Biological Signaling Pathways and Potential Mathematical Network Representations: Biological Discovery through Optimization

by

Juan Fernando Rosas Rubio

A thesis submitted in partial fulfillment of the requirements for the degree of
MASTERS OF SCIENCE
in
INDUSTRIAL ENGINEERING
UNIVERSITY OF PUERTO RICO
MAYAGÜEZ CAMPUS
2015

Approved by:

_____ _____
Saylisse Dávila, PhD Date
Member, Graduate Committee

_____ _____
Jaime Seguel, PhD Date
Member, Graduate Committee

_____ _____
Mauricio Cabrera-Ríos, PhD Date
President, Graduate Committee

_____ _____
Edgar Acuña, PhD Date
Graduate Studies Representative

_____ _____
Viviana Cesaní, PhD Date
Department Director

Abstract

Establishing the role that different genes play in the development of cancer is a daunting task. A step towards this end is the detection of genes that are important in the illness from high-throughput biological experiments. Furthermore, it is safe to say that even with a list of potentially important genes it is highly unlikely that these show expression changes independently. A biological signaling pathway is a more plausible underlying mechanism as favored in the literature. This thesis attempts to build a mathematical network problem through the analysis of microarray experiments. A pre-selection of genes is carried out with a multiple criteria optimization framework previously published by our research group [1]. Afterwards, network optimization methods Traveling Salesperson Problem and Minimum Spanning Tree are proposed to identify potential signaling pathways via the most correlated path among the genes of interest. A case study with lung cancer was conducted where our methodology identified 11 potential biomarkers genes and optimal pathway solutions maximizing correlations between them. Additionally a meta-analysis was undertaken for lung cancer obtaining several gene relationships that may play a key role when smoking habits are taken in consideration. Biological evidencing is provided to assess the effectiveness of the proposed methods.

Resumen

Establecer el papel que diferentes genes juegan en el desarrollo del cáncer es una tarea de gran proporción. Un paso hacia este objetivo es la detección de genes que son importantes en el desarrollo de una enfermedad a partir de experimentos biológicos de alto rendimiento. Es seguro expresar que incluso con una lista de genes potencialmente importantes es poco probable que estos muestran cambios de expresión de forma aislada. Una vía de señalización biológica es un mecanismo subyacente más convincente como es favorecido en la literatura. Esta tesis involucra el análisis experimentos de micro-arreglos para construir un problema de red matemática que modele una ruta de señalización altamente probable. Una pre-selección de genes se lleva a cabo con un marco de optimización de múltiples criterios publicados previamente desarrollado por nuestro grupo de investigación [1]. Posteriormente se emplean métodos de optimización de redes como el problema del agente viajero (traveling salesperson problem) y árbol recubierto mínimo (minimum spanning tree) para identificar potenciales vías de señalización vía la ruta más correlacionada entre genes de interés. Un caso de estudio es realizado con cáncer de pulmón donde nuestra metodología fue capaz de identificar 11 biomarcadores potenciales y soluciones de rutas óptimas maximizando las correlaciones entre dichos genes. Adicionalmente se llevó a cabo una meta-análisis en cáncer de pulmón obteniendo relaciones de genes que pueden jugar un rol clave cuando se consideran hábitos de fumar. Se proveyó evidencia biológica para evaluar la efectividad de los métodos propuestos.

Dedication

This thesis is dedicated to my family and my home country of Honduras. I will always appreciate the guidance and support my parents have given me throughout my life. I love you. Also to my brother and sister, although we may not have to closest of relationships that does not diminish the fact that I love you and my work is also dedicated to you. Lastly my work is dedicated to my country which is constantly in my mind. Today Honduras may not be in the best of circumstances but my hopes are that one day the opportunities that I was blessed with will become accessible to the vast majority people of my country. Honduras, as any other country, has the great potential to be a land of equality, peace, great economic growth although this may not be clear this day. I hope in the near future to be in a position to contribute significantly to this dream. I will be part in one way or another of my childhood dream of seeing Honduras and its people become a great and prosperous nation.

Acknowledgements

Firstly, I am grateful for the opportunities given to me. The two people that I am most grateful to are my parents Romeri Edilect Rubio and Juan Carlos Rosas who have worked hard their entire lives to provide their family with quality education and a comfortable life. They always are to me an inspiration and example to follow. Both come from humble families of Honduras and Peru, but through hard work and devotion are now dedicated professionals. Thank you for support, guidance and tough love that sometimes is needed to push us along.

I must also give praise to the person who was my advisor throughout the process of developing and ultimately defending my thesis, Dr. Mauricio Cabrera. Without his guidance and support this thesis would not have been possible. Thank you for letting form part of your research group and trusting me with the responsibility of carrying out this thesis. Additionally I wanted to thank Dr. Clara Isaza for her help and advice, her knowledge and expertise were key in this research.

I must recognize someone who today has become a very important part of my life, my dear Nitza Garcia. Thank you for giving me the opportunity of being a part of your life and for always supporting me throughout my Master's studies, I love you so very much. I can say without any doubt that I am glad of making the decision to come and study in Puerto Rico, because I was able to meet you. Also I must thank my classmates and friends that accompanied me in our different classes and research group during these two and a half years. Thank you Katia Camacho, Enery Lorenzo, Yaileen Mendez,

Esmeralda Niño, Samuel Bonet, Cesar Salazar, and the many other graduate students who I shared this experience with.

I also want to thank my friends back home in Honduras, Jose A. Oyuela, Ana Acosta, and Ricardo Vicente. Thank you for the years of friendship and hilarious moments we have shared.

The work and dedication of several students of the department of Biology must also be recognized. Thank you Cristina Ortiz, Arlette Marrero, Michael Ortiz and Lynn Perez for the contributions you have done to this thesis.

I want to also thank the members of my graduate committee Dr. Saylisse Davila and Dr. Jaime Seguel for their recommendations and guidance. I must thank the Department of Industrial Engineering for giving the opportunity to come here and for financing my studies throughout this time. I want to also thank Glorimar Lopez for letting me be her teaching assistant for nearly two years, I appreciate you kindness and advice.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Introduction

The study of cancer biomarkers is an important issue due to the role they play in the early detection, diagnosis, and prognosis of cancer. A biomarker can be defined as a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes or pharmacologic responses to a therapeutic intervention [2]. Studying biomarkers and their behavior has great potential to contribute to the discovery and understanding of the origin and evolution of cancer.

The analysis of high throughput experiments such as microarrays can lead to the identification of potential cancer biomarkers. Microarrays can be used to simultaneously analyze thousands of genes. Biomarker genes obtained from these experiments are classified as potential biomarkers due to the fact that the biomarkers identified from microarrays are not experimentally validated yet.

Unlike other diseases such as cystic fibrosis [3] or muscular dystrophy [4], where mutations of one gene can cause disease, no single gene “causes” cancer [5]. Identifying potential cancer biomarker genes plays a critical part in the understanding of cancer, but establishing how these genes possibly interact with one another and how this interaction could possibly contribute to the evolution of cancer has a role that is just as important. One possible and accepted way to better understand these interactions is to identify a signaling pathway among cancer biomarkers. A signaling pathway can be defined as the sequential interaction of products (such as proteins) of different genes which cause an

effect in the behavior of a cell. The abnormal activation of signaling pathways can lead to several different diseases, including cancer. Identifying and understanding these abnormal signaling pathways could possibly contribute to diagnose and fight cancer [6].

This thesis discusses the issue of identifying a potential signaling pathway among a list of potential lung cancer biomarkers through a network representation that leads to optimal configurations that maximize the association values between these genes. The list of potential biomarkers is determined from microarray databases through the application of Multiple Criteria Optimization [1], a strategy proposed by our research group. The network optimization models initially proposed in this work to obtain potential signaling pathways among potential cancer biomarkers are the Traveling Salesperson Problem (TSP) [7] and the Minimum Spanning Tree (MST) [8].

1.2 Objective

The objective of this work is to find a suitable network representation as a proxy to build potential signaling pathways among potential cancer biomarker genes and its optimal solution through the use of network optimization methods. To this end, the study will focus on lung cancer to capitalize on previous work in our research group.

1.3 Motivation

Cancer is a disease with much importance due to its high incidence worldwide and its associated mortality rates. Cancer presents itself in many forms and types, which hinders and complicates advances in research aimed at unveiling its origin and the development of drugs capable of preventing or combating the illness. Biomarkers could have an important role in contributing significantly to our understanding of cancer.

Identifying biomarkers is a crucial step, but assuming that these biomarkers act completely independent of one another is improbable. Biological signaling pathways are more likely models of characterization, although identifying biological signaling pathways from a list of genes of interest presents complications of its own. As an example of the complexity of searching for potential signaling pathways let us take a relatively small list of ten genes of interest: there are (10-1)!=362,880 possible pathways to create a looping path among them. Such cases as the one mentioned before emphasize the importance of efficiently searching the associated high-dimensional space using optimization.

1.4 Scope

The scope of this thesis involves identifying potential signaling pathways in sets of previously selected potential cancer biomarker genes through network optimization models, such as the Traveling Salesperson Problem and the Minimum Spanning Tree. Additionally as a second part of this thesis, the simultaneous analysis of multiple microarray experiments (meta-analysis) is undertaken to identify signaling pathways among multiple sets of databases. These methodologies are applied in the study of lung cancer, to capitalize on previous results from our research group regarding potential biomarkers in this illness.

1.5 Thesis Organization

This thesis is organized as follows: the second chapter is a compilation and overview of relevant literature found regarding the subjects of cancer, biomarkers, signaling pathways, meta-analysis, the discovery or structuring of signaling pathways and general background information on the proposed methodologies of this thesis. The third chapter

emphasizes the proposed methodology, specifically on how the Traveling Salesman Problem (TSP) and Minimum Spanning Tree (MST) can be utilized to construct a signaling pathway from a list of genes of interest identified previously through Multiple Criteria Optimization (MCO). The fourth chapter presents case studies of lung cancer where the proposed methods are applied. Additionally in this chapter an initial meta-analysis is carried out to evaluate different conditions describe in a lung cancer database. A fifth chapter includes biological information gathered from public databases such as KEGG and GeneCards to validate our methodology and to search for opportunities to propose unknown gene relations relevant for lung cancer. The sixth conducts a comparison of our methodology to GeneMANIA, a program that constructs gene networks. The seventh and final chapter establishes the general conclusions of this thesis and lays out the direction for future research.

Chapter 2. Background

2.1 Cancer

The term cancer is used to define an uncontrolled growth of abnormal cells the accumulation of which can interfere with normal tissue functions and is capable of infecting other tissues [9]. When the latter occurs it is called metastasis, which is the principal reason for cancer related deaths [10]. Cancer takes the name of its organ of origin: colon cancer, lung cancer, and so on. There are over 100 different known types of cancer. According to the American Cancer Society, about 1,658,370 new cases of cancer are expected to be diagnosed in 2015 and about 589,430 Americans are expected to die of cancer the same year. This is about 1,620 deaths per day [11]. Lung cancer, which is the main focus of this dissertation, is one of the primary types of cancer related to cancer deaths. For 2015, the expected number of new lung cancer cases is 221,200 and the estimated number of deaths is 158,040 in the United States. [11]. According to statistics from the National Cancer Institute, lung cancer is ranked second among types cancer in estimated new cases and deaths for 2015 in the United States.

2.2 Biomarkers

Biomarkers play an important role in the discovery and understanding of different diseases. Biomarkers can be defined as a characteristic that can be objectively measured and evaluated as an indicator of a physiological as well as a pathological process or pharmacological response to a therapeutic intervention [12]. Biomarkers can be studied and analyzed to track disease progression over time, evaluate the effect of certain drugs and possibly serve as surrogate end points in clinical trials. Specifically when dealing with

cancer, a cancer biomarker is any measurable-specific molecular alteration of a cancer cell either on DNA, RNA, protein or on metabolite level. In this investigation, potential biomarkers genes are identified by applying the methodology proposed by Katia I. Camacho [13] that utilizes Multiple Criteria Optimization. It is based on the differences in genetic expression of genes when comparing control tissues and lung cancer tissues.

2.3 Signaling Pathways

A signaling pathway is a series of actions among molecules in a cell that leads to a certain product or change in a cell [6]. Signaling pathways can trigger the assembly of new molecules, turn genes on or off or spur a cell to move. Signaling pathways have a key role in various functions of a cell, for this reason they are of interest in cancer research. Cancer can derive from an array of different genetic mutations. Studying the pathways that were disrupted by the genetic mutations could possibly narrow the search for improving treatments designed to combat cancer growth and development. As described in the following sections of this thesis, it was pointed out that there are several methods in the literature to identify genes of interest and possible pathways that match these genes. Statistical tests and probability distributions (e.g., Fisher's exact test, hyper-geometric distributions, among others.) are highly utilized first to then make use of specialized database search engines to possibly find matches in already known pathways.

2.4 Meta-analysis

One of the main aims of this thesis is to carry out meta-analysis from different microarray studies. This involves analyzing several thousand set genes to uncover

potential biomarkers through the use of Multiple Criteria Optimization based on Pareto conditions [13] followed by the application of network optimization methods, the Traveling Salesman Problem and Minimum Spanning Tree to identify how these potential biomarkers interact with each other.

According to Glasser, the term Meta-analysis can be used to describe methods for the systematic review of a set of individual studies or patients within each study, with the aim to quantitatively combine their results [14]. There are several different examples and applications of meta-analysis in literature.

Rice, Murphy and Tworoger carried out a systematic search and meta-analysis of articles from different sources (such as PubMed and Web of Science, among others) with the purpose of determining the strength of the association between gynecologic surgeries, tubal ligation and hysterectomy, and ovarian cancer [15]. Their investigation encompassed English-language articles dated between 1969 through March 2011. From this search the authors extracted relative risks and 95% confidence intervals or p-values from selected articles. The authors decided a priori to use a random-effects model to calculate the summary relative risks and 95% confidence intervals. The researchers conducted meta-regression analyses to assess whether effect estimates differed by study design (i.e. case-control versus cohort versus other design) and by population studied (i.e. general population versus BRCA mutation carriers). Additionally, they conducted meta-regression analyses in subsets of the studies to assess whether the effect estimates differed by age at procedure, years since procedure, and for tubal analysis. The researchers were able to conclude that observational epidemiologic evidence strongly

supports that tubal ligation and hysterectomy are associated with a decrease in the risk of ovarian cancer, by approximately 26-30% [15].

Wu, Ye, and Shi made use of meta-analysis for the systematic investigation of the association between dietary vitamin A, retinol intake and blood retinol level and gastric cancer risk [16]. The researchers conducted a literature search in PubMed and Embase for relevant studies. In total, thirty-one studies were included. Either a fixed-effect model or a random-effect model was adopted to pool the study-specific relative risk (RR) according to the heterogeneity. If the heterogeneity was significant, the random effect model was applied, otherwise the fixed-effect model was used. Heterogeneity across studies was tested by the authors with the chi-square test and the I^2 test. The I^2 test quantifies the proportion of total variation across studies due to heterogeneity rather than chance. A proportion≤0.10 in combination with I^2 >50% was taken to signify heterogeneity. The authors analyzed the dose-response relationship using fractional polynomial regression of the inverse variance-weighted data. Comparing the highest with the lowest categories, vitamin A intake significantly reduces gastric cancer risk (pooled RR-0.66, 95% confidence interval: 0.52-0.84), whereas the authors found a marginally inverse association between retinol intake (pooled RR-0.94, 95%CI: 0.87-1) or blood retinol level (pooled RR- 0.87, 95% CI: 0.73-1) and gastric cancer [16].

There has been several studies in the literature involving the application of meta-analysis in microarray experiments to discover outstanding gene expression. An example is the work of Rhodes, Barrette, and Rubin who demonstrate a statistical model for performing meta-analysis of independent microarray datasets of prostate cancer [17]. Their model was implemented on four prostate cancer gene expression data sets. For

each gene in each study, the authors tested the null hypothesis that no relationship exists between the expression values of the gene and the comparison between cancer and benign tissue. The authors then implemented a meta-analysis model to assess the similarity of the findings between studies to ultimately identify sets of over and under-expressed genes in prostate cancer. Their cross validation approach identified a group of genes that were differentially expressed.

The meta-analysis approach has been implemented for several different works as stated before, but in the literature there were no articles found describing the use of the network optimization methods with meta-analysis to identify signaling pathways of interest for a disease. This, in our particular case would be lung cancer. These evidence the novelties of the ideas in this thesis.

2.5 Literature Review

There is an inherent difficulty in analyzing microarrays (and –omics,zsaaa in general) that is linked to the large size of these experiments. Microarray experiments simultaneously measure the expression levels of thousands of genes, generating large amounts of data. The analysis of this data presents a challenge to biologists. Therefore, new tools are needed to derive biological insight from these experiments, including signaling pathways [18] [19].

Currently several experimental methods for determining signaling pathways exist. Signaling pathways, which according to Baxevanis and Ouellette involve many direct protein-protein relationships, can be mapped using protein-protein interaction detection methods [20]. According to the authors, a common experiment depends on co-

purification, using methods known as chromatography and nuclear magnetic resonance. Chromatography is based on the principle that molecular mixture can be decomposed based on component physiochemical properties, such as size or charge [21]. Nuclear magnetic resonance can be used to identify small molecules and proteins directly based on atomic distance measurements and mass respectively [21].

In co-purification, according to Baxevanis and Ouellette proteins that strongly interact will purify as a complex that can be degraded further using harsher purification conditions finally to separate and identify the complex components [20]. The authors state that the definition of a protein complex depends on the purification conditions used. Voet and Voet mention that many more experimental methods exist, but almost all current experiments suffer from observer effect, where the conditions of the experiment disturb the natural biological process. It is only after multiple types of experiments have been performed that a result can be considered reliable [21]. Each experiment involves an investment of economic resources. It can be established that also these experiments have room for errors which can hamper the obtained results. Therefore there exist opportunities to the methods to identify and analyze signaling pathways. The methodology proposed in this thesis, based on network optimization methods, is capable of determining an optimal solution. This overcomes limitations of experimental methods such as the ones mentioned before which obtain local optimal solutions for the problem of identifying signaling pathways.

Additionally, there are also different computational methods for the identification or analysis of signaling pathways. The following table lists software programs found during a recent literature search.

Table 2.1 Pathway Analysis Software [19]

Tool	Method
ArrayXPath	Fisher exact test; Multiple testing correction
Pathway Miner	Fisher exact test
SPIA	p-value, false discovery rate
PathRanker	3M,HME3M
MAPPFinder	Standardized difference score (z) from hypergeometric distribution

The ArrayXPath is a web-based tool for microarray gene expression profile mapping and visualization from biological pathway resources [22]. ArrayXPath automatically recognizes microarray probe identifiers from submitting data and maps onto the pathway database to then calculate the statistical significance of the association of the two data pieces [19]. In spotted microarrays, the probes are oligonucleotides, cDNA, or small fragments of PCR products that correspond to mRNAs. Each probe contains a different, characteristic sequence that is specific to a different group of genes under study. ArrayXPath maps the different identifier sets between microarray probes and node (see Appendix 1) and computes the statistical significance of the association between gene-expression clusters and pathways. It provides an automated annotation of clusters with ranked pathways [22]. ArrayXPath applies Fisher's exact test to evaluate the statistical significance of the correlations. This software analyzes clusters of gene expression and searches already existing pathway databases such as GenMAPP, Kyoto Encyclopedia of Genes and Genomes (KEGG), and BioCarta, among others. The methods proposed in this thesis do not require genes to be analyzed as clusters and identifies a potential pathway solution among previously identified highlighted over or under expressed genes

for a disease of interest, such as in our case study with lung cancer described later along in this work. The methodologies proposed in this work are based on mathematical optimization, which serve as a contrast to the statistical approaches found in literature, such as those in ArrayXPath.

Pathway Miner is another web-based tool that mines gene associations and networks in biological pathway information. The tool permits the analysis and interpretation of pathway information and networks based on the association with databases, suitable for high throughput analysis of gene expression data [19]. Pathway Miner provides two options to analyze genes in a dataset: (1) to search genes based on their associations in metabolic and/or cellular and regulatory pathways from pathway resources and (2) perform a statistical test and rank significant pathways based on their p-values from three different resources KEGG, BioCarta and GenMapp [23]. Pathway Miner applies a one-sided Fisher's exact test, as ArrayXPath does. Similar to several existing methods and software, Pathway Miner relies on statistical procedures. The methodologies presented in this thesis, in contrast, are of deterministic nature.

Signaling pathway impact analysis (SPIA) is a software tool used to measure the actual perturbation on a given pathway under a particular condition [24]. This package provides a technique for pathway analysis based on combination of two types of evidence, the over-representation of differently expressed genes and the perturbation of the pathway as measured by expression changes. For each pathway, a p-value is calculated. With the assumption that the number of differentially expressed (NDE) genes in a pathway follows a hypergeometric distribution, the probability of the number of differentially expressed (P_{NDE}) genes in the given pathway is calculated. The second

probability, probability of pathway perturbation (P_{PERT}), is calculated based on the estimated mount of perturbation in each pathway due to the differential expression of the input gene list [19]. This differential expression is calculated by subtracting the expression levels between all the control samples and case samples of a particular gene. In the particular case of this thesis, these would be the differences in expression levels between control samples and lung cancer samples. SPIA and its probabilistic capability could benefit from the deterministic optimal solution proposed to be elicited in this thesis for comparison or referencing purposes.

PathRanker is a software tool that can identify genetic pathways that dictate the response of metabolic networks to specific experimental conditions [25]. Initially, PathRanker uses a nonparametric pathway extraction method to identify the most correlated paths through a metabolic network. Then, it extracts the defining structure within these top-ranked pathways using both Markov clustering and classification algorithms. Furthermore, detailed node and edge (see Appendix 1) annotations are defined, which enables to track each pathway, not only with respect to its genetic dependencies, but also allow for an analysis of the interacting reactions, compounds and KEGG sub-networks [25]. PathRanker relies on probabilistic clustering, to which a deterministic optimal solution might add a useful point of comparison too.

MAPPFinder is another software that can dynamically link gene-expression data to Gene Ontology (GO) [26] hierarchies. MAPPFinder calculates the percentages of genes measured that meet a user-defined criterion [18]. Such criterion is required for each specific GO node, and for the cumulative total of the number of genes meeting the criterion in a parent GO term combined with all its children, giving a complete picture of

the number of genes associated with a particular GO term. Using this percentage and a z-score, the user ranks the GO terms by their relative amounts of gene-expression changes [27]. As mentioned before, MAPPFinder requires a user-defined criterion, which can be very subjective and possibly affect the convergence of the results. These issues might be circumvented by the methods like those proposed in this work.

The literature review presented here provides evidence of an opportunity to create a deterministic optimization-driven approach to the construction of signaling path proxies, even as a way to contrast results from the stochastic/statistic approaches already available.

2.6 Methodology Background

2.6.1 Traveling Salesperson Problem

The Traveling Salesman Problem (from this point on referred to as the Traveling Salesperson Problem or TSP) is one the most famous combinatorial optimization problems. In its most common interpretation, the TSP tries to construct the shortest tour through *n* cities [28], for a salesperson to visit, usually going back to a preselected base city [29]. In other words, the TSP consists of an optimization problem that searches for a cyclic sequence within a network that minimizes a certain measure (such as costs, distances, among others).

As an example, take the network presented in Figure 2.1. A salesperson has to start traveling from city 1 through each city exactly once and return home to city 1. If the objective were to obtain a route that minimizes the total distances, that is a cycle (see

Appendix1) with minimum total distance, there would be total of (4-1)! =24 possible cycles [30].

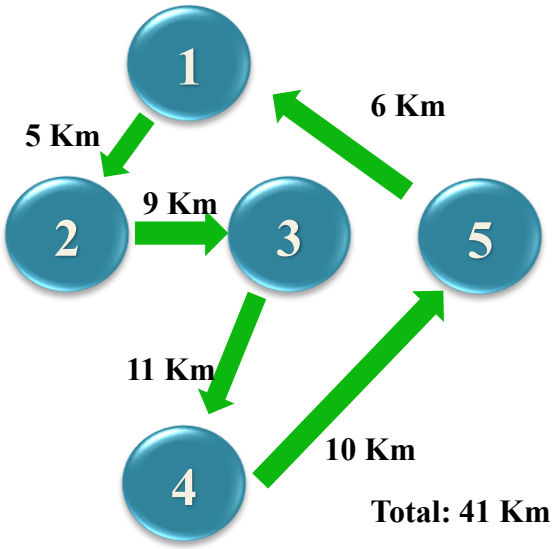

Figure 2.1 Example of the Application of the TSP

The TSP can be modeled as a network optimization problem. Consider that c_{ij} represents the cost of traveling from city *i* to city *j* and let y_{ij} be a binary variable, indicating whether or not the salesperson travels from city *i* to city *j*. Additionally let us define flow variables x_{ij} on each arc (*i,j*) and assume that the salesperson has *n*-1 units available at node 1, which is arbitrarily selected as a "source node", and the salesperson must deliver 1 unit to each of the other nodes or vertices [7]. The model is as follows [7]:

$$Minimize \sum_{(i,j)\in A} c_{ij} y_{ij} \qquad (2.1)$$

$$\sum_{1\le j\le n} y_{ij} = 1 \quad \forall i = 1,2,\dots,n \qquad (2.2)$$

$$\sum_{1 \le i \le n} y_{ij} = 1 \quad \forall\, j = 1,2,\dots,n \qquad (2.3)$$

$$\mathrm{N}x = b \qquad (2.4)$$

$$x_{ij} \le (n-1)y_{ij} \quad \forall (i,j) \in A \qquad (2.5)$$

$$x_{ij} \ge 0 \quad \forall (i,j) \in A \qquad (2.6)$$

$$y_{ij} = 0 \; or \; 1 \quad \forall (i,j) \in A \qquad (2.7)$$

Let $A' = \{(i,j): y_{ij} = 1\}$ and let $A'' = \{(i,j): x_{ij} > 0\}$. The constraints (2.2) and (2.3) imply that exactly one arc of *A'* leaves and enters any node *i*; therefore, *A'* is the union of node disjoint cycles containing all of the nodes of N. In general, any integer solution satisfying (2.2) and (2.3) will be union of disjoint cycles; if any such solution contains more than once cycle; they are referred to as subtours, since they pass through only a subset of nodes [7].

In constraint (2.4) N is a nxm matrix, called the *node-arc incidence matrix* of the minimum cost flow problem. Each column N_{ij} in the matrix corresponds to the variable x_{ij}. The column N_{ij} has a value of $a+1$ in the ith row, and a value of $a-1$ in the jth row; the rest of its entries are zero. Constraint (2.4) ensures that *A"* is connected since we need to send 1 unit of flow from node 1 to every other node via arcs in *A".* The forcing constraints (2.5) imply that *A"* is a subset *A'*. These conditions imply that the arc set *A'* is connected and thus cannot contain subtours [7].

2.6.2 Minimum Spanning Tree

The Minimum Spanning Tree considers an undirected and connected network, where the given information includes some measure of the positive length (e.g., distance,

cost, time, etc.) associated to each link [8]. The MST involves choosing a set of links that have the shortest total length among all sets of links that ensure that the chosen links provide a path (see Appendix 1) between each pair of nodes [29].

An example of the MST is presented in Figure 2.2 where we have five nodes of a network with their potential links and the positive length for each if it is inserted into the network. Enough links must be inserted to satisfy the requirement that there is a path between every pair of nodes. The objective is to satisfy this requirement while at the same time minimizing the total length of the links inserted into the network, in the example in Figure 2.2 the solution is highlighted by the darker and thicker lines. In the case of a spanning tree the total number of possible solutions con be calculated with Cayley's formula n^{n-2} where n is the number of edges or arcs in the graph [31]. In this particular example there is a total of 5^{5-2} =125 possible solutions

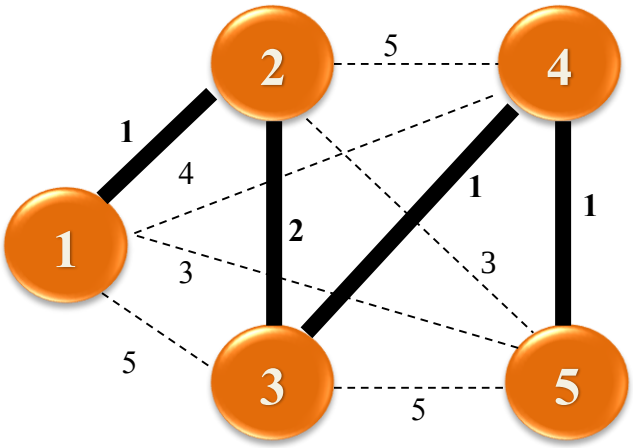

Figure 2.2 Example of a Minimum Spanning Tree (MST)

Chapter 3. Identification of Signaling Pathways through Network Optimization

3.1 Proposed Methodology

How can a signaling pathway related to cancer be identified starting from measurements from hundreds or thousands of genes? As mentioned before in this work, the methods proposed to attempt to obtain an optimal signaling pathway are the Traveling Salesperson Problem and the Minimum Spanning Tree. In order to begin to apply these methods however, the first step is to identify genes of interest based on the differences of expression when comparing control and cancer tissues. Multiple Criteria Optimization based on Pareto functions as described in the works of Lorenzo et al. [32] (see Appendix 3) and Camacho et al. [13]. A MCO program developed for MatLab by Katia Camacho (see Appendix 2) is applied. As a next step, a structuring task using the proposed network models ensue. The networks use statistical linear relationships between genes as measured by Pearson correlation between pairs of genes [33].

Statistical correlation can be defined as a measure of the coordinated behavior between two random variables. It measures the strength or degree of association between two variables, say X and Y. Linear correlation values range from -1 to +1. The closer the linear correlation values are to either +1 or -1, the more intense the correlation can be considered. If we have a series of *n* measurements of X and Y written as x_i and y_i where $i = 1, 2 \ldots n,$ then the sample correlation coefficient can be used to estimate the Pearson correlation *r* between X and Y as follows:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i-1}^{n}(x_i - \bar{x})^2 \sum_{i-1}^{n}(y_i - \bar{y})^2}} \quad (3.1)$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, and s_x and s_y are the sample standard deviations of X and Y [34].

When a linear correlation is positive, this implies a positive association between the two variables being analyzed; for example larger values of X tend to be associated with larger values of Y and smaller values of X tend to be associated with smaller values of Y (see Figure 3.1). When a linear correlation has a negative value this implies a negative or inverse association, that is larger values of X would tend to be associated with smaller values of Y and smaller values of X could tend to be related to larger values of Y (see Figure 3.2) [35].

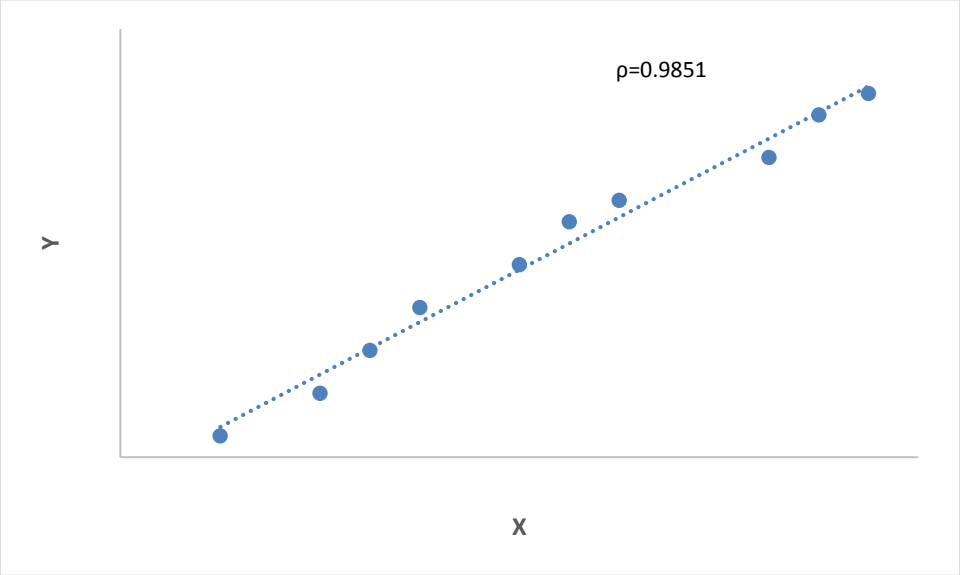

Figure 3.1 Positively Correlated Data

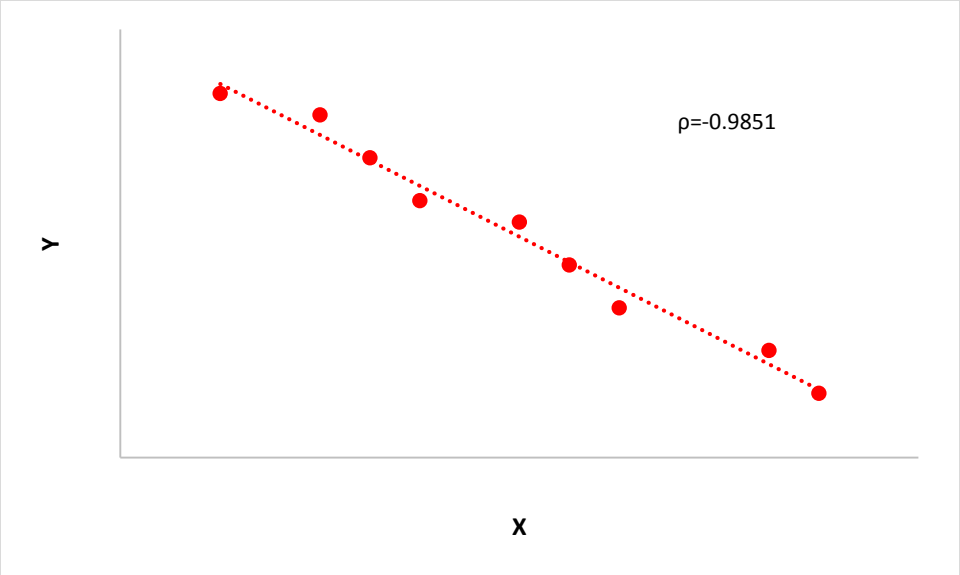

Figure 3.2 Negatively Correlated Data

In this work, as a first approximation, the linear correlations that can be observed among a list of genes considered to be potential biomarkers are used as a base to construct networks such as the one presented in Figure 3.3.

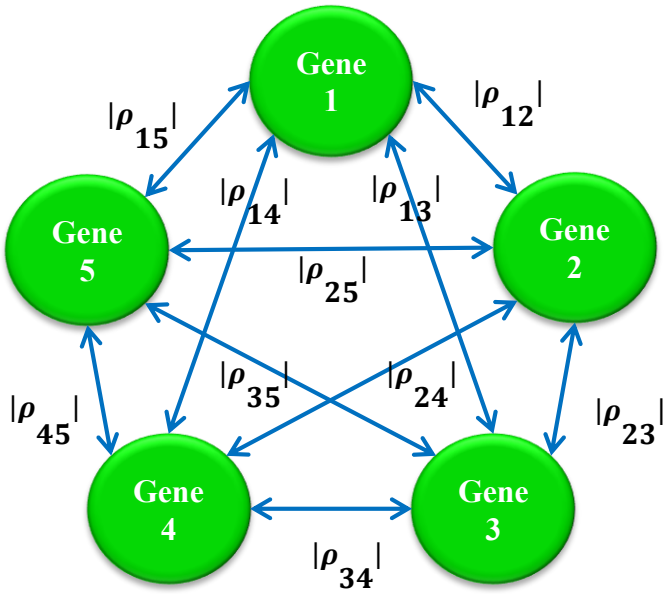

Figure 3.3 Representation of a Potential Sequence of a Signaling Pathway

As it can be observed in Figure 3, there are many possible sequences that can be followed in order to discover a potential signaling pathway even from among a small list of genes. In this example specifically five genes can have (5-1)! = 24 possible solutions. The number of possible solutions grows exponentially as more potential biomarkers are analyzed in order to discover a signaling pathway. Applying the TSP can serve to support the discovery of a signaling pathway among a list of genes of interest with efficiency and efficacy since, if solved optimally, it will arrive to the most correlated cycle path.

3.1.1 Traveling Salesperson Problem (TSP)

The TSP can be used to discover a potential signaling pathway from a network of genes by identifying a sequence that maximizes the linear correlation among the genes [32].

For the TSP, a list of potential biomarkers of interest must first be identified. In the case of this work, this list was pre-selected using the method of Multiple Criteria Optimization [1] developed by our research group from a microarray database of lung cancer containing more than 22,283 genes with a total of 107 control and cancer tissues [36]. Multiple Criteria Optimization aims to identify the best compromise between solutions from a set of possible solutions characterized by at least two possible performance measures in conflict. Once the potential biomarkers are established, the next step is to obtain the differences in genetic expression of each gene when comparing all of its normal tissue samples against all of its cancerous tissue samples. Figure 3.4 graphically exhibits the differences of genetic expression in the case of the gene SPP1 where each point is the level of expression for either a control or cancer tissue. SPP1 was

identified as a potential lung cancer biomarker utilizing multiple criteria optimization based on Pareto optimality conditions.

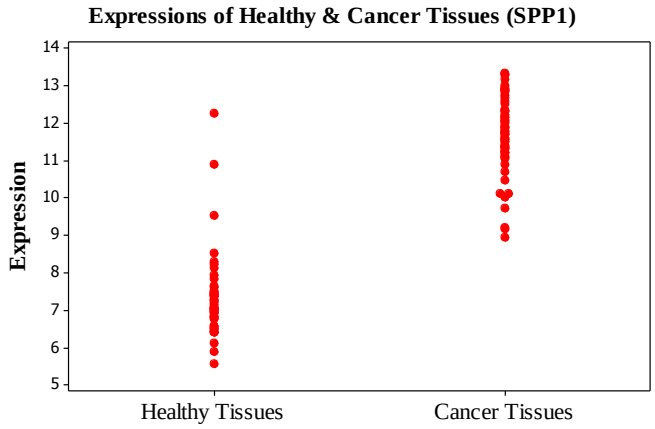

Figure 3.4 Representation of the Expression Differences of Normal and Cancer Tissues

The differences of genetic expression for all the genes considered to be potential biomarkers are used to calculate the linear correlations among each and every one of the genes. The values of the differences serve as input for calculating the linear correlations between all the genes being considered (see Equation 3.1). Figure 3.5 represents the correlation between genes SPP1 and AGER based on the differences in genetic expression of the two genes.

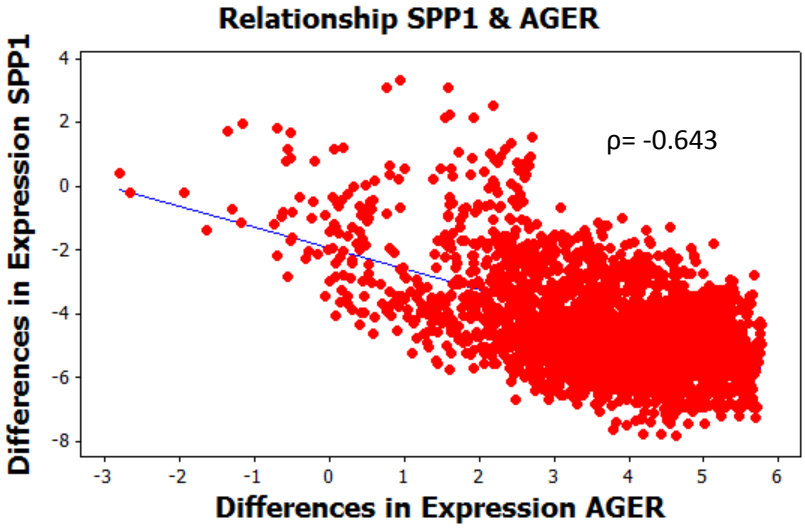

Figure 3.5 Example of Correlation between Potential Biomarkers

The linear correlation values obtained can be associated to the arcs in a network model, such as the one in Figure 3.3. The absolute values of these linear correlations are used due to the fact that the interest is to obtain a sequence that maximizes the intensity of the linear correlations of the analyzed genes. Once the linear correlation values for all the genes are obtained, they are used to construct a matrix such as the one in Table 3.1. The 11 genes listed in Table 3.1 are potential lung biomarkers identified through our methodology described later in this thesis.

Table 3.1 Matrix of Absolute Values of Pairwise Gene Correlations

	AGER	SFTPC	TMEM100	FABP4	SPP1	WIF1	COL11A1	CYP4B1	FCN3	ADH1B	CLDN18
AGER	0	0.643	0.436	0.515	0.531	0.496	0.363	0.348	0.545	0.364	0.567
SFTPC	0.643	0	0.373	0.410	0.410	0.570	0.432	0.550	0.441	0.544	0.550
TMEM100	0.436	0.373	0	0.305	0.270	0.226	0.159	0.239	0.600	0.433	0.319
FABP4	0.515	0.410	0.305	0	0.317	0.257	0.079	0.319	0.201	0.384	0.407
SPP1	0.531	0.410	0.270	0.317	0	0.373	0.375	0.155	0.277	0.273	0.462
WIF1	0.496	0.570	0.226	0.257	0.373	0	0.241	0.390	0.371	0.470	0.446
COL11A1	0.363	0.432	0.159	0.079	0.375	0.241	0	0.236	0.326	0.253	0.265
CYP4B1	0.348	0.550	0.239	0.319	0.155	0.390	0.236	0	0.264	0.497	0.239
FCN3	0.545	0.441	0.600	0.201	0.277	0.371	0.326	0.264	0	0.312	0.370
ADH1B	0.364	0.544	0.433	0.384	0.273	0.470	0.253	0.497	0.312	0	0.404
CLDN18	0.567	0.550	0.319	0.407	0.462	0.446	0.265	0.239	0.370	0.404	0

This matrix serves as input for a program in Matlab developed for our research group by our collaborator J. Rodríguez, who is included in the Acknowledgements section of this thesis. The software is used to obtain an optimal sequence maximizing linear correlations among the eleven genes considered as potential biomarkers.

3.1.2 Minimum Spanning Tree (MST)

Due to the fact that the Traveling Salesperson problem has to make the somewhat restrictive assumption that a signaling pathway behaves as a tour, the decision was made to explore an alternative method. This method was the minimum spanning tree (MST).

The same methodology behind the MST mentioned in Section 2.6.2 is applied as an alternative to the TSP as a means to develop a signaling pathway from a list of genes identified as potential biomarker (see Table 4.1). Starting off from the same matrix developed from the linear correlations of the genes of interest (see Table 3.1) the logic of the MST is applied in order to find a minimal tree with the largest correlation among the nodes of interest.

Chapter 4. Lung Cancer Signaling Pathways

4.1 Case Study: Lung Cancer

According to the National Cancer Institute, lung cancer is the second most common type of cancer and the primary cause of cancer-related death in both men and women in the United States. Also based on analysis conducted by this institution, an estimated $11.9 billion were spent on lung cancer care in 2014 [37].

Reliable information on cancer can be found in specialized repositories such as PubMed. PubMed is a free web literature search service developed and maintained by the National Center for Biotechnology Information (NCBI) [38]. PubMed has served as a primary tool for electronically searching and retrieving biomedical literature in this thesis. For this and previous initial studies on lung cancer, Dr. Clara Isaza from the Department of Pharmacology and Toxicology of the Ponce School of Medicine recommended using lung cancer database GDS3257 for our investigation.

The GDS3257 database was first reported by Landi et al. [36]. The database contains 107 samples, where 49 are control tissue samples and 58 cancerous tissue samples. The subjects involved in this study ranged from the ages of 44-79 years old and had hystollogically confirmed primary adenocarcinoma of the lungs, stages I-IV. Additionally they provided detailed smoking and medical history information, which allows to divide the total of 107 samples into never smokers, former smokers, and current smokers (see Figures 4.1). Each of these samples show the measured relative expression of a total of 22,283 genes.

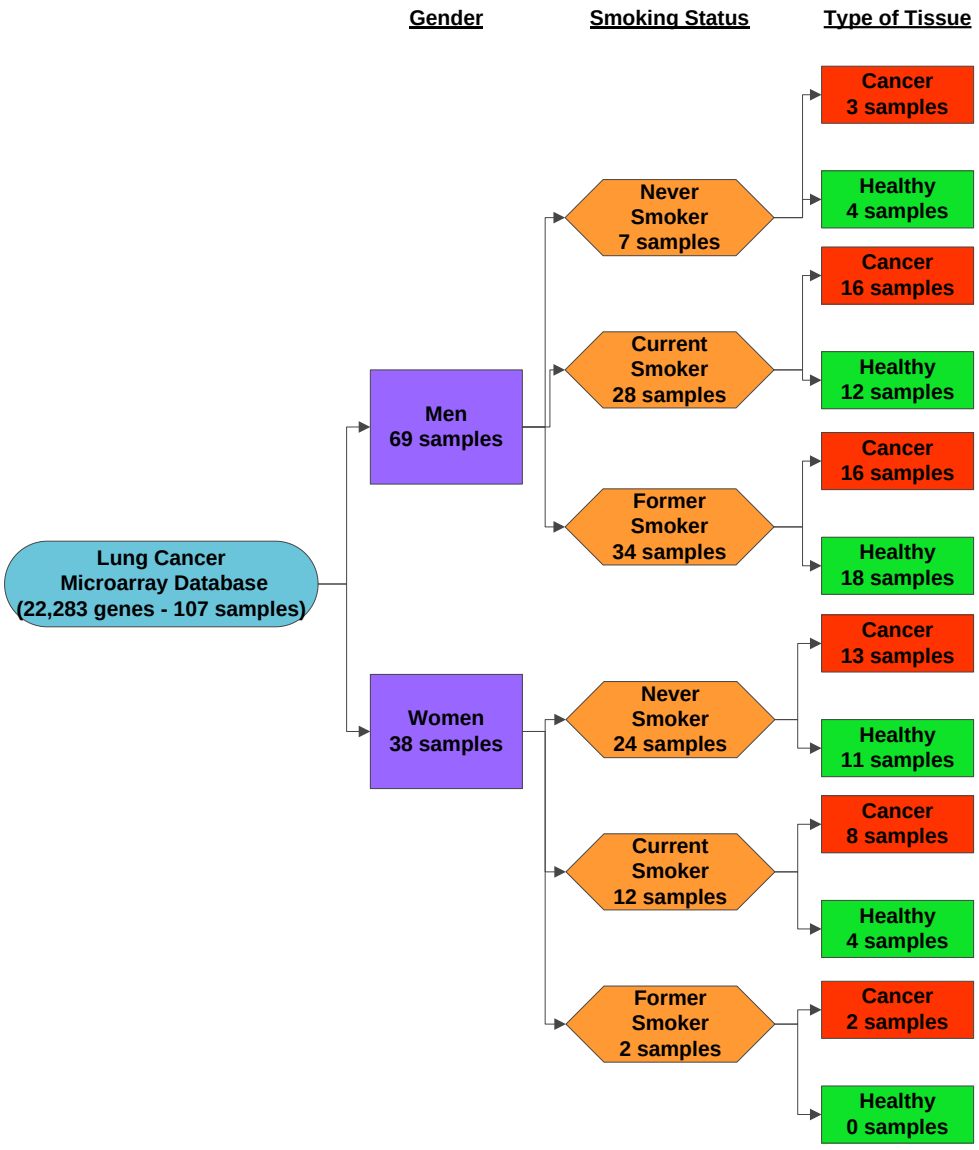

Figure 4.1 Representation Organization of Database (GDS3257)

Both of the network optimization methods discussed in the previous sections of this thesis were applied to identify a potential signaling pathway from among a list of eleven potential biomarkers for lung cancer. These eleven genes were identified by applying a methodology developed previously by our research group based on Multiple Criteria

Optimization [1] from a microarray database of 22,283 genes [36]. This list of potential biomarkers is presented in Table 4.1.

Table 4.1 List of Potential Lung Cancer Biomarkers

Potential Biomarkers	
AGER	COL11A1
SFTPC	CYP4B1
TMEM100	FCN3
FABP4	ADH1B
SPP1	CLDN18
WIF1	

4.1.1 Signaling Pathway Utilizing the TSP

By applying the concept of the TSP, a sequence that maximizes the linear correlations from a large number of possible solutions was obtained with support and contributions from undergraduate student Enery Lorenzo [32]. Figure 4.2 shows the results.

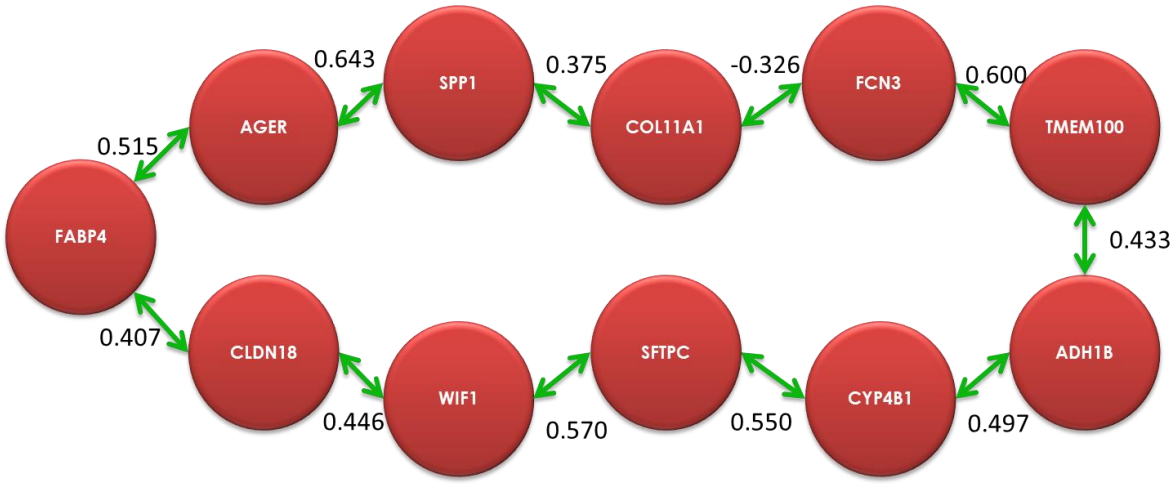

Figure 4.2 TSP Optimal Solution from 11 Potential Biomarkers Obtained from GDS3257 Microarray Database

This sequence was obtained from a number of (11-1)! (around 3.6 million) possible solutions. It represents a solution that maximizes the linear correlations between these 11 genes. This result represents the potential of this work: being able to obtain an optimal

cycling path that maximizes the correlation values from thousands of possible solutions. It must be clarified that due to the fact this is an observational analysis of a microarray database, causality or directionality cannot be established at this point. Additionally the question of how valid is it to assume a signaling pathway behaves as a tour or cycle must be addressed in the future.

4.1.2 Signaling Pathway Utilizing the MST

As an alternative method to the TSP, the MST was applied to obtain a tree that maximizes the linear correlations of the eleven genes that were identified previously by our research group as potential biomarkers. Parting from the list of genes of interest, the MST method was applied and an optimal structure that maximizes these correlation values was obtained (see Figure 4.3).

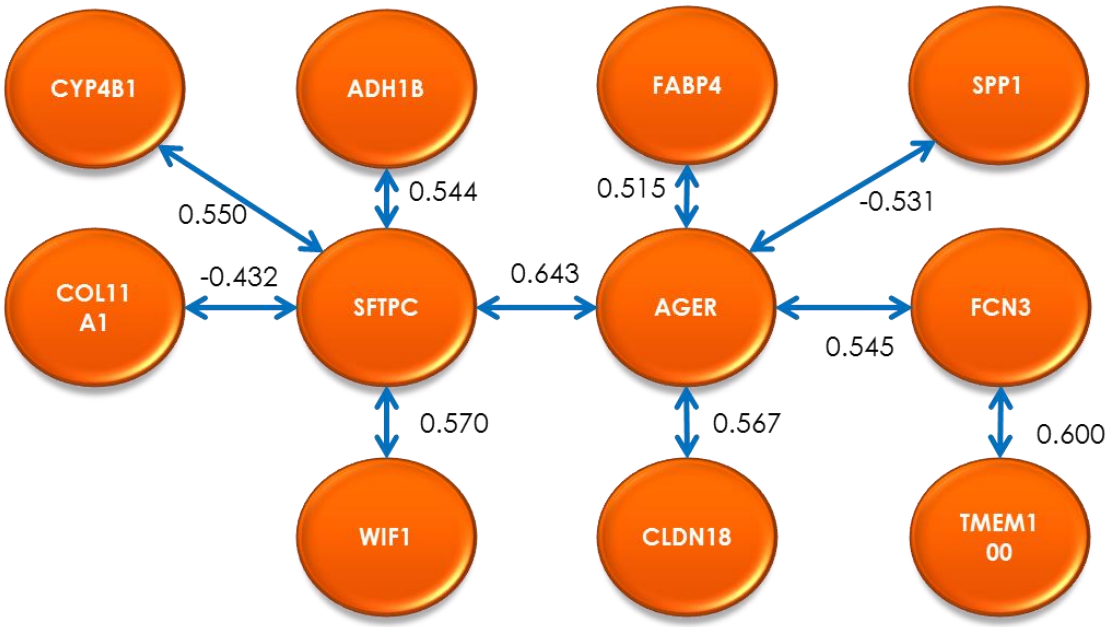

Figure 4.3 MST Optimal Solution from 11 Potential Biomarkers Obtained from GDS3257 Microarray Database

Unlike the solution obtained applying the TSP concept, this solution does not behave as a cycle but as a tree. This could possible hold greater similarity with the natural behavior of a signaling pathway.

Utilizing the combination of methodologies of Multiple Criteria Optimization to determine a list of potential biomarkers from databases and subsequently employing the TSP and MST to identify possibly signaling pathways from these genes of interest, could serve as a viable alternative to existing methods to provide an original analysis pipeline completely driven by optimization procedures. In Section 5.1, a search for biological evidence used to validate and verify the biological significance of the results obtained in this case study is described.

4.1.3 Meta-analysis

This section describes the first steps taken to carry out a meta-analysis in order to discover signaling pathways of interest from the different groups previously described in Section 4.1 between control and cancer tissues groups within GDS3257 database. These groups include: 16 samples Cancer Never-smoker (CNS) vs 24 samples Cancer Current Smoker (CCS), 16 samples Healthy Current Smoker (HCS) vs 24 samples CCS, 16 samples HCS vs 16 samples CNS, 15 samples Healthy Non-smoker (HNS) vs 24 samples CCS, 15 samples HNS vs 16 samples CNS, and 15 samples HNS vs 16 samples HCS.

Initially, MCO was applied to identify genes with high variations in their relative expression levels when utilizing two performance measures, the absolute values of the differences between means and median. For each group comparison the genes were

divided in subgroups of about 7,000 genes in order to facilitate its processing in our MCO program developed by Katia I. Camacho [13] (see Appendix 2). MCO was carried out for each group for ten iterations where genes with the highest variation in expression levels (in Pareto sense) [13] were separated for each iteration.

Once MCO was concluded and a list of genes of interest for each group comparison was obtained, the following step was to proceed and apply the TSP and MST optimization methods to identify potential signaling pathways. For each group comparison the TSP and MST were applied and a signaling pathway was obtained. Figure 4.4 represents the group comparisons conducted.

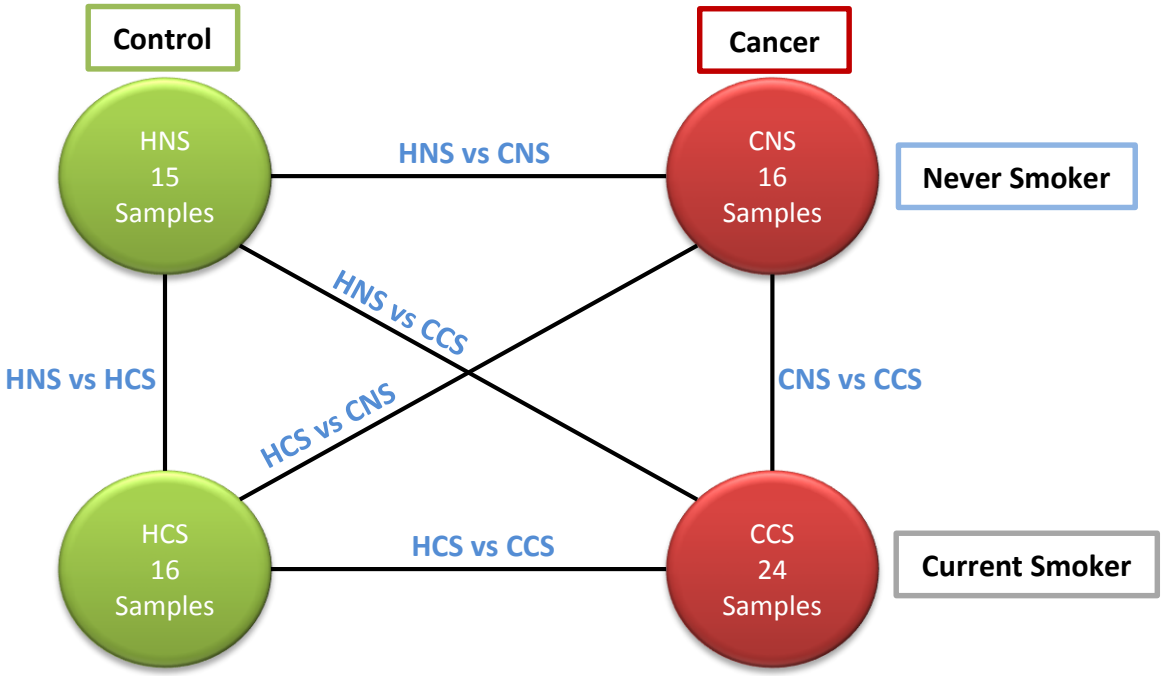

Figure 4.4 Representation of Six Analyses between Four Conditions in Microarray Database GDS3257

Both methods were applied to identify a signaling pathway when analyzing CNS vs CCS. Once MCO was conducted to determine the genes with the largest changes in relative expression, a total of 20 genes were identified as genes of interest for this

particular comparison. The resulting network solution when applying the TSP and MST methods are represented by Figure 4.5 and Figure 4.6 respectively.

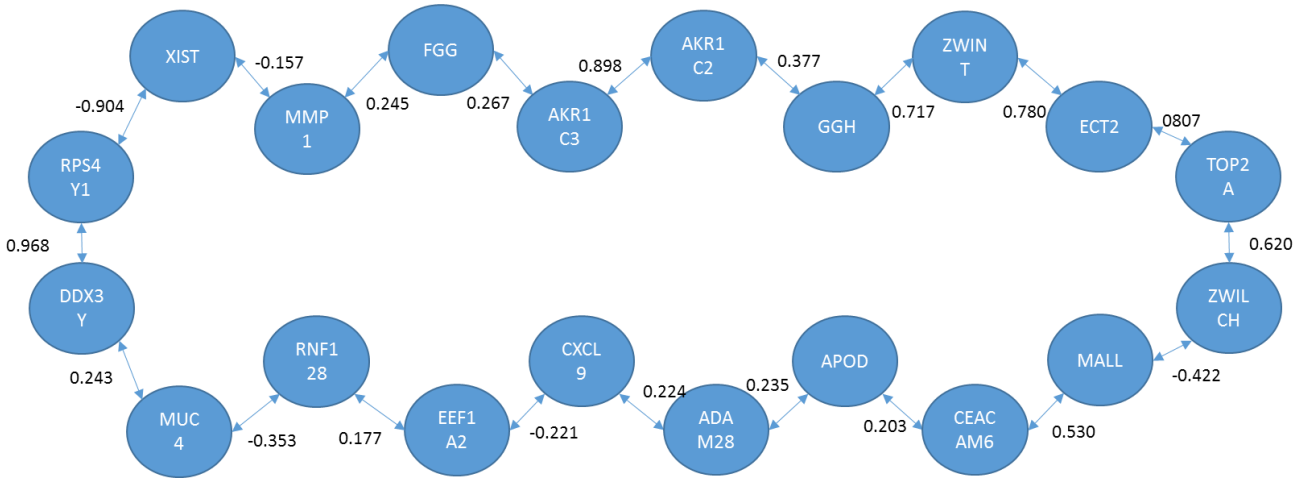

Figure 4.5 Optimal Correlation Cycle Utilizing TSP from 20 Potential Biomarkers from GDS3257 (Cancer Non-smoker vs Cancer Current Smoker)

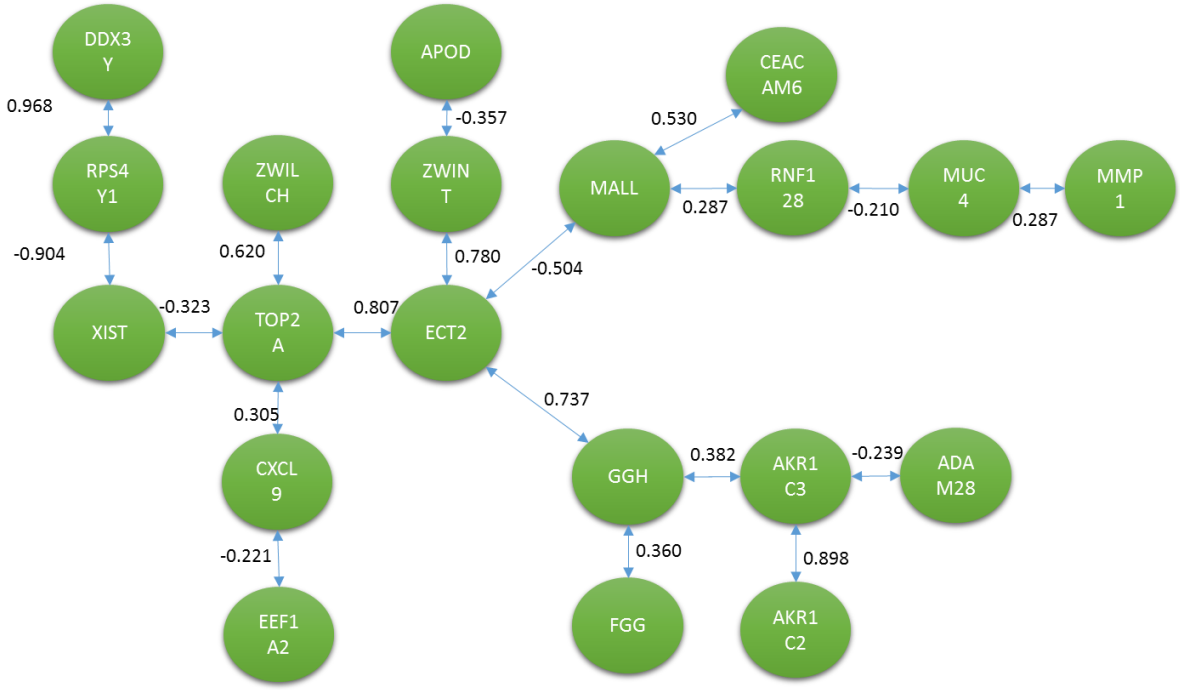

Figure 4.6 Optimal Correlation Tree Utilizing MST from 20 Potential Biomarkers from GDS3257 (Cancer Non-smoker vs Cancer Current Smoker)

Several similarities exist between both networks' solutions (see Tables 4.2, 4.3 and 4.4) which are potentially interesting from a biological perspective. These two networks are the best solution from a total of (20-1)! and 20^{20-2} possible configurations for the TSP and MST respectively.

The next comparison that was conducted is that of HCS vs CCS to obtain a signaling pathway utilizing MCO to identify the genes of interest and afterwards the TSP and MST formulations. Figure 4.7 and Figure 4.8 are the graphical representations of the solutions obtained with both methods.

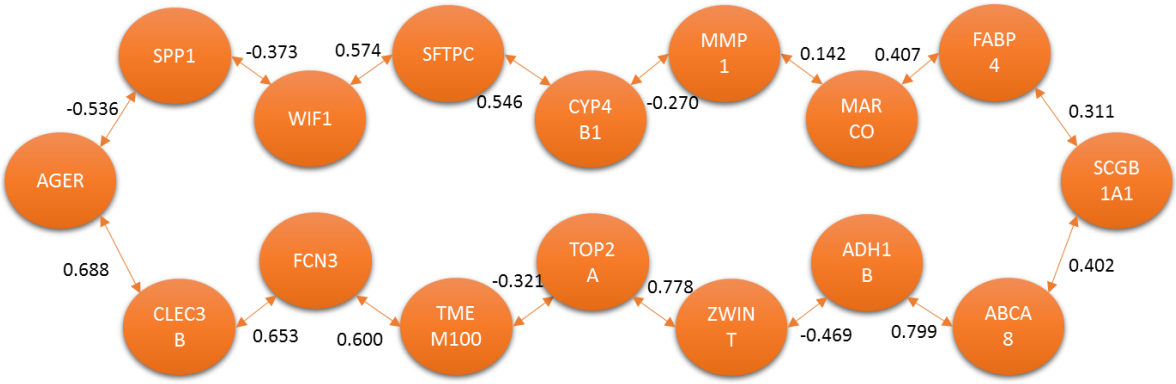

Figure 4.7 Optimal Correlation Cycle Utilizing TSP from 16 Potential Biomarkers from GDS3257 (Healthy Current Smoker vs Cancer Current Smoker)

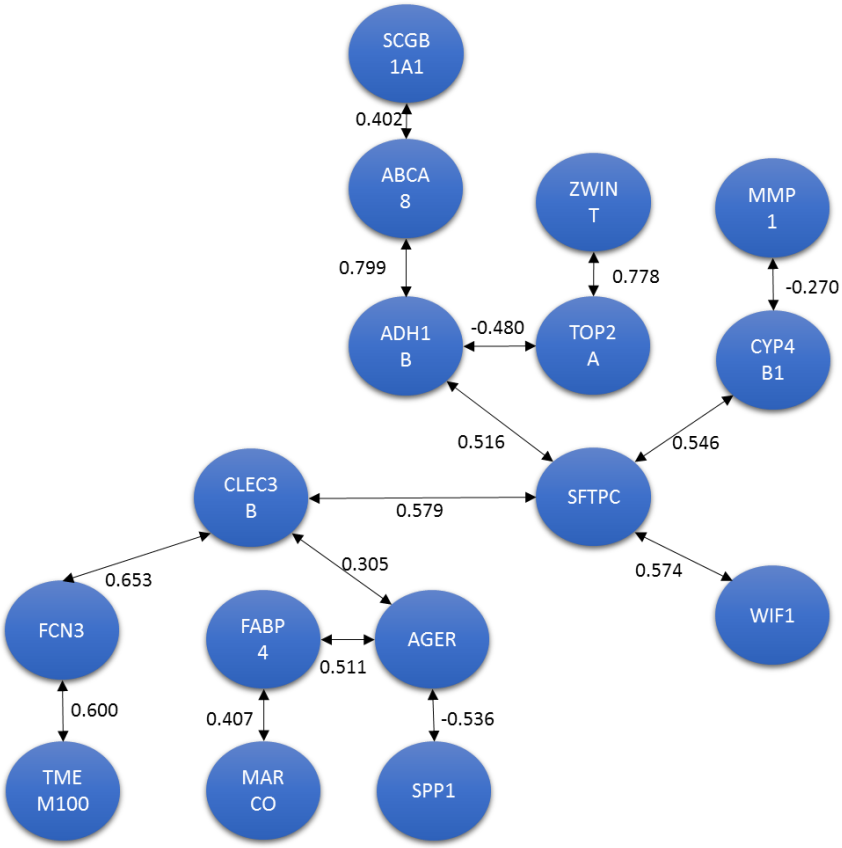

Figure 4.8 Optimal Correlation Tree Utilizing MST from 16 Potential Biomarkers from GDS3257 (Healthy Current Smoker vs Cancer Current Smoker)

Same as previous analysis, some similarities exist between the solutions of the TSP and MST (see Tables 4.2, 4.3 and 4.4). These solutions were constructed from the correlations that were calculated from 16 genes that through MCO were established to be importantly differentially expressed.

The third analysis was the comparison of HCS vs CNS, where a total of 23 genes were considered when carrying out the TSP and MST methods. The resulting signaling pathways are presented in Figure 4.9 and 4.10.

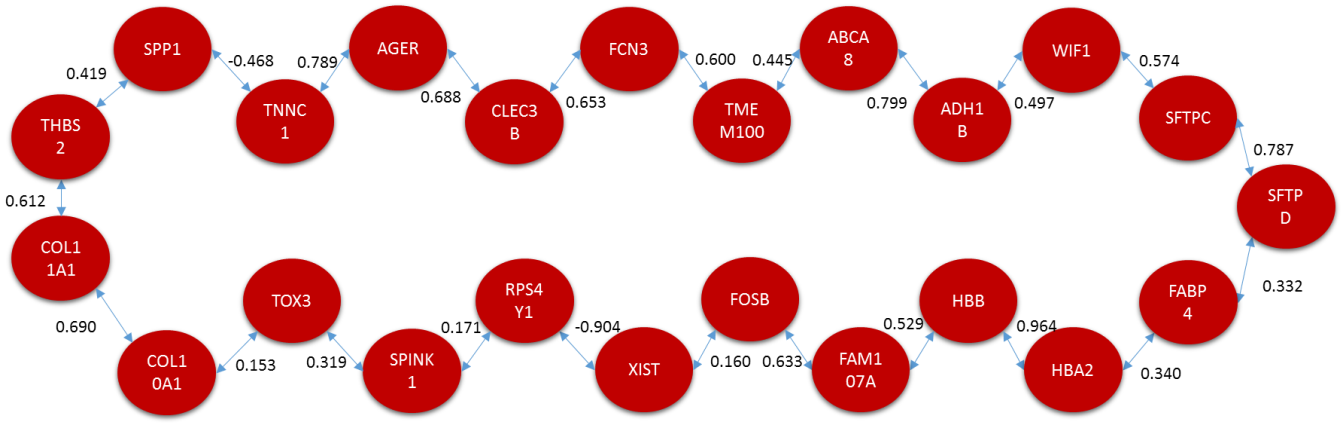

Figure 4.9 Optimal Correlation Cycle Utilizing TSP from 23 Potential Biomarkers from GDS3257 (Healthy Current Smoker vs Cancer Non-smoker)

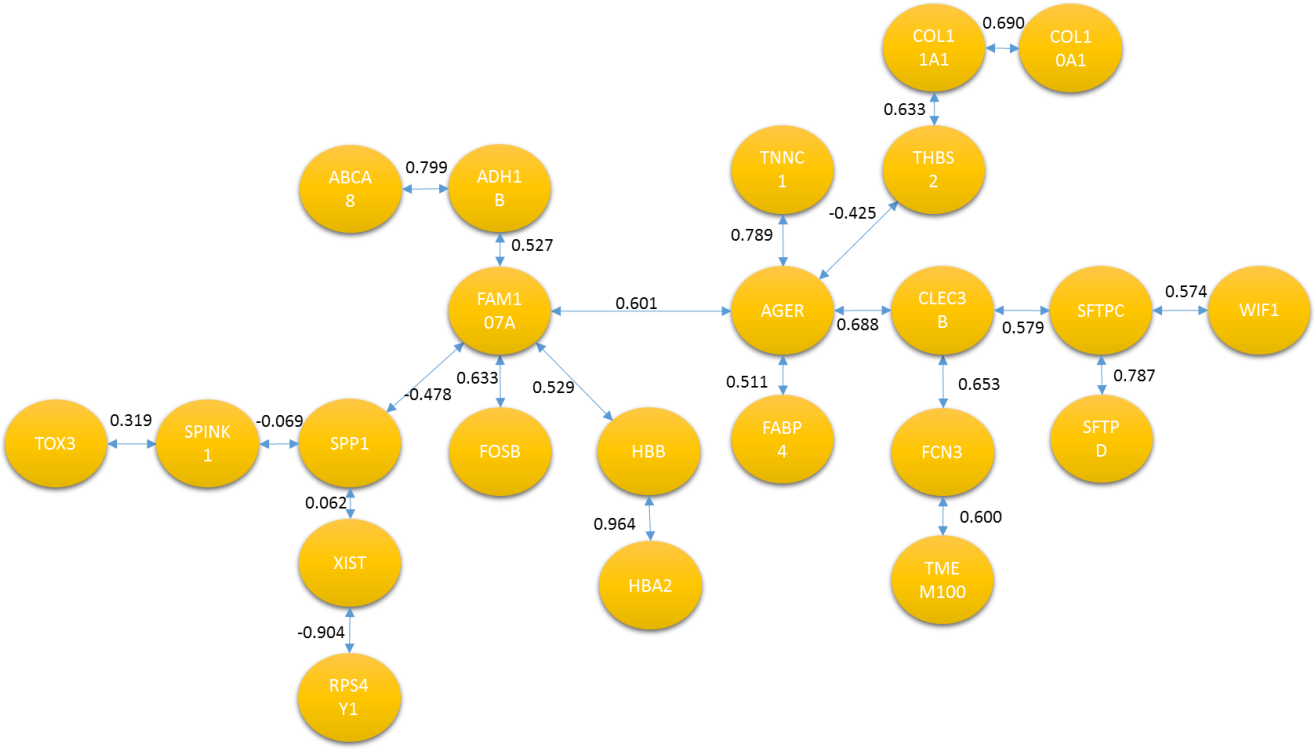

Figure 4.10 Optimal Correlation Tree Utilizing MST from 23 Potential Biomarkers from GDS3257 (Healthy Current Smoker vs Cancer Non-smoker)

The following comparison was done between HNS vs CCS. When MCO was conducted, a total of 17 genes were identified as genes of interest for this comparison in

particular. Once these were identified, TSP and MST were applied to obtain Figure 4.11 and Figure 4.12.

Figure 4.11 Optimal Correlation Cycle Utilizing TSP from 17 Potential Biomarkers from GDS3257 (Healthy Non-smoker vs Cancer Current Smoker)

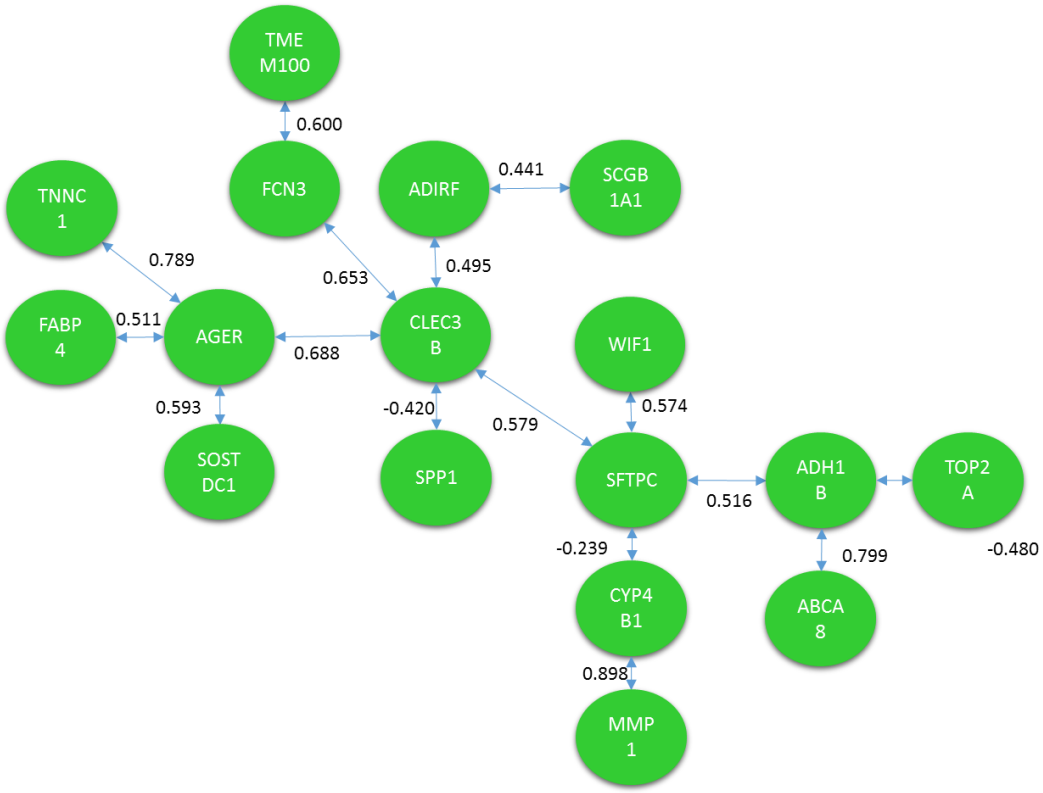

Figure 4.12 Optimal Correlation Tree Utilizing MST from 17 Potential Biomarkers from GDS3257 (Health Non-smoker vs Cancer Current Smoker)

The next analysis involves the search for a signaling pathway when comparing HNS vs CNS. As stated before the TSP and MST are utilized once a list of genes of interest is compiled through the use MCO, which in this case correspond to a total of 18 genes.

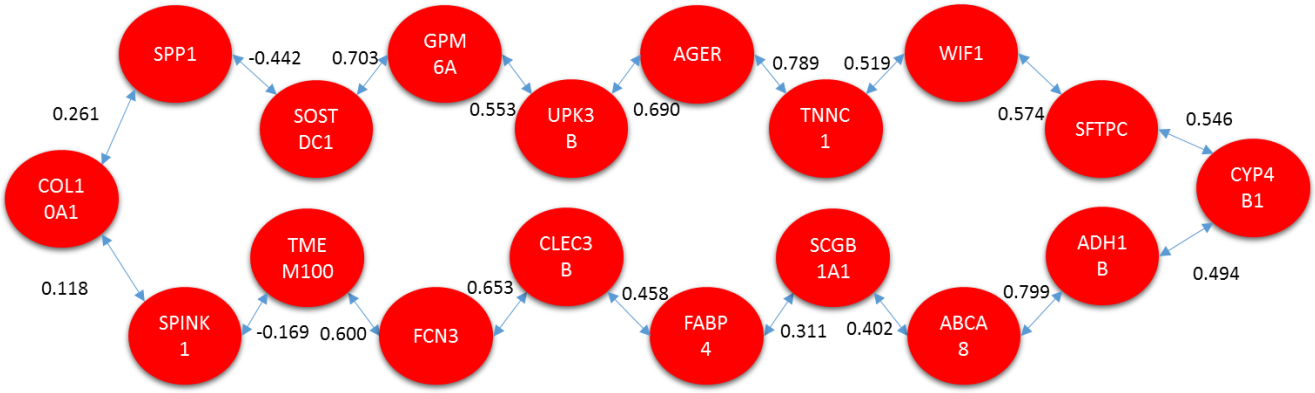

Figure 4.13 Optimal Correlation Cycle Utilizing TSP from 18 Potential Biomarkers from GDS3257 (Health Non-smoker vs Cancer Non-smoker)

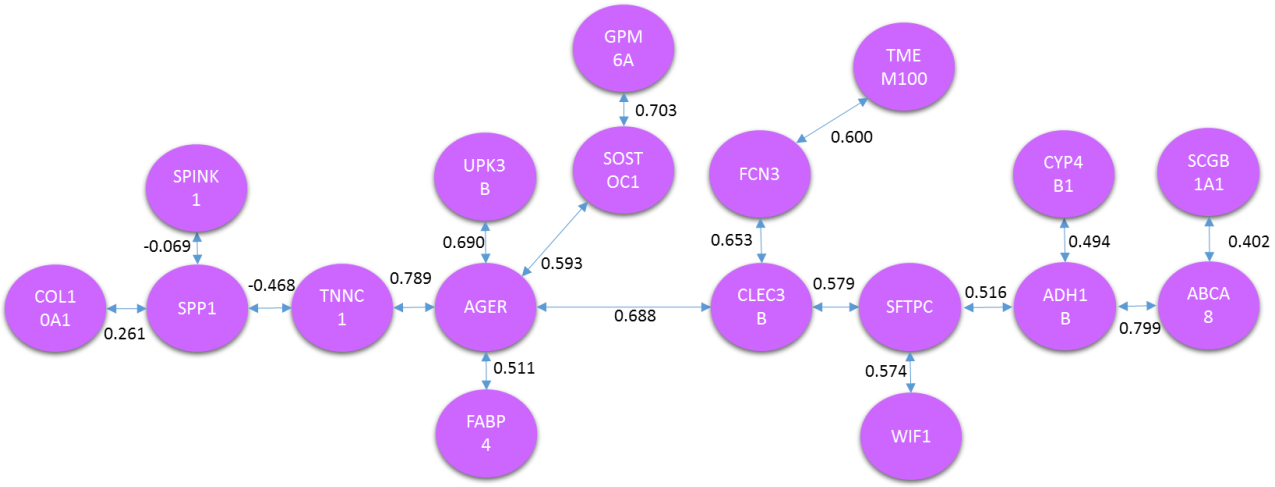

Figure 4.14 Optimal Correlation Tree Utilizing MST (Healthy Non-smoker vs Cancer Non-smoker)

The last analysis carried out for this meta-analysis was that of the comparison between HNS vs HCS. When applying the TSP and MST for this comparison, the networks were to be constructed from 30 genes of interest which heightened the number possible solutions for both methods.

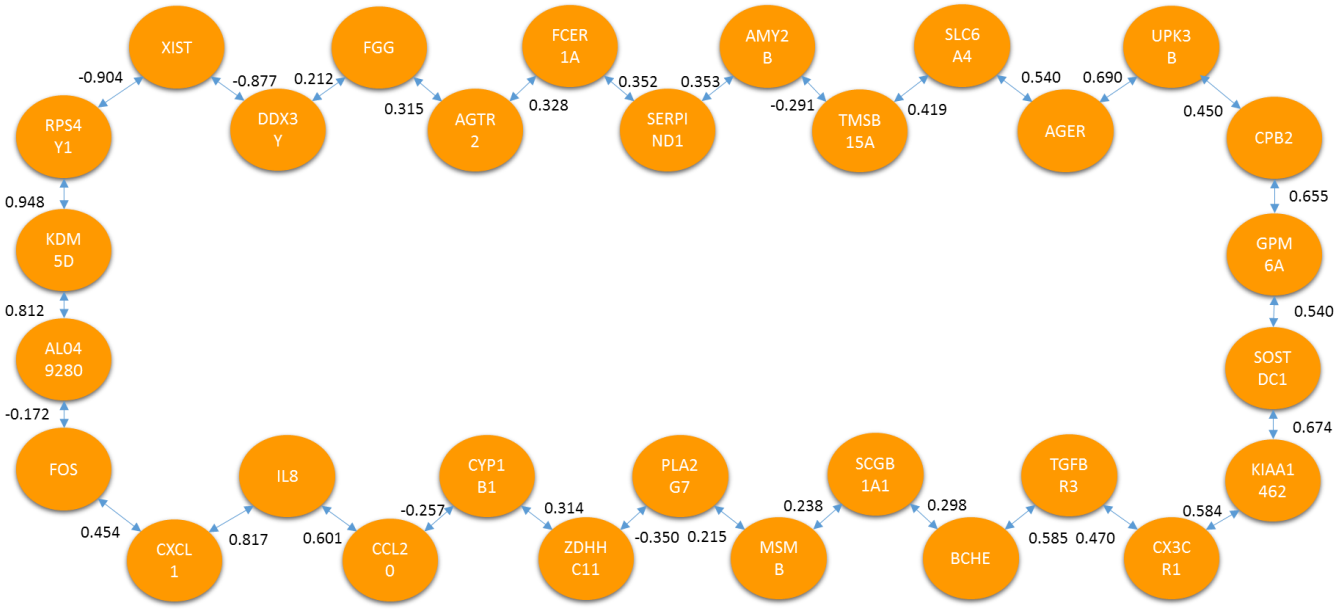

Figure 4.15 Optimal Correlation Cycle Utilizing TSP from 30 Potential Biomarkers from GDS3257 (Healthy Non-smoker vs Healthy Current Smoker)

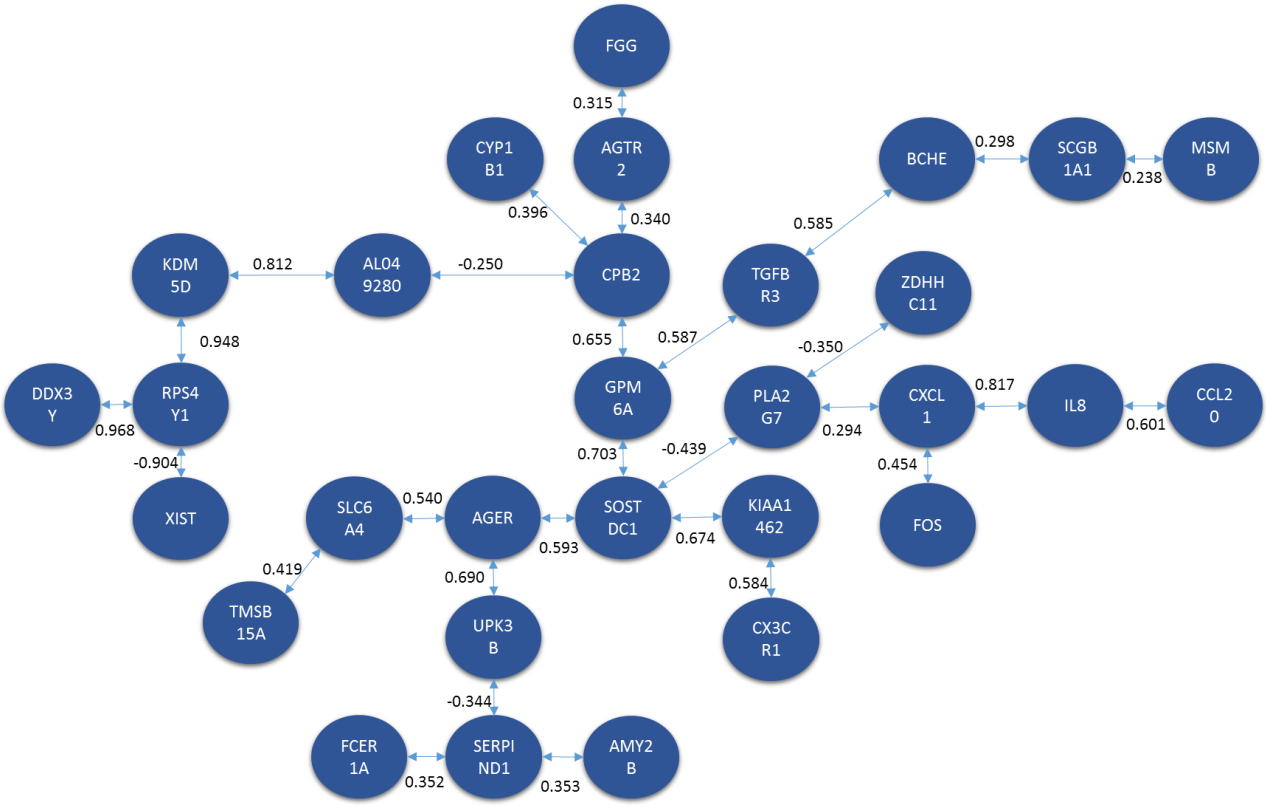

Figure 4.16 Optimal Correlation Tree Utilizing MST from 30 Potential Biomarkers from GDS3257 (Healthy Non-smoker vs Healthy Current Smoker)

As mentioned before, there were several similarities in both the genes established as genes of interest with MCO and gene relationships that form part of the potential signaling constructed by the network optimization methods when comparing HNS, HCS, CNS, and CCS (see Tables 4.2, 4.3 and 4.4). These similarities could prove to be biologically significant once further biological validation can be conducted. Furthermore, additional biological validation will be conducted for each signaling pathway individually. Table 4.2 presents relationships between genes that were observed in several of the resulting pathways when utilizing the TSP.

Table 4.2 Comparison of Gene Relations between Conditions of Microarray Database (TSP)

Gene Relationships in Common between Conditions (TSP)					
CNSvsCCS Optimal Cycle Solution	*HCSvsCCS Optimal Cycle Solution*	*HCSvsCNS Optimal Cycle Solution*	*HNSvsCCS Optimal Cycle Solution*	*HNSvsCNS Optimal Cycle Solution*	*HNSvsHCS Optimal Cycle Solution*
RPS4Y1	AGER	THBS2	SOSTDC1	COL10A1	KDM5D
XIST	SPP1	SPP1	SPP1	SPP1	RPS4Y1
MMP1	WIF1	TNNC1	FABP4	SOSTDC1	XIST
FGG	SFTPC	AGER	ADIRF	GPM6A	DDX3Y
AKR1C3	CYP4B1	CLEC3B	SCGB1A1	UPK3B	FGG
AKR1C2	MMP1	FCN3	ABCA8	AGER	AGTR2
GGH	MARCO	TMEM100	ADH1B	TNNC1	FCER1A
ZWINT	FABP4	ABCA8	TOP2A	WIF1	SERPIND1
ECT2	SCGB1A1	ADH1B	MMP1	SFTPC	AMY2B
TOP2A	ABCA8	WIF1	CYP4B1	CYP4B1	TMSB15A
ZWILCH	ADH1B	SFTPC	SFTPC	ADH1B	SLC6A4
MALL	ZWINT	SFTPD	WIF1	ABCA8	AGER
CEACAM6	TOP2A	FABP4	TNNC1	SCGB1A1	UPK3B
APOD	TMEM100	HBA2	AGER	FABP4	CPB2
ADAM28	FCN3	HBB	CLEC3B	CLEC3B	GPM6A
CXCL9	CLEC3B	FAM107A	FCN3	FCN3	SOSTDC1
EEF1A2	-	FOSB	TMEM100	TMEM100	KIAA1462
RNF128	-	XIST	-	SPINK1	CX3CR1
MUC4	-	RPS4Y1	-	-	TGFBR3
DDX3Y	-	SPINK1	-	-	BCHE
-	-	TOX3	-	-	SCGB1A1
-	-	COL10A1	-	-	MSMB
-	-	COL11A1	-	-	PLA2G7
-	-	-	-	-	ZDHHC11
-	-	-	-	-	CYP1B1
-	-	-	-	-	CCL20
-	-	-	-	-	IL8
-	-	-	-	-	CXCL1
-	-	-	-	-	FOS
-	-	-	-	-	AL049280

Legend
Colored cells represent gene relationships present in multiple optimal cycle solutions for the six different analyses.

Several of the resulting signaling pathways have similar relationships between certain genes in common. Table 4.2 includes the cycles obtained with the TSP details all of these relationships that coincide between the different analysis conducted using the TSP. Each color in Table 4.2 represents a segment solution consisting of two or more adjacent genes that are present in several of the analyzed conditions previously described. Several relationships are present in almost all the signaling pathways such as ABCA8-ADH1B and TMEM100-FCN3-CLEC3B. Biological verification of these relationships could yield important information. This comparison of signaling pathways was also done for the resulting pathways when applying the MST, Table 4.3 summarizes the comparison.

Table 4.3 Comparison of Gene Relationships between Conditions of Microarray Database (MST)

Comparison	**Gene Relationships (MST)**		
	XIST-RPS4Y1	*WIF1-SFTPC-CLEC3B-FCN3-TMEM100*	*AGER-FABP4*
CNSvsCCS	✓		
HCSvsCCS		✓	✓
HCSvsCNS	✓	✓	✓
HNSvsCCs		✓	✓
HNSvsCNS		✓	✓
HNSvsHCS	✓		

As with the results from the TSP, the signaling pathways constructed using the MST had several coupled gene groups that were in common throughout the six different comparisons. Additionally there are some arrangements of genes in common between the TSP and MST, these are included in the following table.

Table 4.4 Comparison of Gene Relationships between Conditions of Microarray Database (TSP-MST)

Comparison	Gene Relationships (TSP-MST)		
	XIST-RPS4Y1	*CLEC3B-FCN3-TMEM100*	*SFTPC-WIF1*
CNSvsCCS	✓		
HCSvsCCS		✓	✓
HCSvsCNS	✓	✓	✓
HNSvsCCs		✓	✓
HNSvsCNS		✓	✓
HNSvsHCS	✓		

The previous tables include several of the segments containing genes in common for the results obtained using the TSP and MST methods. Several of these segments are present when comparing the different conditions of control subjects and subjects with cancer. In the case of the TSP, these include ABCA8-ADH1B, WIF1-SFTPC, TNNC1-AGER and CLEC3B-FCN3-TMEM100 (see Table 4.2). According to our results, these gene relationships have an important role in lung cancer due to their frequency throughout the different conditions. With sufficient biological evidence (see Section 5.2) and the obtained results, several previously unidentified gene relationships can be proposed as relevant to lung cancer and the smoking habits of the patient

Additionally the gene relationships included in Table 4.4 were compared to the gene relationships obtained in our initial lung cancer case study in Sections 4.1.1 and 4.1.2. This was done to identify if any gene relationships were common throughout the results obtained in both our case study and meta-analysis. Two relationships were identified, these are SFTPC-WIF1 and FCN3-TMEM100. These gene relationships recurrently formed part of solutions with both the TSP and MST for both our case study

and meta-analysis. Due to this fact we establish that these relationships are of high interest in terms of lung cancer and of even further investigation.

Chapter 5. Biological Evidence and Validation of Proposed Methodology

5.1 Biological Evidence of Lung Cancer Case Study

This section describes the search of existing biological evidence in literature to analyze and validate the results described in previous sections. With the guidance of Dr. Clara Isaza and contributions by undergraduate students Arlette Marrero and Cristina Ortiz of the Department of Biology of the University of Puerto Rico-Mayagüez, relevant information was collected to validate our method and support its potential.

Based on the signaling pathways proposed by the TSP and MST methods, public databases including KEGG and GeneCards were searched for relevant biological information. Evidence found on the different gene relations contained in both signaling pathways of the TSP (see Figure 4.2) and MST (see Figure 4.3) is presented in the following tables.

Table 5.1 Evidence on Relationships Consisting of Two Genes Included in Proposed Signaling Pathways

Gene Relations	TSP	MST	Biological Evidence		
			Lung Cancer	*Cancer*	*No Evidence Found*
FCN3-COL11A1	✓			✓	
COL11A1-SPP1	✓		✓		
SPP1-AGER	✓	✓	✓		
AGER-FABP4	✓	✓	✓		
FABP4-CLDN18	✓				✓
CLDN18-WIF1	✓			✓	
WIF1-SFTPC	✓	✓		✓	
SFTPC-CYP4B1	✓	✓		✓	
CYP4B1-ADH1B	✓			✓	
ADH1B-TMEM100	✓				✓
TMEM100-FCN3	✓	✓			✓
FCN3-AGER		✓		✓	
CLDN18-AGER		✓			✓
SFTPC-AGER		✓		✓	
SFTPC-ADH1B		✓	✓		
SFTPC-COL11A1		✓			✓

Table 5.1 presents gene relationship consisting of two potential biomarkers that formed part of the solutions obtained with either the TSP or MST. Additionally it indicates whether any previously published studies or information for each group of gene relationship were found linking them to specifically lung cancer, other types of cancer, and lastly if no information was found for the previous categories. The first category serves to validated our methodology, due to the fact that it is capable of obtaining

solutions connecting genes whose relationship have a role in lung cancer previously known. The second category indicates that studies were found involving the particular gene relationship to a type of cancer other than lung cancer. This implies that the relationship does play a role in cancer in general and could be proposed as an important gene relationship for lung cancer. The final category establishes that no previously published evidence was found connecting the gene relationship to either lung cancer or another form of cancer. This emphasizes the potential of our methodology of discovering previously unknown gene relationships that are important in lung cancer or potentially for cancer in general.

Table 5.2 includes information on gene relations consisting of two genes within either signaling pathway. The table presents if any information was found on the adjacent relationship of the genes, if there is a non-adjacent relationship connecting both through one or more genes or if no information was found on the particular gene relation. Table 5.3 includes the same information but for relations including three genes.

Table 5.2 Evidence (from KEGG, GeneCards, among others) on Relations Consisting of Two Genes Included in Proposed Signaling Pathways

2 Genes						
Gene Relations	**Type of Relationship**			**TSP**	**MST**	**Known Pathways Where Gene Relationships are Present**
	Adjacent	***Non-adjacent***	***Not Found***			
FCN3-COL11A1		✓		✓		
COL11A1-SPP1	✓			✓		ERK Signaling, Phospholipase-C Pathway, PI3K-Akt signaling pathway, Degradation of the extracellular matrix, Integrin Pathway
SPP1-AGER	✓			✓	✓	IL-2 Pathway
AGER-FABP4		✓		✓	✓	
FABP4-CLDN18		✓		✓		
CLDN18-WIF1		✓		✓		
WIF1-SFTPC		✓		✓	✓	
SFTPC-CYP4B1		✓		✓	✓	
CYP4B1-ADH1B	✓			✓		Biological Oxidations, Metabolism, Cytochrome P450 - arranged by substrate
ADH1B-TMEM100			✓	✓		
TMEM100-FCN3			✓	✓	✓	
FCN3-AGER		✓			✓	
CLDN18-AGER		✓			✓	
SFTPC-AGER		✓			✓	
SFTPC-ADH1B			✓		✓	
SFTPC-COL11A1		✓			✓	

Table 5.3 Evidence (from KEGG, GeneCards, among others) on Relations Consisting of Three Genes Included in Proposed Signaling Pathways

3 Genes						
Gene Relations	**Type of Relationship**			**TSP**	**MST**	**Known Pathways Where Gene Relationships are Present**
	Adjacent	**Non-adjacent**	**Not Found**			
AGER- SPP1-COL11A1		✓		✓		IL-2 pathway, ERK Signaling, Phospholipase-C Pathway, PI3K-Akt signaling pathway, Degradation of the extracellular matrix, Integrin Pathway
SPP1-AGER-SFTPC		✓			✓	
SPP1-AGER-FABP4		✓		✓	✓	
SPP1-AGER-FCN3		✓			✓	
SPP1-AGER-CLDN1		✓			✓	IL-2 Pathway, Integrin Pathway

Several gene relations were identified to be related either directly adjacent or non-adjacent. SPP1 and AGER (also known as RAGE) is one example of a direct relation that has been previously identified and forms part of the IL-2 pathway. The COL11A1 and SPP1 relation is also documented within several known pathways as described in Table 5.1. Various other gene relations had evidence of being related indirectly to one another through one or more additional genes. Additionally the relation including AGER, SPP1 and COL11A1also is documented within various known pathways (see Table 5.3), as are others.

As seen in Table 5.2, for several gene relationships information was found connecting genes either directly adjacent to one another or non-adjacently, in other words with one or more additional genes in between the relationship. In the case of the TSP,

information was found linking 9 out of the 11 gene relationships included in its results. Similarly, information was found linking 9 out of 11 of the gene relationships included in the MST results. Then in both cases, existing information accounts for 81% of the gene relationships included in our results. The relationships unaccounted for in our initial search, are proposed as important relationships relative to lung cancer once further biological evidence is gathered.

Table 5.4 Gene Relationships in Common between TSP-MST

Common Gene Pairs TSP-MST
FABP4-AGER
AGER-SPP1
FCN3-TMEM100
SFTPC-CYP4B1
SFTPC-WIF1
Common Gene Triads TSP-MST
FABP4-AGER-SPP1
CYP4B1-SFTPC-WIF1

The previous table presents the common gene relationships in the results obtained in our lung cancer case study from the TSP and MST. These consist of five gene pairs and two gene triads. The fact that these gene relationships appear in both results, points to their relevance for the case of lung cancer. As it can be seen in Table 5.2 and Table 5.3, evidence has been found validating several of the gene relationships included in Table 5.4.

The information described in this section provides evidence validating our methodology. More importantly our methodology is capable of providing the initial steps to uncover previously unknown gene interactions that may have a significant role in a disease. This is the case of our studies with lung cancer.

5.2 Biological Evidence of Meta-analysis Results

As described in Section 4.1.3, several gene relations could be observed throughout some of the comparisons between the different conditions of healthy current smoker, healthy non-smoker, cancer current smoker, and cancer non-smoker (see Figure 4.4). Similarly with the previous section, a search for existing biological evidence was conducted based on the gene relations that were repeated in the results when using the TSP (see Table 4.2). With contributions from Dr. Isaza, Cristina Ortiz and Arlette Marrero several publications were identified evidencing several gene relations within our proposed signaling pathway.

In their work, Sun et al. others analyze microarray data containing information on Parkinson's disease [39]. They identified a total of 10 distinctly differentially expressed genes which included the RPS4Y1 and XIST genes that were related in our proposed TSP signaling pathway. According to their analysis, they concluded that along with two other genes, they can be regarded as high confidence distinct gene biomarkers of Parkinson's disease [39].

In the case of the relation of ABCA8 and ADH1B, Liu et al. in their studies of ovarian cancer progression discovered multiple genes dysregulated when comparing ovarian cancer tissues to normal control tissues [40]. ABCA8 and ADH1B were part of six genes that were downregulated. Through their analysis they concluded that high expression of ABCA8, along with ALDH1A2 might predict poor outcome in terms of survival. Additionally they mention ADH1B and ALDH1A2 might be associated with drug resistance [40]. Based

on their conclusions, both ABCA8 and ADH1B have a role in ovarian cancer which could possibly be the similar case in lung cancer according to our results.

These studies are the results of an initial search for biological evidence for the results described in Section 4.1.3. They serve as an initial validation of our methodology when conducting joint analysis of multiple experiments, which involved comparisons among different conditions of lung cancer. With our methodology we are capable of identifying an optimal solution, in terms of maximizing correlations between potential biomarkers of lung cancer. These solutions are not only biologically validated and thus existing, but also uncover potential gene relationships that are important in a disease.

Chapter 6. Methodology Comparison to GeneMANIA Tool

6.1 GeneMANIA

There are several methods or tools for identifying and constructing signaling pathways. Some were described previously in previous sections of this thesis, of which several had issues when attempting to utilize them. One tool without issues when utilizing was the GeneMANIA web-based tool. According to Zuberi et al., GeneMANIA is a web interface for generating hypothesis about gene function, analyzing gene lists and prioritizing for functional analysis [41]. In this section, our methodologies are compared to the GeneMANIA tool.

Data sets utilized by GeneMania are collected from publicly available databases [42]. This includes co-expression data that is collected from Gene Expression Omnibus (GEO), physical and genetic interaction data from BioGRID and predicted protein interaction data is based on orthology from Interologous Interaction Database (I2D). Also pathway and molecular interaction data from Pathway Commons, which contains data from BioGRID, Memorial Sloan-Kettering Cancer Center, Human Protein Reference Database, HumanCyc, among others [42].

According to the authors, a researcher must enter a list of predetermined genes of interest, optionally also select from a list a specific data sets wished to query. GeneMANIA then extends the list with genes that are functionally similar or have shared properties with the initial query genes. Then it displays an interactive association network, which illustrates the relations among the genes and data sets [42]. Based on a query list of genes, it assigns weights to data sets based on how useful they are for each query.

Individual data sets are represented as networks. Each network is assigned a weight primarily based on how well connected genes in the query list are to each other compared with their connectivity to non-query genes.

GeneMANIA is based on a heuristic algorithm that builds a composite functional association network by integrating multiple functional association networks and predicts gene function [43]. The constructed composite network is a weighted sum of individual data sources. Each edge in the composite network is weighted by the corresponding individual data source. Given the composite network, GeneMANIA uses label propagation to score all genes not in the query gene list. The scores are used to rank the genes. The score assigned to each gene reflects how often paths that start at a given gene node end up in one of the query gene nodes and how long and heavily weighted those paths are [42].

6.2 Comparison of TSP & MST to GeneMANIA

The list of 11 genes in Table 4.1 from our lung cancer case study was used to carry out a comparison between our TSP and MST methodologies and GeneMANIA. As mentioned before the program requires the list to be entered in its query list located at its website at [44]. Figure 6.1 displays the web page where the input list of genes must be entered and where the organism to be analyzed must be specified.

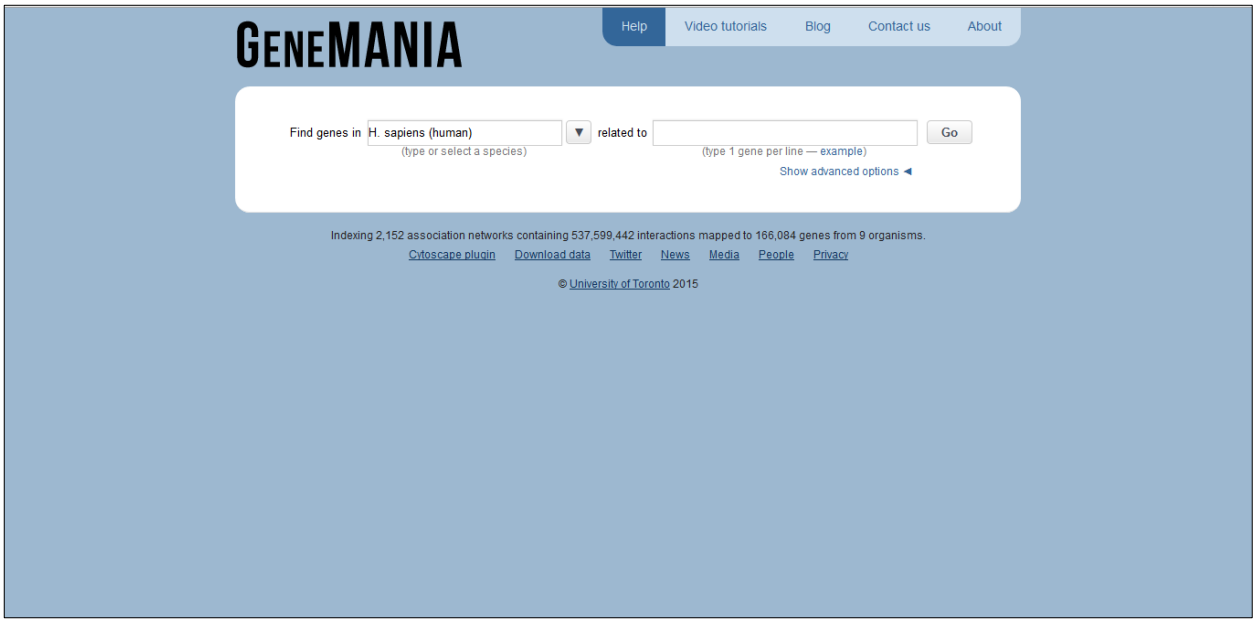

Figure 6.1 GeneMANIA Input Page

Once the input gene list is entered, the following step is to run GeneMANIA to obtain the network trying to associate each of the genes based on existing information in the databases mentioned previously. Figure 6.2 is the resulting network for all eleven genes including additional gene in a gray tone that GeneMANIA associated to our original list.

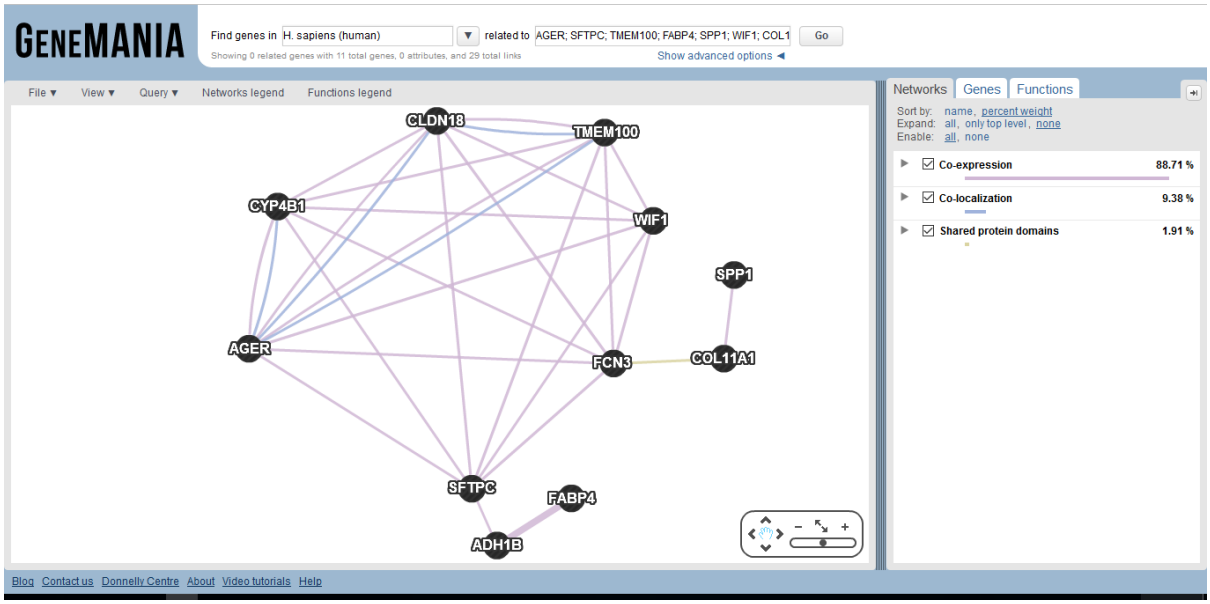

Figure 6.2 Resulting Networks from GeneMANIA

Figure 6.2 is the resulting overlap of three networks that GeneMANIA constructed to relate or connect one gene to another from our original query list. Additionally GeneMANIA included 20 additional genes to our gene query list in order to construct the networks. The maximum total of additional genes that the tool can include is determined by the user from a range of 0 to 100, with a default value of 20.

The network that generated the largest number of relationships and the largest network connecting all of the original eleven genes was that of the Co-expression factor. GeneMANIA identified several studies from its several database sources linking several genes that can be related through co-expression. Each individual study can be accessed by simply moving the cursor on any edge. A window will appear with information on the article of the study which establishes the co-expression of the genes connected by the selected edge. In total there are 24 connections that form part of the network GeneMANIA constructed. It must be noted that GeneMANIA was unsuccessful in including all our original genes in the network. SPP1 and COL11A1 were not connected to the rest of the genes, specifically to the AGER gene. As included in Table 5.3, several documented signaling pathways link these genes which evidence the importance of this relationship. Additionally GeneMANIA did not relate the following gene relationships for which biological evidence was found: SPP1-AGER, CYP4B1-ADH1B.

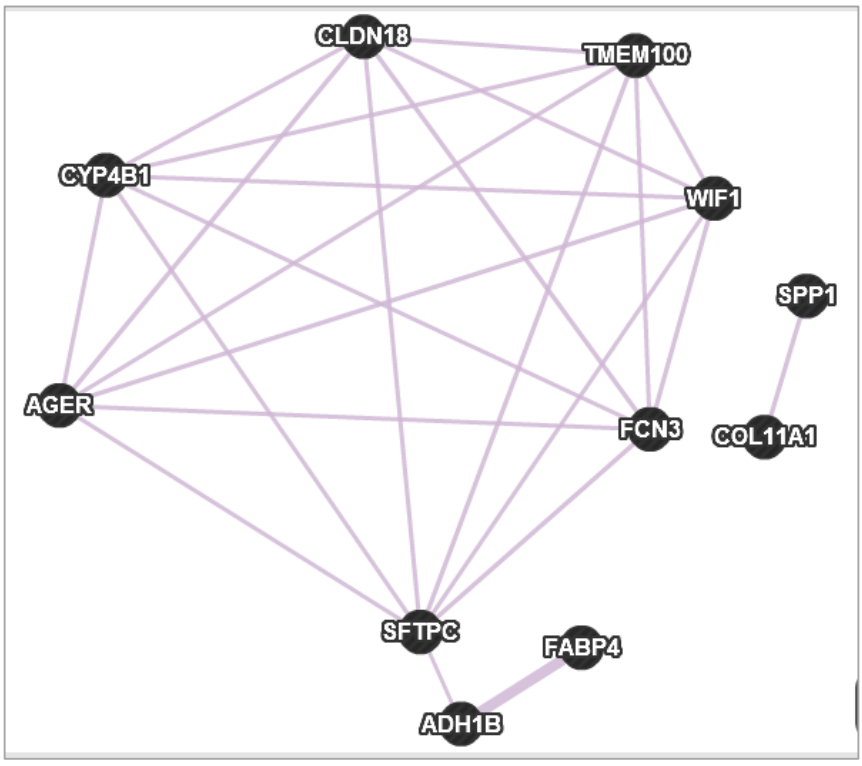

Figure 6.3 Co-expression Network Constructed by GeneMANIA

The second network constructed by GeneMANIA containing the some genes of our query list is the category of Co-localization. As can be seen in Figure 6.4, GeneMANIA did not include five o4f our eleven genes as part of the resulting network. In this category GeneMANIA only connected 4 genes of our query list with a total of 4 connections. It was not successful in directly relating all the genes to each other.

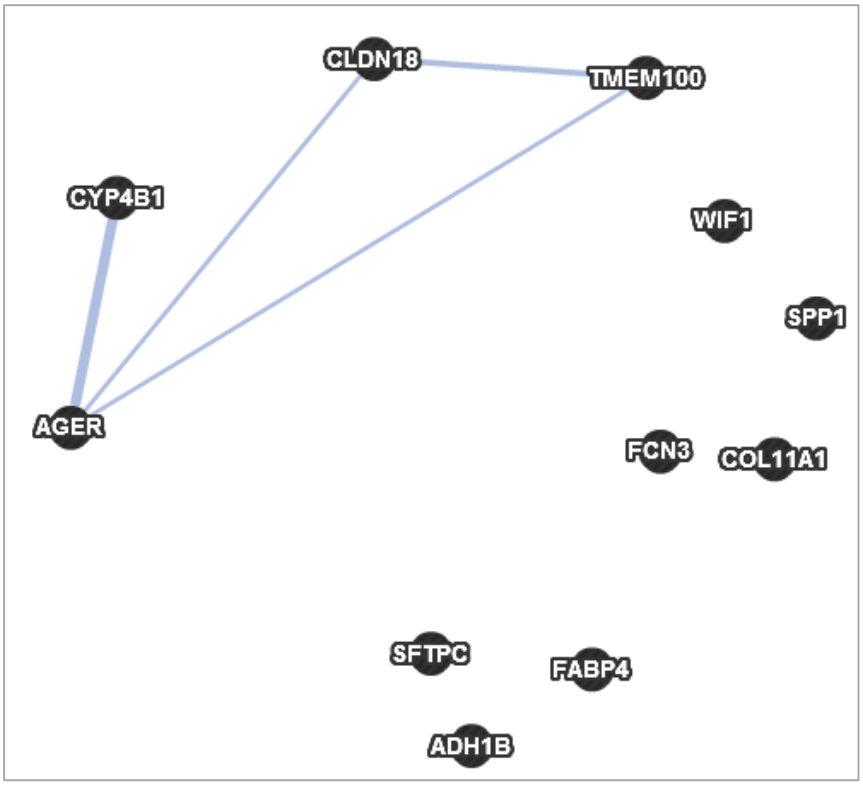

Figure 6.4 Co-localization Network Constructed by GeneMANIA

The third and last network generated by GeneMANIA can be seen in Figure 6.5. This network is categorized as genes that share protein domains. This network only links two of our eleven genes from our lung cancer study with only one connection. As in the other categories it was unsuccessful in linking all of our original query list of genes.

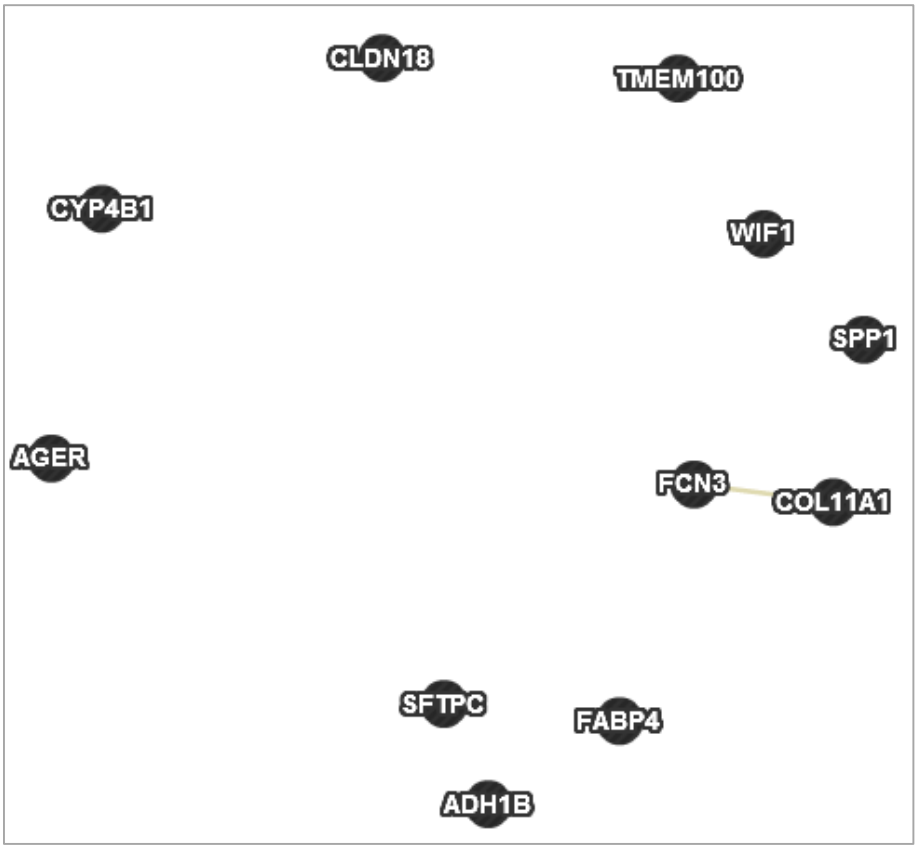

Figure 6.5 Shared Protein Domains Network Constructed by GeneMANIA

GeneMANIA is a versatile web tool capable of searching existing studies from multiple databases in an attempt of establishing relationships between a query lists of genes. In our case study of lung cancer, it generated three networks based on three separate categories. Only in the first category of co-expression, GeneMANIA was successful in establishing a network linking all eleven of our genes of interest. None of these three networks generated by GeneMANIA was in the category of pathways.

GeneMANIA relies on published analysis to generate networks, our methodology does not have this dependence for generating signaling pathways. Therefore our method possesses an interesting potential for determining previously undiscovered pathways or gene relations important in a disease, in our case lung cancer. Also due to its

deterministic nature, our methodology is able to obtain a global optimal solution with either the TSP or MST that maximizes the correlation between potential biomarkers. Both methods were capable of obtaining solutions that include all 11 genes that were importantly expressed in lung cancer. Additionally our methodology included gene relationships for which biological evidence was found directly relating them such as the case of SPP1-AGER, CYP4B1-ADH1B, and other previously described in this section.

Chapter 7. Conclusions and Future Work

Identifying cancer biomarkers is an important step in the diagnosis, prognosis, and prevention of this disease; but determining how these biomarkers are related or interact is just as important. A signaling pathway among these biomarkers is, then, a worthy aim. Uncovering signaling pathways of potential biomarkers could further the understanding of the origins and the evolution of cancer, if validation succeeds.

The initial step to identifying a potential signaling pathway related to lung cancer, is to first apply the Multiple Criteria Optimization (MCO) methodology proposed by Katia I. Camacho [13]. This allows the identification of genes with an important change in genetic expression from a microarray database. Once these potential biomarker genes are identified they were used to construct our mathematical representation where our optimization methods are applied.

As a first approach to structure a network of potential biomarkers the linear correlations that exist among these genes are used. Once the network of a preselected list of potential biomarkers is constructed, the Traveling Salesperson Problem can be applied to obtain an optimal sequence that maximizes the linear correlations. This sequence represents the potential signaling pathway. In Section 5.1 biological information was gathered evidencing the potential of our methodology and of the important gene relations in our proposed signaling pathway.

Additionally the MST formulation is used here as an alternative method to the TSP to discover a signaling pathway from the same list of preselected potential biomarkers. The

MST provides an optimal tree, a representation that must be contrasted to that of TSP's tour.

In this thesis the first steps towards meta-analysis have been laid out as described in Section 4.1.3. Currently there is work being carried out for the simultaneous analysis of multiple databases to identify genes of interest and then construct signaling pathways with the TSP and MST.

Sections 5.1 and 5.2 of this thesis include the exploration for biological evidence to assess the resulting signaling pathways constructed when comparing the different conditions of the microarray database described in both our lung cancer case study and meta-analysis. Several groupings of genes were common throughout several of the conditions which this could have a biological significance worthy of even further biological validation. Public databases such as KEGG and GeneCards provided information which was used for this purpose. Evidence was found validating that several gene relationships included in our results have a documented relationship within previously published signaling pathways. Additional studies were obtained that established that several gene relationships indeed played a role in lung cancer, other cancer types, and other diseases (Parkinson's disease, Alzheimer's disease, among others).

As described in Section 5.1, several gene relationships that formed part of the TSP and MST optimal solutions were present in both solutions. Table 5.4 lists these groups of gene relationships. The fact that they are included in both optimal solutions, brings emphasis on these relationships as possibly forming an even more important and relevant role in lung cancer. Similarly in Section 5.2, several relationships were common

throughout the solutions of the six different condition comparisons conducted. Table 4.4 lists these gene relationships which not only could be highly relevant when comparing control to lung cancer patients, but also when taking in consideration the smoking habits of patients.

Our methodology is capable of obtaining optimal solutions that maximize the correlations between potential biomarkers identified previously through Multiple Criteria Optimization. By applying the TSP and MST it is capable of finding optimal solutions from a large number of possible solutions. This presents its potential to serve as a starting point for experimental analysis, which involves large sums of resources. In conjunction with the work of Katia I. Camacho [13], our methodology identifies potential biomarkers and proposes important relationships between them that can play an important role in a disease such as lung cancer. Additionally there are plans to explore and apply the methods discussed in this thesis to analyze other -omics experiments, such as microRNA databases.

References

[1] M. L. Sanchez, C. Isaza, M. Cabrera Ríos, J. Perez-Morales, C. Rodriguez-Padilla and J. Castro, "Identification of Potential Biomarkers from Microarray Experiments Using Multiple Criteria Optimization," *National Center for Biotechnology Information,* 2013.

[2] K. Strimbu and J. A. Tavel, "What are Biomarkers?," *National Institute of Health,* 2011.

[3] "National Heart, Lung, and Blood Institute," 23 December 2013. [Online]. Available: www.nhlbi.nih.gov/health-topics/topics/cf/causes.

[4] "Genetics Home Reference," February 2012. [Online]. Available: http://ghr.nlm.nih.gov/condition/duchenne-and-becker-muscular-dystrophy.

[5] B. Vogelstein and K. W. Kinzler, "Cancer Genes and the Pathways They Control," *Nature Medicine,* 2004.

[6] "National Human Genome Research Institute," National Institute of Health, 12 June 2012. [Online]. Available: http://www.genome.gov/27530687.

[7] R. K. Ahuja, T. L. Magnanti and J. B. Orlin, Network Flows: Theory, Algorithms, and Applications, Prentice Hall, 1993.

[8] S. Pettie, "On the Shortest Path and Minimum Spanning Tree Problems," UMI, 2003.

[9] "National Cancer Institute, Understanding cancer and related topics [electronic resource]: understanding gene testing. Bethesda, Md.," National Cancer Institute, 2011.

[10] M. Cabrera-Ríos, M. L. Sanchez and C. E. Isaza, Identification of Potential Cancer Biomarkers from Microarray Data: A parameter-free novel tool for meta-analysis of microarray databases, VDM Verlag, 2011.

[11] American Cancer Society, "Cancer Facts & Figures 2015," Atlanta: American Cancer Society, 2015.

[12] K. K. Jain, The Handbook of Biomarkers, Humana Press, 2010.

[13] K. I. Camacho, "Optimization-Driven Meta-analysis: The Simultaneous Search for Cancer Biomarkers with Multiple Microarrays Experiments," 2014.

[14] S. P. Glasser, Essentials of Clinical Research, Springer, 2014.

[15] M. S. Rice, M. A. Murphy and S. S. Tworoger, "Tubal ligation, hysterectomy and ovarian cancer:," *Journal of Ovarian Research,* 2012.

[16] Y. Wu, Y. Ye, Y. Shi and P. Li, "Association between vitamin A, retinol intake and blood retinol level and gastric cancer risk: A meta-analysis," *Clinical Nutrition,* 2014.

[17] D. R. Rhodes, T. R. Barrette and M. A. Rubin, "Meta-Analysis of Microarrays: Interstudy Validation of Gene Expression Profiles Reveals Pathway Dysregulation in Prostate Cancer," *Cancer Research,* pp. 4427-4433, 2002.

[18] S. W. Doniger, N. Salomonis and K. D. Dahlquist, "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data," *Genome Biology,* 2003.

[19] A. H. M. Sallah, "A Review on Pathway Analysis Software Based on Microarray Data Interpretation," *International Journal of Bio-Science and Bio-Technology,* vol. 5, 2013.

[20] A. D. Baxevanis and B. F. Ouellette, Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Wiley, 2005.

[21] D. Voet and J. G. Voet, Biochemistry, Wiley, 2004.

[22] H.-J. Chung, C. H. Park and M. R. Han, "ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using Scalable Vector Graphics," *Nucleic Acids Research,* vol. 33, no. W621–W626, 2005.

[23] R. Pandey, R. K. Guru and D. W. Mount, "Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data," *Bioinformatics,* vol. 20, 2004.

[24] A. L. Tarca, S. Draghici and P. Khatri, "A novel signaling pathway impact analysis," *Bioinformatics,* vol. 25, 2008.

[25] T. Hancock, I. Takigawa and H. Mamitsuka, "Mining metabolic pathways through gene expression," *Bioinformatics,* vol. 26, 2010.

[26] M. Ashburner, C. Ball and J. Blake, "Gene Ontology: tool for the unification of biology," *Nature Genetics,* 2000.

[27] S. W. Doniger, N. Salomonis and K. D. Dahlquist, "MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data," *Genome Biology,* 2003.

[28] G. Laporte, "A Concise Guide to the Traveling Salesman Problem," *Journal of Operational Research Society,* 2010.

[29] F. S. Hillier and G. J. Lieberman, Introduction to Operations Research, McGraw Hill, 2005.

[30] W. L. Winston, Operations Research Applications and Algorithms, Thomson, 2004.

[31] B. Y. Wu and K.-M. Chao, Spanning Trees and Optimization Problems, Chapmann & Hall/CRC Press, 2004.

[32] E. Lorenzo, K. Camacho-Caceres, A. J. Ropelewski, J. Rosas, M. Ortiz-Mojer, L. Perez-Marty, J. Irizarry, V. Gonzalez, J. A. Rodriguez, M. Cabrera-Ríos and C. Isaza, "An Optimization-Driven Analysis Pipeline to Uncover Biomarkers and Signaling Paths: Cervix Cancer," *microarrays,* pp. 287-310, 2015.

[33] S. Simon, "Stats: What is a Correlation (Pearson Correlation)?," *Children's Mercy Hospital and Clinics,* p. www.childrenmercy.org/stats/definitions/correlation.htm, 2005.

[34] D. Francis, A. J. Coats and D. G. Gibson, "How high can a correlation coefficient be?," *International Journal of Cardiology,* 1999.

[35] L. L. Lapin, Probability and Statistics for Modern Engineering, PWS-Kent, 1990.

[36] Landi, "Gene Expression Signature of Cigarette Smoking and Its Role in Lung Adenocarcinoma Development and Survival," *PLoS ONE,* vol. Vol.3, no. no.2, 2008.

[37] "National Cancer Institute," National Institutes of Health, 5 November 2014. [Online]. Available: http://www.cancer.gov/.

[38] Z. Lu, "PubMed and beyond: a survey of web tools for searching biomedical literature," *Database,* 2011.

[39] A.-G. Sun, J. Wang, Y. Shan, W.-J. Yu and X. Li, "Identifying distinct candidate genes for early Parkinson's diseas by analysis of gene expression in whole blood," *Neuroendocrinology Letters,* 2014.

[40] X. Liu, Y. Gao, B. Zhao and X. Li, "Discovery of Microarray-Identified Genes Associated with Ovarian Cancer Progression," *International Journal of Oncology,* 2015.

[41] K. Zuberi, M. Franz, H. Rodrigue and J. Montojo, "GeneMANIA Prediction Server 2013 Update," *Nucleic Acids Research,* 2013.

[42] D. Warde-Farley, S. Donaldson, O. Comes, K. Zuberi and R. Badrawi, "The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic Acids Research,* 2010.

[43] M. Molina-Navarro, J. C. Triviño, L. Martinez-Dolz and F. Lago, "Functional Networks of Nucleocytolasmic Transport-Related Genes Differentiate Ischemic and Dilated Cardiomyopathies. A New Therapeutic Opportunity," *PLoS ONE,* 2014.

[44] "GeneMANIA," University of Toronto, 2015. [Online]. Available: www.genemania.org.

[45] D. Q. Nykamp, "Math Insight," [Online]. Available: http://mathinsight.org/definition/network_edge.

[46] C. H. Papadimitriou and K. Steiglitz, Combinatorial Optimization: Algorithms and Complexity, Prentice-Hall, 1982.

[47] "GeneCards: Human Gene Database," Weizmann Institute of Science, [Online]. Available: http://www.genecards.org/. [Accessed 2015].

[48] S. Kumari, J. Nie, H.-S. Chen and H. Ma, "Evaluation of Gene Association Methods for Coexpression Network Construction and Biological Knowledge Discovery," *PLoS One,* 2012.

[49] N. Chok, "Pearson's versus spearman's and kendall's correlation coefficient for continuous data," *University of Pittsburgh,* 2010.

[50] J. Rodgers and W. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician,* 1988.

Appendix 1

- **Node:** a *node* (or vertex) of a network is a point or area where two lines or edges intersect or branch off.
- **Edge:** an *edge* (or link) of a network (or graph) is one of the connections between the *nodes* (or vertices) of the network [45].
- **Walk:** a *walk* is a subgraph of *G* consisting of a sequence of nodes or vertices and arcs i_1-a_1-i_2-a_2-…-i_{r-1}-a_{r-1}-i_r satisfying the property that for all $1 \leq k \leq r-1$, either $a_k = (i_k, i_{k+1}) \in A$ or $a_k = (i_{k+1}, i_k) \in A$ [7].
- **Path:** a *path* is a *walk* without any repetition of nodes. The *walk* shown in Figure 0.1a) is a path but the *walk* shown in Figure 0.1b) is not because it repeats node 2 twice [7].
- **Cycle:** a *cycle* can be defined as a closed *walk* with no repeated nodes other than its first and last one, Figure 0.2 presents an example of a *cycle* [46].

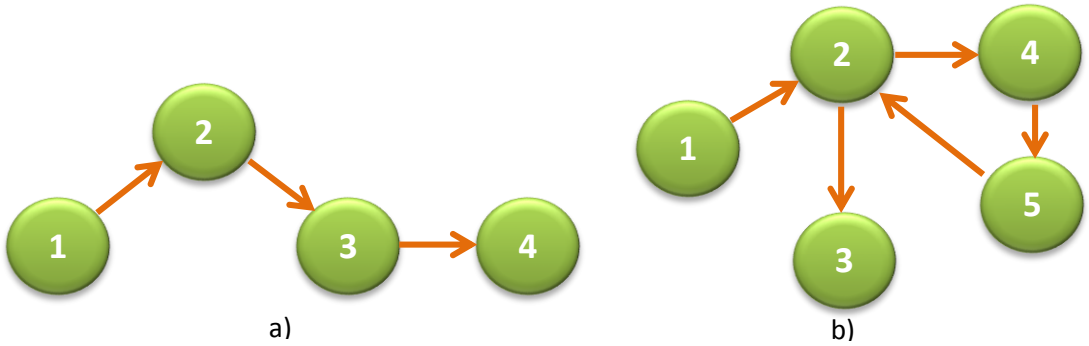

Figure 0.1Examples of Walks

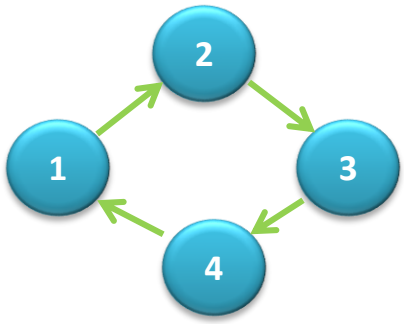

Figure 0.2 Example of a Cycle

Appendix 2

```
%Análisis de frontera Pareto de cinco criterios
%Autor: Katia I Camacho Cáceres

dataT = load('data5Criteria.txt'); %Cargar la data
[x,y] = size(dataT); % data completa x=num filas, y=num columnas
data = dataT(:,2:end); %se toma solo las columnas de los criterios
[n,m]=size(data); %n=num filas (k=PM), m=num columnas (j = criterios)
c1 = 1000*ones(n,n,m);   % matriz primera condición con j criterios
for j=1:m
    for a=1:n
        for b=1:n
            if data(a,j) == data(b,j) %condición 1.1
                c1(a,b,j)=0;
            elseif data(a,j)<data(b,j)
                c1(a,b,j)=-1;
            end
        end
    end
end

% Procedimiento para sumar c1 para cinco criterios
c2=zeros(n,n);  %matriz segunda condición
for a=1:n
    for b=1:n
        if c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5)==0
            c2(a,b)=2500;
        elseif c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5)==1000
            c2(a,b)=2500;
        elseif c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5)==2000
            c2(a,b)=2500;
        elseif (c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5))==3000
            c2(a,b)=2500;
        elseif (c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5))==4000
            c2(a,b)=2500;
        elseif (c1(a,b,1)+c1(a,b,2)+c1(a,b,3)+c1(a,b,4)+c1(a,b,5))==5000
            c2(a,b)=5000;
        end
    end
end

% Procedimiento para encontrar conjunto dominado cd y no dominado cnd
cnd = zeros(x,y); %matriz del conjunto no dominado
cd = zeros(x,y);  % matriz del conjunto dominado
i=0; %contador para cd
j=0; %contador para cnd
for a=1:x
    sumfila=sum(c2(a,:));
     if sumfila>=5000; % conjunto dominado
        i=i+1;
        cd(i,:)=dataT(a,:);
    else  % conjunto no dominado
        j=j+1;
        cnd(j,:)=dataT(a,:);
```

```
    end
end

index = 1:x;
disp([round(index') cd]);
disp([round(index') cnd]);
%Mostrar el Conjunto no dominado en un notepad, con los datos de
%Posicion,f1,f2, f3, f4, f5c
disp('   Conjunto no dominado    ');
cnd=cnd(1:j,:);
filecnd = fopen('cnd5CriteriaBio.txt','w');
fprintf(filecnd,'%6s   %12s   %12s   %12s   %12s
%12s\r\n','Posicion','F1','F2','F3', 'F4', 'F5');
fprintf(filecnd,'%6.4f   %12.4f   %12.4f   %12.4f   %12.4f
%12.4f\r\n',cnd');
fclose(filecnd);
```

Appendix 3

Gene	Function [47]
AGER	Acts as a mediator of both acute and chronic vascular inflammation in conditions such as atherosclerosis and in particular as a complication of diabetes. AGE/RAGE signaling plays an important role in regulating the production/expression of TNF-alpha, oxidative stress, and endothelial dysfunction in type 2 diabetes. Interaction with S100A12 on endothelium, mononuclear phagocytes, and lymphocytes triggers cellular activation, with generation of key proinflammatory mediators. Interaction with S100B after myocardial infarction may play a role in myocyte apoptosis by activating ERK1/2 and p53/TP53 signaling (By similarity). Receptor for amyloid beta peptide. Contributes to the translocation of amyloid-beta peptide (ABPP) across the cell membrane from the extracellular to the intracellular space in cortical neurons. ABPP-initiated RAGE signaling, especially stimulation of p38 mitogen-activated protein kinase (MAPK), has the capacity to drive a transport system delivering ABPP as a complex with RAGE to the intraneuronal space.
SFTPC	This gene provides instructions for making a protein called surfactant protein-C (SP-C). This protein is one of four proteins (each produced from a different gene) in surfactant, a mixture of certain fats (called phospholipids) and proteins that lines the lung tissue and makes breathing easy. Without normal surfactant, the tissue surrounding the air sacs in the lungs (the alveoli) sticks together after exhalation (because of a force called surface tension), causing the alveoli to collapse. As a result, filling the lungs with air on each breath becomes very difficult, and the delivery of oxygen to the body is impaired. Surfactant lowers surface tension, easing breathing and avoiding lung collapse. The SP-C protein helps spread the surfactant across the surface of the lung tissue, aiding in the surface tension-lowering property of surfactant.
TMEM100	Plays a role during embryonic arterial endothelium differentiation and vascular morphogenesis through the ACVRL1 receptor-dependent signaling pathway upon stimulation by bone morphogenetic proteins, such as GDF2/BMP9 and BMP10. Involved in the regulation of nociception, acting as a modulator of the interaction between TRPA1 and TRPV1, two molecular sensors and mediators of pain signals in dorsal root ganglia (DRG) neurons. Mechanistically, it weakens their interaction, thereby releasing the inhibition of TRPA1 by TRPV1 and increasing the single-channel open probability of the TRPA1-TRPV1 complex.
FABP4	Lipid transport protein in adipocytes. Binds both long chain fatty acids and retinoic acid. Delivers long-chain fatty acids and retinoic acid to their cognate receptors in the nucleus
SPP1	Acts as a cytokine involved in enhancing production of interferon-gamma and interleukin-12 and reducing production of interleukin-10 and is essential in the pathway that leads to type I immunity.

WIF1	Binds to WNT proteins and inhibits their activities. May be involved in mesoderm segmentation.
COL11A1	May play an important role in fibrillogenesis by controlling lateral growth of collagen II fibrils
CYP4B1	Cytochromes P450 are a group of heme-thiolate monooxygenases. In liver microsomes, this enzyme is involved in an NADPH-dependent electron transport pathway. It oxidizes a variety of structurally unrelated compounds, including steroids, fatty acids, and xenobiotics.
FCN3	May function in innate immunity through activation of the lectin complement pathway. Calcium-dependent and GlcNAc-binding lectin. Has affinity with GalNAc, GlcNAc, D-fucose, as mono/oligosaccharide and lipopolysaccharides from S.typhimurium and S.minnesota
ADH1B	
CLDN18	Plays a major role in tight junction-specific obliteration of the intercellular space, through calcium-independent cell-adhesion activity.

Appendix 4

Microarrays **2015**, *4*, 1-x manuscripts; doi:10.3390/microarrays40x000x

OPEN ACCESS

microarrays

ISSN 2076-3905

www.mdpi.com/journal/microarrays

Article

An Optimization-Driven Analysis Pipeline to Uncover Biomarkers and Signaling Paths: Cervix Cancer

Enery Lorenzo, Katia Camacho-Caceres, Alexander J. Ropelewski, Juan Rosas, Michael Ortiz-Mojer, Lynn Perez-Marty, Juan Irizarry, Valerie Gonzalez, Jesús A Rodríguez, Mauricio Cabrera-Rios and Clara Isaza *

[1] Bio IE Lab, The Applied Optimization Group at UPRM, Industrial Engineering Dept., University of Puerto Rico at Mayaguez; E-Mail: enery.lorenzo@upr.edu, katia.camacho@upr.edu, juan.rosas1@upr.edu, michael.ortiz6@upr.edu, juan.irizarry4@upr.edu, valerie.gonzalez9@upr.edu, jesusandres.rodriguez@upr.edu, mauricio.cabrera1@upr.edu

[2] Pittsburgh Supercomputing Center; E-Mails: ropelews@psc.edu

[3] Dept. of Pharmacology and Toxicology, Ponce School of Medicine; E-Mails: cisaza@psm.edu

* Author to whom correspondence should be addressed; E-Mail: cisaza@psm.edu; Tel.: +1-787-840-2575 ext. 2198

Academic Editor:

Received: / Accepted: / Published:

Abstract: Establishing how a series of potentially important genes might relate to each other is relevant to understand the origin and evolution of illnesses such as cancer. High-throughput biological experiments have played a critical role in providing information in this regard. A special challenge, however, is that of trying to conciliate information from separate microarray experiments to build a potential genetic signaling path. This work proposes a two-step analysis pipeline based on optimization to approach meta-analysis aiming to build a proxy for a genetic signaling path.

Keywords: Traveling Salesman Problem; Signaling Pathways; Cancer Biology

---

Introduction

Technology advancement has accelerated the capability to generate large amounts of biological data. The capability to translate these data into usable knowledge has, however, grown at a much slower rate. The technologies used to generate these data are often rendered obsolete by newer ones before the data already available are fully analyzed and taken to their full potential for biological and medical advancement. Microarrays constitute a technology of this sort: one used to generate a large number of experiments, many of which will be greatly under-utilized. The analysis of microarrays, however, still holds a large potential for the discovery of genetic biomarkers for all types of cancer, as well as elicit their signaling pathways. Extracting this kind of knowledge from microarray experiments has historically been considered challenging, largely due to two main difficulties: (i) the use of incommensurable units across different experiments and (ii) the lack of analysis techniques that converge to a consistent set of biomarkers. These two difficulties propagate uncertainty to the task of determining a reliable signaling pathway. To this end, this work proposes a two-step pipeline that involves (1) a meta-analysis strategy, based on multiple-criteria optimization, which circumvents both of the main difficulties described previously to detect highly differentially expressed genes; and (2) a method, based on integer programming to find the most correlated path among the genes from the previous step. The central hypothesis is that there is a strong signal of relative expression in microarrays that is effectively discoverable through mathematical optimization.

It is critical that the detection of genetic cancer biomarkers through meta-analysis can be carried out faster, more consistently and more accurately in order to shorten the lead-time from data generation to data interpretation and knowledge application. The simultaneous meta-analysis of multiple experiments via optimization and the subsequent identification of the highest correlated genetic path described in this work offer these capabilities. Microarray data already in repositories can be readily analyzed and, prospectively, new high-throughput biological technologies could be fully utilized earlier in the fight against cancer. The gap between raw data and applicable

biomedical/medical knowledge can be reduced significantly; especially when considering that historic biological data will now be able to be brought into perspective to design new experiments and focus on more precise aspects of exploration.

2. Method

The proposed analysis pipeline has two sequential stages: 1) Meta-analysis for detection of highly differentially expressed genes and 2) Finding the most correlated path. These are explained next.

2.1. Stage 1) Meta-analysis for detection of highly differentially expressed genes

Meta-analysis involves the joint study of multiple databases to obtain conclusions that apply across all of them. Meta-analysis can help detect potential genetic cancer biomarkers through the study of microarray databases. However, to this end, a series of difficulties are apparent: (a) Microarray experiments that are publicly available use different technologies, platforms and, most of the times, different scales. Incommensurability renders many meta-analyses efforts unfeasible [1] due to the inability to make comparisons across all experiments of interest. Even when the same units are used, often time, data normalization is required for comparability. (b) There is not an efficient, systematic method to carry out meta-analysis. Most of the studies analyze one particular database and try to generalize the results to other databases or analyze several databases separately and try to make sense of all the independent results [2]. (c) The issue of having a large number of measurements and genes generally results in large number of significant genes that must be validated [3]. (d) Meta-analysis of microarrays –and of high throughput biological experiments in general – is a laborious process that is often outpaced by the development of technology to generate ever-larger data sets. That is, data generation capabilities are larger and grow faster than our abilities to make sense and translate these data into usable knowledge. (e) Large repositories of public data generated through costly microarray experiments could go underanalyzed and underutilized in the fight for cancer when the researchers´ attention shifts to the next high-throughput technology. The problem of making sense of large quantities of data, however, will persist.

2.2. Multiple Criteria Optimization

Multiple Criteria Optimization (MCO) is a field from Engineering Mathematics that deals with making decisions in the presence of multiple performance measures in conflict i.e. decisions where optimizing one criterion results in moving away from optimality in at least another criterion. Because of the presence of conflict, an MCO problem does not find a single best solution but rather a set of best compromising solutions in light of the performance measures under analysis. The best compromises define solutions called Pareto-Efficient (or simply Efficient, for short) that define the Efficient Frontier of the MCO problem at hand. A typical multiple criteria optimization with two conflicting performance measures (objectives), PMs, can be visualized as in Figure 1. In this

figure, a set of seven candidate points characterized by their values on both performance measures are shown. The performance measure represented in the x-axis is to be maximized while the performance measure in the y-axis is to be minimized in this example. The problem is to find those candidate points that dominate all of the other points in both performance measures. In the face of conflict, this will result in a group of candidates in the southeast extreme of the set in Figure 1, solutions 3 and 5. These are Pareto-efficient solutions and, when all of them are accounted for, they integrate the Efficient Frontier of the MCO problem. In this example, it can be noted that among efficient solutions, an improvement in one performance measure can only come strictly at the detriment of another one: moving from solution 5 to solution 3 will result in an improvement in the performance measure associated to the vertical direction, but in a loss in the performance measure associated to the horizontal direction. Note that the general problem involves at least two performance measures to be optimized, where only the case with two performance measures has a convenient graphical representation. An MCO problem, however, can include as many dimensions (or performance measures) as necessary.

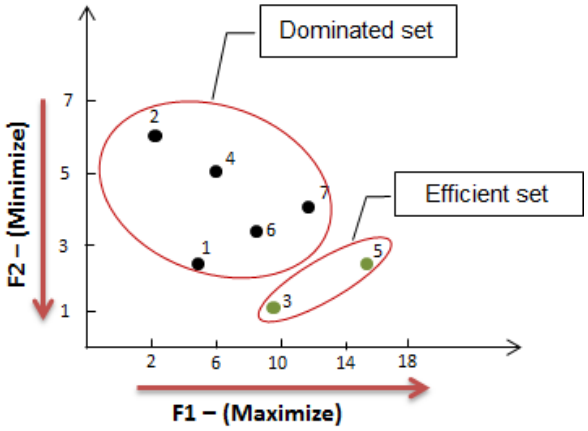

Figure 1. Representation of a multiple criteria optimization problem with two performance measures.

The general mathematical formulation of an unconstrained MCO problem is as follows:

$Find\ \boldsymbol{x}\ to$

(P1)

$Minimize\ f_j(\boldsymbol{x}) \quad j = 1,2,\dots,J$

The MCO problem in P1 can be discretized onto a set K with $|K|$ points in the space of the decision variables so as to define particular solutions $\boldsymbol{x^k}$, $(k = 1,2,\dots,|K|)$ which can in turn be evaluated in the J performance measures to result in values $f_j(\boldsymbol{x^k})$. That is, the k^{th} combination of values for the decision variables evaluated in the j^{th} objective function. The illustrative example in Figure 1 follows this discretization with $J = 2$ performance measures and $|K| = 7$ solutions.

The MCO formulation under such discretization is, then as follows:

$Find\ \boldsymbol{x^k}\ (k \in K)\ to$

(P2)

$Minimize\ f_j(\boldsymbol{x^k}) \quad j = 1,2,\dots,J$

The solutions to P2 are, then, the Pareto-efficient solutions of the discretized MCO problem. Considering formulation P2, a particular combination $\boldsymbol{x^0}$ with evaluations $f_j(\boldsymbol{x^0})$ will yield a Pareto-Efficient solution to P2 if and only if no other solution $\boldsymbol{x^\psi}$ exists that meets two conditions, from this point on called Pareto-optimality conditions:

$f_j(\boldsymbol{x^\psi}) \leq f_j(\boldsymbol{x^0})\ \forall j$

(Condition 1)

$f_j(\boldsymbol{x^\psi}) < f_j(\boldsymbol{x^0}) \quad in\ at\ least\ one\ j$

(Condition 2)

Conditions (1) and (2) imply that no other solution $\boldsymbol{x^\psi}$ dominates the solution under evaluation, $\boldsymbol{x^0}$, in all performance measures simultaneously.

On previous publications [3**,** 4] our group has demonstrated that if a set of candidate solutions evaluated by multiple performance measures is available, it is possible to determine a series of best compromises between all criteria through a technique called Data Envelopment Analysis (DEA). The idea behind DEA is to use an optimization model to compute a relative efficiency score for each particular solution with respect to the rest of the candidate solutions. The resulting best compromises, identified through their efficiency score, form the envelope of the solution set, therefore the name Data Envelopment Analysis. These solutions are indeed Pareto-efficient solutions of the problem under analysis.

The DEA linear programming formulations proposed by Banks, Charnes and Cooper [5] are shown below:

Find $\boldsymbol{\mu},\boldsymbol{\nu},\mu_0^+,\mu_0^-$ to

Maximize $\boldsymbol{\mu}^T\mathbf{Y}_0^{\max} + \mu_0^+ - \mu_0^-$

Subject to

$$\boldsymbol{\nu}^T\mathbf{Y}_0^{\min} = 1$$

$$\boldsymbol{\mu}^T\mathbf{Y}_j^{\max} - \boldsymbol{\nu}^T\mathbf{Y}_j^{\min} + \mu_0^+ - \mu_0^- \le 0 \quad j = 1,...,n$$

$$\boldsymbol{\mu}^T \ge \varepsilon \cdot \mathbf{1}$$

$$\boldsymbol{\nu}^T \ge \varepsilon \cdot \mathbf{1}$$

$$\mu_0^+,\mu_0^- \ge 0$$

(P3)

Find $\boldsymbol{\nu},\boldsymbol{\mu},\nu_0^+,\nu_0^-$ to

Minimize $\boldsymbol{\nu}^T\mathbf{Y}_0^{\min} + \nu_0^+ - \nu_0^-$

Subject to

$$\boldsymbol{\mu}^T\mathbf{Y}_0^{\max} = 1$$

$$\boldsymbol{\nu}^T\mathbf{Y}_j^{\min} - \boldsymbol{\mu}^T\mathbf{Y}_j^{\max} + \nu_0^+ - \nu_0^- \ge 0 \quad j = 1,...,n$$

$$\boldsymbol{\nu}^T \ge \varepsilon \cdot \mathbf{1}$$

$$\boldsymbol{\mu}^T \ge \varepsilon \cdot \mathbf{1}$$

$$\nu_0^+,\nu_0^- \ge 0$$

(P4)

where $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are column vectors containing multipliers to be optimally determined together with scalar variables μ_0^+ and μ_0^- in the first case and together with ν_0^+ and ν_0^- in the second case; $\mathbf{Y}_j^{\min}$ and $\mathbf{Y}_j^{\max}$ are column vectors containing the values of the *j*th combination of performance measures to be minimized and maximized respectively; and ε is a scalar usually set to a value of $1x10^{-6}$.

Model P3 is called the BCC Input Oriented Model and Model P4 is called the BCC Output Oriented Model. Both models are applied to each of the *n* candidate solutions. A particular solution with an objective function score of 1 (i.e. an efficiency score of 1) using both formulations is in the envelope of the set and is considered to be an efficient solution to the MCO problem. The BCC model is just one of many possible DEA formulations, albeit a very powerful one. This model's mathematical linear nature provides it with the capability of finding efficient solutions associated with the data set under analysis through a series of piece-wise linear segments. Nonlinear behavior is, then, approached with tractability and with the certainty that at least the efficient solutions lying in the convex part of the frontier are being found. Figure 2 shows an MCO problem solved through with DEA, specifically with the BCC model.

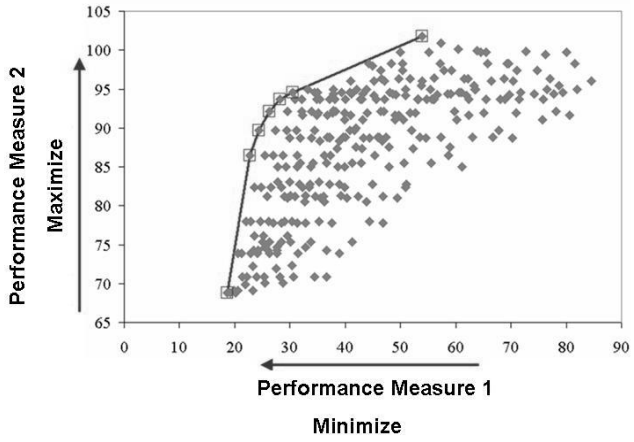

Figure 2. Multiple Criteria Optimization Problem solved using Data Envelopment Analysis (BCC model). The efficient solutions are identified through the use of piecewise-linear segments.

DEA has several advantages including: (i) computational efficiency owing to its linear optimization structure; (ii) objectivity and consistency of results, which follows from not requiring the adjustment of parameters or assigning weights to the different performance measures by the user, and (iii) capability of analyzing several microarray experiments with incommensurate units. Appendix A discusses the volcano plot, a widely used tool to detect differentially expressed genes, to illustrate how the analyst can bias the results. On the other hand, one limitation of DEA is that of depending on a series of local linear approximations, as shown in Figure 2. Every time that a linear segment is superimposed over the set under analysis, there are genes lying in the nonconvex part of the set frontier that escape detection. These genes could be potential biomarkers, however. In order to circumvent the said disadvantage, the authors proposed that DEA be applied successively 10 times, each time removing the genes found in a particular iteration from the set for subsequent analyses. This strategy results in 10 frontiers, as seen in Figure 3. The number of efficient frontiers is, admittedly, an arbitrary number at this point, thus further refinement is necessary in this aspect.

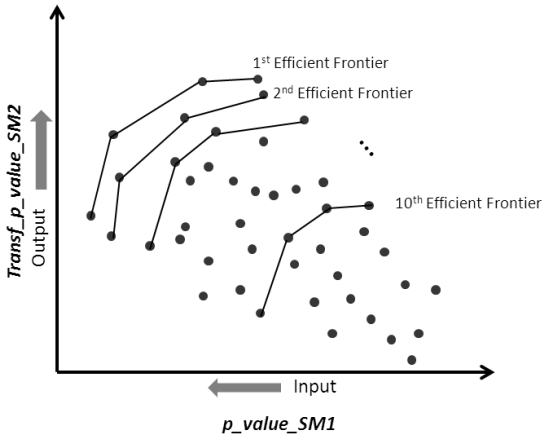

Figure 3. A case with genes characterized by two performance measures. Referring to this figure, and following the proposed method, at this point it is recommended to identify the first 10 efficient frontiers. This can be easily done by identifying the genes in the first efficient frontier through DEA, then removing them from the set and continue with a second DEA iteration.

At the end of Stage 1, the analyst is left with a set of differentially expressed genes that can be investigated to establish their role in the condition or illness under study, cancer in this case. This set of genes in the proposed method, however, will be used to determine how these are maximally correlated in Stage 2.

2.3. Stage 2) Finding the most correlated path

It is proposed that the most correlated path among the list of candidate genes identified in the previous stage can be found optimally. To this end, the optimization problem identified in the literature as the Travelling Salesman Problem (TSP), is introduced here as a viable model.

The TSP is generally stated as follows: a salesman needs to visit n cities and needs to minimize the travel distance starting and finishing in the city of origin. Each city must be visited only once. The solution, then, is a tour. In n cities, there is a total of n! tours. If a particular city of origin is selected a priori, then the number of tours is (n-1)!. In our case, the objective is to find the tour among n genes of interest that maximizes the sum of the absolute values of pairwise correlations. This tour would then be interpreted as a surrogate for a biological pathway, defined as "a series of actions among molecules in a cell" [6], and more specifically for a genetic signaling pathway. A biological pathway "can provide clues about what goes wrong when a disease strikes." [6]

As a first approximation, it is proposed that the absolute values of linear correlation coefficients computed among a list of genes of potential biomarkers be used to construct networks such as the one presented in Figure 3, where the TSP can be readily applied. The idea of using a linear statistical correlation is, indeed, widely used in the literature as a means to uncover genetic coexpression. This information, in turn, should

help cancer researchers in understanding the disease as well as look for targeted treatments. The paper by Kumari, et al. (2012, PLOS) has studied different coexpression measurements, recommending to carry out a preliminary study to determine the most appropriate one for different objectives. It is, then convenient at this point to resort to the use of the Pearson correlation coefficient as a starting point in this work.

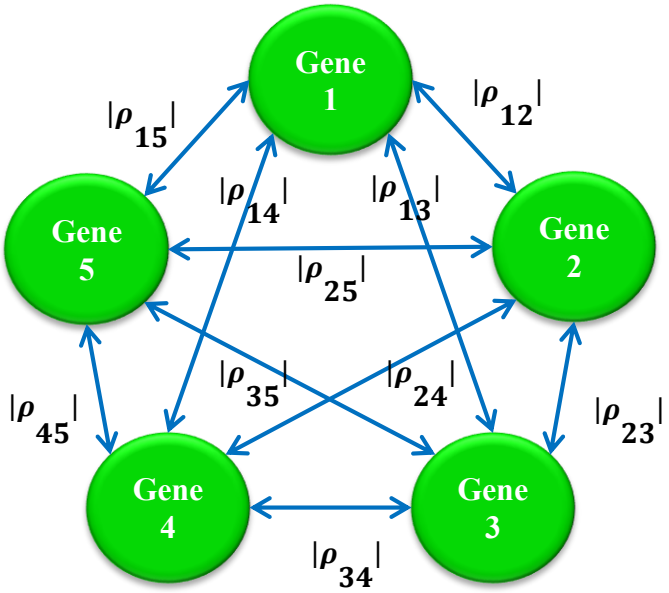

Figure 3. Representation of the many options for a cyclic path for 5 genes.

The TSP can, indeed, be understood as an optimization problem. Consider that c_{ij} represents the cost of traveling from city i to city j and let y_{ij} be a binary variable, indicating whether or not the salesman travels from city i to city j. Additionally let us define flow variables x_{ij} on each arc (i,j) and assume that the salesman has n-1 units available at node 1, which is arbitrarily selected as a "source node", and he must deliver 1 unit to each of the other nodes [7]. The optimization model is as follows:

$$Minimize \sum_{(i,j)\in A} c_{ij}y_{ij} \qquad (P5a)$$

$$\sum_{1\le j\le n} y_{ij} = 1 \quad \forall i = 1,2,\dots,n \qquad (P5b)$$

$$\sum_{1\le i\le n} y_{ij} = 1 \quad \forall\, j = 1,2,\dots,n \qquad (P5c)$$

$$Nx = b \qquad (P5d)$$

$$x_{ij} \leq (n-1)y_{ij} \quad \forall (i,j) \in A \qquad (P5e)$$

$$x_{ij} \geq 0 \quad \forall (i,j) \in A \qquad (P5f)$$

$$y_{ij} = 0 \; or \; 1 \quad \forall (i,j) \in A \qquad (P5g)$$

Following the description in [7], let *A'* = {(*i*,*j*): y_{ij} =1} and let *A''* ={(*i*,*j*): x_{ij} >0}. The constraints (P5b) and (P5c) imply that exactly one arc of *A'* leaves and enters any node *i*; therefore, *A'* is the union of node disjoint cycles containing all of the nodes of N. In general, any integer solution satisfying (P5b) and (P5c) will be union of disjoint cycles; if any such solution contains more than once cycle; they are referred to as subtours, since they pass through only a subset of nodes.

In constraint (P5d) N is an *nxm* matrix, called the *node-arc incidence matrix* of the minimum cost flow problem. Each column N_{ij} in the matrix corresponds to the variable x_{ij}. The column N_{ij} has a +1 in the *i*th row, a -1 in the *j*th row; the rest of its entries are zero. Constraint (P5d) ensures that *A"* is connected since we need to send 1 unit of flow from node 1 to every other node via arcs in *A"*. The forcing constraints (P5e) imply that *A"* is a subset *A'*. These conditions imply that the arc set *A'* is connected and thus cannot contain subtours [7].

The TSP is known to be a hard problem to solve to optimality; however, with a manageable number of entities (nodes) optimality is well within reach. In our group's experience it has been possible to obtain the optimal TSP tour with a list of up to 100 genes in less than 1 hour of computing time in a personal computer. The Branch & Bound –an exact algorithm- was used to this end, as coded in Matlab. An exact algorithm is defined as one capable to arrive to a global optimal solution -provided that one exists- with certainty. Although it is also possible to use heuristics to approach the TSP, it must be understood that a heuristic method by definition does not provide certainty on arriving to a global optimal solution.

Referring back to Figure 3, it should be now apparent that in n genes associated to the nodes in the network, it is possible to obtain pairwise correlations to connect all genes among them resulting in a fully connected network. This network, in turn, can be mathematically translated into formulation P5a-P5g to identify the most correlated path. Thus, at the end of this stage, the most correlated path among all candidate genes from Stage 1, will be available as a proxy for a signalling

path. The application of this two-stage analysis pipeline is demonstrated next in the context of cervix cancer.

3. Results for Cervix Cancer

3.1. Stage 1

In order to demonstrate the proposed analysis pipelines, this section presents results in cervix cancer previously published in [3]. The database used for this study was introduced in [8] and contained 8 healthy tissues and 25 cervical cancer tissues, all of them with expression level readings for 10,692 genes from a cDNA microarray. The list of 28 potential biomarkers after applying Data Envelopment Analysis (DEA) is shown in Table 1. The genes in this list were cross validated for agreement in the direction of expression change in an independent database associated to [9]. As described previously, these genes were extracted from the first 10 frontiers of the analysis. The role of the selected genes in cancer was previously discussed in a previous publication of our group [3]. The fourth column of Table 1 summarizes the types of cancer that where the particular genes were found to be involved following such results.

Table 1. List 28 genes found through DEA as being differentially expressed in cervix cancer and cross validated for the direction of expression change [3].

Gene Probe	**Gene Name**	**Sign of expression change from healthy tissues to cancer tissues**		**Examples of cancer types where the gene is involved**	**Reference**
		Database 1 [8]	**Database 2 [9]**		
202575_at	CRABP2	-	-	Head and Neck, Breast	[10,11]
205402_x_at	PRSS2	-	-	Colorectal, Gastric Tumorigenesis	[12,13]
218677_at	S100A14	-	-	Esophageal squamous cell carcinoma cells, oral squamous cell carcinoma	[14,15]
202096_s_at	TSPO	-	-	Thyroid, Breast	[16,17]
212249_at	PIK3R1	-	-	Endometrial, Colorectal	[18,19]
212567_s_at	MAP4	-	-	Breast, non small cell lung carcinomas	[20,21]

211366_x_at	CASP1	-	-	Cervical squamous carcinoma cells	[22]
212889_x_at	GADD45GIP1	-	-	SKOV3 and HeLa cell lines	[23]
206626_x_at	SSX1	-	-	Prostate, multiple myeloma	[24,25]
213450_s_at	ICOSLG	-	-	Metastatic melanoma, ductal pancreatic adenocarcinoma	[26,27]
220405_at	SNTG1	-	-		
208032_s_at	GRIA3	-	-	Pancreatic	[28]
205690_s_at	BUD31	-	-		
206543_at	SMARCA2	-	-	Prostate, Skin	[29,30]
212291_at	HIPK1	+	+	Acute myeloid leukemia	[31,32]
211615_s_at	LRPPRC	+	+	Lung adenocarcinoma cell lines, oesophageal squamous cell carcinoma, stomach, colon, mammary and endometrial adenocarcinoma, and lymphoma	[33]
222027_at	NUCKS1	+	+	Breast	[34]
205362_s_at	PFDN4	+	+	Colorectal	[35]
211929_at	HNRNPA3	+	+	Non-small cell lung cancer	[36]
203738_at	C5orf22	+	+		
201794_s_at	SMG7	+	+		
200607_s_at	RAD21	+	+	Breast	[37]
201011_at	RPN1	+	+	Hematologic malignancies	[38]
201761_at	MTHFD2	+	+	Bladder, breast	[39,40]

203880_at	COX17	+	+	Non-small cell lung cancer	[41]
212255_s_at	ATP2C1	+	+	Breast, Cervical	[42,43]
205112_at	PLCE1	+	+	Gastric adenocarcinoma, colorectal	[44,45]
201663_s_at 201664_at	SMC4	+	+	Breast, cervical	[46,47,48]

3.2. Stage 2

Correlation is used in this project as a proxy for inhibitory or excitatory behavior between differences in the expression levels of two genes. As a first step, the linear correlation values between potential biomarkers are obtained. The following step was to arrange the correlation values in a matrix. To construct this matrix, first the differences between control and cancer tissues had to be calculated for each gene. Then, the absolute values of the correlation coefficients were calculated among each pair of genes based on these differences and stored in the said matrix. The absolute correlation values were consequently associated to the arcs in a fully connected graph with nodes representing potential biomarker genes. The resulting graph made possible the use of the formulation of the TSP. The optimal solution to this particular TSP is the tour among the genes of interest with the largest possible correlation, or similarly, the most correlated cyclic path as shown in Figure 4. It must be recalled at this point that there are a total of $28! \approx 3.04 \times 10^{29}$ ways in which a cyclic path can be drawn among the 28 genes.

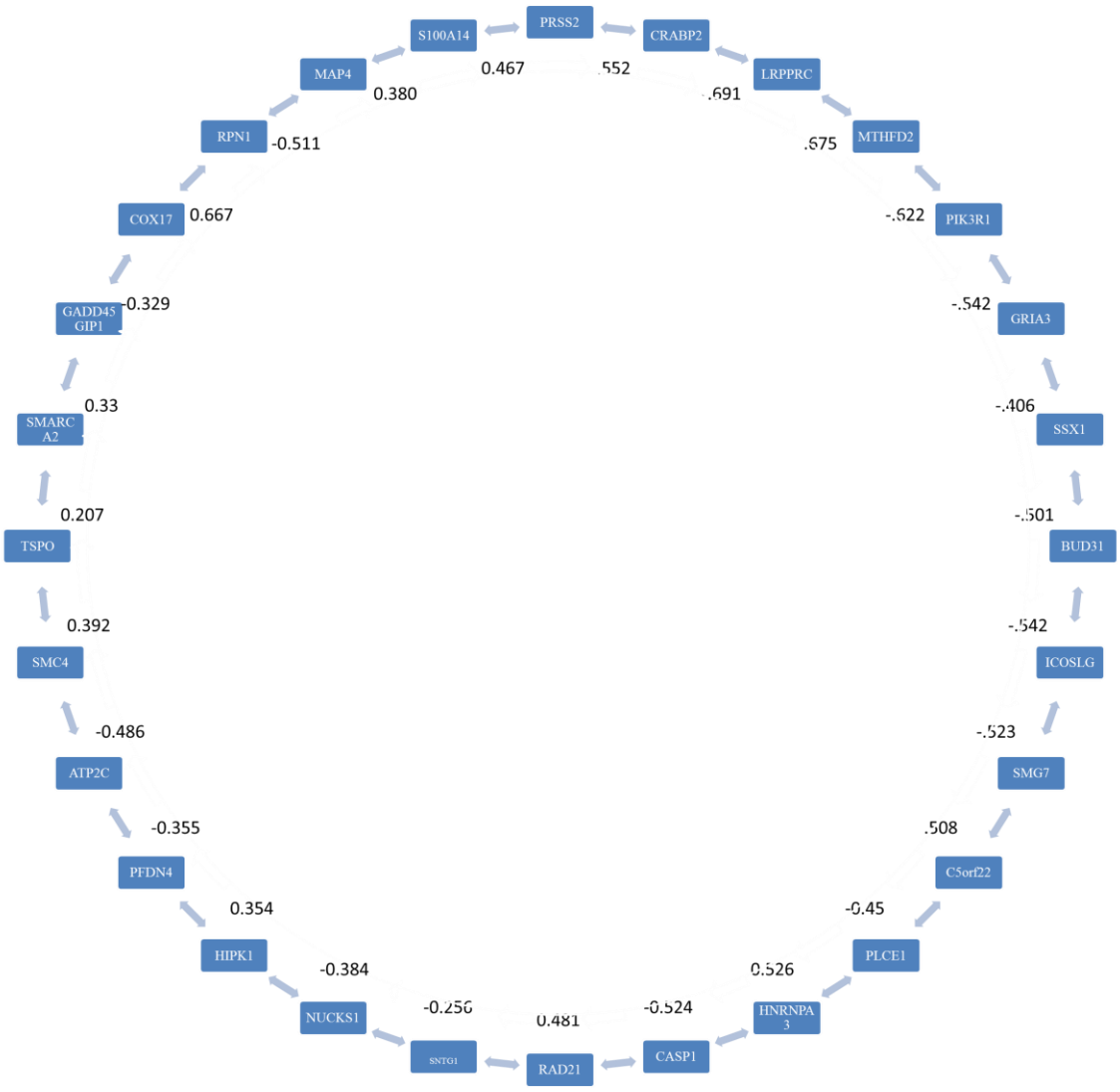

Figure 4: Highest Correlated Cyclic Path among the 28 genes identified in Stage 1.

The TSP formulation allows a wide range of analyses. In this case, the idea was to test the stability of the TSP solutions. In order to do so, TSP solutions were obtained using increasing numbers of potential biomarkers in the list of genes presented in Table 1 following the increasing order of the efficient frontier in which these were found. Starting with 5 genes, each time 5 more genes were introduced until the list was depleted on each case. Path segments that persisted across both databases were identified. Furthermore, path segments that persisted along the entire study were deemed the most stable. The results of this study were then matched against known biological pathways publicly available in the Kyoto Encyclopedia of Genes and Genomes (KEGG) [49]. A python script was written to make this process more efficient. This script is provided in Appendix B. Table 2 summarizes the results for each progressive analysis that introduced five genes at a

time. As shown in Table 2, (LRPPRC with MTHFD2) and (RPN1 with COX17) are adjacent in the correlated cyclic path when the optimal solution is obtained for 25 and 28 genes. In addition, gene S100A14 is adjacent to TSPO when the optimal solution for 5, 15, 20 and 25 genes is found.

Table 2. Adjacent genes in the solutions for the correlated cyclic path found adding five genes at a time.

Number of Genes	Adjacent Genes
5	(CRABP2 with PRSS2) and (S100A14 with TSPO)
10	(PIK3R1 with MAP4) and (GADD45GIP1 with ICOSLG)
15	(SSX1 with BUD31), (ICOSLG with SNTG1), and (S100A14 with TSPO)
20	(LRPPRC with C5orf22) and (S100A14 with TSPO)
25	(S100A14 with TSPO), (SSX1 with GRIA3), (LRPPRC with MTHFD2), (RAD21 with BUD31), and (RPN1 with COX17)
28	(LRPPRC with MTHFD2) and (RPN1 with COX17)

A search for biological pathways in KEGG databases was conducted, however not every gene could be linked to a pathway. When comparing the known biological pathways with the obtained optimal solutions, for database GSE 7803 [8] and GSE 9750 [9], the only genes that appeared adjacent in the correlated cyclic path for both were COX17 with RPN1 and the only KEGG pathway common to both has the identifier 01100 that corresponds to the collection of Methabolic pathways. On the other hand for database GSE 7803 [8], medium correlation was observed between genes HNRPA3 with BUD31, and both gene products are present in KEGG pathway 03040 that corresponds to the splisosome. For database GSE 9750 [9], PLCE1 is adjacent to PIK3R1, both gene products share the following KEGG pathways: 04012 that corresponds to the ErbB signaling pathway, 04015 Ras signaling pathway, 04015 Rap1 signaling pathway, 04066 HIF-1 signaling pathway, 04070 Phosphatidylinositol signaling system, 04370 VEGF signaling pathway, 04650 Natural killer cell mediated cytotoxicity, 04660 T cell receptor signaling pathway, 04664 Fc epsilon RI signaling pathway, 04666 Fc gamma R-mediated phagocytosis, 04670 Leukocyte transendothelial migration, 04722 Neurotrophin signaling pathway, 04750 Inflammatory mediator regulation of TRP channels, 04919 Thyroid hormone signaling pathway, 05169 Epstein-Barr virus infection, 05200 Pathways in cancer, 05200 Pathways in cancer, 05214 Glioma, 05223 Non-small cell lung cancer, and the 05231 KEGG pathway that corresponds to Choline metabolism in cancer.

Table 3. Selected genes localization

Gene	Location
HIPK1	1p13.2
NUCKS1	1q32.1
SMG7	1q25.3
CRABP2	1q21.3
S100A14	1q21.1
HNRNPA3	2q31.2
LRPPRC	2p21
MTHFD2	2p13.1
SMC4	3q26.1
ATP2C	3q22.1
RPN1	3q21.3
MAP4	3p21.31
COX17	3q13.33
C5orf22	5p13.3
PIK3R1	5q13.1
BUD31	7q22.1
PRSS2	7q34
SNTG1	8q11.21
RAD21	8q24.11
SSX1	Xp11.23
GRIA3	Xq25
PFDN4	20q13.2
CASP1	11q22.3
PLCE1	10q23.33
ICOSLG	21q22.3
GADD45G	19p13.2
SMARCA2	9p22.3
TSPO	22q13.31

In cancer there are chromosomal physical changes that produce gains or losses of certain genes. To explore if the position of the genes in the cyclic path could also provide information about these chromosomal changes, the location of each gene was consider (this information was obtained from [50]). This information is listed in table 3. All chromosomes in table 3 have been reported as having changes in cervical cancer, in regions close to the ones where the selected genes belong. It is interesting to note that some of the genes that are neighbors in the cyclic path are also neighbors in their genetic localization.

HIPK1, NUCKS1, SMG7 and CRABP2 are all in chromosome 1, the first two genes of the list are adjacent in the cyclic path and the others are scatter through the cycle. Reported changes in chromosome 1 in

cervical cancer include: gains in the 1p region [51-53], increment on the 1q32.1-32.2 genes expression [54], aneusomy of the chromosome [55] among others.

Three genes are in chromosome 2, HNRNPA3, LRPPRC and MTHFD2. There are several changes in chromosome 2 related to cervical cancer, for example reduced expression of genes in 2p has been reported [56], it has also been reported that deletions of the 2q33-q37 are common in cervical carcinoma [57] as well as loss of heterozygosity at 2q35-q37.1 [58].

COX17, RNP1, MAP4, and SMC4 (separated by three genes from the group), and ATP2C (adjacent to SMC4) are all in chromosome 3. Changes in chromosome 3 have been extensively reported for cervical cancer. Gain of chromosome 3q has been reported in pre-cancer and cancer of the cervix (these are some of the reports: [59-62]) while loss of 3p12-p14 has also been observed [63] and loss of heterozygosity on chromosome 3p has been also reported in this cancer [64].

C5orf22 and PIK3R1 are both in chromosome 5. Chromosome 5 is known to have alterations in cervical cancer [62, 66-67]. BUD31 and PRSS2 belong to chromosomes 7, there are known changes of this chromosome in cervical cancer [68-70]. SNTG1 and RAD21 are in chromosome 8, examples of reported changes in this chromosome can be found in: [71-74]. Genes SSX1 and GRIA3 are both in X chromosome. Examples of the association of changes in chromosome X in cervical cancer can be found in [75-77]. Genes PFDN4, CASP1, PLCE1, ICOSLG, GADD45G, SMARCA2, and TSPO are located in different chromosomes, and there are reports for changes in each one of these chromosomes in cervix cancer, for examples the reader is refer to: [9,62,78-91].

The results suggest that the chromosomal gains and losses known for cervical cancer could include bigger regions. It is clear that true experimental validation is critical to further support the results of the proposed pipeline analysis at this point. It is also important, however, to notice its potential for biological discovery. Every time that a biological pathway is discovered, it basically is a problem of selecting a path by systematically choosing pairs of genes with scientific basis. If a mathematical point of view is adopted, this practice implies that the solution is built heuristically as opposed to optimally. This insight has important implications for the adoption of optimization methods in Medicine and Biology.

4. Conclusions

This work proposes a pipeline analysis based on optimization to facilitate the discovery of genetic signaling paths related to cancer and also could provide information about expanded chromosomal regions that are compromised for cases to be studied. In this instance, the method was applied to cervix cancer. The potential of the proposed method is significant if the detection of a biological pathway is understood as a combinatorial problem similar to the Traveling Salesman Problem, for which an optimal solution exists. If positively verified, this point of view

could also imply that current biological pathways might have room for improvement to fully capture the signal in microarray experiments, and thus open the possibility of further discovery in the understanding –and fight- against cancer.

Acknowledgments

This work was made possible thanks to the National Institutes of Health (NIH) MARC Grant: T36-GM-095335 Bioinformatics Programs at Minority Institutions. It was also partially supported by BioSEI UPRM grant 330103080301and PRLSAMP. The collaboration of Jesús Andrés Rodríguez in the coding of the TSP is gratefully acknowledged.

Author Contributions

E Lorenzo first investigated the use of the TSP as a proxy to signaling paths under the mathematical advice of M Cabrera-Ríos and the biolgocal advice of CE Isaza. She coded the KEGG search routine under the supervision of A Ropelewski. K Camacho Cáceres coded the matlab tool for the first stage, while JA Rodríguez did it for the second stage. J Rosas provided the mathematical treatment of the TSP. M Ortiz-Mojer, L Pérez-Marty, J Irizarry & V González supported the generation of literature-based biological evidence presented in this work. M Cabrera-Ríos and CE Isaza codirect and coadvise the research group.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Fierro AC, Vandenbussche F, Engelen K, Van de Peer Y, Marchal K, Meta Analysis of Gene Expression Data within and Across Species, Curr Genomics, 9:8 (2008) 525-534
2. Owzar K, Barry WT, Jung SH, Statistical Considerations for Analysis of Microarray Experiments, Clin Transl Sci, 4:6 (2011) 466-477
3. Sánchez-Peña ML, Isaza CE, Pérez-Morales J, Rodríguez-Padilla C, Castro JM, Cabrera-Ríos M. Identification of potential biomarkers from microarray experiments using multiple criteria optimization. Cancer Medicine. 2013;2(2):253–65
4. Watts-Oquendo E, Sánchez-Peña M, Isaza CE, Cabrera-Ríos M. Potential colon cancer biomarker search using more than two performance measures in a multiple criteria optimization approach. P R Health Sci J. 2012 Jun;31(2):59–63.
5. Charnes A, Cooper WW, Lewin AY, Seiford LM. Data Envelopment Analysis: Theory , Methodology and Applications. Boston MA, USA: Kluwer AcademicPublishers. 1993
6. "National Human Genome Research Institute," National Institute of Health, 12 June 2012. [Online]. Available: http://www.genome.gov/27530687.

7. R. K. Ahuja, T. L. Magnanti and J. B. Orlin, Network Flows: Theory, Algorithms, and Applications, Prentice Hall, 1993
8. Zhai Y, Kuick R, Nan B, Ota I, Weiss SJ, Trimble CL, Fearon, Cho KR Gene Expression Analysis of Preinvasive and Invasive Cervical Squamous Cell Carcinomas Identifies HOXC10 as a Key Mediator of Invasion. Cancer Res 2007, 67:10163-10172
9. Scotto L, Narayan G, Nandula SV, Arias-Pulido H, Subramaniyam S, Schneider A, Kaufmann AM, Wright JD, Pothuri B, Mansukhani M, Murty1 VV.,Identification of Copy Number Gain and Overexpressed Genes on Chromosome Arm 20q by an Integrative Genomic Approach in Cervical Cancer: Potential Role in Progression. GENES, CHROMOSOMES & CANCER 2008, 47:755–765
10. Calmon MF, Rodrigues RV, Kaneto CM, Moura RP, Silva SD, Mota LD, Pinheiro DG, Torres C, de Carvalho AF, Cury PM, Nunes FD, Nishimoto IN, Soares FA, da Silva AM, Kowalski LP, Brentani H, Zanelli CF, Silva WA Jr, Rahal P, Tajara EH, Carraro DM, Camargo AA, Valentini SR Epigenetic silencing of CRABP2 and MX1 in head and neck tumors. Neoplasia 2009, 11:1329-1339.
11. Geiger T, Madden SF, Gallagher WM, Cox J, Mann M Proteomic portrait of human breast cancer progression identifies novel prognostic markers. Cancer Res. 2012, 72:2428-2439.
12. Williams SJ, Gotley DC, Antalis TM Human trypsinogen in colorectal cancer. Int J Cancer 2001, 93:67-73.
13. Rajkumar T, Vijayalakshmi N, Gopal G, Sabitha K, Shirley S, Raja UM, Ramakrishnan SA Identification and validation of genes involved in gastric tumorigenesis. Cancer Cell Int. 2010, 10:45.
14. Chen H, Yuan Y, Zhang C, Luo A, Ding F, Ma J, Yang S, Tian Y, Tong T, Zhan Q, Liu Z Involvement of S100A14 Protein in Cell Invasion by Affecting Expression and Function of Matrix Metalloproteinase (MMP)-2 via p53-dependent Transcriptional Regulation. J Biol Chem. 2012, 287:17109-17119.
15. Sapkota D, Bruland O, Costea DE, Haugen H, Vasstrand EN, Ibrahim SO S100A14 regulates the invasive potential of oral squamous cell carcinoma derived cell-lines in vitro by modulating expression of matrix metalloproteinases, MMP1 and MMP9. Eur J Cancer 2011, 47:600-610.
16. Klubo-Gwiezdzinska J, Jensen K, Bauer A, Patel A, Costello J, Burman K, Wartofsky L, Hardwick MJ, Vasko VV The expression of translocator protein in human thyroid cancer and its role in the response of thyroid cancer cells to oxidative stress. J Endocrinol. 2012 May 29. [Epub ahead of print]
17. Mukherjee S, Das SK Translocator protein (TSPO) in breast cancer. Curr Mol Med. 2012, 12:443-457.
18. Cheung LW, Hennessy BT, Li J, Yu S, Myers AP, Djordjevic B, Lu Y, Stemke-Hale K, Dyer MD, Zhang F, Ju Z, Cantley LC, Scherer SE, Liang H, Lu KH, Broaddus RR, Mills GB High Frequency of PIK3R1 and PIK3R2 Mutations in Endometrial Cancer Elucidates a Novel Mechanism for Regulation of PTEN Protein Stability. Cancer Discov. 2011, 1:170-185.
19. Nowakowska-Zajdel E, Mazurek U, Ziółko E, Niedworok E, Fatyga E, Kokot T, Muc-Wierzgoń M Analysis of expression profile of gene encoding proteins of signal cascades activated by insulin-like growth factors in colorectal cancer. Int J Immunopathol Pharmacol. 2011, 24:781-787.
20. Chen X, Wu J, Lu H, Huang O, Shen K Measuring β-tubulin III, Bcl-2, and ERCC1 improves pathological complete remission predictive accuracy in breast cancer. Cancer Sci. 2012, 103:262-268.

21. Cucchiarelli V, Hiser L, Smith H, Frankfurter A, Spano A, Correia JJ, Lobert S Beta-tubulin isotype classes II and V expression patterns in nonsmall cell lung carcinomas. Cell Motil Cytoskeleton 2008, 65:675-685.
22. Arany I, Ember IA, Tyring SK All-trans-retinoic acid activates caspase-1 in a dose-dependent manner in cervical squamous carcinoma cells. Anticancer Res. 2003, 23:471-473.
23. Nakayama K, Nakayama N, Wang TL, Shih IeM NAC-1 controls cell growth and survival by repressing transcription of Gadd45GIP1, a candidate tumor suppressor. Cancer Res. 2007, 67:8058-8064.
24. Smith HA, Cronk RJ, Lang JM, McNeel DG Expression and immunotherapeutic targeting of the SSX family of cancer-testis antigens in prostate cancer. Cancer Res. 2011, 71:6785-6795.
25. Van Duin M, Broyl A, de Knegt Y, Goldschmidt H, Richardson PG, Hop WC, Van der Holt B, Joseph-Pietras D, Mulligan G, Neuwirth R, Sahota SS, Sonneveld P Cancer testis antigens in newly diagnosed and relapse multiple myeloma: prognostic markers and potential targets for immunotherapy. Haematologica 2011, 96:1662-1669.
26. Fu T, He Q, Sharma P The ICOS/ICOSL pathway is required for optimal antitumor responses mediated by anti-CTLA-4 therapy. Cancer Res. 2011, 71:5445-5454.
27. Tjomsland V, Spångeus A, Sandström P, Borch K, Messmer D, Larsson M Semi mature blood dendritic cells exist in patients with ductal pancreatic adenocarcinoma owing to inflammatory factors released from the tumor. PLoS One 2010, 5:e13441.
28. Ripka S, Riedel J, Neesse A, Griesmann H, Buchholz M, Ellenrieder V, Moeller F, Barth P, Gress TM, Michl P Glutamate receptor GRIA3--target of CUX1 and mediator of tumor progression in pancreatic cancer. Neoplasia 2010, 12:659-667.
29. Sun A, Tawfik O, Gayed B, Thrasher JB, Hoestje S, Li C, Li B Aberrant expression of SWI/SNF catalytic subunits BRG1/BRM is associated with tumor development and increased invasiveness in prostate cancers. Prostate 2007, 67:203-213.
30. Moloney FJ, Lyons JG, Bock VL, Huang XX, Bugeja MJ, Halliday GM Hotspot mutation of Brahma in non-melanoma skin cancer. J Invest Dermatol. 2009, 129:1012-1015.
31. Mougeot JL, Bahrani-Mougeot FK, Lockhart PB, Brennan MT Microarray analyses of oral punch biopsies from acute myeloid leukemia (AML) patients treated with chemotherapy. Oral Surg Oral Med Oral Pathol Oral Radiol Endod. 2011, 112: 446-452
32. Aikawa Y, Nguyen LA, Isono K, Takakura N, Tagata Y, Schmitz ML, Koseki H, Kitabayashi I Roles of HIPK1 and HIPK2 in AML1- and p300-dependent transcription, hematopoiesis and blood vessel formation. EMBO J. 2006, 25:3955-3965.
33. Tian T, Ikeda JI, Wang Y, Mamat S, Luo W, Aozasa K, Morii E Role of leucine-rich pentatricopeptide repeat motif-containing protein (LRPPRC) for anti-apoptosis and tumourigenesis in cancers. Eur J Cancer 2012 Feb 10. [Epub ahead of print]
34. Ziółkowski P, Gamian E, Osiecka B, Zougman A, Wiśniewski JR Immunohistochemical and proteomic evaluation of nuclear ubiquitous casein and cyclin-dependent kinases substrate in invasive ductal carcinoma of the breast. J Biomed Biotechnol. 2009, 2009:919645.

35. Miyoshi N, Ishii H, Mimori K, Nishida N, Tokuoka M, Akita H, Sekimoto M, Doki Y, Mori M Abnormal expression of PFDN4 in colorectal cancer: a novel marker for prognosis. Ann Surg Oncol. 2010, 17:3030-3036.
36. Boukakis G, Patrinou-Georgoula M, Lekarakou M, Valavanis C, Guialis A Deregulated expression of hnRNP A/B proteins in human non-small cell lung cancer: parallel assessment of protein and mRNA levels in paired tumour/non-tumour tissues. BMC Cancer 2010, 10:434.
37. Atienza JM, Roth RB, Rosette C, Smylie KJ, Kammerer S, Rehbock J, Ekblom J, Denissenko MF Suppression of RAD21 gene expression decreases cell growth and enhances cytotoxicity of etoposide and bleomycin in human breast cancer cells. Mol Cancer Ther. 2005, 4:361-368.
38. Shimizu S, Suzukawa K, Kodera T, Nagasawa T, Abe T, Taniwaki M, Yagasaki F, Tanaka H, Fujisawa S, Johansson B, Ahlgren T, Yokota J, Morishita K Identification of breakpoint cluster regions at 1p36.3 and 3q21 in hematologic malignancies with t(1;3)(p36;q21). Genes Chromosomes Cancer 2000, 27:229-238.
39. Andrew AS, Gui J, Sanderson AC, Mason RA, Morlock EV, Schned AR, Kelsey KT, Marsit CJ, Moore JH, Karagas MR Bladder cancer SNP panel predicts susceptibility and survival. Hum Genet. 2009, 125:527-539.
40. Xu X, Qiao M, Zhang Y, Jiang Y, Wei P, Yao J, Gu B, Wang Y, Lu J, Wang Z, Tang Z, Sun Y, Wu W, Shi Q Quantitative proteomics study of breast cancer cell lines isolated from a single patient: discovery of TIMM17A as a marker for breast cancer. Proteomics 2010, 10:1374-1390.
41. Suzuki C, Daigo Y, Kikuchi T, Katagiri T, Nakamura Y Identification of COX17 as a therapeutic target for non-small cell lung cancer. Cancer Res. 2003, 63:7038-7041.
42. Grice DM, Vetter I, Faddy HM, Kenny PA, Roberts-Thomson SJ, Monteith GR Golgi calcium pump secretory pathway calcium ATPase 1 (SPCA1) is a key regulator of insulin-like growth factor receptor (IGF1R) processing in the basal-like breast cancer cell line MDA-MB-231. J Biol Chem. 2010, 285:37458-3766.
43. Wilting SM, de Wilde J, Meijer CJ, Berkhof J, Yi Y, van Wieringen WN, Braakhuis BJ, Meijer GA, Ylstra B, Snijders PJ, Steenbergen RD Integrated genomic and transcriptional profiling identifies chromosomal loci with altered gene expression in cervical cancer. Genes Chromosomes Cancer 2008, 47:890-8905.
44. Wang M, Zhang R, He J, Qiu L, Li J, Wang Y, Sun M, Yang Y, Wang J, Yang J, Qian J, Jin L, Ma H, Wei Q, Zhou X Potentially functional variants of PLCE1 identified by GWASs contribute to gastric adenocarcinoma susceptibility in an eastern Chinese population. PLoS One 2012, 7:e31932.
45. Danielsen SA, Cekaite L, Ågesen TH, Sveen A, Nesbakken A, Thiis-Evensen E, Skotheim RI, Lind GE, Lothe RA Phospholipase C isozymes are deregulated in colorectal cancer--insights gained from gene set enrichment analysis of the transcriptome. PLoS One 2011, 6:e24419.
46. Zhai Y, Kuick R, Nan B, Ota I, Weiss SJ, Trimble CL, Fearon, Cho KR Gene Expression Analysis of Preinvasive and Invasive Cervical Squamous Cell Carcinomas Identifies HOXC10 as a Key Mediator of Invasion. Cancer Res 2007, 67:10163-10172.

47. Chang H, Jeung HC, Jung JJ, Kim TS, Rha SY, Chung HC Identification of genes associated with chemosensitivity to SAHA/taxane combination treatment in taxane-resistant breast cancer cells. Breast Cancer Res Treat. 2011, 125:55-63.
48. Kulawiec M, Safina A, Desouki MM, Still I, Matsui S, Bakin A, Singh KK Tumorigenic transformation of human breast epithelial cells induced by mitochondrial DNA depletion. Cancer Biol Ther. 2008, 7:1732-1743.
49. KEGG: Kyoto Encyclopedia of Genes and Genomes http://www.genome.jp/kegg/
50. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics 1998, 14:656-664.
51. Wang J, Tai LS, Tzang CH, Fong WF, Guan XY, Yang M 1p31, 7q21 and 18q21 chromosomal aberrations and candidate genes in acquired vinblastine resistance of human cervical carcinoma KB cells. Oncol Rep. 2008 19:1155-1164.
52. Lee M, Nam ES, Jung SH, Kim SY, Lee SJ, Yoon JH, Lee NW, Jeon S, Choi JS, Cho CH, Moon Y, Chung YJ, Kwon Y 1p36.22 region containing PGD gene is frequently gained in human cervical cancer. J Obstet Gynaecol Res. 2014, 40:545-553.
53. Wilting SM, Steenbergen RD, Tijssen M, van Wieringen WN, Helmerhorst TJ, van Kemenade FJ, Bleeker MC, van de Wiel MA, Carvalho B, Meijer GA, Ylstra B, Meijer CJ, Snijders PJ Chromosomal signatures of a subset of high-grade premalignant cervical lesions closely resemble invasive carcinomas. Cancer Res. 2009, 69:647-655.
54. Wilting SM, de Wilde J, Meijer CJ, Berkhof J, Yi Y, van Wieringen WN, Braakhuis BJ, Meijer GA, Ylstra B, Snijders PJ, Steenbergen RD Integrated genomic and transcriptional profiling identifies chromosomal loci with altered gene expression in cervical cancer. Genes Chromosomes Cancer 2008, 47:890-905.
55. Cortés-Gutiérrez EI1, Dávila-Rodríguez MI, Muraira-Rodríguez M, Said-Fernández S, Cerda-Flores RM Association between the stages of cervical cancer and chromosome 1 aneusomy. Cancer Genet Cytogenet. 2005, 159:44-47.
56. Kozlowski L, Filipowski T, Rucinska M, Pepinski W, Janica J, Skawronska M, Poznanski J, Wojtukiewicz MZ Loss of heterozygosity on chromosomes 2p, 3p, 18q21.3 and 11p15.5 as a poor prognostic factor in stage II and III (FIGO) cervical cancer treated by radiotherapy. Neoplasma. 2006, 53:440-443
57. Rao PH1, Arias-Pulido H, Lu XY, Harris CP, Vargas H, Zhang FF, Narayan G, Schneider A, Terry MB, Murty VV Chromosomal amplifications, 3q gain and deletions of 2q33-q37 are the frequent genetic changes in cervical carcinoma. BMC Cancer. 2004 Feb 6;4:5.
58. Edelmann J, Richter K, Hänel C, Hering S, Horn LC X chromosomal and autosomal loss of heterozygosity and microsatellite instability in human cervical carcinoma. Int J Gynecol Cancer. 2006, 16(3):1248-1253.
59. Thomas LK, Bermejo JL, Vinokurova S, Jensen K, Bierkens M, Steenbergen R, Bergmann M, von Knebel Doeberitz M, Reuschenbach M Chromosomal gains and losses in human papillomavirus-associated

neoplasia of the lower genital tract - a systematic review and meta-analysis. Eur J Cancer. 2014, 50:85-98.

60. Wright TC, Compagno J, Romano P, Grazioli V, Verma Y, Kershnar E, Tafas T, Kilpatrick MW Amplification of the 3q chromosomal region as a specific marker in cervical cancer. Am J Obstet Gynecol. 2015 Feb 4. pii: S0002-9378(15)00115-5 [Epub ahead of print].
61. Policht FA, Song M, Sitailo S, O'Hare A, Ashfaq R, Muller CY, Morrison LE, King W, Sokolova IA Analysis of genetic copy number changes in cervical disease progression. BMC Cancer 2010, 10:432.
62. Luhn P, Houldsworth J, Cahill L, Schiffman M, Castle PE, Zuna RE, Dunn ST, Gold MA, Walker J, Wentzensen N Chromosomal gains measured in cytology samples from women with abnormal cervical cancer screening results. Gynecol Oncol. 2013, 130:595-600.
63. Lando M, Wilting SM, Snipstad K, Clancy T, Bierkens M, Aarnes EK, Holden M, Stokke T, Sundfør K, Holm R, Kristensen GB, Steenbergen RD, Lyng H Identification of eight candidate target genes of the recurrent 3p12-p14 loss in cervical cancer by integrative genomic profiling. J Pathol. 2013, 230:59-69.
64. Kozlowski L, Filipowski T, Rucinska M, Pepinski W, Janica J, Skawronska M, Poznanski J, Wojtukiewicz MZ Loss of heterozygosity on chromosomes 2p, 3p, 18q21.3 and 11p15.5 as a poor prognostic factor in stage II and III (FIGO) cervical cancer treated by radiotherapy. Neoplasma. 2006, 53:440-443.
65. Johnson LG, Schwartz SM, Malkki M, Du Q, Petersdorf EW, Galloway DA, Madeleine MM Risk of cervical cancer associated with allergies and polymorphisms in genes in the chromosome 5 cytokine cluster. Cancer Epidemiol Biomarkers Prev. 2011, 20:199-207.
66. Scotto L, Narayan G, Nandula SV, Subramaniyam S, Kaufmann AM, Wright JD, Pothuri B, Mansukhani M, Schneider A, Arias-Pulido H, Murty VV Integrative genomics analysis of chromosome 5p gain in cervical cancer reveals target over-expressed genes, including Drosha. Mol Cancer. 2008, 7:58.
67. Huang FY, Chiu PM, Tam KF, Kwok YK, Lau ET, Tang MH, Ng TY, Liu VW, Cheung AN, Ngan HY Semi-quantitative fluorescent PCR analysis identifies PRKAA1 on chromosome 5 as a potential candidate cancer gene of cervical cancer. Gynecol Oncol. 2006, 103:219-225.
68. Schrevel M, Gorter A, Kolkman-Uljee SM, Trimbos JB, Fleuren GJ, Jordanova ES Molecular mechanisms of epidermal growth factor receptor overexpression in patients with cervical cancer. Mod Pathol. 2011, 24:720-728.
69. Thein A, Trková M, Fox M, Parrington J The application of comparative genomic hybridization to previously karyotyped cervical cancer cell lines. Cancer Genet Cytogenet. 2000, 116:59-65.
70. Mian C, Bancher D, Kohlberger P, Kainz C, Haitel A, Czerwenka K, Stani J, Breitenecker G, Wiener H Fluorescence in situ hybridization in cervical smears: detection of numerical aberrations of chromosomes 7, 3, and X and relationship to HPV infection. Gynecol Oncol. 1999, 75:41-46.
71. Ferber MJ, Eilers P, Schuuring E, Fenton JA, Fleuren GJ, Kenter G, Szuhai K, Smith DI, Raap AK, Brink AA Positioning of cervical carcinoma and Burkitt lymphoma translocation breakpoints with

respect to the human papillomavirus integration cluster in FRA8C at 8q24.13. Cancer Genet Cytogenet. 2004, 154:1-9.

72. Sokolova I, Algeciras-Schimnich A, Song M, Sitailo S, Policht F, Kipp BR, Voss JS, Halling KC, Ruth A, King W, Underwood D, Brainard J, Morrison L Chromosomal biomarkers for detection of human papillomavirus associated genomic instability in epithelial cells of cervical cytology specimens. J Mol Diagn. 2007, 9:604-611.
73. Bhattacharya N, Singh RK, Mondal S, Roy A, Mondal R, Roychowdhury S, Panda CK Analysis of molecular alterations in chromosome 8 associated with the development of uterine cervical carcinoma of Indian patients. Gynecol Oncol. 2004, 95:352-362.
74. Seng TJ, Low JS, Li H, Cui Y, Goh HK, Wong ML, Srivastava G, Sidransky D, Califano J, Steenbergen RD, Rha SY, Tan J, Hsieh WS, Ambinder RF, Lin X, Chan AT, Tao Q The major 8p22 tumor suppressor DLC1 is frequently silenced by methylation in both endemic and sporadic nasopharyngeal, esophageal, and cervical carcinomas, and inhibits tumor cell colony formation. Oncogene. 2007, 26:934-944.
75. Dellas A, Torhorst J, Gaudenz R, Mihatsch MJ, Moch H DNA copy number changes in cervical adenocarcinoma. Clin Cancer Res. 2003, 9:2985-2991
76. Marzano R, Corrado G, Merola R, Sbiroli C, Guadagni F, Vizza E, Del Nonno F, Carosi M, Galati M M, Sperduti I, Cianciulli AM Analysis of chromosomes 3, 7, X and the EGFR gene in uterine cervical cancer progression. Eur J Cancer. 2004, 40:1624-1629.
77. Hopman AH, Smedts F, Dignef W, Ummelen M, Sonke G, Mravunac M, Vooijs GP, Speel EJ, Ramaekers FC Transition of high-grade cervical intraepithelial neoplasia to micro-invasive carcinoma is characterized by integration of HPV 16/18 and numerical chromosome abnormalities. J Pathol. 2004, 202:23-33.
78. Tabach Y, Kogan-Sakin I, Buganim Y, Solomon H, Goldfinger N, Hovland R, Ke XS, Oyan AM, Kalland KH, Rotter V, Domany E Amplification of the 20q chromosomal arm occurs early in tumorigenic transformation and may initiate cancer. PLoS One. 2011, 6(1):e14632.
79. Lorenzetto E, Brenca M, Boeri M, Verri C, Piccinin E, Gasparini P, Facchinetti F, Rossi S, Salvatore G, Massimino M, Sozzi G, Maestro R, Modena P YAP1 acts as oncogenic target of 11q22 amplification in multiple cancer subtypes. Oncotarget. 2014, 5:2608-2621.
80. Kehrmann A, Truong H, Repenning A, Boger R, Klein-Hitpass L, Pascheberg U, Beckmann A, Opalka B, Kleine-Lowinski K Complementation of non-tumorigenicity of HPV18-positive cervical carcinoma cells involves differential mRNA expression of cellular genes including potential tumor suppressor genes on chromosome 11q13. Cancer Genet. 2013, 206:279-292.
81. Mazumder Indra D, Mitra S, Roy A, Mondal RK, Basu PS, Roychoudhury S, Chakravarty R, Panda CK Alterations of ATM and CADM1 in chromosomal 11q22.3-23.2 region are associated with the development of invasive cervical carcinoma. Hum Genet. 2011, 130:735-748.

82. Huang KF, Lee WY, Huang SC, Lin YS, Kang CY, Liou CP, Tzeng CC Chromosomal gain of 3q and loss of 11q often associated with nodal metastasis in early stage cervical squamous cell carcinoma. J Formos Med Assoc. 2007, 106:894-902.
83. Rizvi MM, Alam MS, Mehdi SJ, Ali A, Batra S Allelic loss of 10q23.3, the PTEN gene locus in cervical carcinoma from Northern Indian population. Pathol Oncol Res. 2012, 18:309-313.
84. Wang S, Li Y, Han F, Hu J, Yue L, Yu Y, Zhang Y, He J, Zheng H, Shi S, Fu X, Wu H Identification and characterization of MARVELD1, a novel nuclear protein that is down-regulated in multiple cancers and silenced by DNA methylation. Cancer Lett. 2009, 282:77-86.
85. Poignée M, Backsch C, Beer K, Jansen L, Wagenbach N, Stanbridge EJ, Kirchmayr R, Schneider A, Dürst M Evidence for a putative senescence gene locus within the chromosomal region 10p14-p15. Cancer Res. 2001, 61:7118-7121.
86. Amiel A, Kolodizner T, Fishman A, Gaber E, Klein Z, Beyth Y, Fejgin MD Replication pattern of the p53 and 21q22 loci in the premalignant and malignant stages of carcinoma of the cervix. Cancer. 1998, 83:1966-1971.
87. Simpson S, Woodworth CD, DiPaolo JA Altered expression of Erg and Ets-2 transcription factors is associated with genetic changes at 21q22.2-22.3 in immortal and cervical carcinoma cell lines. Oncogene. 1997, 14:2149-2157.
88. Lennerz JK, Perry A, Mills JC, Huettner PC, Pfeifer JD Mucoepidermoid carcinoma of the cervix: another tumor with the t(11;19)-associated CRTC1-MAML2 gene fusion. Am J Surg Pathol. 2009, 33:835-843.
89. Miyai K, Furugen Y, Matsumoto T, Iwabuchi K, Hirose S, Kinoshita K, Fujii H Loss of heterozygosity analysis in uterine cervical adenocarcinoma. Gynecol Oncol. 2004, 94:115-120.
90. Engelmark MT, Ivansson EL, Magnusson JJ, Gustavsson IM, Wyöni PI, Ingman M, Magnusson PK, Gyllensten UB Polymorphisms in 9q32 and TSCOT are linked to cervical cancer in affected sib-pairs with high mean age at diagnosis. Hum Genet. 2008, 123:437-443.
91. Jee KJ, Kim YT, Kim KR, Aalto Y, Knuutila S Amplification at 9p in cervical carcinoma by comparative genomic hybridization. Anal Cell Pathol. 2001, 22:159-163.

Appendix A

Stage 1 of the proposed analysis pipeline involves the use of Multiple Criteria Optimization (MCO). MCO does not require the adjustment of parameters by the users to detect differentially expressed genes, thus it preserves the objectivity and the consistent convergence of the analysis. As a comparative example, a tool like the volcano plot Figure A1, would require for the user to define different cutoff values to select different genes, thereby biasing the analysis.

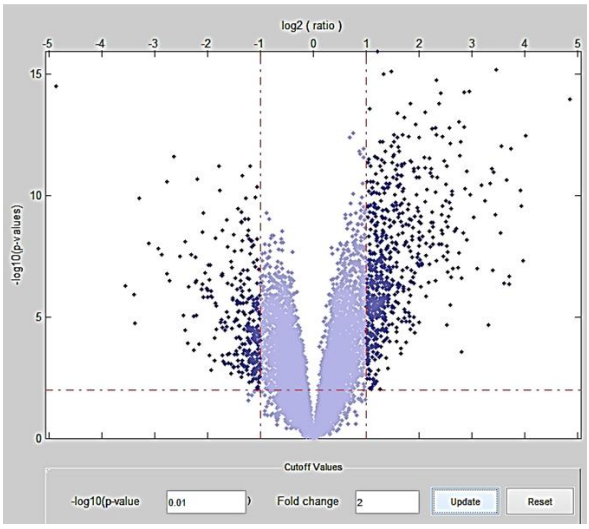

Figure A1. A volcano plot. Two cutoff values must be set by the user to decide upon the significance of different genes: one for fold change (x-axis) and one for p-value (y-axis).

Besides the evident result of choosing different sets of genes when the analyst picks different sets of cutoff values, this decision also greatly affects the number of genes that are deemed to change their expression significantly, as shown in Table A1.

Table A1: An example of how the number of genes deemed significant changes when choosing different cutoff values for fold change and p-value.

P-value	Fold change	Differentially expressed genes (number)	Number of genes Overexpressed	Number of genes Underexpressed
10^{-2}	2	934	645	289
10^{-2}	8	29	23	6

10^{-2}	24	2	1	1
10^{-7}	2	649	516	133
10^{-7}	8	27	22	5
10^{-7}	24	2	1	1
10^{-12}	2	130	121	9
10^{-12}	8	12	11	1
10^{-12}	24	2	1	1

Appendix B

A Python script was written to gather information from KEGG. This is provided in its three parts below.

Part 1

```
#!/opt/python2.6/bin/python

# Import Comma Separated Value Library ...
import csv
import sys
import urllib2

# Open file to be read
ifile  = open('results.csv', "rb")

# Create the reader object (in order to read from CSV file)
reader = csv.reader(ifile)

# Create file to be output
ofile  = open('details.csv', 'w')

# Create writer in order to write to output file
writer = csv.writer(ofile, delimiter=',', quotechar='"', quoting=csv.QUOTE_MINIMAL)

# Function in charge of extracting each genes in the database file
def detailPathwayExtractor():
    for row in reader:
        if (len(row)==1):
            path=row[1]
```

```
                try:
                    url_to_go_to = "http://rest.kegg.jp/get/" + path
                    print url_to_go_to
                    handle = urllib2.urlopen(url_to_go_to)
#read contenent
                    content = handle.read()
                    for row_of_file in content.split("\n"):
                        if row_of_file.split() != []:
                            print row_of_file.split()
                            writer.writerow(row_of_file.split())

# Run the pathway extractor function
                except IOError:
                    print "can't open file"

detailPathwayExtractor()

# Close both opened files
ifile.close()
ofile.close()
```

Part2

```
#!/opt/python2.6/bin/python


# Import Comma Separated Value Library ...

import csv

import sys

import urllib2


# Open file to be read

ifile  = open('test.csv', "rb")


# Create the reader object (in order to read from CSV file)

reader = csv.reader(ifile)
```

```
# Create file to be output
ofile  = open('results.csv', 'w')


# Create writer in order to write to output file
writer = csv.writer(ofile, delimiter=',', quotechar='"', quoting=csv.QUOTE_MINIMAL)


# Function in charge of extracting each genes in the database file
def pathwayExtractor():
      for row in reader:
            if (len(row)==4):
                  HSA=row[3]
                  try:
                        url_to_go_to = "http://rest.kegg.jp/link/pathway/hsa:" + HSA
                        print url_to_go_to
                        handle = urllib2.urlopen(url_to_go_to)
#read contenent
                        content = handle.read()
                        for row_of_file in content.split("\n"):
                              if row_of_file.split() != []:
                                    print row_of_file.split()
                                    writer.writerow(row_of_file.split())


# Run the pathway extractor function
                  except IOError:
                        print "can't open file"
```

```
pathwayExtractor()



# Close both opened files

ifile.close()

ofile.close()
```

Part 3

```
#!/opt/python2.6/bin/python



# Import Comma Separated Value Library



import csv

import sys

import urllib2




# Open file to be read

ifile  = open('results.csv', "rb")



# Create the reader object (in order to read from CSV file)

reader = csv.reader(ifile)



# Create file to be output

ofile  = open('details.csv', 'w')



# Create writer in order to write to output file

writer = csv.writer(ofile, delimiter=',', quotechar='"', quoting=csv.QUOTE_MINIMAL)
```

```
# Function in charge of extracting each genes in the database file
def detailPathwayExtractor():
    for row in reader:
        if (len(row)==1):
            path=row[1]
            try:
                url_to_go_to = "http://rest.kegg.jp/get/" + path
                print url_to_go_to
                handle = urllib2.urlopen(url_to_go_to)
#read contenent
                content = handle.read()
                for row_of_file in content.split("\n"):
                    if row_of_file.split() != []:
                        print row_of_file.split()
                        writer.writerow(row_of_file.split())



# Run the pathway extractor function
            except IOError:
                print "can't open file"
detailPathwayExtractor()
# Close both opened files
ifile.close()
ofile.close()
```