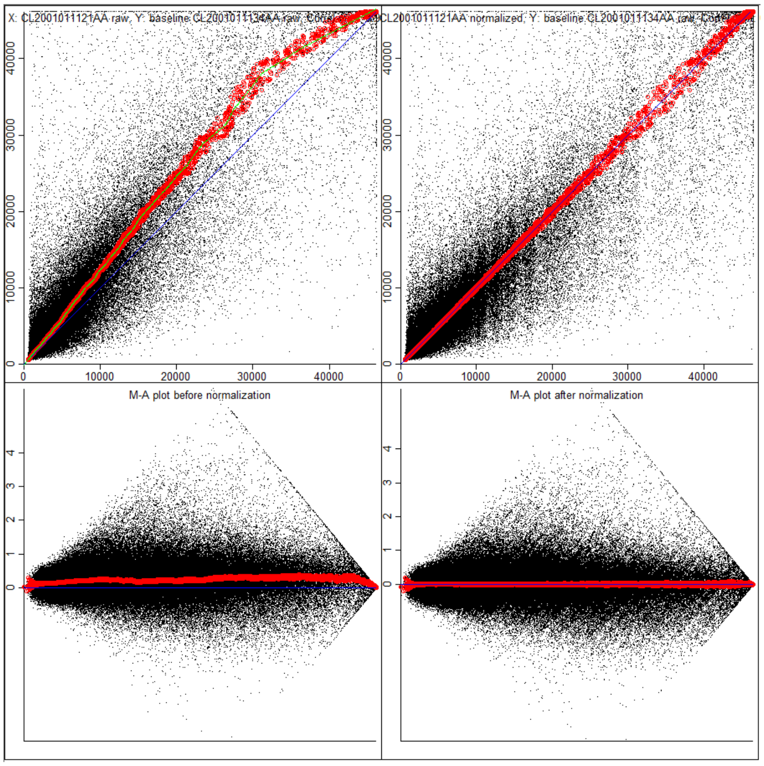


Figure B.16 Normalization result of CL2001011121AA

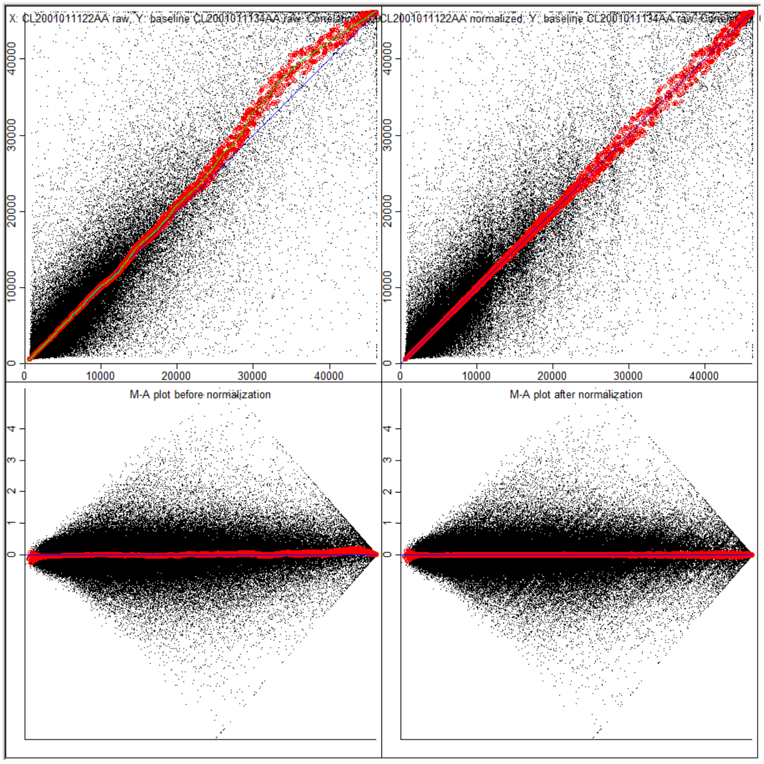


Figure B.17 Normalization result of CL2001011122AA

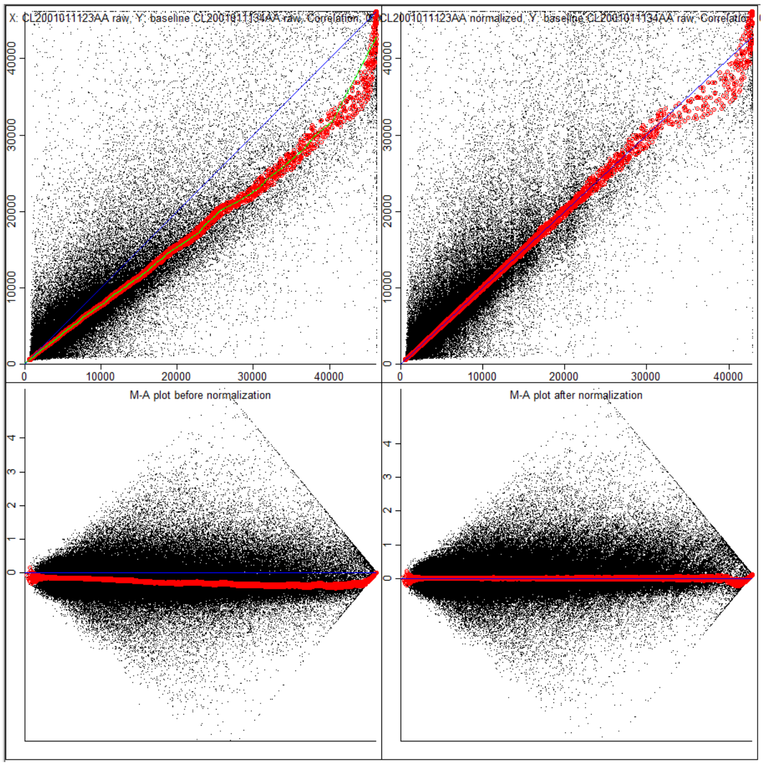


Figure B.18 Normalization result of CL2001011123AA

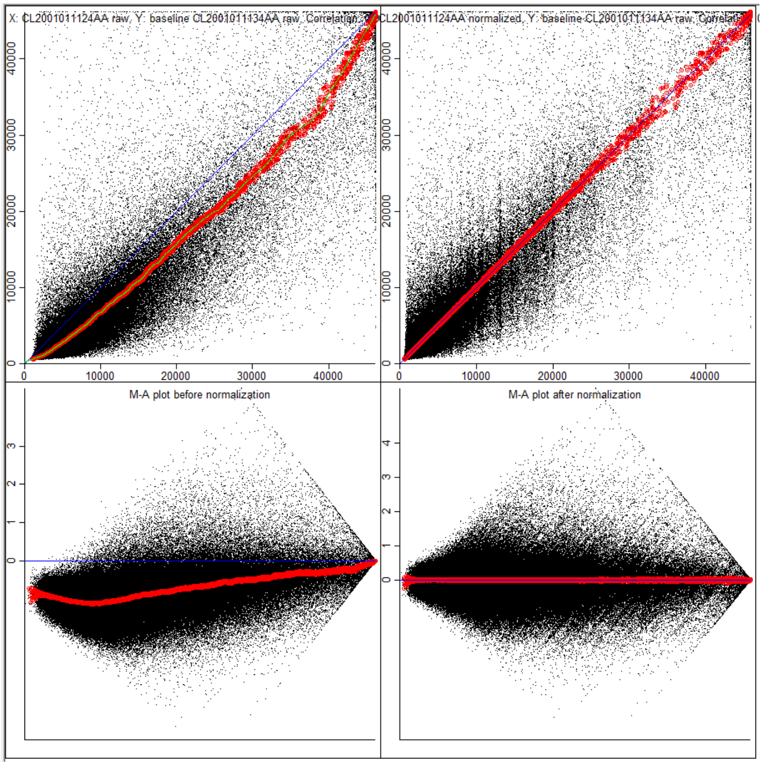


Figure B.19 Normalization result of CL2001011124AA

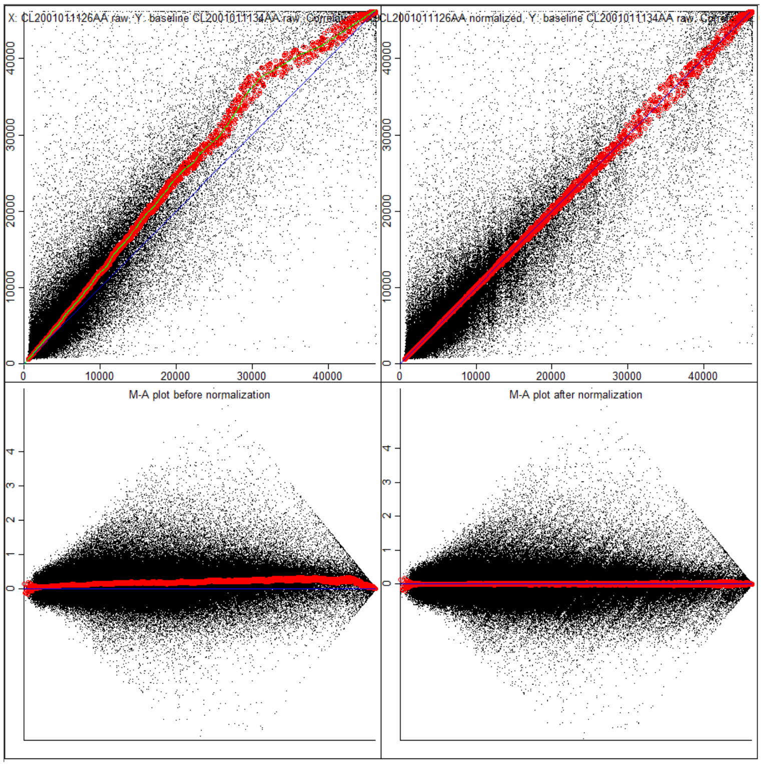


Figure B.20 Normalization result of CL2001011126AA

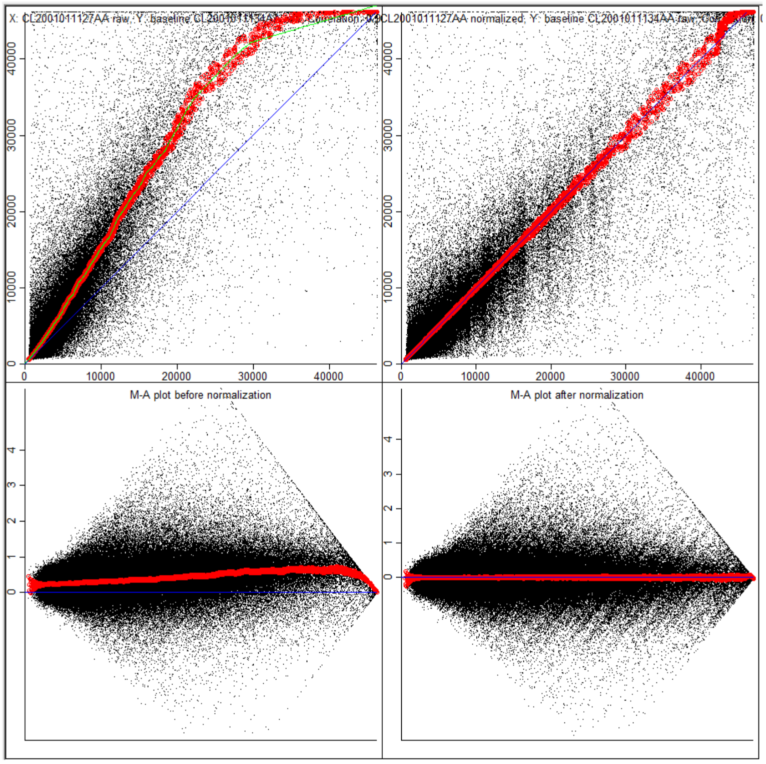


Figure B.21 Normalization result of CL2001011127AA

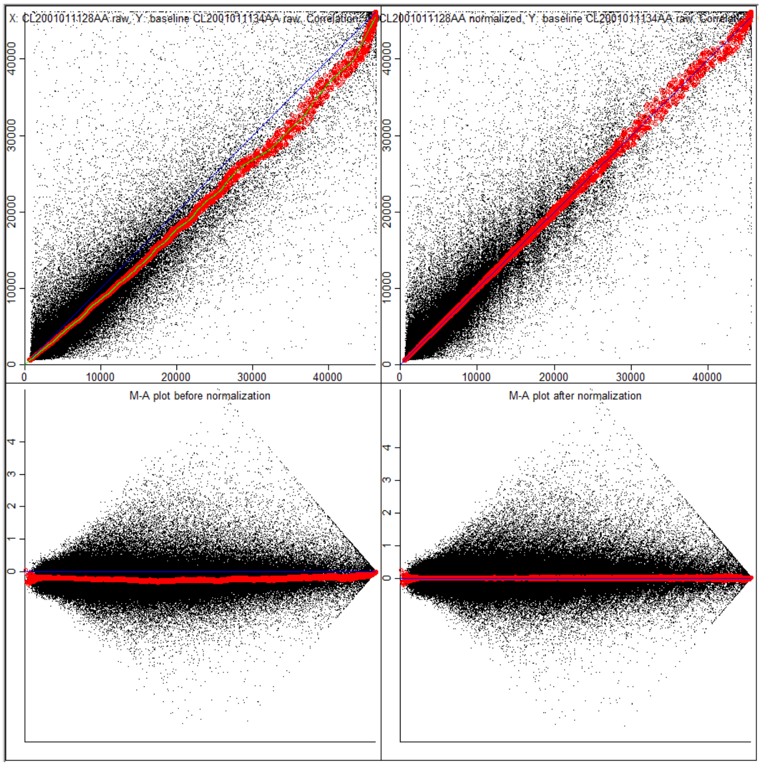


Figure B.22 Normalization result of CL2001011128AA

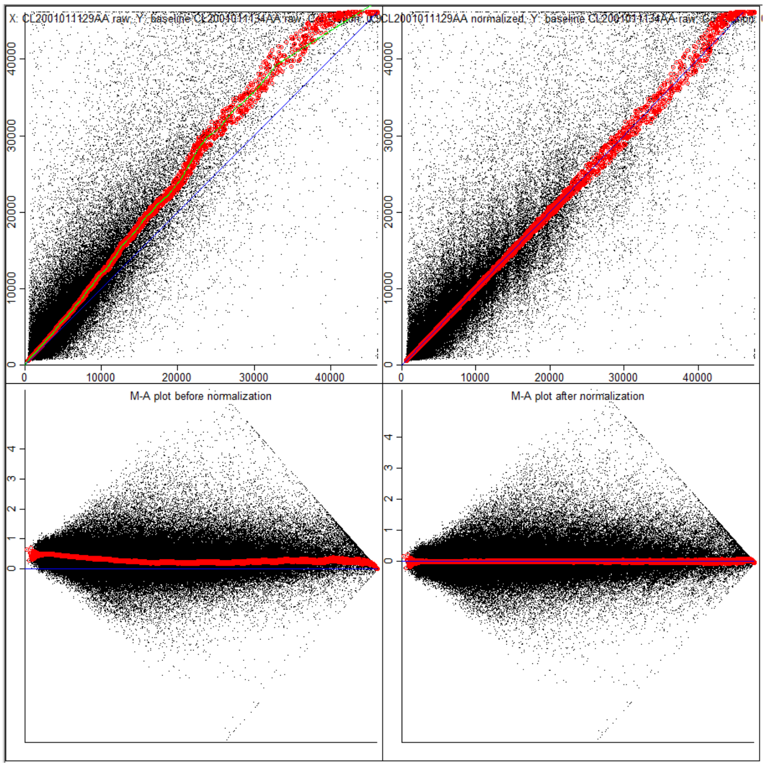


Figure B.23 Normalization result of CL2001011129AA

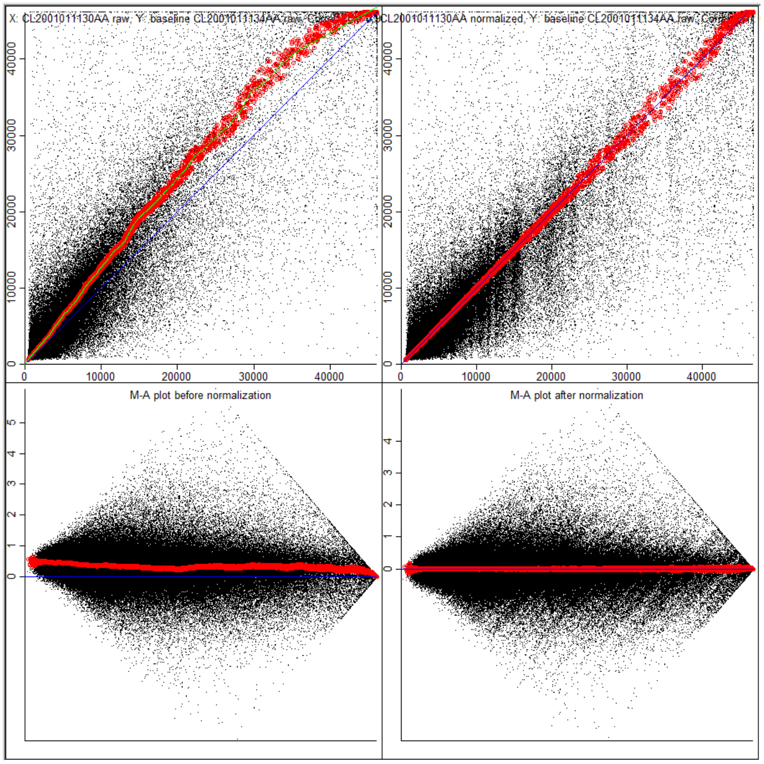


Figure B.24 Normalization result of CL2001011130AA

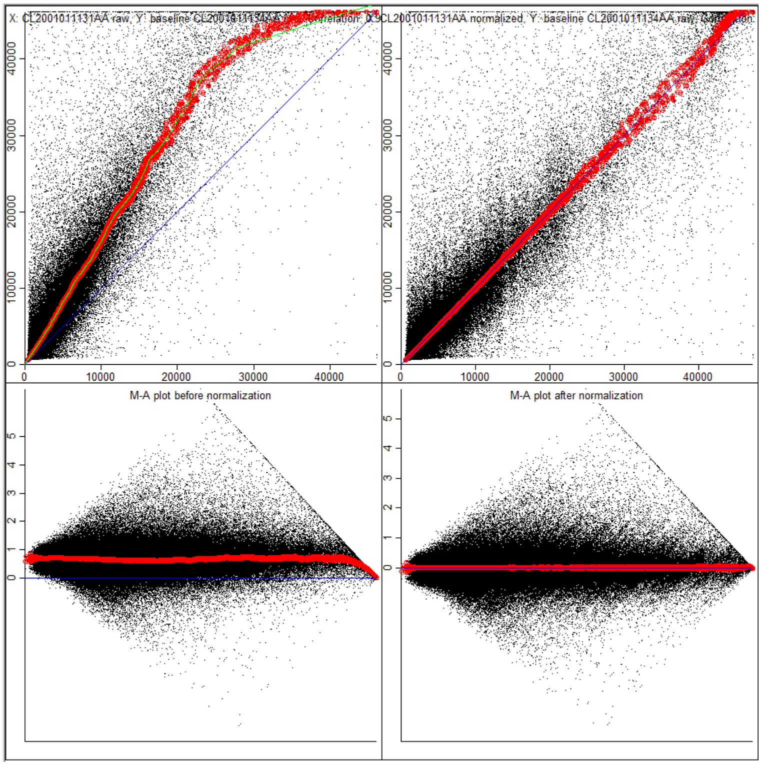


Figure B.25 Normalization result of CL2001011131AA

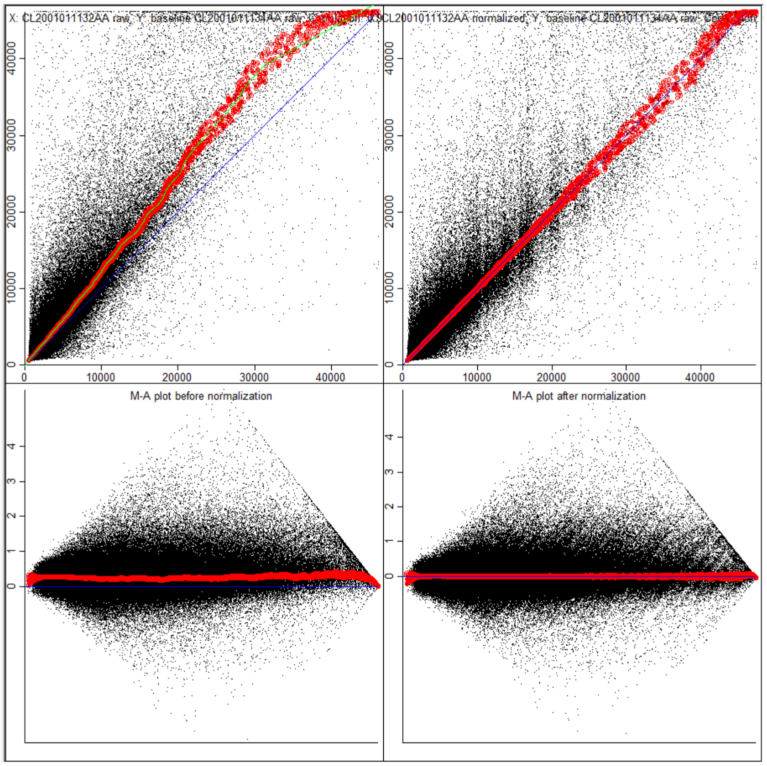


Figure B.26 Normalization result of CL2001011132AA

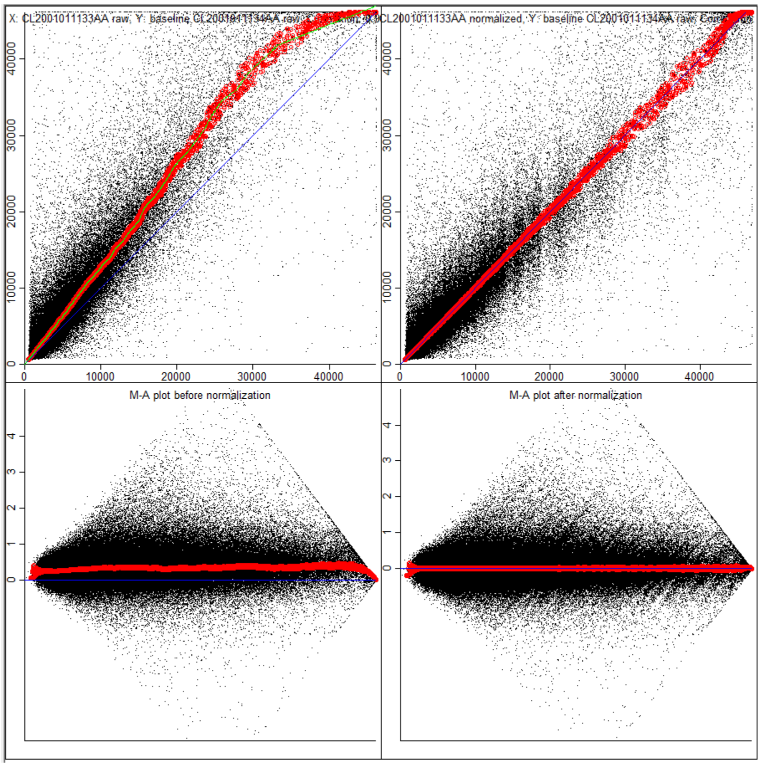


Figure B.27 Normalization result of CL2001011133AA

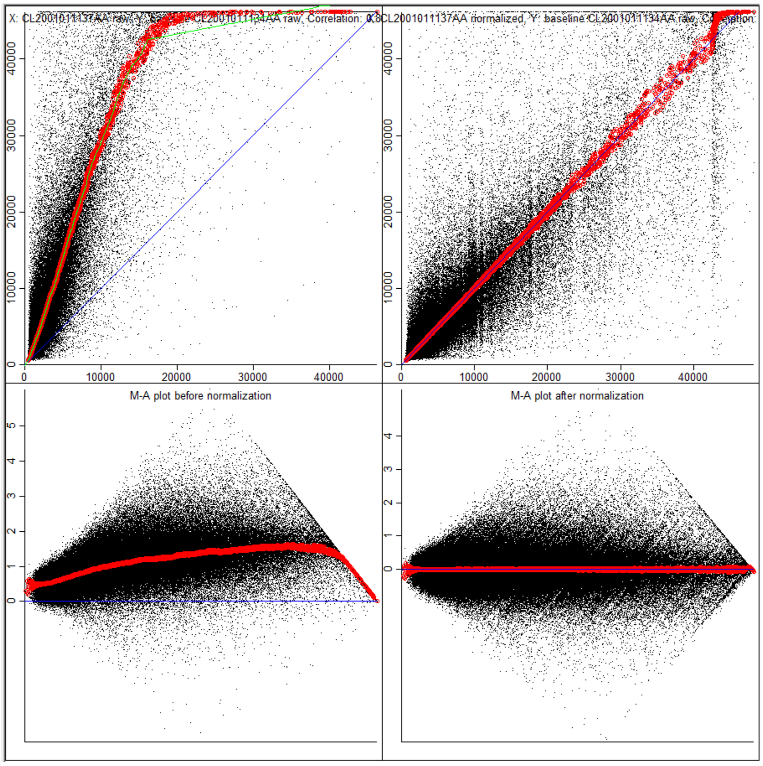


Figure B.28 Normalization result of CL2001011137AA

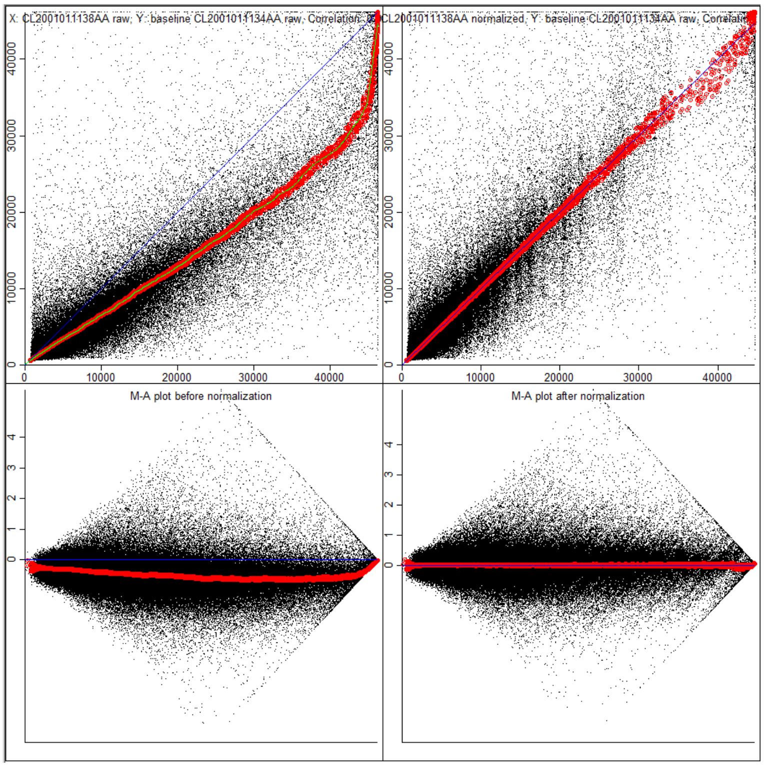


Figure B.29 Normalization result of CL2001011138AA

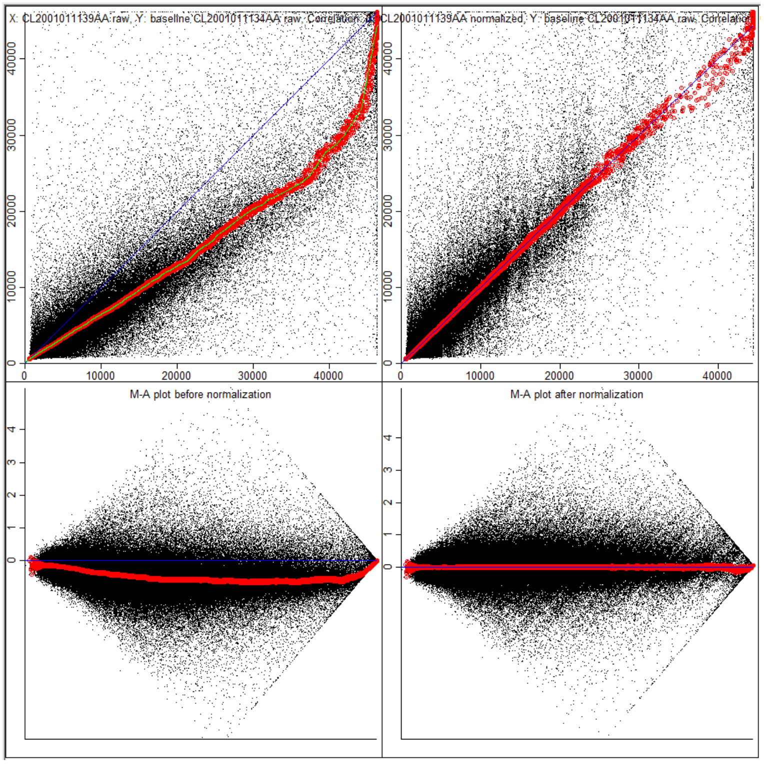


Figure B.30 Normalization result of CL2001011139AA

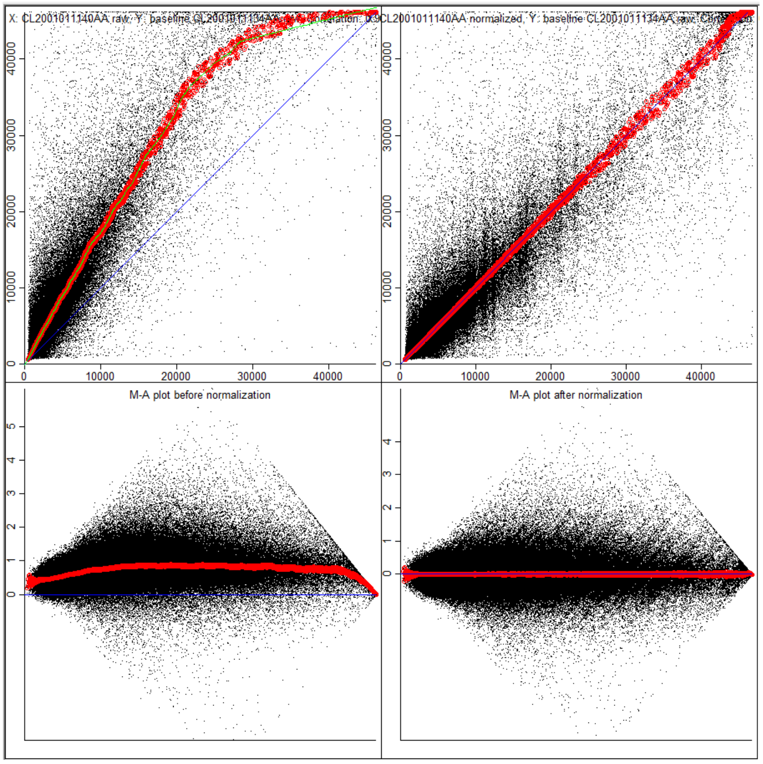


Figure B.31 Normalization result of CL2001011140AA

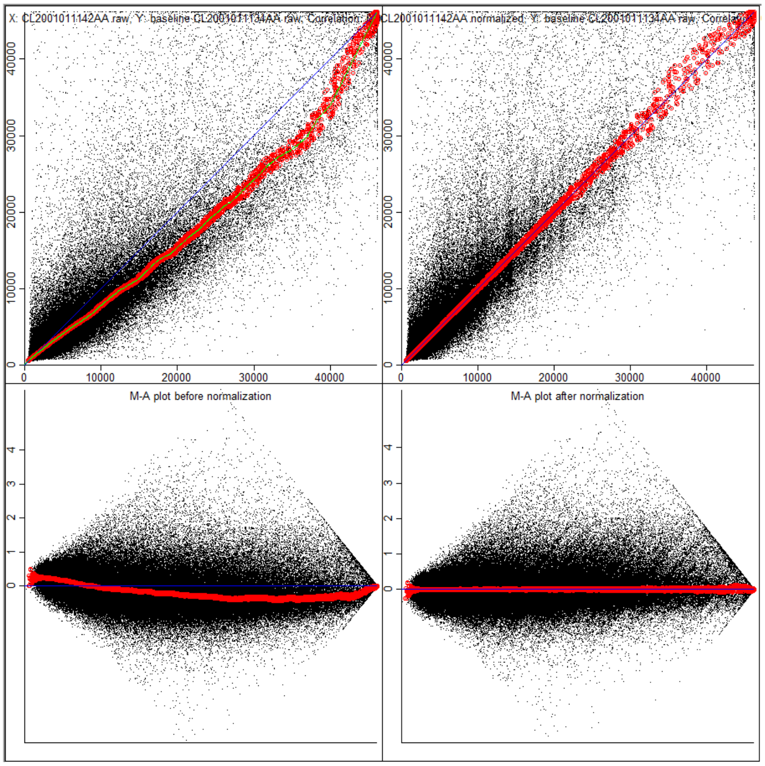


Figure B.32 Normalization result of CL2001011142AA

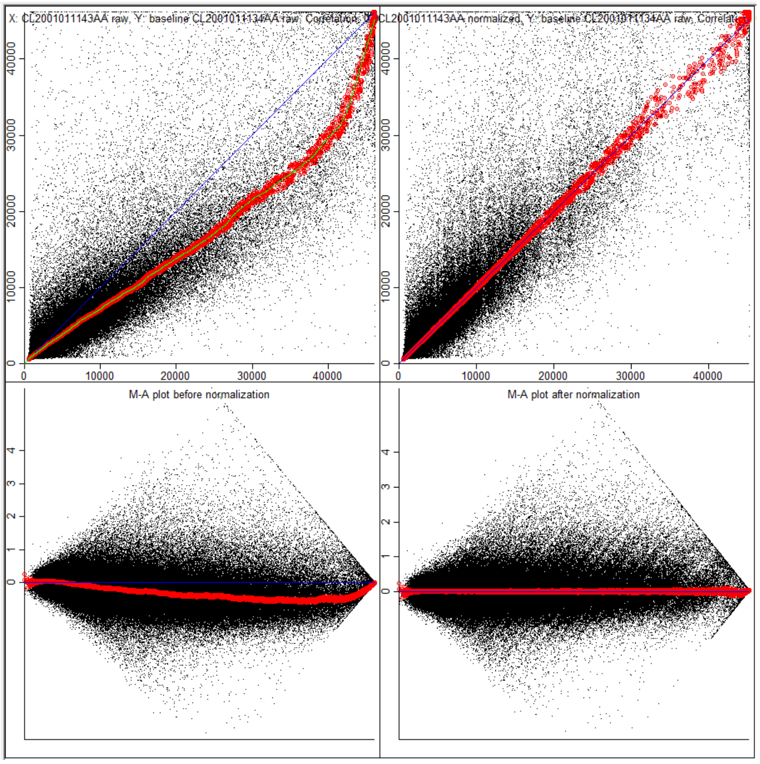


Figure B.33 Normalization result of CL2001011143AA

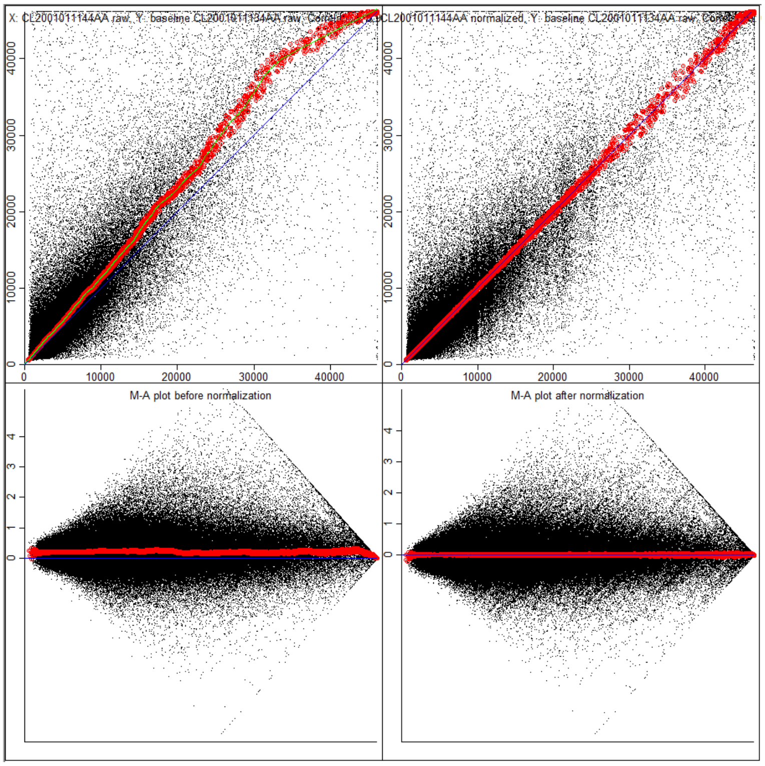


Figure B.34 Normalization result of CL2001011144AA

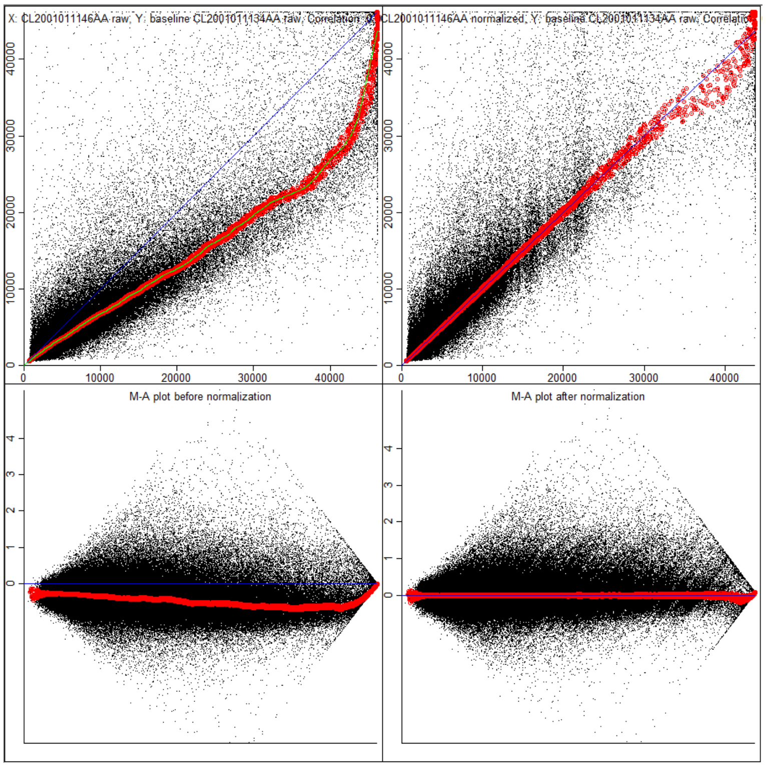


Figure B.35 Normalization result of CL2001011146AA

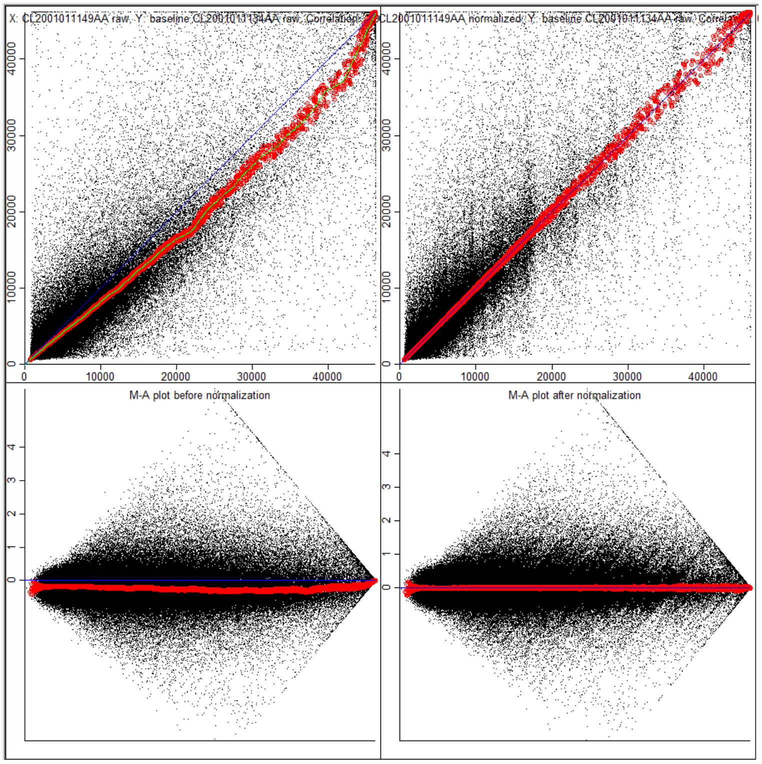


Figure B.36 Normalization result of CL2001011149AA

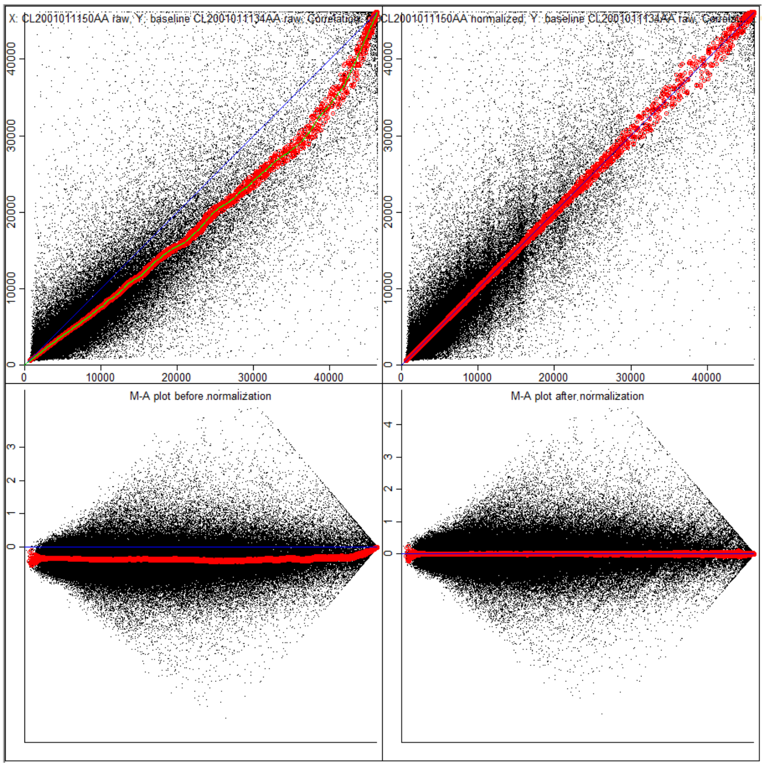


Figure B.37 Normalization result of CL2001011150AA

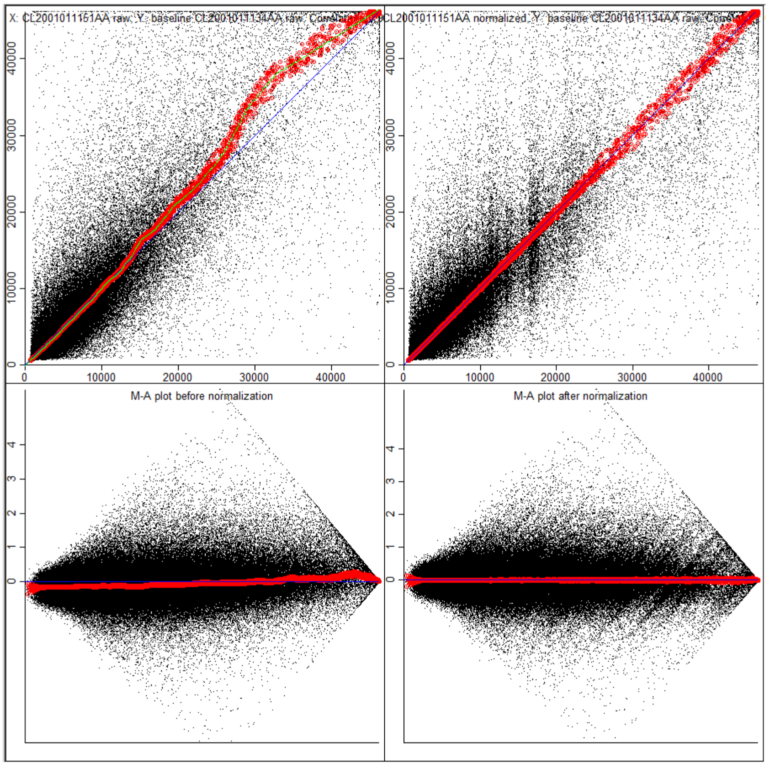


Figure B.38 Normalization result of CL2001011151AA

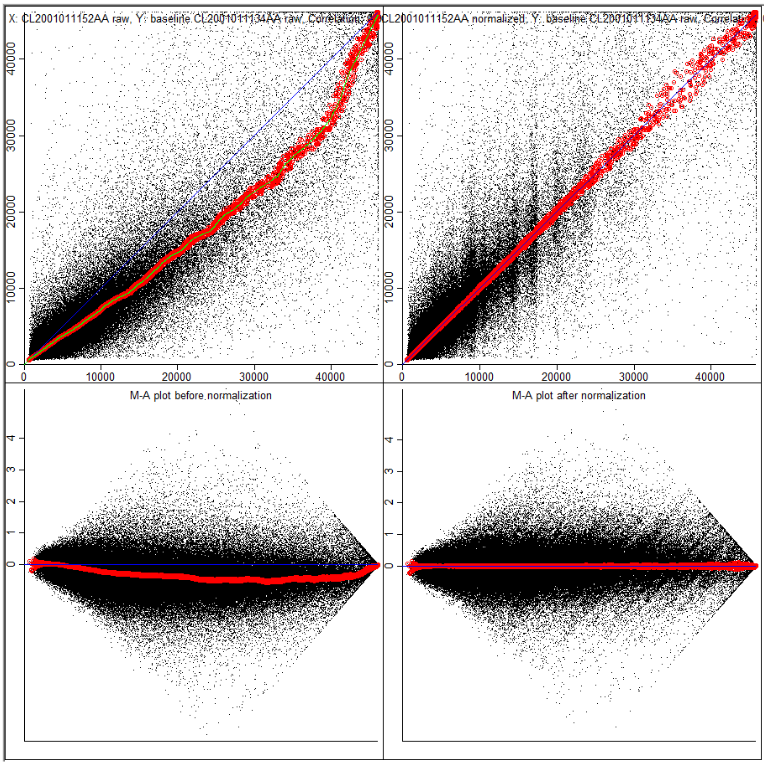


Figure B.39 Normalization result of CL2001011152AA

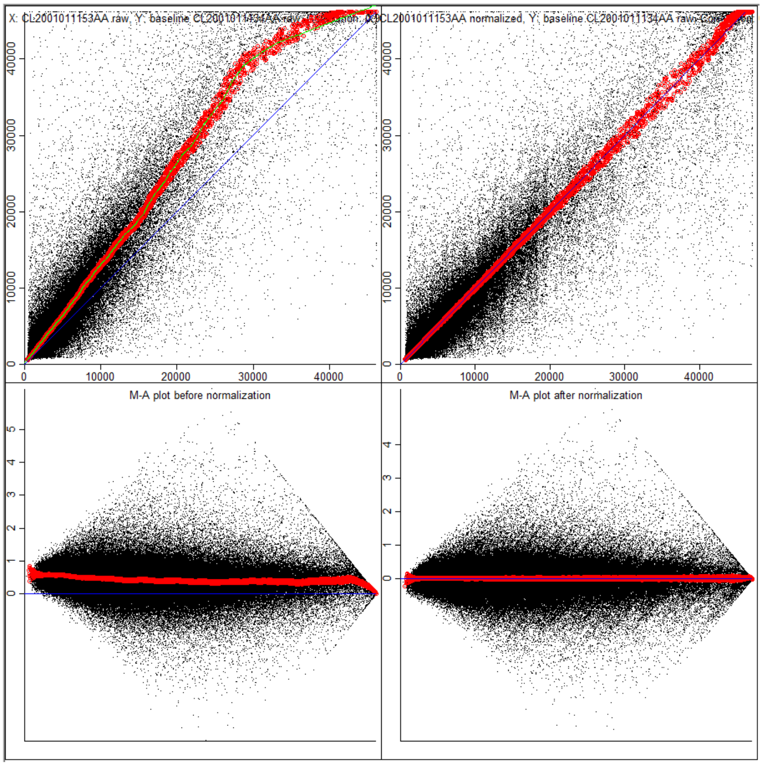


Figure B.40 Normalization result of CL2001011153AA

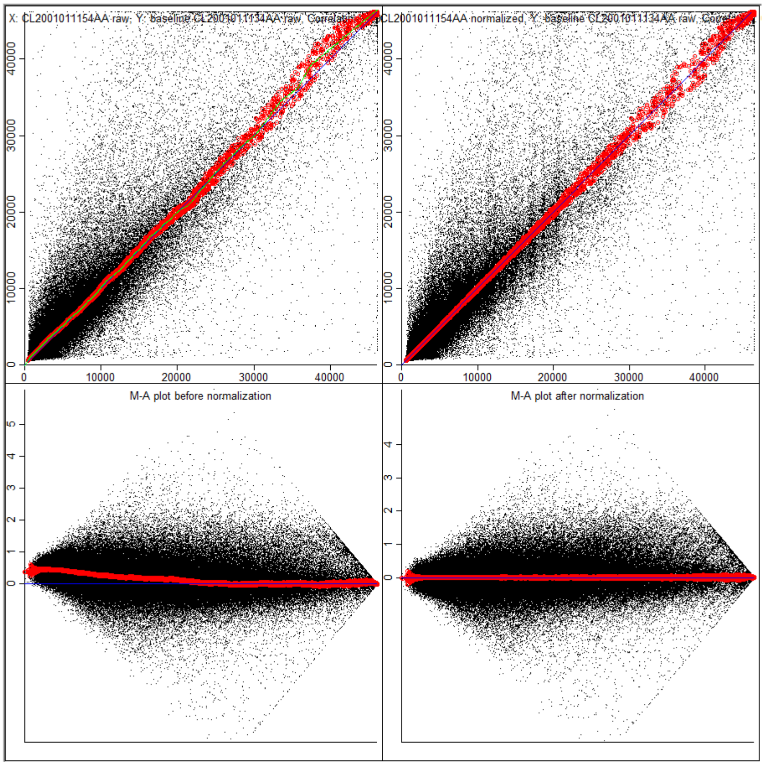


Figure B.41 Normalization result of CL2001011154AA

Annexure – C

CONNECTING THE DOTS

The heading of this Annexure is taken from the first story told by Steve Jobs, ex CEO of Apple Computer and Pixar Animation Studios, on June 12, 2005 at Stanford University, where he said:

“Again, you can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future. You have to trust in something - your gut, destiny, life, karma, whatever. This approach has never let me down, and it has made all the difference in my life.”

These motivational words give me direction during thesis writing. The initial goal of my doctoral research is to study contemporary applications of database management in bioinformatics, but as the work progressed, new research interests were also evolved. Henceforth, development of gene ontology data mining tool using contemporary bioinformatics and systems biology of cancer became the central issue. The work progressed in four distinct but logically connected parts, each with a different focus, different results, and directed by different supervisors. As such, it was anticipated that contemporary bioinformatics and computational systems biology applications will expand in innovative directions. Such new disciplines have fuzzy boundaries due to rapidly evolving topics. Existing scientific communities join the new field, exert various influences and shape it in sometimes unexpected ways. These directions will be dictated by biological research questions of high clinical relevance, and by progress in investigation technologies.

While connecting the dots, recent genomic studies disclose association between human genetics and plant genetics in the book “What a plant knows: a field guide to the senses” [355] where researcher predicts that genome wide similarity studies on plants genes could unveil biological complexities of cancer besides lead to behavioural studies in animals. Such similarity studies have great impact and can accelerate medical research initiatives where plants can be used for clinical trials on cellular level. In the current year 2014, published studies on sea anemone [356, 357] found that the complex interaction network of genes in anemone bear a resemblance to the genes present in the animals and therefore sea anemone can be treated as half plants and half animals.

The ideas presented in this thesis have not become mainstream so far, and still matters of current research but we can anticipate that they will soon bring important insights in all living forms leading to plants animal continuum gene based studies. This requires a holistic-integrative approach and meta analysis across different omics databases. Scientists are cognisant of ongoing research and assets required in their respective field, yet in the form of modular pieces of a puzzle as illustrated in Figure C.1. New breed of researchers who believe in open source philosophy for the betterment of society are needed to integrate biological pieces of puzzle that can be easily accessible. HPC infrastructure such as cluster, cloud or grid computing is definitely required for data storage, data transfer, data computation and access control. Such massively parallel infrastructure allows many tasks to be run simultaneously in order to reduce the effective computation time. Efficient bioinformatics definitely relies on strong information and technology support. To conclude, integration of many layers of gene based information is necessary and really important to understand the holy grail of living systems as shown in Figure C.1.

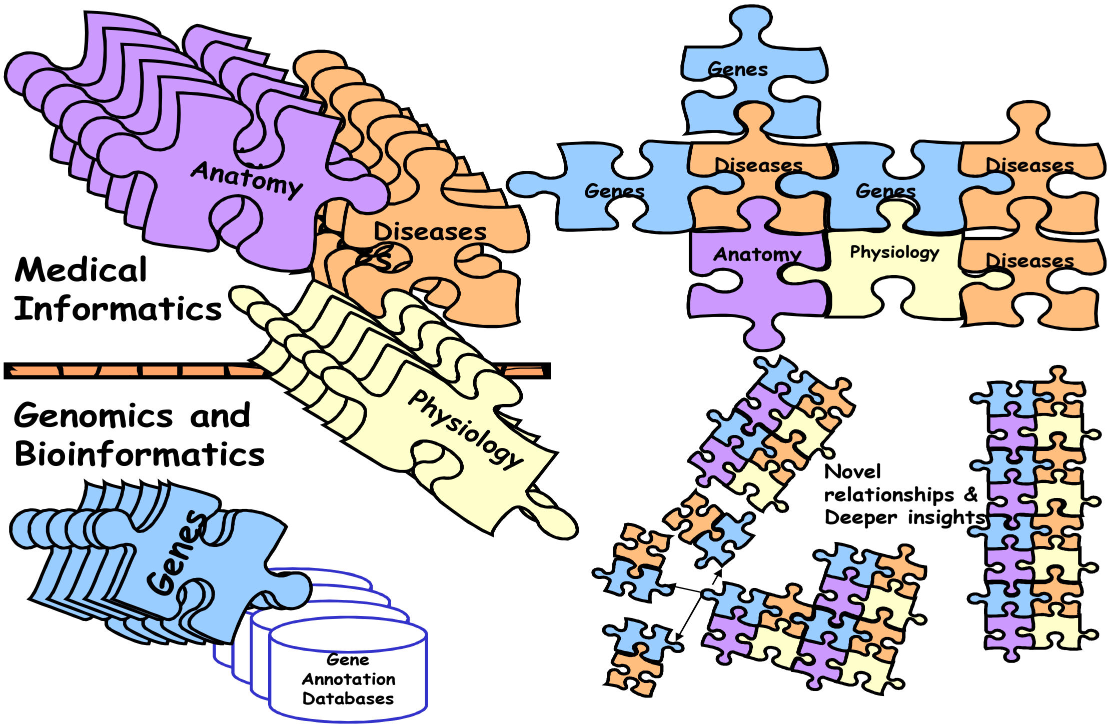


Figure C.1 Linking biological pieces of puzzle for novel relationships and deeper insights