ON BOOLEAN MODELING OF GENE REGULATORY NETWORKS FOR IMPROVED CANCER COMBINATORIAL THERAPY DESIGN AND TRANSCRIPTOME ASSEMBLIES FOR PACIFIC WHITELEG SHRIMP

A Dissertation

by

OSAMA ALI ARSHAD

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Aniruddha Datta
Committee Members,	Shankar P. Bhattacharyya
	Ulisses Braga-Neto
	Charles D. Johnson
Head of Department,	Miroslav M. Begovic

May 2017

Major Subject: Electrical Engineering

ABSTRACT

Cancer cells are known to exhibit atypical metabolic characteristics. While alterations in tumor cell metabolism are necessary for the sustained uncontrolled cell growth that characterizes cancer, it is also a vulnerability which can be exploited to design therapies that preferentially target cancer cells. We develop a testable theoretical framework for cancer therapy design which is used to elucidate a role for the metabolism targeting anti-diabetic drug Metformin as part of a combination cocktail therapy that could potentially provide better and less toxic clinical outcomes.

Castration-resistant prostate cancer is an advanced form of prostate cancer with limited treatment options where patients become refractory to surgical or medical castration. We use Boolean logic modeling of the key signaling pathways implicated in the development and progression of this malignancy to simultaneously test various combinations of agents for their efficacy in attenuating cancer growth and design targeted therapies for the management of the disease. Furthermore, stochastic computational modeling is utilized to identify potentially vulnerable components in the network that may serve as viable candidates for drug development.

Finally, we present novel transcriptome assemblies and functional annotations for Pacific whiteleg shrimp, a non-model organism of significant economic import that lacks solid transcriptome and genome references. In addition, as evaluating the quality of de novo transcriptome assemblies has proven to be challenging, we propose a pipeline comprising multiple quality check metrics that in unison provide a clear evaluation of assembly performance.

DEDICATION

To my family

ACKNOWLEDGEMENTS

First and foremost I would like to express the sincerest of gratitude to my adviser Professor Aniruddha Datta for his invaluable mentorship throughout the doctoral program. I am also indebted to Dr. Charles Johnson for his support and guidance. I would also like to thank Professor Ulisses Braga-Neto and Professor Shankar Bhattacharyya for serving on my dissertation committee. Finally, I would like to acknowledge my friends and colleagues in the genomic signal processing lab, past and present, particularly Dr. Sriram Sridharan, Dr. Anwoy Mohanty, Bibhu Mishra, Priya Venkat, Hyundoo Jeong and Xingde Jiang.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

The work was supported by a dissertation committee consisting of Professor Aniruddha Datta (advisor), Professor Ulisses Brag-Neto and Professor Shankar Bhattacharyya of the Department of Electrical & Computer Engineering and Dr. Charles Johnson of the Department of Molecular & Environmental Plant Sciences.

The analyses in Section 4 were conducted in part by Hyundoo Jeong of the Department of Electrical & Computer Engineering and Dr. Noushin Ghaffari of AgriLife Genomics and Bioinformatics Services.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

This work was funded in part by the National Science Foundation under grants ECCS-1068628 and ECCS-1404314.

Graduate study was supported by a research assistantship from the Texas Engineering Experiment Station (TEES)-AgriLife Center for Bioinformatics and Genomics Systems Engineering, Texas A&M University.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION[1]

The uncontrolled cell proliferation that is characteristically associated with cancer is usually accompanied by alterations in the genome and cell metabolism. Indeed, the phenomenon of cancer cells metabolizing glucose using a less efficient anaerobic process even in the presence of normal oxygen levels, termed the "Warburg effect", is currently considered to be one of the hallmarks of cancer. Diabetes, much like cancer, is defined by significant metabolic changes. Recent epidemiological studies have shown that diabetes patients treated with the anti-diabetic drug Metformin, have significantly lowered risk of cancer as compared to patients treated with other anti-diabetic drugs. In Section 2, we utilize a Boolean logic model of the pathways commonly mutated in cancer to not only investigate the efficacy of Metformin for cancer therapeutic purposes but also demonstrate how Metformin in concert with other cancer drugs could provide better and less toxic clinical outcomes as compared to using cancer drugs alone.

Prostate cancer is one of the most prevalent cancers in males in the United States and amongst the leading causes of cancer related deaths. A particularly virulent form of this disease is castration-resistant prostate cancer (CRPC), where patients no longer respond to medical or surgical castration. CRPC is a complex, multifaceted

[1]Parts of this section are reprinted with permission from O. A. Arshad, P. S. Venkatasubramani, A. Datta and J. Venkatraj "Using Boolean Logic Modeling of Gene Regulatory Networks to Exploit the Links between Cancer and Metabolism for Therapeutic Purposes", *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 1, pp. 399-407, 2016, doi: 10.1109/JBHI.2014.2368391 © 2016 IEEE; Osama A. Arshad and Aniruddha Datta, "Towards Targeted Combinatorial Therapy Design for the Treatment of Castration-resistant Prostate Cancer", *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, Seattle, WA, Oct 2-5 2016, doi: 10.1145/2975167.2985671 © 2016 ACM; and Noushin Ghaffari, Osama A. Arshad, Hyundoo Jeong, John Thiltges, Michael F. Criscitiello, Byung-Jun Yoon, Aniruddha Datta, Charles D. Johnson, "De novo Transcriptome Assemblies and Annotation for Pacific Whiteleg Shrimp", *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, Atlanta GA, 3-5 Dec, 2014, doi:10.1109/GlobalSIP.2014.7032342 © 2014 IEEE.

and heterogeneous malady with limited standard treatment options. The growth and progression of prostate cancer is a complicated process that involves multiple pathways. In Section 3, the signaling network comprising the integral constituents of the signature pathways involved in the development and progression of prostate cancer is modeled as a combinatorial circuit. The failures in the gene regulatory network that lead to cancer are abstracted as faults in the equivalent circuit and the Boolean circuit model is then used to design therapies tailored to counteract the effect of each molecular abnormality and to propose potentially effective combinatorial therapy regimens. Furthermore, stochastic computational modeling is utilized to identify potentially vulnerable components in the network that may serve as viable candidates for drug development. The results presented herein can aid in the design of scientifically well-grounded targeted therapies that can be employed for the treatment of prostate cancer patients.

RNA-Seq, the high-throughput sequencing of expressed messenger RNA, has emerged as the technology of choice for transcriptome studies. For non-model organisms, those that lack a reference genome, the RNA-Seq reads must be assembled de novo. De novo transcriptome assembly techniques provide an incredible opportunity to establish molecular level knowledge about organisms. In Section 4, we present novel transcriptome assemblies and functional annotations for Pacific whiteleg shrimp (Litopenaeus vannamei), a species of great importance in global mariculture, that lacks solid transcriptome and genome references. We examine the new Pacific whiteleg transcriptome assemblies via multiple metrics, and provide further annotations for the species.

2. UTILIZING BOOLEAN LOGIC MODELING OF GENE REGULATORY NETWORKS TO EXPLOIT THE LINKS BETWEEN CANCER AND METABOLISM FOR THERAPEUTIC PURPOSES[1]

2.1 Introduction

Cancer has traditionally been described as a condition that evolves via a multi-step process accumulating mutations with six essential genetic alterations or hallmarks leading to changes in cell physiology including genomic instability and increased mutability [34]. However, recently aerobic glycolysis or the "Warburg Effect", where cancer cells switch to glycolysis (an event common in normal cells when there is a lack of oxygen) even when oxygen is available, which is an alteration in the metabolic phenotype, has been added as a seventh hallmark [95]. Type 2 diabetes mellitus (T2D) on the other hand has been described as a group of metabolic diseases in which a person has high blood sugar, either because the body does not produce enough insulin, or the cells do not respond to the insulin that is produced. There is thus a metabolic shift in T2D where abundant blood glucose, a key biological fuel essential for fast proliferating cancer cells, is available. Added to this scenario are the hyperinsulinemic [93] environment in T2D and increased gene expression of glycolytic enzymes in cancer [74], permitting an ideal microenvironment for tumor cells. Consequently, the connection of metabolic alterations in both diseases has juxtaposed these two conditions at the clinical, biological and genetic levels [30, 77, 87].

Epidemiological studies have also demonstrated a positive association between T2D and the risk of cancer and cancer-related mortality [17]. Moreover, the diabetic

[1] © 2016 IEEE. Reprinted, with permission, from O. A. Arshad, P. S. Venkatasubramani, A. Datta and J. Venkatraj "Using Boolean Logic Modeling of Gene Regulatory Networks to Exploit the Links between Cancer and Metabolism for Therapeutic Purposes", *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 1, pp. 399-407, 2016.

drug Metformin has been shown to have a direct anti-cancer activity against breast and other cancers [86]. We exploit the association between T2D and cancer to develop a testable theoretical framework for cancer therapy design involving Metformin and chemotherapeutic drugs.

2.2 Biological background

Cancer cells exhibit unique metabolic characteristics. In this section we discuss the features of tumor cell metabolism and look at how targeting altered tumor cell metabolism through Metformin might afford a therapeutic opportunity.

2.2.1 Cancer cell metabolism

Cancer cells are known to exhibit characteristic alterations in their metabolic activity [12, 19, 45, 50]. Metabolism in normal cells differs in aerobic and anaerobic conditions. In the presence of oxygen, non-malignant cells convert glucose to pyruvate through a multi-step process called glycolysis. The pyruvate that is produced is transported to the mitochondria, the power house of the cell. The mitochondria then oxidize the pyruvate via a process called oxidative phosphorylation (OXPHOS) to generate adenosine triphosphate (ATP). ATP is the energy currency of the cell and is capable of storing large amounts of energy in its phosphoanhydride bonds. When the supply of oxygen is limited however, the cells shunt the pyruvate away from the mitochondria and convert it to lactate. Otto Warburg observed in the 1920s an anomalous characteristic of tumor cell metabolism that cancer cells even in the presence of oxygen opt for the latter route i.e. irrespective of the extracellular levels of oxygen, cancer cells continue to metabolize glucose to lactate instead of utilizing mitochondrial OXPHOS. This peculiar characteristic of cancer cell metabolism is called "aerobic glycolysis" or the "Warburg effect" [91].

On the face of it, the Warburg effect is counterintuitive as it is a highly inefficient

method for energy production: for every molecule of glucose, glycolysis generates 2 molecules of ATP whereas OXPHOS produces 34 molecules. Cancer cells presumably have a high demand for energy so the metabolic switch to aerobic glycolysis does not seem rational. There are a number of reasons for this switch, the most important of which is that it allows cancer cells to divert the intermediate bimolecular products of the glycolytic chain towards biosynthetic pathways. Normal cells, in the quiescent state produce energy as efficiently as possible. However, a cancer cell is tuned to incessant growth and proliferation. Towards this end, the tumor cells instead of metabolizing glucose with the goal of efficient energy generation, divert the nutrients towards anabolic processes that will provide the necessary substrates for cell growth. Aerobic glycolysis allows cells to divert intermediates towards biomass accumulation. In order to make up for the inefficiency of aerobic glycolysis, the cancer cells take up much larger amounts of glucose. Indeed, enhanced glucose uptake and the accompanying increased glycolytic flux is a universal metabolic alteration in cancer, and forms the basis of the positron emission tomography (PET) scan technique for cancer detection. Another advantage of aerobic glycolysis is that it confers better survivability in an oxygen starved (hypoxic) environment, conditions which are common in tumor tissue [64].

One of the principal mechanisms behind the switch to aerobic glycolysis is the constitutive expression of the Hypoxia Inducible Factor 1 (HIF-1), which is a transcription factor ordinarily activated by hypoxic stress [11, 45]. HIF-1 drives many of the metabolic adaptations in cancer [19, 44]. First it increases the uptake of glucose by upregulating the glucose transporters. Second, it increases glycolytic flux by activating enzymes in the glycolytic pathway. Third, it shunts pyruvate away from the mitochondria (pyruvate enters the tri-carboxylic acid (TCA) cycle in the mitochondria through conversion to acetyl-CoA. This reaction is catalyzed by pyru-

vate dehydrogenase (PDH). HIF-1 inhibits PDH by activating pyruvate dehydrogenase kinase 1 thereby slowing the entry of pyruvate into the TCA cycle). Finally, HIF-1 activates the enzyme lactate dehydrogenase A (LDHA) which catalyzes the conversion of pyruvate to lactate, and upregulates monocarboxolate transporter 4 (MCT-4) to discharge the lactate into the extracellular matrix (ECM). This leads to the acidification of the ECM milieu which in turn promotes metastasis. Therefore, HIF initiates a transcriptional cascade which acts as a key driver of the metabolic adaptation of cancer cells.

In addition to the Warburg effect, another major facet of the metabolic reprogramming in cancer cells is increased "de novo fatty acid synthesis", a process where the cells synthesize the requisite lipids in-house rather than relying on the circulating exogenous supply from the blood stream as normal cells do [78]. The endogenous synthesis of lipids requires citrate which is derived from the TCA cycle. To sustain the TCA cycle, the depleted citrate is replenished through a process called anaplerosis [19]. Sterol Regulatory Element Binding Protein (SREBP) is a master transcriptional regulator of genes involved in de novo lipid and sterol biosynthesis [24]. Enhanced expression of SREBP has been shown to correlate with breast cancer progression [78].

Thus, fundamental alterations in tumor cell metabolism include aerobic glycolysis and de novo lipid synthesis with HIF-1 and SREBP as key markers of the metabolic reprogramming that takes place in cancer cells.

2.2.2 Metformin and cancer

Altered tumor metabolism is of paramount importance for sustained uncontrolled cell growth, a vulnerability that can be exploited for therapeutic intervention [45,78]. In the context of targeting cancer metabolism, the widely used anti-diabetic drug

Metformin has garnered attention for its potential anti-cancer properties suggesting a role for this drug in cancer therapy and prevention [8,22,69,70]. Tumor cells have a voracious appetite for glucose. Metformin suppresses hepatic gluconeogenesis which reduces glucose levels thereby diminishing the tumor fuel supply. The principal mechanism of action of Metformin is the activation of adenosine monophosphate activated protein kinase (AMPK), the cellular energy sensor which when activated switches on ATP-generating pathways and diminishes energy consuming biosynthetic processes thereby curtailing proliferation. Activation of AMPK by Metformin, phosphorylates the tuberous sclerosis complex (TSC) which in turn inhibits the mammalian target of rapamycin (mTOR) complex, the master stimulator of protein synthesis and cell growth.

In view of the central role of energy metabolism in cell proliferation, we investigate the therapeutic value of Metformin for cancer treatment. Moreover, there are several other potential benefits of adding Metformin to cancer therapy regimens. It is an FDA approved stable oral agent with a long history of use, is widely available, has an extremely low toxicity profile and is very inexpensive [54].

2.3 Therapy design

Cancer is an umbrella term for a set of diseases characterized by a break down in cell cycle control that allows cells to escape the usual controls on cell proliferation and survival. In [47], the authors take the view that in essence, it is a disease that results in aberrant signaling caused by breakdown(s) in the normal signaling pathway of a given cell, and therefore it can be meaningfully treated or managed through remedying the effect of such breakdown(s). By adopting a similar approach, we consider the signaling pathways commonly mutated in cancer, map the biological pathway information to a digital circuit which is then used to determine the possible

fault locations and devise an appropriate therapeutic scheme.

2.3.1 Pathway model

Mutations in the PI3K/AKT/mTOR and Ras/MEK/ERK (MAPK) signaling pathways are common in breast cancer malignancies with frequent genetic alteration in several key players from these pathways [37, 79]. These pathways are activated via growth factor receptor tyrosine kinases and regulate cell metabolism, survival and growth. A schematic representation of these pathways is given in [47]. To this schematic, we add the pathway segment incorporating the transcription factors HIF-1 and SREBP. This leads us to our signaling pathway model of Fig. 2.1. The black and red lines in the diagram indicate relationships which are activating and inhibitory respectively.

The red boxes show breast cancer drugs and their points of intervention in the pathway. The cancer drugs in our model are "targeted molecular therapies", agents which act with great specificity on particular molecules in the signal transduction network known to be important in cancer [6, 48, 62, 76]. The cancer drugs are Lapatinib (a dual tyrosine kinase inhibitor of EGFR and ERBB2) [6,48,62,76], Trastuzumab (a monoclonal anti-body targeting ERBB2) [6, 48, 62, 76], Cixutumumab (anti-IGF1R monoclonal antibody) [59], U0126 (MAPK pathway inhibitor targeting MEK) [23,27] and LY294002 (PI3K/Akt pathway inhibitor targeting PIK3CA) [48, 62, 76].

Genes exhibit switch-like on/off behavior and thus a gene regulatory network (GRN) can be modeled with a Boolean circuit [47]. The marginal interactions amongst genes represented by the signal transduction network, can be translated to an equivalent Boolean network. For example if either of two genes say A or B can activate a third gene C, then this component of the GRN can be represented by an OR gate with inputs A and B and output C. Using such a procedure outlined in [47]

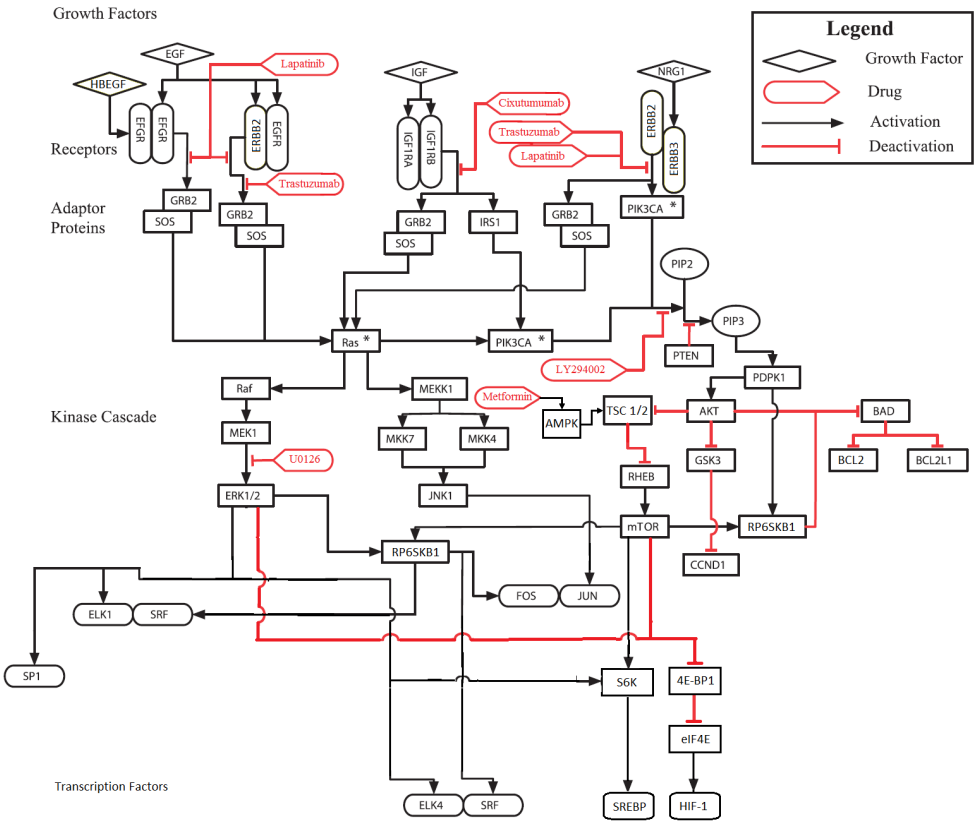


Figure 2.1: A schematic diagram of signaling pathways commonly mutated in breast cancer.

for translating the interactions of different genes in a signaling network to a logic circuit, we can model the pathways in Fig. 2.1. with a Boolean circuit, to arrive at Fig. 2.2. In this circuit, there are nine outputs, 6 of which are transcription factors (marked in yellow) and the remaining (which are not colored) reflect the activation status of some key proteins from our target signaling pathways.

2.3.2 Fault locations and drug intervention points

In normal cells, cell division is under extremely tight control and cells only divide to form further cells if they receive external signals to do so. These external

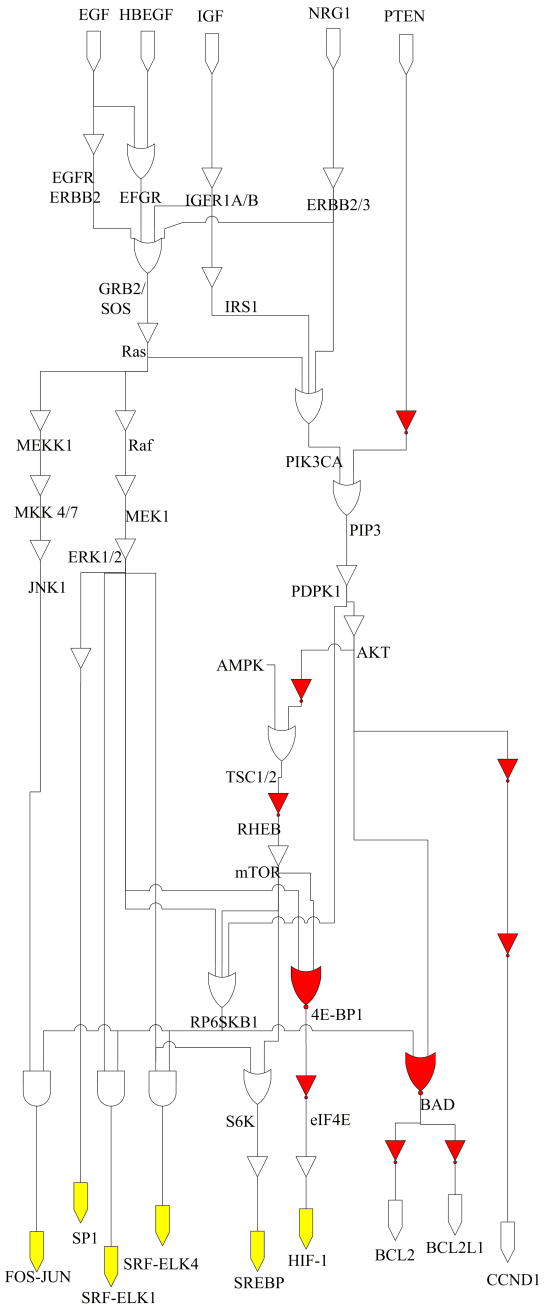


Figure 2.2: Boolean circuit model.

signals that stimulate a cell to divide are called growth factors or mitogens. Cancer is characterized by a breakdown in cell signaling in which cells are set-free from the

usual controls on cell-cycle progression and continue to grow and proliferate even in the absence of mitogenic signaling. Such abnormalities in the signaling network can be modeled as stuck-at faults where a point in the network is permanently fixed (stuck) at a particular value of either 1 (stuck-at-1 fault) or zero (stuck-at-0 fault) corresponding respectively to the constitutive (perpetual) activation or inactivation of a gene [47]. For example, in cancerous cells, the proto-oncogenes can get mutated to become oncogenes or a tumor suppressor gene can lose its braking function. For instance, if the PIK3CA proto-oncogene, a gene frequently mutated in breast cancer [37] mutates to PIK3CA oncogene, the encoded PIK3CA oncoprotein can become constitutively active and start perpetually signaling to the downstream proteins. In that case, even if there is no mitogenic signaling from the outside, the cell will be stimulated to divide. Such constitutive activation of PIK3CA can be modeled as a stuck-at-1 fault. Similarly, mutation in the PTEN tumor suppressor can cause a cell to ultimately undergo uncontrolled cell division, and possibly turn cancerous. Such a fault that renders PTEN inactive, corresponds to a stuck-at-0 fault. Thus, cancer is a disease of aberrant cell signaling caused by failures in the signaling pathways which can be represented as stuck-at faults in the network. For simplicity, we consider single stuck-at faults. From the network in Fig. 2.1, we identify 27 possible fault locations illustrated in Fig. 2.3 with the "stuck-at-1" faults in black numerals and the "stuck-at-0" faults in red.

The intervention points in the Boolean circuit for the cancer drugs are shown in Fig. 2.4. Since the cancer drugs of Fig. 2.1, with the exception of Metformin, break or stop the effect of the kinase (to which they bind) on the molecules further downstream of the signaling cascade, a drug of this type can be modeled as an inverted input to an "AND" gate at the point of intervention [47]. As discussed in sub-section 2.2.2, Metformin however acts as an activator of the AMPK, overcoming the dysregulation

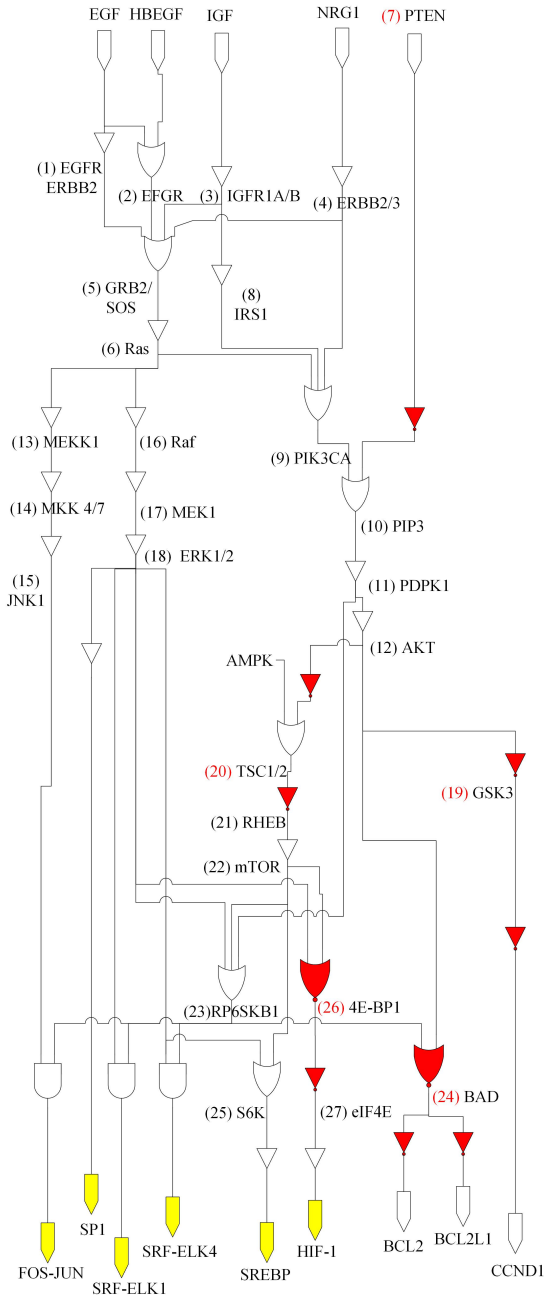


Figure 2.3: Possible fault locations.

of AMPK in cancer. Thus, in order to incorporate the effect of Metformin in our Boolean network, its action is modeled via the activation of AMPK.

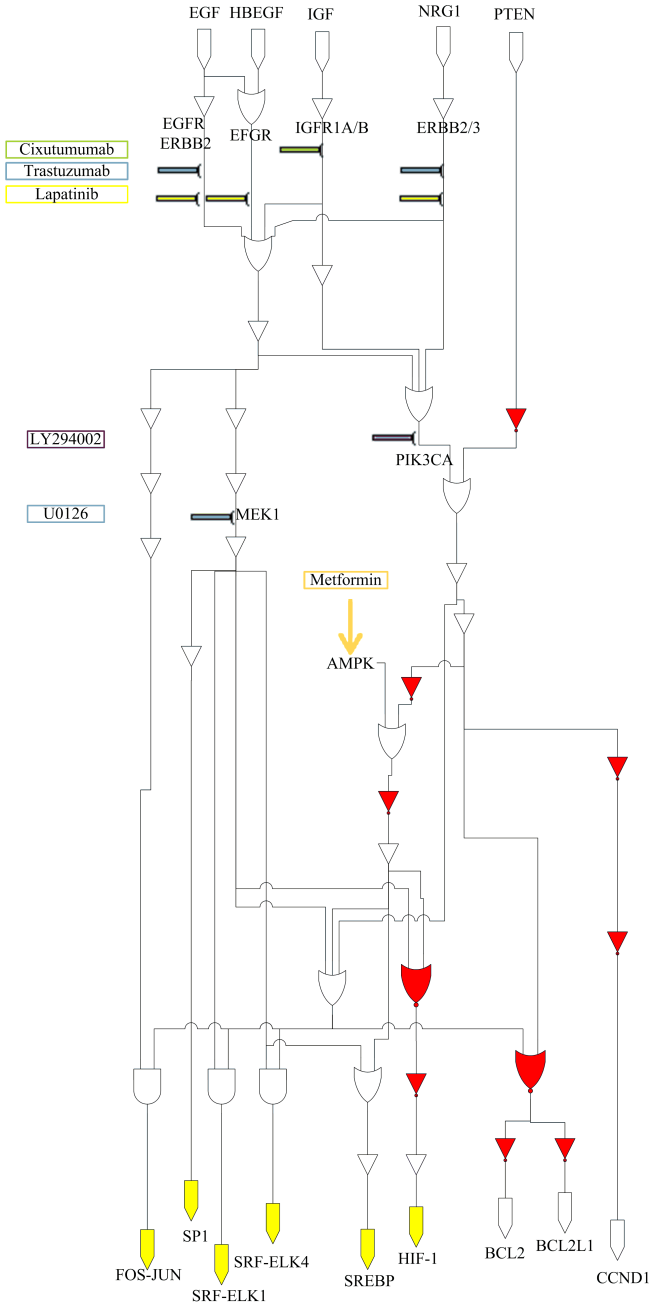


Figure 2.4: Drug intervention locations.

2.3.3 Fault classification

In this sub-section we group the faults identified in Fig. 2.3 into different classes of equivalent faults based on their output to the non-proliferative input. The input

and output vectors are defined as:

$$INPUT = [\text{EGF}, \text{HBEGF}, \text{IGF}, \text{NRG1}, \text{PTEN}]$$

$$OUTPUT = [\text{FOS-JUN}, \text{SP1}, \text{SRF-ELK1}, \text{SRF-ELK4}, \text{SREBP}, \text{HIF-1}, \text{BCL2}, \text{BCL2L1}, \text{CCND1}]$$

Each input can take on binary values. We set the input vector to [00001] which corresponds to the growth factors being absent and the tumor suppressor PTEN being active i.e. non-proliferation. For each of the 27 faults that may cause cancer, the output is tabulated in Table 2.1a. Fault location zero corresponds to the fault-free case.

Based on the outputs, the faults can be grouped together into classes of equivalent faults. Faults which produce identical output for the same input test vector are equivalent. From the outputs in Table 2.1a, the sets of equivalent faults for the test vector [00001] are shown in Table 2.1b.

2.3.4 Simulation results for drug intervention

The total number of drugs is six, so we define a binary drug vector of length six with each component having a value of either 1 if the corresponding drug is applied, and zero if it is not. Thus, there are a total number of $2^6 = 64$ possible drug vectors representing all the possible drug combinations. For each of the 27 faults that may cause cancer, the simulation determines the output for every drug combination and maps the output to a real number indicating the extent of proliferation. The drug vector is defined as follows:

$$DRUG\ VECTOR = [\text{Metformin}, \text{Lapatinib}, \text{Trastuzumab}, \text{Cixutumumab}, \text{U0126}, \text{LY294002}]$$

Table 2.1: Fault classification.

(a) Output vector for all single stuck-at faults with input = [00001].

Fault Location	FOS-JUN	SP1	SRF-ELK1	SRF-ELK4	SREBP	HIF-1	BCL2	BCL2L1	CCND1
0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1
7	0	0	0	0	1	1	1	1	1
8	0	0	0	0	1	1	1	1	1
9	0	0	0	0	1	1	1	1	1
10	0	0	0	0	1	1	1	1	1
11	0	0	0	0	1	1	1	1	1
12	0	0	0	0	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0
16	0	1	1	1	1	1	1	1	0
17	0	1	1	1	1	1	1	1	0
18	0	1	1	1	1	1	1	1	0
19	0	0	0	0	0	0	0	0	1
20	0	0	0	0	1	1	1	1	0
21	0	0	0	0	1	1	1	1	0
22	0	0	0	0	1	1	1	1	0
23	0	0	0	0	0	0	1	1	0
24	0	0	0	0	0	0	1	1	0
25	0	0	0	0	1	0	0	0	0
26	0	0	0	0	0	1	0	0	0
27	0	0	0	0	0	1	0	0	0

(b) Equivalent fault groups for input = [00001].

Output	Equivalent Fault Groups
000000000	0, 13, 14, 15
111111111	1, 2, 3, 4, 5, 6
000011111	7, 8, 9, 10, 11, 12
011111110	16, 17, 18
000000001	19
000011110	20, 21, 22
000000110	23, 24
000010000	25
000001000	26, 27

The input and output vectors are defined as before in the previous sub-section. Again, for the simulation we set the input vector to [00001] i.e. a non-proliferative input. For this input, we expect all the outputs comprising proliferative transcription factors, metabolic adaptation markers, cell-cycle progression and anti-apoptotic proteins to be de-activated or turned off and indeed, for the fault-free scenario, this is certainly the case. However, faults in the signaling network, can cause a proliferative (non-zero) output even for the non-proliferative input. Our objective is to nullify the effect of the faults by targeted drug intervention to produce an output as close to [000000000] and as far away from the extremely proliferative output [111111111] as possible.

To quantify the degree of abnormal behavior, we define a transformation as in [47] to map the $2^9 = 512$ output vectors to the continuous real number scale. Since the first six components of the output vector are transcription factors, and the remaining are the activation status of some proteins, these two groups of outputs have different biological significance and are encoded separately. Defining N_1 to be the number of active transcription factors and N_2 to be the number of active remaining outputs, the transformation below maps in a many-to-one fashion the output vector to a real number:

$$Output = [a, b, c, d, e, f, g, h, i]$$

$$N_1 = [a + b + c + d + e + f]$$

$$N_2 = [g + h + i]$$

$$P = N_1 \times N_2$$

$$S = N_1 + N_2$$

$$\psi(output) = \alpha P + (1 - \alpha)S$$

where $\alpha \in (0, 1)$ is a free design parameter that defines the convex combination of the sum and product of N1 and N2. It determines the relative weights assigned to the sum and product. Since there is no obvious reason to assign greater weight to the sum or product term relative to the other, we assign equal weights by selecting a value for α that is right in the middle of the parameter space. Hence, α is chosen to be 0.5 for the simulation. Thus, we have $\alpha = 0.5(N_1N_2 + N_1 + N_2)$. Note that this is a non-linear cost function with a product term N_1N_2. As a consequence, this transformation produces a higher cost (for the same number of total active output components), when genes from both groups in the output vector are expressed compared to the situation when only one group in isolation is trying to drive proliferation. This makes sense as we expect that the extent of proliferation will be higher when both sets of genes, which when deregulated play important complementary roles towards unchecked proliferation, are simultaneously active. It is pertinent to point out however, that the results that follow are not predicated on the exact definition of the mapping.

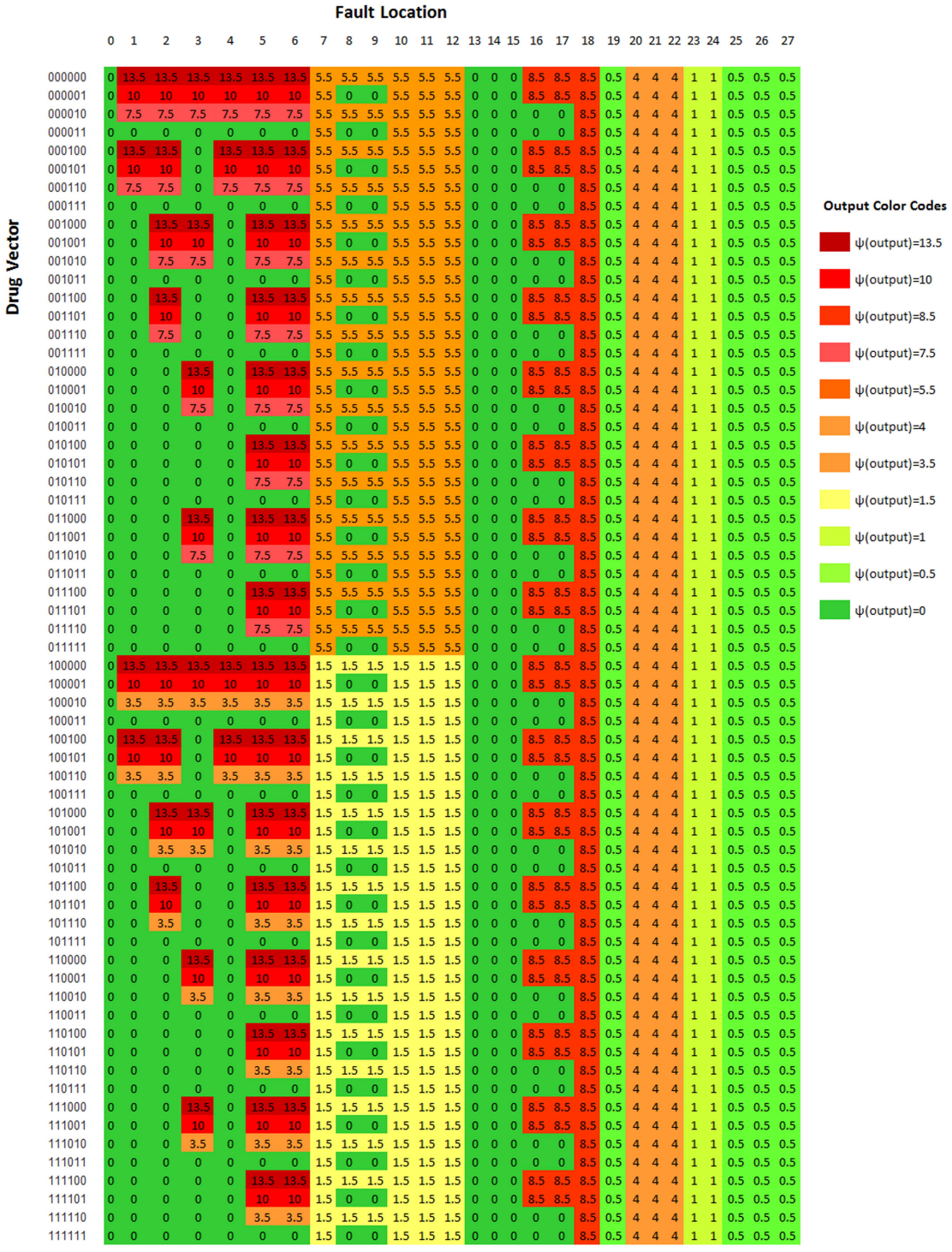


Figure 2.5: Drug vector response.

The values of the function ψ for all possible faults and drug combinations are shown in Fig. 2.5. with the fault locations and drug vectors along the horizontal and vertical directions respectively. The outputs are color-coded on a scale with red representing extreme proliferation and green non-proliferation. The color codes used are listed on the right side of Fig. 2.5. Again fault location zero corresponds to the fault-free case.

We would like to drive as many of the faults towards green (non-proliferation) as possible using as few of the cancer drugs as we can since these drugs have toxic side effects. From Fig. 2.5 we can immediately see the benefit of Metformin. It mitigates the effect of faults 7-12, which are faults in the insulin/insulin-like growth factor (IGF) signaling pathway and the PI3K/AKT pathway, the pathway involved in regulating cell metabolism. From the previous sub-section, these faults are classified in the same fault group and it is for this class of faults that we expect Metformin to show therapeutic benefit i.e. Metformin should ameliorate the effect of faults that lead to the de-regulation or hyper-activation of the PI3K/AKT/mTOR axis. Indeed, Metformin has been shown to overcome the dysregulation of the PI3K pathway by suppression of mTOR through AMPK activation in breast cancer cells. A number of studies have indicated the anti-tumorigenic effects of Metformin in multiple cancer cell lines including breast cancer with the use of Metformin as an anti-cancer agent now being evaluated in clinical trials [15, 22, 31, 69, 81]. Moreover, one of the most common ways in which the PI3K/AKT/mTOR pathway can be deregulated in breast cancer is the loss of PTEN, the negative regulator of this cascade, an event found in up to 40% of breast tumors [68]. Metformin has been shown to delay the onset of tumors in PTEN-negative mice [38]. Experimental studies have also demonstrated the benefit of Metformin in combination with chemotherapeutic agents and provided a rationale for Metformin as part of combination therapy for breast cancer [36]. All

in all, our simulation results with respect to the therapeutic benefit of Metformin for cancer seem to be in concordance with the literature.

Note that no drug vector has any effect on fault 18. This makes sense as fault 18 corresponds to a mutation in the ERK1/ERK2 protein. This fault is downstream of all the drugs in our pathway model, so no drug combination is able to counteract the effect of this particular fault. In addition, the faults 13-15 produce an all zero output i.e. these faults are "undetectable" as they generate the same output as the fault-free case. Since we are only concerned with faults that can induce cancer, we do not have to worry about these particular faults as they produce a non-proliferative output.

The best two-drug vector in terms of driving faults towards green is 100010, the combination of Metformin and U0126. However, considering cancer drugs only (i.e. excluding Metformin) the best drug vector is 000011, i.e. the drug combination of U0126 and LY294002 is the best two drug combination of cancer drugs. If to this combination we add Metformin, we see we get an even better result as more of the faults are driven towards green. This better outcome is obtained at minimal additional cost as in contrast to cancer drugs, Metformin is inexpensive and does not have adverse side-effects. We conclude that incorporating Metformin in the mix for cancer therapy can lead to improved outcomes.

Therefore, it seems that U0126 and LY294002 along with Metformin should be a potent combination therapy for breast cancer. We thus propose that a cancer combination therapy of U0126 or some other MEK (mitogen-activated protein kinase) inhibitor and LY294002 along with Metformin can lead to better therapeutic results. The exact same cancer drug combination of U0126 and LY294002 has been proposed as a therapeutic approach in the prevention and treatment of human melanoma [7]. Furthermore, recently a similar drug combination that targets the MAPK and PI3K

pathways has shown promising results for Rhabdomyosarcoma [71]. Since these very signaling pathways are the most frequently deregulated signaling cascades in human breast cancer [37, 79], it is reasonable to expect the above drug combination to be effective for breast cancer malignancies. The fact that our model prediction regarding the therapeutic potential of Metformin is in consonance with the literature and that the proposed cancer drug combination has shown promising results in other cancers with frequent mutations in the same pathways, suggests that this particular cancer therapeutic regimen warrants further investigation. In any case, our results indicate at a minimum that incorporating the metabolism-targeting drug Metformin in the cancer therapy cocktail should give better outcomes compared with the use of cancer drugs alone.

Thus, by computer simulation of the Boolean logic equivalent model of the critical gene regulatory pathways of breast cancer, we have been able to demonstrate the benefit of Metformin use in cancer therapy which suggests a role for this drug in combination with cancer drugs. The theoretical results presented herein regarding the potential benefits of including the metabolism targeting drug Metformin as part of a combination cocktail therapy for cancer ultimately need to be validated via actual experiments on cancer cell lines.

The work presented here has some limitations. One of the major impediments to the success of therapeutic intervention in cancer is the presence of feedback signaling. Any attempt to counter the deregulation of a particular signaling pathway by administering targeted therapy is counteracted in cancer by exploiting the redundancy in the cellular signaling network through the compensatory activation of feedback loops which ultimately limit the potency of any attempted therapeutic intervention [55]. Furthermore, a tumor population is generally heterogeneous in that it is comprised of a number of different subpopulations that harbor distinct mutations with the result

that no two cancers are completely alike [60]. As a consequence different subpopulations require different treatments. These issues of tumor heterogeneity and feedback signaling are complex research problems in their own right but need to be addressed i.e. a comprehensive system model needs to incorporate these considerations. Some preliminary progress in this direction has been made [61, 82].

3. TOWARDS TARGETED COMBINATORIAL THERAPY DESIGN FOR THE TREATMENT OF CASTRATION-RESISTANT PROSTATE CANCER

3.1 Background

Prostate cancer is the most common noncutaneous male malignancy and one of the leading causes of cancer mortality in the western world [80]. The growth and progression of prostate cancer is stimulated by androgens [26]. Androgens are male sex steroid hormones that are responsible for the development of male characteristics. Testosterone is the most important androgen in men. The effects of androgens are mediated through the androgen receptor (AR) [10]. The androgen receptor is a nuclear receptor, which is activated in response to the binding of androgens. Upon activation, it mediates transcription of target genes that modulate growth and differentiation of prostate epithelial cells. In malignant prostate cells, androgen signaling is deregulated and the homeostatic balance between the rate of cell proliferation and programmed cell death is lost. As prostate cancer relies on androgens for growth, the main line of treatment focuses on abrogating the action of androgens. Androgen deprivation therapy (ADT) in the form of surgical or medical castration is the cornerstone of treatment for prostate cancer [20]. Initially, androgen ablation induces significant regression of the tumor. However, the response to ADT is temporary and prostate cancer invariably stops responding to this treatment regimen, leading to a clinical condition that is known as hormone-refractory prostate cancer, androgen-independent prostate cancer or castration-resistant prostate cancer (CRPC). CRPC is a more aggressive and typically lethal phenotype where the tumor continues to grow in spite of the very low levels (<50 ng/ml) of circulating serum testosterone. Standard treatment options are limited and palliative docetaxel-based

chemotherapy is generally used for patients who have become refractory to hormone treatment. However, median survival time for patients following first-line chemotherapeutic treatment is just eighteen to twenty-four months [49]. There is therefore a clear rationale for advances in alternative therapeutics in order to evolve and expand the landscape of treatment options for malignant forms of prostate cancer that recur after abatement.

Over recent years, there has been a significant effort towards furthering our understanding of the molecular mechanisms underpinning tumor development, growth and progression. It is now appreciated that in spite of castrate levels of androgens, the cancer cells are able to maintain persistent androgen receptor signaling through a variety of contributory mechanisms including AR gene amplification that results in overexpression of AR, gain-of-function mutations in AR which enable promiscuous activation of the receptor through other steroids or even in the absence of ligand binding, changes in AR co-activators and the expression of AR splice variants [14]. This compensatory response allows cancer cells to survive in a low testosterone environment and the reactivated AR signaling axis continues to play a role after neoplastic transformation. Additionally, certain androgen-independent cellular signaling pathways that promote proliferation and inhibit apoptosis, have been critically implicated as drivers of continued progression of prostate cancer. Hence, accumulating evidence indicates that the growth and progression of prostate cancer is a complicated process that involves interaction between multiple pathways. Advances in our knowledge of the biology of prostate cancer has led to the development of a number of novel therapies designed to target signaling pathways involved in disease progression. With the exception of certain androgen synthesis and AR signaling antagonists that have received regulatory approval, these advanced agents are under various stages of clinical trials [67].

Castration-resistant prostate cancer is a complex malady. Given the inherent complexity of the CRPC signaling cascade, there is no one dominant molecular driver across all tumors and hence no single drug can act as a "magic bullet" by being uniformly effective for treating the malignancy [3, 4]. At best, limited benefit will be derived from targeting a single molecule. Rational combinations of signal-modulating therapeutic agents have higher likelihood of yielding better outcomes. While there are several drugs being tested on cell lines, most of these studies focus on a single pharmaceutical agent and very few of those experiments involve trying out drug combinations. Furthermore, prostate cancer is a markedly heterogeneous disease, with different tumors varying in their composition and makeup. In other words, different tumors will harbor different malfunctions in the signaling pathways. Thus, tailored targeted therapies based on individual tumor characteristics are required to maximize the potential benefits from treatment.

Mathematical and computational modeling plays a pivotal role in systems biology in elucidating biological insights from large-scale biomolecular signaling networks that are not amenable to straightforward intuitive interpretation. A diverse array of formalisms have been proposed in this domain as suitable representations for complex multicomponent networks such as cellular signaling pathways [88]. Amongst these frameworks, Boolean network models [35, 43] have emerged as an extremely useful parameter-free approach to capture the qualitative behavior of extensive genetic networks wherein knowledge of kinetic parameters is scarce. Boolean logic models have been successfully applied to study biological signaling networks and cellular processes [5, 89], for instance the cell cycle [25], apoptosis [75], the T-cell survival network [97], hypoxia stress response pathways [83] and the gene regulatory network regulating cortical area development [29]. In this section, we use Boolean logic modeling of the key signaling pathways implicated in the development and progression

of prostate cancer to simultaneously test various combinations of agents for their efficacy in attenuating cancer growth and design targeted therapies for the management of the disease. In addition, we attempt to delineate components in the signaling network that can be pharmacologically manipulated to therapeutic advantage.

3.2 Boolean modeling of prostate cancer signaling

Cellular processes such as growth and division are regulated by an interconnected network of molecules referred to as signaling pathways. Key cellular signal transduction pathways known to play a major role in cell survival, growth, differentiation and the development of castration-resistance in prostate cancer are the Androgen Receptor (AR), PI3K/AKT/mTOR and Mitogen-Activated Protein Kinase (MAPK) pathways. The aberrant behavior of prostate cancer cells is characterized by dysfunction in these selective oncogenic signaling pathways promoting malignant characteristics. These pathways play a role in a diverse range of essential physiological cellular processes such as differentiation, survival, proliferation, protein synthesis and metabolism. Malfunctions in these pathways are common in prostate cancer malignancies. For example, approximately 70% of advanced prostate cancers have genomic alterations in the PI3K/AKT/mTOR pathway [13]. These three pathways are the most frequently over-activated pathways increasing survival of cancer cells and promoting cancer progression [96]. A schematic representation of these pathways is given in [40–42].

In the context of methodologies that are applied to model cellular signal transduction networks, Boolean networks are probably the simplest where the state of each node in the network is either active (on) or inactive (off). In a Boolean network, the nodes are the genes and the edges represent the interaction amongst the genes. Since the molecules in a gene-regulatory-network (GRN) exhibit switch-like

behavior, genes may be regarded as binary devices where a gene can be considered to be active if it is being transcribed and inactive if it is not. Moreover, the relationships amongst the genes may be represented by means of logical functions. Thus, a GRN is amenable to such a representation. The Boolean formalism is analogous to a digital circuit where logic gates can be used to represent the regulatory relationships amongst the nodes and the activation level of the nodes is indicated by binary logic. The biological interactions amongst the various nodes (genes) represented in the gene regulatory network of prostate cancer can therefore be translated to an equivalent Boolean circuit [92]. Let's say either gene X or Y can activate a third gene Z, then we can model this component of the signaling network with an OR gate with two inputs, namely X and Y and with output Z. Thus, the prostate cancer signaling network can be mapped to the combinational circuit shown in Fig. 3.1. This digital logic circuit represents our multi-input multi-output (MIMO) systems model of the prostate cancer signaling transduction network. Each node is assigned a numeric label in parentheses. These labels also serve to enumerate the fault locations with stuck-at-one and stuck-at-zero faults in black and red numerals respectively. The dotted arrows indicate the intervention points for the respective drugs. These pharmacologic agents are highly specific pathway inhibitors. These reagents modulate growth-factor receptors and the downstream pathways abnormally activated in CRPC by targeting with great specificity certain signaling nodes in the network.

Cancer is a disease of abnormal cell signaling caused by a breakdown in the normal signaling pathways leading to the loss of cell cycle control and uncontrolled cell proliferation. These abnormalities in the signaling network can be represented as stuck-at faults [2]. A stuck-at fault is said to occur when a line in the network is permanently set to a fixed value of one (stuck-at-one fault) or zero (stuck-at-zero fault) with the result that the state of the line is stuck at the faulty value and no

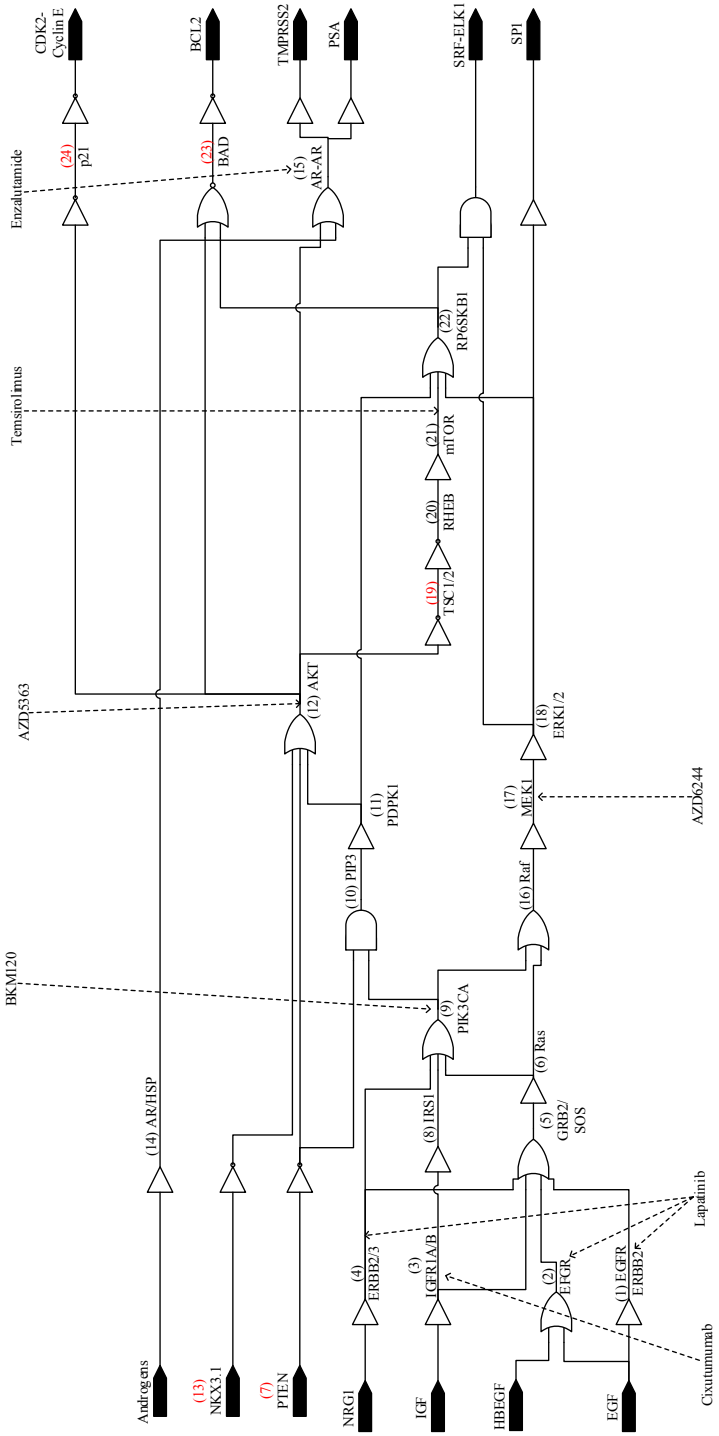


Figure 3.1: Combinational circuit model of prostate cancer signaling pathways.

longer depends on the state of the signaling network upstream that drives that line i.e. the faulty line has a constant (1/0) value independent of other signal values in the circuit. A stuck-at-fault can occur either at the input or output of a gate. An example of a stuck-at-fault is given in Fig. 3.2. Suppose the input vector is <abcd>= 1100. In this case, the output is 0. However, if there is a stuck-at-one fault at the output of the NAND gate (at the location marked with a cross) with the same input vector as before, the output of the faulty circuit is one instead of zero. This notion of stuck-at-faults has immediate biological relevance: on account of mutations or other structural abnormalities, a gene might become dysfunctional and hence stuck at a particular state irrespective of the signals that it is receiving from surrounding genes [47]. These biological defects can be abstracted as stuck-at faults. For instance, as discussed earlier, a diverse array of mechanisms engender persistent AR signaling in CRPC even with castrate serum levels of androgen. This constitutive (permanent) activation of the androgen receptor where the receptor remains active i.e. continues to signal downstream even in the absence of androgens can be represented as a stuck-at-1 fault. By the same token, the inactivation in cancer of a tumor suppressor, which acts as a molecular brake on cell growth in a normal cell, can be represented as a stuck-at-0 fault. From our Boolean circuit model, we can explicitly enumerate the different locations where a fault can occur. These fault locations are numbered in Fig. 3.1 with the stuck-at-0 and stuck-at-1 faults in red and black numerals respectively. There is a total number of 24 possible fault locations.

The objective is to counteract the effect of these faults by targeted drug intervention, so we incorporate the drugs in our model. The drug intervention points are illustrated in Fig. 3.1 which are the locations of the molecules that these prostate cancer drugs are known to target. Since the drugs inhibit the activity of their target

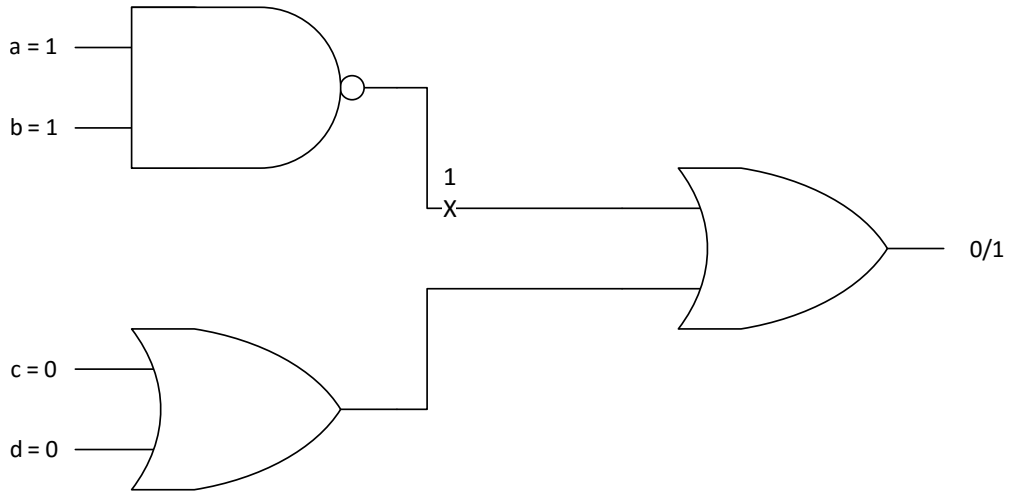


Figure 3.2: Circuit with stuck-at fault.

i.e. the main mechanism of action of the anti-cancer drugs is to cut off downstream signaling, their action is incorporated in our model as an inverted input to an AND gate with the result that whenever the drug is applied, the gene that it targets is turned off.

3.3 Simulation for fault mitigation with drug intervention

We can now use our Boolean model to test different combination therapies in terms of their efficacy in mitigating the effects of the faults. For each fault, we would like to intervene with the best possible drug combination i.e. we want to determine which set of drugs would be most effective in attempting to nullify the effect of that fault, thereby providing us with a targeted therapy based on the tumor signature. Define, the input vector as follows:

$$INPUT = [\text{EGF}, \text{HBEGF}, \text{IGF}, \text{NRG1}, \text{PTEN}, \text{NKX3.1}, \text{Androgens}]$$

The first four components of this vector are growth factors, which are external signals that stimulate a cell to grow and replicate. The next two input components, namely PTEN and NKX3.1 are tumor suppressors which act as molecular brakes on cell division. The last input vector component consists of the external hormones that stimulate the AR pathway in a normal prostate cell. The input vector is set to be [0000110]. This corresponds to all the external signals that stimulate cell growth being absent and the molecular brakes being active i.e. this input vector corresponds to a non-proliferative input which produces a non-proliferative output in the fault-free case. The output vector is defined to be:

$$OUTPUT = [\text{SP1}, \text{SRF-ELK1}, \text{PSA}, \text{TMPRSS2}, \text{BCL2}, \text{CDK2-CyclinE}]$$

The output vector consists of key markers of cell growth and proliferation in prostate cancer. In the fault-free scenario, a non-proliferative input to the regulatory network should produce a non-proliferative output characterized by the all-zero vector. However, faults in the network will produce a non-zero (proliferative) output even when the input is non-proliferative. The objective is to drive the faulty network's output as close as possible to that of the fault-free circuit i.e. towards the all-zero vector through targeted drug intervention. Define, the drug vector as:

$$DRUG\ VECTOR = [\text{Lapatinib}, \text{Cixutumumab}, \text{AZD6244}, \text{BKM120}, \text{AZD5363}, \text{Temsirolimus}, \text{Enzalutamide}]$$

Each component of the drug vector is one if the corresponding drug is applied and is zero otherwise i.e. the i^{th} bit of the drug vector is one if the drug is selected and zero if it is not. Thus, for example, the drug vector [0010010] represents the combination of AZD6244 and Temsirolimus. Since, the total number of drugs is seven, the number

of possible drug combinations is 128. The objective is to determine the best possible therapy for each fault. Each fault represents a different molecular abnormality and hence a tumor with a different profile.

For each of the faults, the problem is to find the drug selection that can rectify the fault i.e. change the faulty output to the correct output. If that is not possible, the best drug vector will drive the output as close as possible to the fault-free output. A simple metric that can be used as a distance measure to determine how far the output vector is from the fault-free vector is Hamming distance. Faults that produce an output vector with a greater Hamming distance from the correct output have more of the proliferative genes active and presumably a greater proliferative effect. Since the correct output is the all-zero vector, the Hamming distance of the output vector from the correct output is simply the Hamming weight of the output vector (for binary vectors Hamming weight is equivalent to the L_1-norm). For each fault, we determine the output under every possible drug vector. The best therapy for that fault is the drug vector that produces the output with the smallest Hamming weight. In addition, since the drugs have deleterious side-effects, we would like to choose a drug combination with the fewest number of drugs. Thus, the best targeted therapy for each of the cancer-inducing faults is the one that under the presence of the fault, produces the best output with the smallest Hamming weight with the minimal number of drugs. The best therapy for each of the faults is shown in table 3.1 with the drug vector defined as above. Note that for certain faults, no drug vector can improve the output. Such faults are said to be untestable since no test (drug vector in this case) can rectify the fault. This is because there are no drugs on the fan-out of these genes. However, all these faults with the exception of fault 18 are minimally proliferative as they produce a faulty output with the least possible Hamming weight of one.

Table 3.1: Best therapy for each fault.

Fault Location	Drug Vector
1	1000000
2	1000000
3	0100000
4	1000000
5	0011000
6	0011000
7	0000100
8	0001000
9	0001000
10	0000100
11	0000100
12	0000100
13	0000100
14	0000001
15	0000001
16	0010000
17	0010000
18	0000000
19	0000010
20	0000010
21	0000010
22	0000000
23	0000000
24	0000000

Thus, there are many locations in the gene regulatory network of prostate cancer where malfunctions can occur resulting in a cancer that is different, requiring a specific targeted therapy. The table facilitates arriving at such a therapy as it maps each malfunction to an appropriate set of drugs. The look-up table can be used to devise therapies that have a higher likelihood of success since they are tailored specifically to the molecular abnormalities in critical pathways and thereby facilitates an individualized approach to therapy design.

To determine the best combination therapy across all faults, for each drug com-

bination we determine the sum of the Hamming weights of the output vector across all possible combinations of faults and choose the drug combination that yields the smallest total. In order to keep the computation tractable, we restrict the number of possible faults in any fault combination to be no more than three i.e. up to three genes can be faulty simultaneously. We constrain the cardinality of the drug vector to be less than or equal to three, in essence limiting the number of drugs in the combination to three since on account of the harmful side-effects of the drugs, administering four or more cancer drugs simultaneously might not be prudent.

For the Boolean network (BN) of Fig. 3.1, let N, M and P be the total number of primary inputs, primary outputs and fault locations respectively, then N=7, M=6 and P=24. Let $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$ be the input and output vectors respectively where $\mathcal{X}$ and $\mathcal{Z}$ represent the space of all binary vectors of dimensions N and M respectively. Let $\mathbf{x}^* = [0, 0, 0, 0, 1, 1, 0]$ be the input vector corresponding to the non-proliferative input.

Let D represent the total number of drug combinations (vectors) with no more than three drugs in any combination, then $D = \sum_{k=0}^{3} \binom{7}{k}$. Denote each drug vector in the drug space as $\mathbf{d_i}$ with $i = 0, \ldots, D-1$ ($\mathbf{d_0}$ is the all-zero drug vector meaning no drug is applied). Let $\mathcal{D}$ be this space of drug vectors.

Let C be the total number of fault combinations with no more than three faults in any combination, then $C = \sum_{k=0}^{3} \binom{P}{k}$. Assign each fault combination in the fault space a label f_j with $j = 0, \ldots, C-1$ (f_0 represents the fault-free case). Let $\mathcal{F}$ be this set of faults.

Let $\boldsymbol{\psi}$ denote the mapping from a given input vector, drug combination and fault combination to an output vector: $\mathbf{x} \in \mathcal{X}, \mathbf{d} \in \mathcal{D}, f \in \mathcal{F} \xrightarrow{\boldsymbol{\psi}} \mathbf{z} \in \mathcal{Z}$ i.e. $\boldsymbol{\psi}$ represents the output of the BN for a given input $\mathbf{x}$ when a drug combination $\mathbf{d}$ is applied under fault scenario f. Let ψ_i be the i$^{\text{th}}$ component of this M-dimensional vector $\boldsymbol{\psi}$.

The best drug vector $\mathbf{d_i},\ i \in \{0, 1, \ldots, D-1\}$ for each single fault $f_j,\ j \in \{1, 2, \ldots, P\}$ is the vector of smallest Hamming weight that minimizes $\|\boldsymbol{\psi}(\mathbf{x}^*, \mathbf{d_i}, f_j)\|_1$.

The optimal drug combination across all faults is:

$$\mathbf{d_i^*} = \arg\min_{\mathbf{d_i}} \sum_{j=1}^{C-1} \|\boldsymbol{\psi}(\mathbf{x}^*, \mathbf{d_i}, f_j)\|_1$$

$\mathbf{d_i^*}$ is determined by exhaustive enumeration by explicitly searching for the drug combination that for a non-proliferative input, minimizes the sum of Hamming weights (L_1-norms) of the output vector across all possible combinations of faults. This gives the drug cocktail of AZD6244, AZD5363 and Enzalutamide as a combination therapy for advanced prostate cancer. In a recent study, the drug combination of AZD5363 and Enzalutamide has demonstrated an impressive response in prostate cancer models [85]. Moreover, AZD6244 in partnership with an AKT pathway inhibitor (analogous to AZD5363), has been proposed as a strategy for the treatment of CRPC [65]. Thus, we propose that the aforementioned drug triad which represents a horizontal blockade approach, wherein combination therapy is used for the concerted pharmacologic inhibition of multiple compensatory pathways, as a therapeutic modality that may attenuate prostate cancer survival and growth.

3.4 Node vulnerability assessment

In electronic circuits, reliability refers to the probability of a circuit functioning as intended i.e. producing the correct output. Reliability assessment is used to determine the vulnerability of a circuit to faults. A number of different techniques have been proposed for reliability analysis in digital circuits [16]. Recently, in [33] a scalable, efficient and accurate simulation-based framework based on stochastic computations was introduced for logic circuit reliability evaluation. In biological systems, dysfunctions in nodes in the signaling network cause deviation from normative be-

havior. Reliability assessment methodologies can be leveraged on Boolean network models of pathways to determine the vulnerability of the network to the dysfunction of each node [1, 99]. In this section we conduct a stochastic logic based vulnerability analysis of the prostate cancer signal transduction network in order to discover the most vulnerable nodes thereby allowing us to prioritize such segments in the network whose perturbation has the greatest potential to yield the most clinical benefit.

In stochastic logic, signal probabilities are encoded in random binary bit streams (the signal probability of a node corresponds to the likelihood of that node having logic value one). For example, the binary sequence 0110010100 of length ten encodes the probability 0.4 since the proportion of ones in this sequence is $\frac{4}{10}$. In practice, the length of the stochastic sequences typically used is much larger. Since the biological literature is devoid of precise ligand binding probabilities, each primary input is assumed equally likely to be 0 or 1 i.e. all primary input signal probabilities are taken to be 0.5.

Stochastic logic often makes use of Bernoulli sequences for the random binary streams where each bit in the stream is generated independently from a Bernoulli random variable with a probability of one equal to p. The use of probabilistic sequences inevitably introduces stochastic fluctuations which implies that the result produced is non-deterministic. These fluctuations can be significantly reduced by representing the initial input probabilities by non-Bernoulli sequences [52] defined as random permutations of sequences containing a fixed number of ones and zeros. For a given probability p and sequence length L, a non-Bernoulli sequence contains a fixed number pL of ones, with the positions of the ones determined by a random permutation. Thus, for example, to represent the probability 0.5 by a non-Bernoulli stream of length 10, we could randomly permute the sequence 1111100000 which has five ones (instead of generating each bit from a Bernoulli random variable with

$p = 0.5$ as would have been done to represent the same probability by a Bernoulli sequence). We use non-Bernoulli sequences of random permutations of fixed number of ones and zeros in order to encode the initial input probabilities.

A logic circuit operating on stochastic bit streams (see Fig. 3.3 for an example), accepts as input random sequences representing the probability of each input being one and produces ones and zeros like any digital circuit [100] i.e. a stochastic logic circuit uses Boolean gates to operate on sequences of random bits. Each bit-stream represents a stochastic number interpreted as the probability of seeing a one in an arbitrary position. Thus, the computations performed by such a circuit are probabilistic in nature. The output bit stream produced can be decoded as the probability of the output being one by counting the number of ones in the stream and dividing by its length.

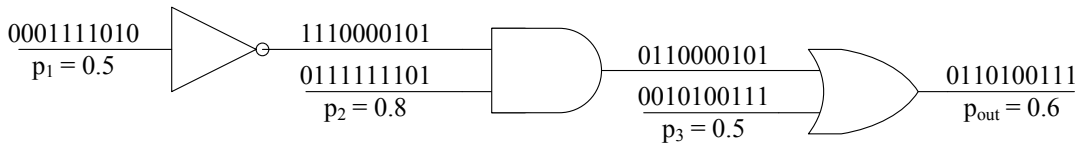


Figure 3.3: A stochastic logic circuit.

The vulnerability of a node is defined as the probability that the system produces incorrect output if that particular node is dysfunctional (faulty) i.e. it is the probability that the output of the network is different when that node is dysfunctional and is the complement of reliability. The procedure to determine the node vulnerabilities is illustrated in Fig. 3.4 is as follows. We generate non-Bernoulli sequences of length L=1,000,000 in which exactly half of the bits are set to one at each of the seven

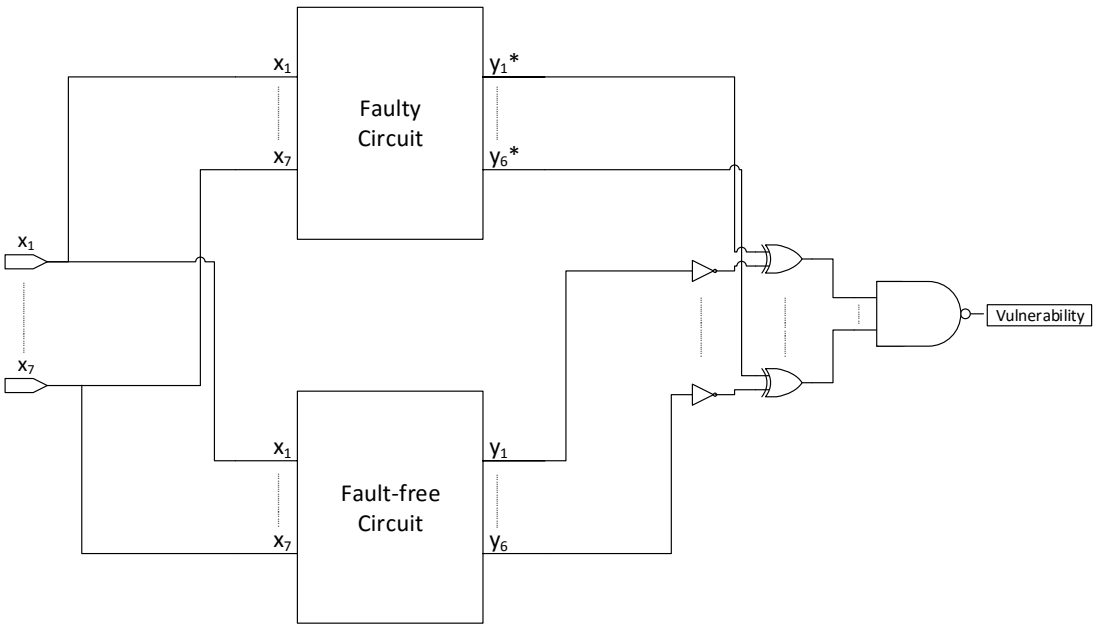


Figure 3.4: Stochastic architecture for computation of node vulnerability.

initial inputs (denoted by x_1 to x_7 in Fig. 3.4). The input stochastic sequences are propagated through both the original error-free circuit and the circuit in which the node of interest is dysfunctional. As discussed in the previous section, the dysfunction of a node is represented by a corresponding stuck-at fault of the requisite type at the particular location. This produces two sets of stochastic bit streams, one at each of the primary outputs of the fault-free circuit and the other at the primary outputs of the unreliable circuit (represented by y_1 to y_6 and y_1^* to y_6^* respectively in Fig. 3.4). The proportion of ones in the output bit stream encodes the output signal probabilities i.e. the probability of the output being one. Since the reliability of the circuit under the fault is the probability that the circuit output is same as that of the fault-free circuit, the sequence encoding the output reliability can be obtained from the output sequence of the faulty circuit by comparing it to the output sequence of

the fault-free circuit and setting each bit to one whenever the corresponding bits in the sequences are the same and zero if they are different. The proportion of ones in this resulting sequence will then correspond to the reliability of that output. Thus, we can obtain the stochastic sequence representing the reliability of each output by taking the XOR of each output bit stream of the faulty circuit with the complement of the corresponding output bitstreams of the fault-free circuit. For a circuit with multiple primary outputs as is the case here, the stochastic sequence encoding the joint output reliability can be obtained by taking the stochastic AND of the outputs of the XOR gates as the stochastic AND operation on the output of XOR gates produces a one only if all the corresponding bits at each XOR gate are one i.e. if all the corresponding bits in the respective outputs of the fault-free and faulty circuit are same. We then take the complement of the bit stream at the output of this AND gate to obtain the stream that encodes vulnerability. This bit stream can then be decoded to determine the node vulnerability with the proportion of ones in this stream equivalent to the vulnerability of the node.

Let $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}$ represent input non-Bernoulli sequences of length L with each sequence represented as a vector of length L whose i$^{\text{th}}$ component is equal to the i$^{\text{th}}$ bit in the sequence. Define the $L \times N$ matrix $X = \begin{pmatrix} \mathbf{x}_1^\top & \mathbf{x}_2^\top & \ldots & \mathbf{x}_N^\top \end{pmatrix}$. Thus, each row of this matrix contains the corresponding bits of each of the primary input streams. The vulnerability v_j of node $j \in \{1, 2, \ldots, P\}$ is given by:

$$v_j = \frac{1}{L} \sum_{k=1}^{L} \left(\prod_{i=1}^{M} \psi_i(\mathbf{x} = [X_{k1}, \ldots, X_{kN}], \mathbf{d_0}, f_j) \oplus \psi_i'(\mathbf{x} = [X_{k1}, \ldots, X_{kN}], \mathbf{d_0}, f_0) \right)'$$

where $'$ is the bit-complement operator and $\oplus$ is the binary XOR operator.

The procedure for computing the vulnerability of a node described above and

depicted in Fig. 3.4 is summarized as follows:

1. Generate non-Bernoulli streams encoding input probabilities at each of the primary inputs.

2. Propagate the input binary streams through the fault-free circuit and obtain a random bit sequence for each output.

3. Propagate the same input binary streams through the circuit with a stuck-at fault at the location of the node whose vulnerability we want to determine and again obtain a random bit sequence for each output.

4. XOR each primary output sequence from the faulty circuit obtained in step 3 with the complement of the corresponding primary output sequence from the fault-free circuit.

5. AND all the sequences obtained from each XOR gate. Take the complement of the stream so obtained. The vulnerability of the node is the fraction of ones in the resulting bit stream.

Thus, in a nutshell, the node vulnerabilities are obtained by propagating the initial input stochastic bit streams encoding the input probabilities through both the faulty and fault-free circuit, comparing the respective outputs obtained from each and decoding probabilities from the resulting streams. The vulnerability values so obtained are given in Table 3.2.

Vulnerability assessment can be used to identify candidates for targeted drug development. Nodes whose vulnerabilities are higher should be presumably better targets for drugs since potentially therapeutic benefit is more likely for nodes which are more vulnerable. We observe that the AR-mediated signaling axis remains a

Table 3.2: Node vulnerabilities.

Node	Vulnerability (%)
1	6.25
2	6.25
3	6.25
4	6.25
5	6.25
6	6.25
7	24.98
8	6.25
9	6.25
10	24.98
11	24.98
12	24.98
13	24.98
14	12.47
15	12.47
16	6.25
17	6.25
18	6.25
19	1.57
20	1.57
21	1.57
22	1.57
23	1.57
24	24.98

valid target. Furthermore, we see that dysfunction in the AKT nexus and the loss of tumor-suppressors have higher vulnerability values so drugs that attempt to alleviate these aberrations should be effective in attenuating tumor growth. The design of anti-cancer therapeutics directed at the loss of tumor suppressors has been difficult [21]. Additionally, AKT-selective drug development is challenging due to its homology with other kinases [9]. These complications notwithstanding, accelerated development of novel agents that target these aberrations is warranted. In contrast, the vulnerabilities for certain nodes such as those in the mTOR axis are low in-

dicating that they might not be attractive targets for drug development. Indeed, marginal clinical activity has been observed for mTOR inhibition with agents such as everolimus and temsirolimus failing to impact tumor proliferation in men with prostate cancer [20, 73]. Finally, in terms of the key pathways implicated in the disease we see that castration-resistant prostate cancer shows most vulnerability on aggregate to dysfunction in the AKT pathway. In a study it was demonstrated that the AKT pathway dominates AR signaling in CRPC [39].

4. DE NOVO TRANSCRIPTOME ASSEMBLIES AND ANNOTATION FOR PACIFIC WHITELEG SHRIMP[1]

4.1 Introduction

RNA-Seq, the sequencing of expressed messenger RNA by leveraging next generation deep-sequencing (NGS) technology, enables the rapid profiling of the transcriptome [57, 90]. For non-model species, the reference genome is not available and the massive amounts of short reads generated by RNA-Seq must therefore be assembled de novo by aligning them against each other into a set of putative transcripts. The goal of de novo transcriptome assembly is to reconstruct full-length transcripts from the short reads [56]. Accurate de novo assembly is a fundamental first step towards the reliable annotation of important non-model organisms for which little to no genomic information is available. Clearly, putting together a transcriptome from millions of short reads with sequencing errors in the absence of a reference is a computationally challenging problem. Moreover, in comparison to genome assembly, transcriptome assembly presents additional specific challenges [63] (for instance on account of alternative splicing, a single gene can code for multiple transcripts). To address the unique challenges of de novo transcriptome assembly, a number of software programs have been developed specifically for this purpose. There is though, no universal best transcriptome assembly program. Thus, it is important to understand the power and limitations of each algorithm, and choose the appropriate tool

[1]Parts of this section are reprinted with permission from Noushin Ghaffari, Osama Arshad, Hyundoo Jeong, John Thiltges, Michael Criscitiello, Byung-Jun Yoon, Aniruddha Datta, Charles Johnson, "Examining De Novo Transcriptome Assemblies via a Quality Assessment Pipeline", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, doi:10.1109/TCBB.2015.2446478 © 2015 IEEE and "De novo Transcriptome Assemblies and Annotation for Pacific Whiteleg Shrimp", *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, Atlanta GA, 3-5 Dec, 2014, doi:10.1109/GlobalSIP.2014.7032342 © 2014 IEEE.

accordingly.

Pacific whiteleg shrimp (Litopenaeus vannamei) is a prawn native to the eastern Pacific Ocean from southern Sonora Mexico to northern Peru, and largely farmed in the United States. Litopenaeus vannamei is a decapod (e.g., crabs, lobster, shrimp) crustacean of great interest as the dominant species in the global aquaculture industry. Whiteleg shrimp has great potential to provide food security. It however, suffers from pandemics caused by viruses [53]. Limited genomic information is available for this organism and there is a pressing need for enriching the current transcriptomic knowledge for this species.

In this study we used three leading transcriptome assembly algorithms namely SOAPdenovo-Trans [94], Trans-ABySS [72] and Trinity [32] to reconstruct the transcriptome of L. vannamei. We evaluate the assemblies across a number of metrics. Furthermore, the output transcripts from each assembly algorithm are annotated. Fig. 4.1 depicts the pipeline used to assess the quality of the resultant assemblies and annotate the transcripts. The results presentented herein will serve to improve the available transcriptome knowledge for an important non-model species.

4.2 RNA-Seq dataset

The RNA-Seq dataset consists of paired-end reads sequenced using Illumina HiSeq technology from samples obtained from shrimp abdominal muscle, hepatopancreas, gills and pleopods. The reads from the different samples were pooled together. The total number of reads is 199,528,356 in each read-pair with each read 100 base-pairs (bp) in length representing a total of about 40 Gigabases of sequence data.

4.3 Transcriptome assembly

Shrimp RNA-Seq reads [28] are assembled into contigs (contiguous sequences) or transcripts using three state-of-the-art de novo transcriptome assemblers: 1-

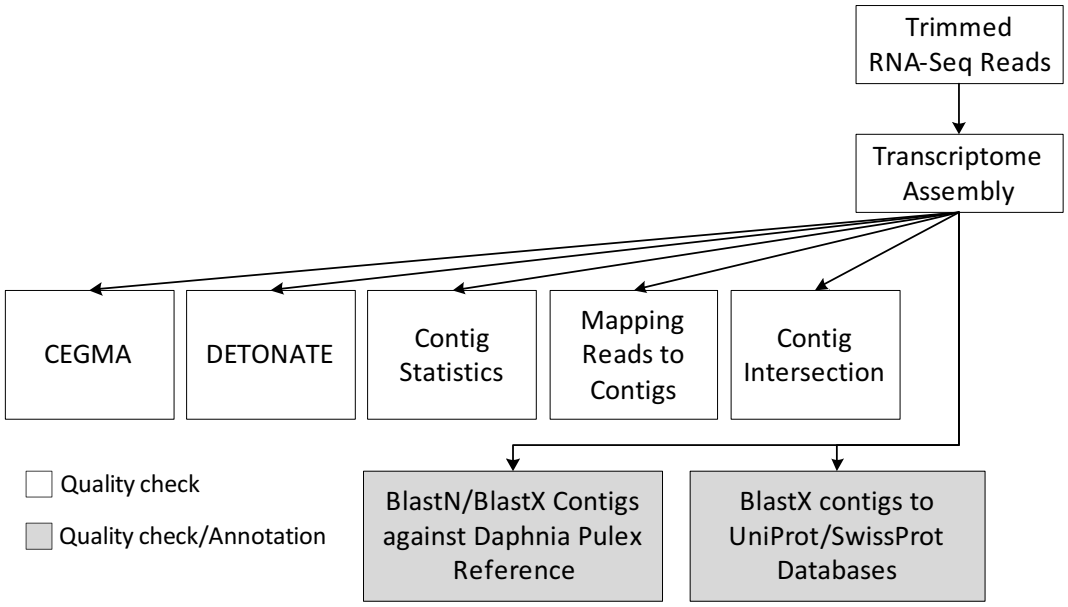


Figure 4.1: Transcriptome assembly, quality assessment, and annotation pipeline for Pacific whiteleg shrimp.

SOAPdenovo-Trans (release 1.03), 2- Trans-ABySS (version 1.5.1), 3- Trinity (release r2013-02-25). Default parameters were used for each program. It is customary to discard the short contigs of an assembly. Transcripts shorter than 200 base pairs (bp) are filtered out from the SOAPdenovo-Trans and Trans-ABySS assemblies (Trinity by default will only include transcripts longer than 200 bp in the final assembly).

4.4 Assembly statistics

Basic metrics of an assembly pertain to the size of the output along with statistics related to the length of the contigs [84] and include the total size in base pairs (span) of the assembly, number of assembled transcripts, length of the largest contig and the mean, median and N50 (defined as the contig size such that all the contigs equal to or greater than that size account for half of the total assembled bases [63]) of the contig length. These statistics for the three transcript assemblies are shown in Table

4.1.

We see from Table 4.1 that Trans-ABySS produces the largest number of transcripts whilst SOAPdenovo-Trans produces the fewest transcripts. However, SOAP produces longer transcripts as the mean and median contig length of its assembly is longer than those of the others. Trinity, produces somewhat fewer transcripts than Trans-ABySS but it creates the assembly with the largest span. Moreover, its N50 value is higher than those of the other assemblies.

Table 4.1: Standard assembly metrics.

	SOAPdenovo-Trans	**Trans-ABySS**	**Trinity**
Total number of contigs	62,514	119,772	110,474
Length of largest contig (bp)	30,864	17,067	31,344
Assembly size (bp)	74,156,520	105,766,302	125,657,935
Mean contig length (bp)	1,186	883	1,137
Median contig length (bp)	503	479	429
N50 (bp)	2,596	1,498	2,701
GC Content (%)	41.34	42.61	44.12

4.5 Evaluation of completeness of assemblies

The Core Eukaryotic Genes Mapping Approach (CEGMA) pipeline (version 2.5) [66] was used to evaluate the "completeness" of the assemblies. The pipeline defines a set of 248 highly-conserved proteins that are present in a wide variety of eukaryotes. Completeness refers to how many of these 248 core eukaryotic genes (CEGs) are present in the assemblies and is one of the metrics that can be used to assess the quality of an assembly. The number of CEGs represented in each assembly is shown in Table 4.2. We see that all three assemblers produce fairly complete assemblies preserving a very high number of CEGs. Trinity, in particular, performed very well,

conserving 246 CEGs.

Table 4.2: CEGMA evaluation.

	SOAPdenovo-Trans	**Trans-ABySS**	**Trinity**
No. of 248 ultra-conserved CEGs present	239	241	246
Completeness (percentage of CEGs present)	96.37	97.18	99.19

4.6 DETONATE evaluation

Recently, a transcriptome assembly evaluation methodology and corresponding software package called DETONATE (DE novo TranscriptOme rNa-seq Assembly with or without the Truth Evaluation) [51] has been developed. DETONATE is based on a probabilistic model and evaluates assemblies with or without a reference. The two components of the package are RSEM-VAL and REF-EVAL. RSEM-EVAL is a reference-free approach that only relies on the input RNA-Seq reads used to create the assembly and the assembly itself. REF-EVAL needs a reference and provides more information about the assembly than currently available tools. In this study, we used RSEM-EVAL since the ground-truth transcriptome of L. Vannamei is under construction. RSEM-EVAL is a model-based approach and provides a score to evaluate the assembly. The score is the log joint probability of the assembly and the reads, under the defined model.

We computed the RSEM-EVAL scores for the three assemblies. The scores are shown in Table 4.3. Higher scores correspond to better asssemblies. Based on RSEM-EVAL scores, we have the following rank ordering of the assemblers in terms of assembly quality: Trinity > Trans-ABySS > SOAPdenovo-Trans.

Table 4.3: RSEM-EVAL score for different assemblies.

	RSEM-EVAL score
SOAPdenovo-Trans	-18,898,212,410
Trans-ABySS	-13,589,448,379
Trinity	-4,295,090,084

4.7 Mapping of reads to assembled transcriptome

Another metric that can be used to evaluate the quality of an assembly is the alignment rate of the input reads against the assembled contigs. The greater the number of reads that map to the contigs, the better is the quality of the assembly. A "good" assembler will preserve the input information and utilize as much of the reads as possible to re-construct transcripts [98].

The input reads were mapped back to the assembled transcriptome using the program bowtie2 [46] with default parameters. The read-mapping rate (percentage of reads assembled into contigs) for each assembler is shown in Table 4.4. One can observe that Trinity and Trans-ABySS transcripts have higher read-mapping rates compared to SOAPdenovo-Trans.

Table 4.4: Read-mapping rate.

	SOAPdenovo-Trans	**Trans-ABySS**	**Trinity**
No. of input reads	199,528,356		
No. of reads mapped to contigs	151,233,987	181,585,589	179,204,712
Read mapping rate (%)	75.80	91.01	89.81

4.8 BLAST against Daphnia pulex references

Assembled transcripts of non-model organisms are generally validated by a similarity search with the reference sequences of a related organism that is well characterized [84]. The assembled transcripts of whiteleg shrimp were analyzed for sequence conservation against the references of a related species Daphnia pulex (water flea) obtained from the Joint Genome Institute (JGI) [18] using BLAST [58], a program designed to perform homology searches. The BLASTX and BLASTN tools were used to find similarities between the contigs and D. pulex proteins and transcripts/CDS, respectively.

An important parameter in a BLAST search is the "Expect (E)-value" which defines the expected number of hits by chance, with the closer the E-value threshold to zero, the more significant the matches [58]. We filtered our BLASTN and BLASTX results for two significance levels of 1E-4 and 1E-10 respectively. The number of transcripts with a BLASTX hit against the Daphnia pulex protein data set is shown in Fig. 4.2.

BLASTN of the contigs from each assembly against the Daphnia pulex transcript and CDS sequences was also carried out, and the results are presented in Fig. 4.3. Trans-ABySS transcripts have the highest absolute number of hits against the reference proteins (BLASTX), while the Trinity transcripts have the largest number of hits against the reference transcripts (BLASTN). The relative proportion of the contigs with a significant hit (ratio of contigs with a significant hit to the total number of contigs in the assembly) is fairly consistent across all three assemblers.

We also compared the protein homologies with Daphnia pulex for the three transcriptome assemblies. The homologous proteins in Daphnia pulex found from BLASTX for each assembly were intersected with the other assemblies to find the

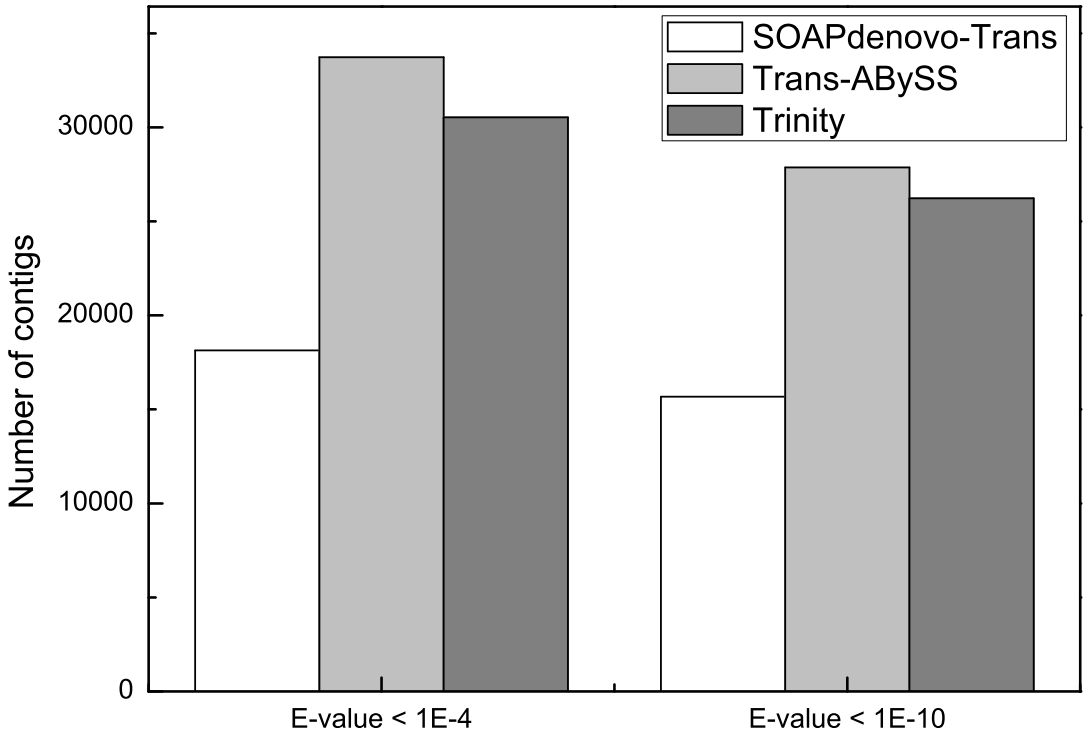


Figure 4.2: Contig BLASTX hits against Daphnia pulex protein database.

shared proteins. Fig. 4.4 shows the results as Venn diagrams that depict the unique and identical significant protein hits amongst the assemblies. There is a very high degree of concordance in the homologous proteins found in Daphnia pulex from all three assemblies.

4.9 BLAST against UniProt/SwissProt databases

The protein sequences and functional annotations for the assemblies were assigned by employing the BLASTX tool (BLAST+ version 2.2.29). BLASTX queries all six open reading frames (ORF) of a sequence against the protein database. We used the latest version of UniProt/SwissProt databases (released July 2014) as the reference. The BLASTX results for each assembly were sorted based on the most appearing protein hits. Comparing the protein hit frequencies shows that Trinity and Trans-

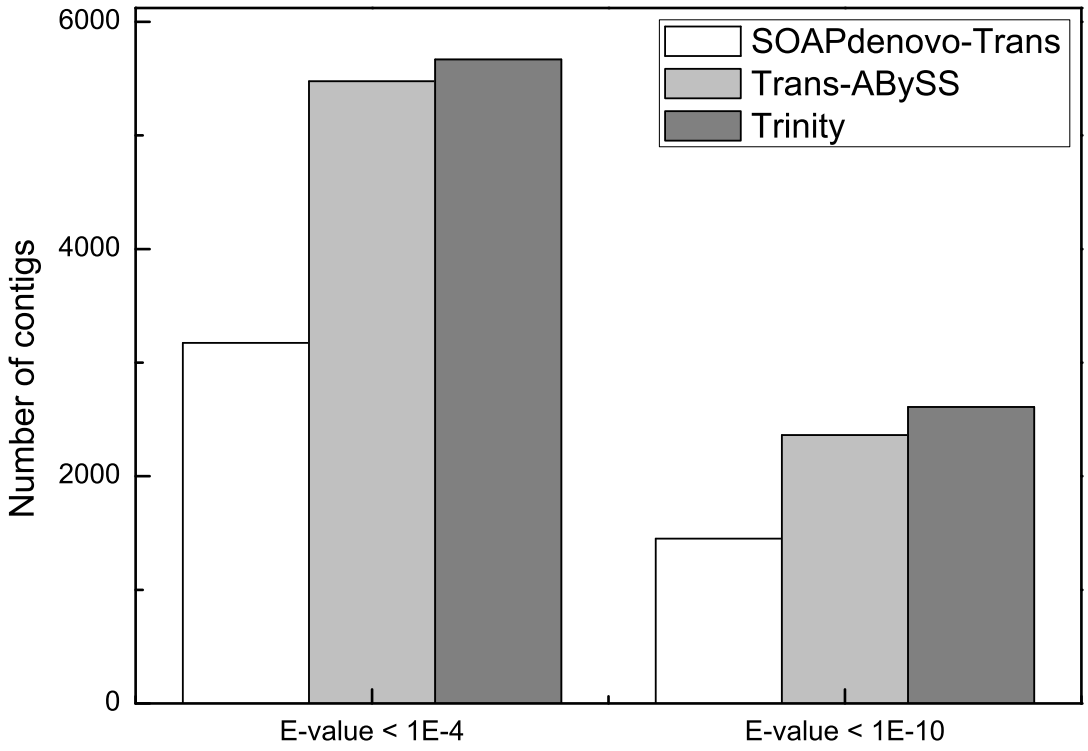
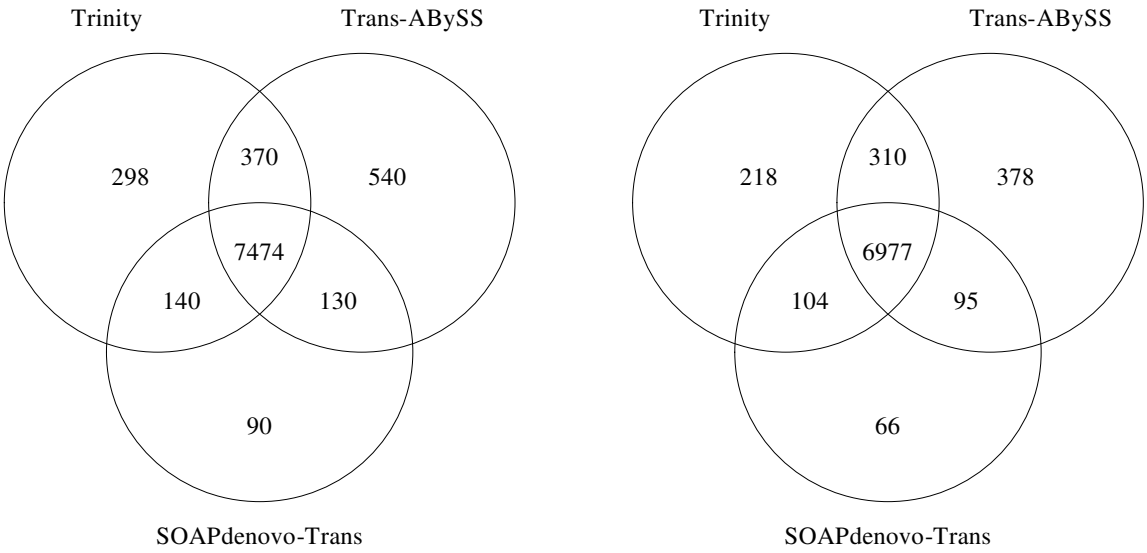


Figure 4.3: Contig BLASTN hits against Daphnia pulex transcript database.

ABySS have more similar BLASTX results. Fig. 4.5 breaks down the appearance of protein hits for different assemblies (with X in the figure defined as the total number of hits). Trans-ABySS has 86,122 hits, Trinity 81,122 and SOAPdenovo-Trans 44,972. The variation of the protein hits is likely caused by the larger total contigs that Trans-ABySS and Trinity algorithms produced. It should be noted that if the goal of the de novo transcriptome assembly study is to investigate the functional annotations of the species, the total number of contigs will play an important role.

Intersecting multiple BLASTX results on different assemblies can increase the confidence on the accuracy of the reported protein. We selected the proteins that appeared at least 15 times in the SOAPdenovo-Trans results and cross-checked them with Trinity and Trans-ABySS results. More than 99% of those proteins were in



(a) E-value < 1E-10 (b) E-value < 1E-10

Figure 4.4: Comparision of Daphnia pulex protein homologs found by BLASTX search from each assembly for different E-value thresholds.

common between the three assemblies. Our recommendation is to rely on the protein hits that are shared between at least two assembly BLASTX results.

4.10 Contigs intersection

The assembly methods differ in how they handle the challenges of reconstructing a complete transcriptome from short RNA-Seq reads. However, it is important to ensure that their final product is similar. To examine the resemblance, we used BLAST (version 2.29+) to match similar contigs between assemblies. A BLAST database was made from each assembly, and BlastN compared each database/assembly pair. Hits with $\geq$90% identity were considered matches, and were reported. Fig. 4.6 shows the results. The figure confirms that each of the three assemblers shares the majority of its transcripts with the other two. Trinity has the most unique transcripts when its contigs set is intersected with the other two contig sets. SOAPdenovo-Trans has

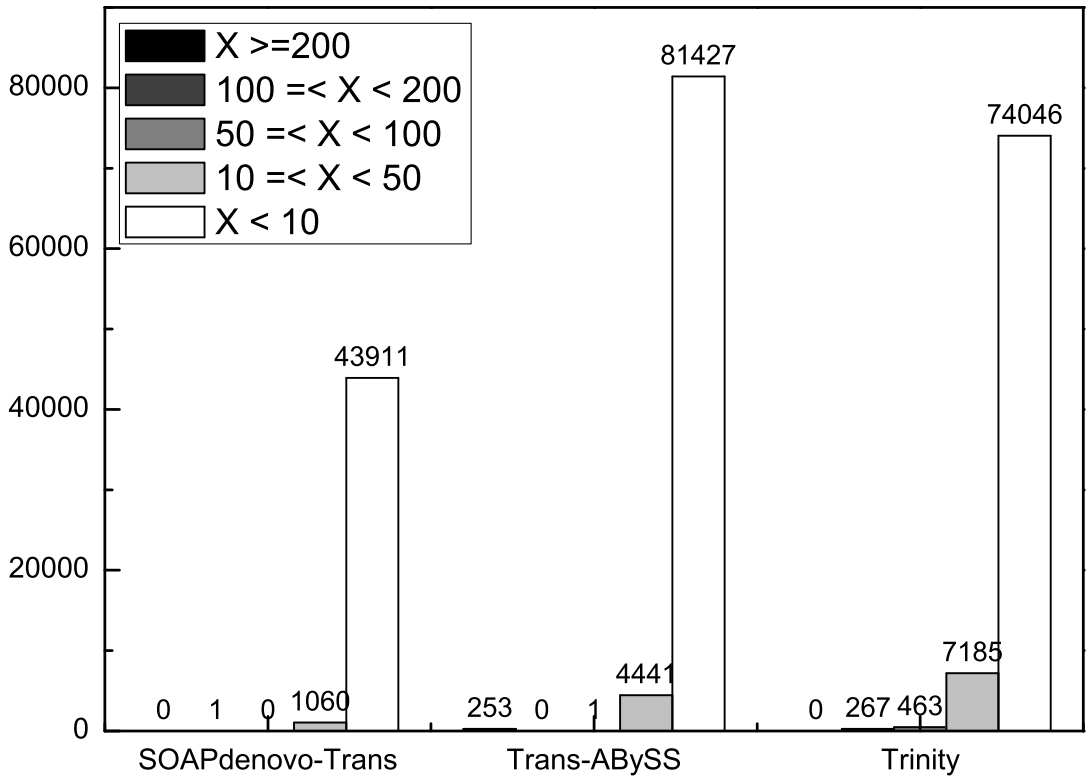


Figure 4.5: BLASTX search results for assemblies against UniProt and SwissProt databases.

the maximum contigs shared. However, it has the minimum total contigs among all three assemblers. Overall, the intersection of contigs indicates great compatibility among the three assembled contig sets.

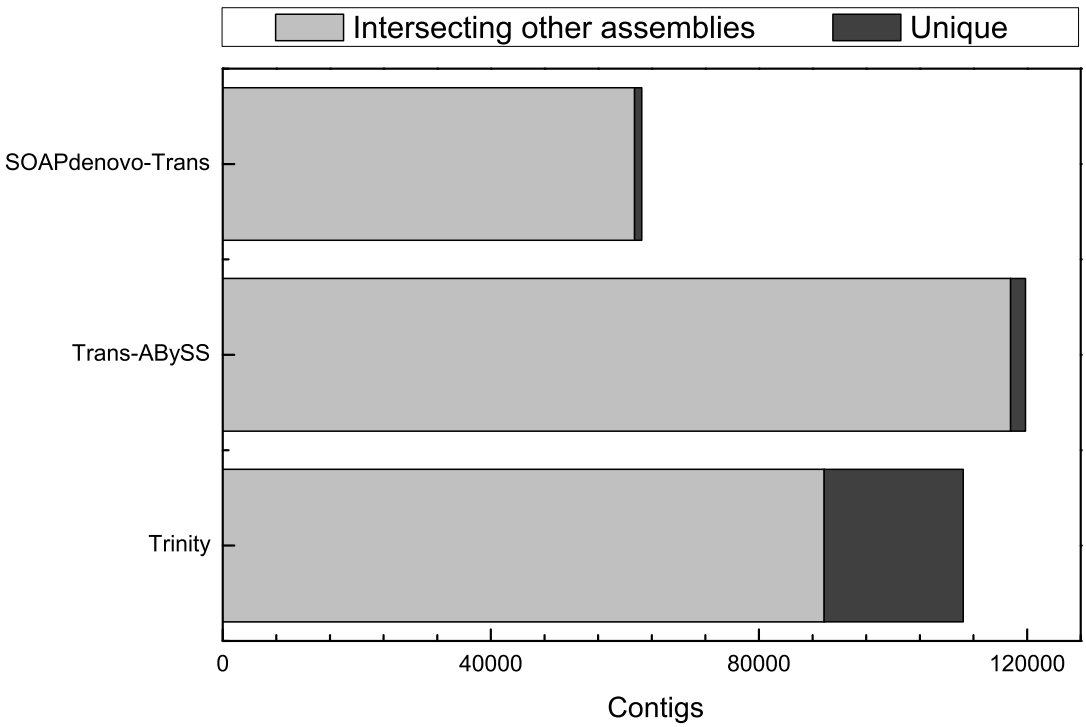


Figure 4.6: Intersection of each contig set with other assemblies' contigs.

5. SUMMARY[1]

Cancer cells are known to show atypical metabolic characteristics: an "Achilles' heel" [19] that provides a therapeutic opportunity. We have investigated via simulations the benefit of targeting tumor cell metabolism by using the anti-diabetic drug Metformin. The biological pathways involved in cell growth and metabolic regulation were mapped to a Boolean network. The equivalent digital circuit was used to identify locations at which faults could occur and categorize them into equivalent classes based on their output. We incorporated the drug intervention points into our model allowing us to test different combination therapies in terms of their efficacy in mitigating the effects of faults in the network. We have shown that incorporating Metformin in the therapeutic regimen can lead to better outcomes. We predict that a combination therapy of Metformin and cancer drugs will lead to improved cancer therapy design. One of our long term objectives is to experimentally validate such predictions using cancer cell lines.

Castration-resistant prostate cancer is a hormone refractory phenotype of significant morbidity and mortality in the prostate cancer disease continuum where patients no longer respond to androgen ablation therapy. The biomolecular network representing the signaling pathways involved in the pathogenesis of this lethal malignancy is translated to a digital circuit. The locations of possible malfunctions in the digital circuit are identified and computer simulation of the equivalent model

[1]Parts of this section are reprinted with permission from O. A. Arshad, P. S. Venkatasubramani, A. Datta and J. Venkatraj "Using Boolean Logic Modeling of Gene Regulatory Networks to Exploit the Links between Cancer and Metabolism for Therapeutic Purposes", *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 1, pp. 309-407, 2016, doi: 10.1109/JBHI.2014.2368391© 2016 IEEE and Noushin Ghaffari, Osama Arshad, Hyundoo Jeong, John Thiltges, Michael Criscitiello, Byung-Jun Yoon, Aniruddha Datta, Charles Johnson, "Examining De Novo Transcriptome Assemblies via a Quality Assessment Pipeline", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, doi:10.1109/TCBB.2015.2446478 © 2015 IEEE.

is used to predict effective therapies that mitigate the effect of different faults. A prospectively attractive combinatorial therapeutic strategy for the constellation of abnormalities is to leverage an AR axis targeted agent in conjunction with reciprocal inhibitors of other dysregulated pathways that are fundamental in coordinately driving oncogenesis. Proof of principle of clinical use for the proposed regimen remains to be demonstrated. A reliability (vulnerability) analysis methodology of digital circuits premised on stochastic logic modeling is utilized to quantify the vulnerability of the network to the dysfunction in discrete components in the signaling cascade thereby identifying key variables as targets for intervention that conceivably might be exploited by a new generation of novel therapeutics. These findings can contribute to the development of new rational approaches for the possible treatment of androgen-refractory prostate cancer. There is however a paucity of companion predictive biomarkers that can be used for the stratification of patients based on molecular aberrations in order to prescribe the apposite treatment. Furthermore, the histological and clinical heterogeneity of CRPC and the inherent redundancy along with the presence of feedback loops in pathways whose molecular underpinnings in the context of the disease induction and development are not yet fully understood, tender any potential translation into objective clinical efficacy of therapeutic implications derived from computations fraught with challenges.

New de novo transcriptome assembly and annotation methods provide an incredible opportunity to study the transcriptome of organisms that lack an assembled and annotated genome. There are currently a number of de novo transcriptome assembly methods, but it has been difficult to evaluate the quality of these assemblies. In order to assess the quality of the transcriptome assemblies, we composed a workflow of multiple quality check measurements that in combination provide a clear evaluation of the assembly performance. We presented novel transcriptome assemblies

and functional annotations for Pacific whiteleg shrimp (Litopenaeus vannamei), a mariculture species with great national and international interest, and no solid transcriptome/genome reference. We examined Pacific whiteleg transcriptome assemblies via multiple metrics, and provide an improved gene annotation. Our investigations show that assessing the quality of an assembly purely based on the assembler's statistical measurements can be misleading; we propose a hybrid approach that consists of statistical quality checks and further biological-based evaluations.

REFERENCES

[1] A Abdi, M B Tahoori, and E S Emamian. Fault diagnosis engineering of digital circuits can identify vulnerable molecules in complex cellular pathways. *Science Signaling*, 1(42), 2008.

[2] M Abramovici, M A Breuer, and A D Friedman. *Digital systems testing and testable design*. Wiley-IEEE Press, New York, 1994.

[3] N Agarwal, G Sonpavde, and C N Sternberg. Novel molecular targets for the therapy of castration-resistant prostate cancer. *European Urology*, 61(5):950–960, 2012.

[4] R Aggarwal and C J Ryan. Castration-resistant prostate cancer: targeted therapies and individualized treatment. *The Oncologist*, 16(3):264–275, 2011.

[5] R Albert and J Thakar. Boolean modeling: a logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 6(5):353–369, 2014.

[6] R H Alvarez, V Valero, and G N Hortobagyi. Emerging targeted therapies for breast cancer. *Journal of Clinical Oncology*, 28(20):3366–3379, 2010.

[7] B Bedogni, S M Welford, A C Kwan, J Ranger-Moore, K Saboda, and M B Powell. Inhibition of phosphatidylinositol-3-kinase and mitogen-activated protein kinase kinase 1/2 prevents melanoma development and promotes melanoma regression in the transgenic TPRas mouse model. *Molecular Cancer Therapeutics*, 5(12):3071–3077, 2006.

[8] C Belda-Iniesta, O Pernía, and R Simó. Metformin: a new option in cancer treatment. *Clinical and Translational Oncology*, 13(6):363–367, 2011.

[9] R L Bitting and A J Armstrong. Targeting the PI3K/Akt/mTOR pathway in castration-resistant prostate cancer. *Endocrine-related Cancer*, 20(3):R83–R99, 2013.

[10] L K Boyd, X Mao, and Y J Lu. The complexity of prostate cancer: genomic alterations and heterogeneity. *Nature Reviews Urology*, 9(11):652–664, 2012.

[11] M C Brahimi-Horn, J Chiche, and J Pouyssegur. Hypoxia signalling controls metabolic demand. *Current Opinion in Cell Biology*, 19(2):223–229, 2007.

[12] J R Cantor and D M Sabatini. Cancer cell metabolism: one hallmark, many faces. *Cancer Discovery*, 2(10):881–898, 2012.

[13] B S Carver, C Chapinski, J Wongvipat, H Hieronymus, Y Chen, S Chandarlapaty, V K Arora, C Le, J Koutcher, H Scher, P T Scardino, and N Rosen. Reciprocal feedback regulation of PI3K and androgen receptor signaling in PTEN-deficient prostate cancer. *Cancer Cell*, 19(5):575–586, 2011.

[14] Y Chen, N J Clegg, and H I Scher. Anti-androgens and androgen-depleting therapies in prostate cancer: new agents for an established target. *The Lancet Oncology*, 10(10):981–991, 2009.

[15] Y K Choi and K Park. Metabolic roles of AMPK and metformin in cancer cells. *Molecules and Cells*, 36(4):279–287, 2013.

[16] M R Choudhury and K Mohanram. Reliability analysis of logic circuits. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 28(3):392–405, 2009.

[17] D H Cohen and D LeRoith. Obesity, type 2 diabetes, and cancer: the insulin and IGF connection. *Endocrine-related Cancer*, 19(5):F27–F45, 2012.

[18] J K Colbourne, M E Pfrender, D Gilbert, W K Thomas, A Tucker, T H Oakley, S Tokishita, A Aerts, G J Arnold, M Kumar Basu, et al. The ecoresponsive genome of Daphnia pulex. *Science*, 331(6017):555–561, 2011.

[19] R J DeBerardinis, J J Lum, G Hatzivassiliou, and C B Thompson. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell Metabolism*, 7(1):11–20, 2008.

[20] C L Derleth and Y Y Evan. Targeted therapy in the treatment of castration-resistant prostate cancer. *Oncology*, 27(7):620–628, 2013.

[21] L M Dillon and T W Miller. Therapeutic targeting of cancers with loss of PTEN function. *Current Drug Targets*, 15(1):65–79, 2014.

[22] R J O Dowling, P J Goodwin, and V Stambolic. Understanding the benefit of metformin use in cancer treatment. *BMC Medicine*, 9(33), 2011.

[23] J V Duncia, J B Santella, C A Higley, W J Pitts, J Wityak, W E Frietze, F W Rankin, J-H Sun, R A Earl, A C Tabaka, et al. MEK inhibitors: the chemistry and biological activity of U0126, its analogs, and cyclization products. *Bioorganic & Medicinal Chemistry Letters*, 8(20):2839–2844, 1998.

[24] K Düvel, J L Yecies, S Menon, P Raman, A I Lipovsky, A L Souza, E Triantafellow, Q Ma, R Gorski, S Cleaver, et al. Activation of a metabolic gene regulatory network downstream of mTOR complex 1. *Molecular Cell*, 39(2):171–183, 2010.

[25] 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*,

22(14):e124–e131, 2006.

[26] J Feng, S L Zheng, W Liu, W B Isaacs, and J Xu. Androgen receptor signaling in prostate cancer: new twists for an old pathway. *Journal of Steroids and Hormonal Science*, 2011.

[27] H Fukazawa, K Noguchi, Y Murakami, and Y Uehara. Mitogen-activated protein/extracellular signal-regulated kinase kinase (MEK) inhibitors restore anoikis sensitivity in human breast cancer cell lines with a constitutively activated extracellular-regulated kinase (ERK) pathway. *Molecular Cancer Therapeutics*, 1(5):303–309, 2002.

[28] N Ghaffari, A Sanchez-Flores, D Ryan, K D Garcia-Orozco, P L Chen, A Ochoa-Leyva, A A Lopez-Zavala, J S Carrasco, C Hong, L G Brieba, et al. Novel transcriptome assembly and improved annotation of the whiteleg shrimp (Litopenaeus vannamei), a dominant crustacean in global seafood mariculture. *Nature Scientific Reports*, 2014.

[29] C E Giacomantonio and G J Goodhill. A Boolean model of the gene regulatory network underlying mammalian cortical area development. *PLoS Computational Biology*, 6(9), 2010.

[30] E Giovannucci, D M Harlan, M C Archer, R M Bergenstal, S M Gapstur, L A Habel, M Pollak, J G Regensteiner, and D Yee. Diabetes and cancer: a consensus report. *CA: a Cancer Journal for Clinicians*, 60(4):207–221, 2010.

[31] A M Gonzalez-Angulo and F Meric-Bernstam. Metformin: a therapeutic opportunity in breast cancer. *Clinical Cancer Research*, 16(6):1695–1700, 2010.

[32] M G Grabherr, B J Haas, M Yassour, J Z Levin, D A Thompson, I Amit, X Adiconis, L Fan, R Raychowdhury, Q Zeng, et al. Full-length transcriptome

assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7):644–652, 2011.

[33] J Han, H Chen, J Liang, P Zhu, Z Yang, and F Lombardi. A stochastic computational approach for accurate and efficient reliability evaluation. *IEEE Transactions on Computers*, 63(6):1336–1350, 2014.

[34] D Hanahan and R A Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[35] T Helikar, J Konvalina, J Heidel, and J A Rogers. Emergent decision-making in biological signal transduction networks. *Proceedings of the National Academy of Sciences*, 105(6):1913–1918, 2008.

[36] H A Hirsch, D Iliopoulos, P N Tsichlis, and K Struhl. Metformin selectively targets cancer stem cells, and acts together with chemotherapy to block tumor growth and prolong remission. *Cancer Research*, 69(19):7507–7511, 2009.

[37] A Hollestelle, F Elstrodt, J H A Nagel, W W Kallemeijn, and M Schutte. Phosphatidylinositol-3-OH kinase or RAS pathway mutations in human breast cancer cell lines. *Molecular Cancer Research*, 5(2):195–201, 2007.

[38] X Huang, S Wullschleger, N Shpiro, V A McGuire, K Sakamoto, Y L Woods, W Mcburnie, S Fleming, and D R Alessi. Important role of the LKB1–AMPK pathway in suppressing tumorigenesis in PTEN-deficient mice. *Biochemical Journal*, 412(2):211–221, 2008.

[39] M Kaarbø, Ø L Mikkelsen, L Malerød, S Qu, V H Lobert, G Akgul, T Halvorsen, G M Mælandsmo, and F Saatcioglu. PI3K-AKT-mTOR pathway is dominant over androgen receptor signaling in prostate cancer cells. *Analytical Cellular Pathology*, 32(1-2):11–27, 2010.

[40] M Kanehisa and S Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[41] M Kanehisa, S Goto, M Furumichi, M Tanabe, and M Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(suppl 1):D355–D360, 2010.

[42] M Kanehisa, S Goto, M Hattori, K F Aoki-Kinoshita, M Itoh, S Kawashima, T Katayama, M Araki, and M Hirakawa. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(suppl 1):D354–D357, 2006.

[43] S A Kaufmann. *The origins of order: self-organization and selection in evolution*. Oxford University Press, New York, 1993.

[44] J Kim, I Tchernyshyov, G L Semenza, and C V Dang. HIF-1-mediated expression of pyruvate dehydrogenase kinase: a metabolic switch required for cellular adaptation to hypoxia. *Cell Metabolism*, 3(3):177–185, 2006.

[45] G Kroemer and J Pouyssegur. Tumor cell metabolism: cancer's Achilles' heel. *Cancer Cell*, 13(6):472–482, 2008.

[46] B Langmead and S L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, 2012.

[47] R Layek, A Datta, M Bittner, and E R Dougherty. Cancer therapy design based on pathway logic. *Bioinformatics*, 27(4):548–555, 2011.

[48] R J Lee, A C Armstrong, and A M Wardley. Emerging targeted combinations in the management of breast cancer. *Breast Cancer: Targets and Therapy*, 5:61–72, 2013.

[49] S Leo, C Accettura, and V Lorusso. Castration-resistant prostate cancer: targeted therapies. *Chemotherapy*, 57(2):115–127, 2010.

[50] A J Levine and A M Puzio-Kuter. The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. *Science*, 330(6009):1340–1344, 2010.

[51] B Li, N Fillmore, Y Bai, M Collins, J A Thomson, R Stewart, and C N Dewey. Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biology*, 15(12):553, 2014.

[52] J Liang and J Han. Stochastic Boolean networks: an efficient approach to modeling gene regulatory networks. *BMC Systems Biology*, 6(113), 2012.

[53] Donald V Lightner. The penaeid shrimp viral pandemics due to IHHNV, WSSV, TSV and YHV: history in the Americas and current status. *Proceedings of the 32nd Joint UJNR Aquaculture Panel Symposium, Davis and Santa Barbara, California, USA*, pages 17–20, 2003.

[54] B Liu, Z Fan, S M Edgerton, X Deng, I N Alimova, S E Lind, and A D Thor. Metformin induces unique biological and molecular responses in triple negative breast cancer cells. *Cell Cycle*, 8(13):2031–2040, 2009.

[55] J S Logue and D K Morrison. Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy. *Genes & Development*, 26(7):641–650, 2012.

[56] J Martin, V M Bruno, Z Fang, X Meng, M Blow, T Zhang, G Sherlock, M Snyder, and Z Wang. Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads. *BMC Genomics*, 11(663), 2010.

[57] Jeffrey A Martin and Zhong Wang. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10):671–682, 2011.

[58] S McGinnis and T L Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32(suppl. 2):W20–W25, 2004.

[59] K P McKian and P Haluska. Cixutumumab. *Expert Opinion on Investigational Drugs*, 18(7):1025–1033, 2009.

[60] C E Meacham and S J Morrison. Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337, 2013.

[61] A K Mohanty, A Datta, and V Venkatraj. A model for cancer tissue heterogeneity. *IEEE Transactions on Biomedical Engineering*, 61(3):966–974, 2014.

[62] R Munagala, F Aqil, and R C Gupta. Promising molecular targeted therapies in breast cancer. *Indian Journal of Pharmacology*, 43(3):236–245, 2011.

[63] M Mundry, E Bornberg-Bauer, M Sammeth, and P G D Feulner. Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach. *PLoS One*, 7(2), 2012.

[64] I Papandreou, R A Cairns, L Fontana, A L Lim, and N C Denko. HIF-1 mediates adaptation to hypoxia by actively downregulating mitochondrial oxygen consumption. *Cell Metabolism*, 3(3):187–197, 2006.

[65] H Park, Y Kim, J-W Sul, I G Jeong, H-J Yi, J B Ahn, J S Kang, J Yun, J J Hwang, and C-S Kim. Synergistic anticancer efficacy of MEK inhibition and dual PI3K/mTOR inhibition in castration-resistant prostate cancer. *The Prostate*, 75(15):1747–1759, 2015.

[66] G Parra, K Bradnam, and I Korf. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9):1061–1067, 2007.

[67] J C Patel, B L Maughan, A M Agarwal, A M Batten, T Y Zhang, and N Agarwal. Emerging molecularly targeted therapies in castration refractory prostate cancer. *Prostate Cancer*, 2013.

[68] G Pérez-Tenorio, L Alkhori, B Olsson, M A Waltersson, B Nordenskjöld, L E Rutqvist, L Skoog, and O Stål. PIK3CA mutations and PTEN loss correlate with similar prognostic factors and are not mutually exclusive in breast cancer. *Clinical Cancer Research*, 13(12):3577–3584, 2007.

[69] M A Pierotti, F Berrino, M Gariboldi, C Melani, A Mogavero, T Negri, P Pasanisi, and S Pilotti. Targeting metabolism for cancer treatment and prevention: metformin, an old drug with multi-faceted effects. *Oncogene*, 32(12):1475–1487, 2013.

[70] B J Quinn, H Kitagawa, R M Memmott, J J Gills, and P A Dennis. Repositioning metformin for cancer prevention and treatment. *Trends in Endocrinology & Metabolism*, 24(9):469–480, 2013.

[71] J Renshaw, K R Taylor, R Bishop, M Valenti, A D H Brandon, S Gowan, S A Eccles, R R Ruddle, L D Johnson, F I Raynaud, et al. Dual blockade of the PI3K/AKT/mTOR (AZD8055) and RAS/MEK/ERK (AZD6244) pathways synergistically inhibits rhabdomyosarcoma cell growth in vitro and in vivo. *Clinical Cancer Research*, 19(21):5940–5951, 2013.

[72] G Robertson, J Schein, R Chiu, R Corbett, M Field, S D Jackman, K Mungall, S Lee, H M Okada, J Q Qian, et al. De novo assembly and analysis of RNA-Seq data. *Nature Methods*, 7(11):909–912, 2010.

[73] D Sarker, A H M Reid, T A Yap, and J S de Bono. Targeting the PI3K/AKT pathway for the treatment of prostate cancer. *Clinical Cancer Research*, 15(15):4799–4805, 2009.

[74] U G A Sattler, F Hirschhaeuser, and W F Mueller-Klieser. Manipulation of glycolysis in malignant tumors: fantasy or therapy? *Current Medicinal Chemistry*, 17(2):96–108, 2010.

[75] R Schlatter, K Schmich, I A Vizcarra, P Scheurich, T Sauter, C Borner, M Ederer, I Merfort, and O Sawodny. ON/OFF and beyond-a Boolean model of apoptosis. *PLoS Computational Biology*, 5(12), 2009.

[76] C M Schlotter, U Vogt, H Allgayer, and B Brandt. Molecular targeted therapies for breast cancer treatment. *Breast Cancer Research*, 10(4), 2008.

[77] S Schott, A Schneeweiss, and C Sohn. Breast cancer and diabetes mellitus. *Experimental and Clinical Endocrinology & Diabetes*, 118(10):673–677, 2010.

[78] A Schulze and A L Harris. How cancer metabolism is tuned for proliferation and vulnerable to disruption. *Nature*, 491(7424):364–373, 2012.

[79] R J Shaw and L C Cantley. Ras, PI(3)K and mTOR signalling controls tumour cell growth. *Nature*, 441(7092):424–430, 2006.

[80] R L Siegel, K D Miller, and A Jemal. Cancer statistics 2015. *CA: a Cancer Journal for Clinicians*, 65(1):5–29, 2015.

[81] C W Song, H Lee, R P M Dings, B Williams, J Powers, T Dos Santos, B Choi, and H J Park. Metformin kills and radiosensitizes cancer cells and preferentially kills cancer stem cells. *Scientific Reports*, 2(362), 2012.

[82] S Sridharan, R Layek, A Datta, and J Venkatraj. Boolean modeling and fault diagnosis in oxidative stress response. *BMC Genomics*, 13(Suppl 6):S4, 2012.

[83] S Sridharan, R Varghese, V Venkatraj, and A Datta. Hypoxia stress response pathways: Modeling and targeted therapy. *IEEE Journal of Biomedical and Health Informatics*, in press.

[84] Shawn T. O′ Neil and Scott J Emrich. Assessing de novo transcriptome assembly metrics for consistency and utility. *BMC Genomics*, 14(465), 2013.

[85] P Toren, S Kim, T Cordonnier, C Crafter, B R Davies, L Fazli, M E Gleave, and A Zoubeidi. Combination AZD5363 with enzalutamide significantly delays enzalutamide-resistant prostate cancer in preclinical models. *European Urology*, 67(6):986–990, 2015.

[86] A Vazquez-Martin, C Oliveras-Ferraros, S Cufí, B Martin-Castillo, and J A Menendez. Metformin and energy metabolism in breast cancer: from insulin physiology to tumour-initiating stem cells. *Current Molecular Medicine*, 10(7):674–691, 2010.

[87] P Vigneri, F Frasca, L Sciacca, G Pandini, and R Vigneri. Diabetes and cancer. *Endocrine-related Cancer*, 16(4):1103–1123, 2009.

[88] N Vijesh, S K Chakrabarti, and J Sreekumar. Modeling of gene regulatory networks: a review. *Journal of Biomedical Science and Engineering*, 6(2):223–231, 2013.

[89] R Wang, A Saadatpour, and Reka Albert. Boolean modeling in systems biology: an overview of methodology and applications. *Physical Biology*, 9(5), 2012.

[90] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10:57–63, 2009.

[91] O Warburg. On the origin of cancer cells. *Science*, 123(3191):309–314, 1956.

[92] S Watterson, S Marshall, and P Ghazal. Logic models of pathway biology. *Drug Discovery Today*, 13(9):447–456, 2008.

[93] P J Wysocki and B Wierusz-Wysocka. Obesity, hyperinsulinemia and breast cancer: novel targets and a novel role for metformin. *Expert Review of Molecular Diagnostics*, 10(4):509–519, 2010.

[94] Y Xie, G Wu, J Tang, R Luo, J Patterson, S Liu, W Huang, G He, S Gu, S Li, et al. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30(12):1660–1666, 2014.

[95] S J Yeung, J Pan, and M-H Lee. Roles of p53, MYC and HIF-1 in regulating glycolysis-the seventh hallmark of cancer. *Cellular and Molecular Life Sciences*, 65(24):3981–3999, 2008.

[96] H. F. Yuen, O Abramcyzk, G. Montgomery, K K Chan, Y H Huang, T Sasazuki, S Shirasawa, S Gopesh, K W Chan, D Fennell, P Janne, M El-Tanani, and J T Murray. Impact of oncogenic driver mutations on feedback between the PI3K and MEK pathways in cancer cells. *Bioscience Reports*, 32(4):413–422, 2012.

[97] R Zhang, M V Shah, J Yang, S B Nyland, X Liu, J K Yun, R Albert, and T P Loughran. Network model of survival signaling in large granular lymphocyte leukemia. *Proceedings of the National Academy of Sciences*, 105(42):16308–16313, 2008.

[98] Q-Y Zhao, Y Wang, Y-M Kong, D Luo, X Li, and P Hao. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*, 12(Suppl 14):S2, 2011.

[99] P Zhu, H M Aliabadi, H Uludağ, and J Han. Identification of potential drug targets in cancer signaling pathways using stochastic logical models. *Scientific Reports*, 6, 2016.

[100] P Zhu and J Han. Asynchronous stochastic boolean networks as gene network models. *Journal of Computational Biology*, 21(10):771–783, 2014.