Commonalities in Genetic Signatures and Signaling Pathways in Neurological Disorders

by

Nicole Janice Ortiz-Rodriguez

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

2017

Approved by:

_____	_____
Jaime Seguel, PhD Member, Graduate Committee	Date
_____	_____
Mayra Méndez, PhD Member, Graduate Committee	Date
_____	_____
Mauricio Cabrera-Rios, PhD President, Graduate Committee	Date
_____	_____
Celia Colón-Rivera, PhD Representative of Graduate Studies	Date
_____	_____
Viviana Cesaní, PhD Chairperson of the Department	Date

Abstract

Mathematical optimization is used to detect potential genetic differentially expressed genes and propose probable signaling structures common to Alzheimer (AD), Parkinson (PD) and Huntington's Disease (HD). The characterizations of these three affections have been elusive in the literature, although their impact in society is projected to increase in the next decades worldwide. There are studies in the literature for each individual illness, but this work novelty is the study and comparison of all three disorders together. The search for the most correlated path among differentially expressed genes is carried out using network optimization formulations: Travelling Salesman Problem (TSP) and Minimum Spanning Tree (MST). For both, a set of differentially expressed genes is identified previously through multiple criteria optimization; a correlation coefficient is used to link every pair of genes.A cost model was developed with the information of the high yearly cost of the neurological diseases discussed in this work.

Resumen

Utilización de optimización matemática para detectar genes potenciales con diferenciación genética de expresión y proponer estructuras probables de señalización comunes en la enfermedad de Alzheimer (AD), Parkinson (PD) y Huntington (HD). Las caracterizaciones de estas afecciones son escasas en la literatura, su impacto en la sociedad proyecta aumentar en las próximas décadas mundialmente. Hay estudios para cada enfermedad individual, sin embargo la novedad es el estudio y la comparación de los tres trastornos juntos. La búsqueda de la vía más correlacionada entre los genes expresados diferencialmente se realiza utilizando formulaciones de optimización de redes: TSP ("Traveling Salesman Problem") y MST ("Minimum Spanning Tree"). Para ambos, un conjunto de genes expresados diferencialmente se identifica previamente mediante la optimización de criterios múltiples; un coeficiente de correlación se utiliza para vincular cada par de genes. Con la información del costo anual de las enfermedades neurológicas se desarrolló un modelo de costo.

Acknowledgements

This work was possible thanks to NIH MARC Assisting Bioinformatics Efforts at Minority Schools project 2T36GM008789.

I am really grateful with all the support and knowledge received from my advisor, professor, mentor and friend, the mastermind behind this work, Dr. Mauricio Cabrera-Rios. Thanks for teaching me all I needed to complete this work, for mentoring me throughout this process and give me the opportunity to work with you. Without your help, this could not be possible.

I also want to thank, Dra. Clara Isaza, for all your guidance, patience and good vibes. Your help and background knowledge was critical to the success of this work, you are a true asset in our team.

Thanks to my committee members, Dra. Mayra Mendez and Dr. Jaime Seguel for helping to make this work possible, with all your advice, feedback and knowledge I learned a lot. Your inputs and help were critical in the development of this research.

I want to thanks two really good undergraduate students, Janice Garcia and Rebecca Betances, for all the support and biological advice provided.

Table of Contents

Table List

Table of Figures

1. Introduction

This project proposes the study of commonalities among three important neurological conditions: PD, AD and HD. The characterizations of these three conditions have been elusive in the literature, although their impact in society is projected to dramatically increase in the next decades worldwide. In United States, current estimates of people affected with these diseases add up to 5,300,000 for AD, 1,000,000 for PD, and 30,000 for HD [1-3]. In addition, it must also be noted that Caribbean-Americans are 1.5 times as likely to suffer dementia as White Americans [4]. The high incidence of these conditions in Puerto Rico calls for accelerating their understanding and characterization.

There is a large amount of publicly-available data associated to mRNA microarrays and microRNA arrays, as well as several other types of high throughput biological experiments related to AD, PD and HD with the potential to be analyzed in a coordinated manner [5, 6]. Also, the three diseases have been reported as induced by protein aggregation and as neurodegenerative in nature. In PD, one of the characteristics is the aggregation of α-synuclein protein. In AD, Beta amyloids aggregates outside the cells and tau-filaments inside the cells. In HD the aggregation of huntingtin is an important characteristic. In PD and AD it is known that some genetic information could be hereditary in addition to spontaneous mutations that relate to both illnesses. For HD it is known that it is caused by at least one gene and that it is hereditary.

Our research group has specialized in designing analysis strategies based on mathematical optimization that facilitate the simultaneous analysis of multiple experimental data without the manipulation of parameters by the user. In this sense, our strategies offer consistent convergence

to a manageable number of key pieces of information for each disorder that can be correlated with convenience in search for commonalities. Figure 1 is a scheme of the proposed research design for this project. Moving horizontally from left to right, the first stage involves the simultaneous analysis of multiple microarrays (or microRNA arrays) to detect potentially important genes or regulatory molecules. This first stage will be executed with our originally-designed multiple criteria optimization approach. The second stage will use the lists of potentially important genes from the previous stage to determine the optimally correlated circular path among them as a proxy for a biological signaling path. Moving vertically downwards in Figure 1, these two stages will be carried out for all three conditions (AD, PD and HD). In Stage 3, these analyses will then allow establishing commonalities in terms of differentially expressed genes and potential signaling paths. These commonalities could be common genes, similar levels of expression, common pathways, among others. Biological and medical literature will be used to marshal evidence of similar mechanisms among illnesses and to propose mechanisms that have eluded discovery to date.

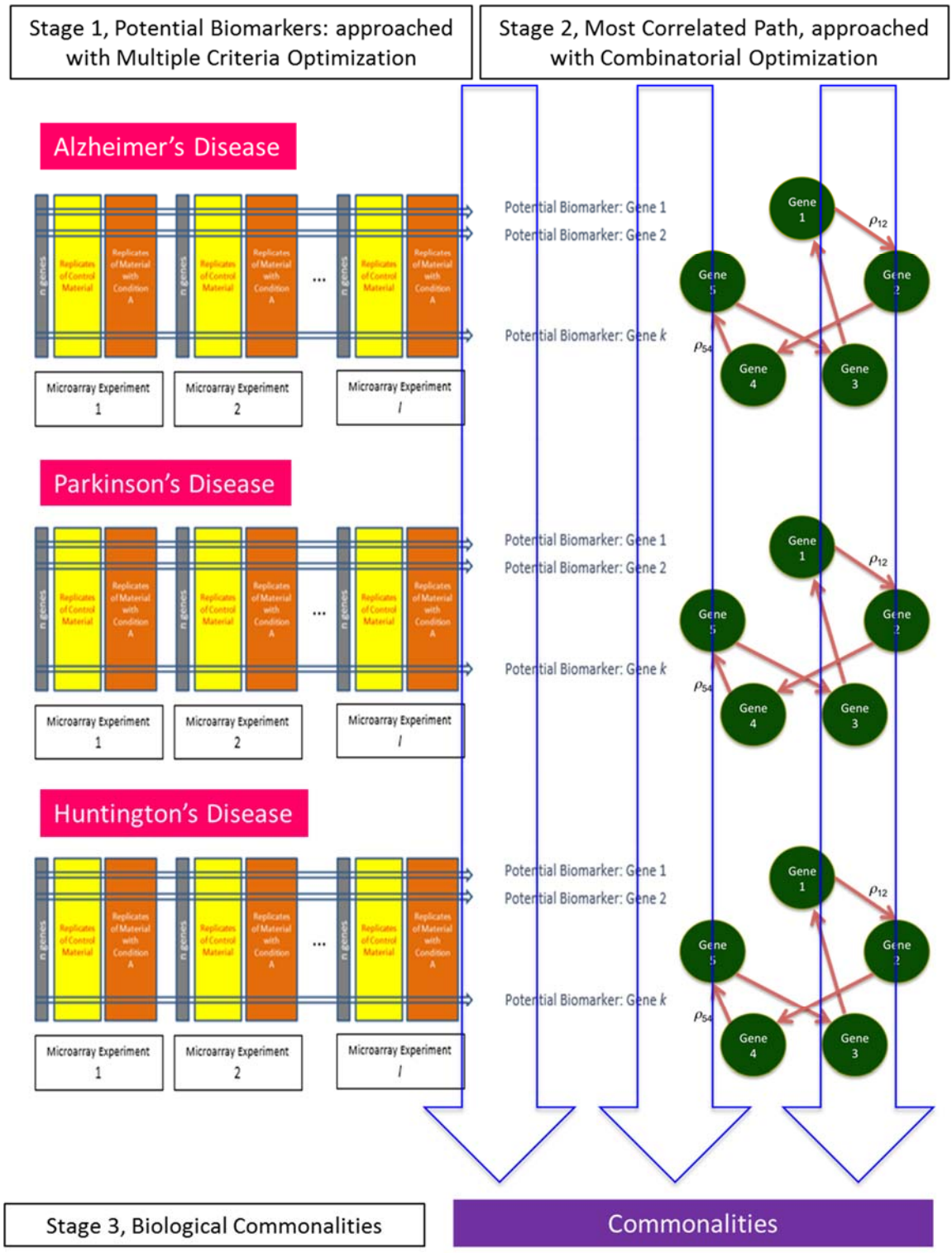


Figure 1: Research Design for Project

This work's goal is to characterize commonalities between the three neurological disorders. Below, the generalities of each illness are discussed individually and then, commonalities between them are presented.

1.1 Parkinson's disease

Parkinson's disease is a neurodegenerative brain disorder that progresses slowly in most people. Most people's symptoms take years to develop, and they live for years with the disease [7].

The person's brain slowly stops producing a neurotransmitter called dopamine. With less and less dopamine, a person has less and less ability to regulate their movements, body and emotions. Dopamine is a chemical that relays messages between the substantia nigra and other parts of the brain to control movements of the human body. Dopamine helps humans to have smooth, coordinated muscle movements. When approximately 60 to 80% of the dopamine-producing cells are damaged, and do not produce enough dopamine, the motor symptoms of Parkinson's disease appear. This process of impairment of brain cells is called neurodegeneration [8].

There are four main symptoms of PD. First symptom is tremor, which means shaking or trembling, they may affect hands, arms, or legs. Other symptoms include stiff muscles, slow movement and/or problems with balance or walking [9]. In time, PD affects muscles all through the body, so it can lead to problems like trouble swallowing or constipation. In the later stages of the disease, a person with PD may have a fixed or blank expression, trouble speaking, and other problems. Some people also lose mental skills (dementia). People usually start to have symptoms between the ages of fifty

and sixty. But sometimes symptoms start earlier [9]. Currently there is not cure for Parkinson's disease.

Parkinson's disease has five stages [10]. The first stage is recognized when people has mild symptoms that generally do not interfere with daily activities. In this stage tremor and other movement symptoms occur on one side of the body only, also some changes in posture, walking and facial expression may be noticed. In the second stage of the disease, the tremor, rigidity and other movement symptoms affect both sides of the body; also walking problems and poor posture may become apparent. Completing day-to-day tasks becomes more difficult and may take longer. In the third stage or mid stage, loss of balance and slowness of movements are hallmarks. Falls are also very common. Symptoms significantly impair activities of daily living such as dressing and eating. The stage four symptoms get severe and become very limiting. The person needs help with activities of daily living and is unable to live alone. The stage five is the most advanced and debilitating stage of Parkinson's disease. Stiffness in the legs may make it impossible to stand or walk, making it a requirement to use a wheelchair or remain confined to a bed. The person may experience hallucinations and delusions [10].

This work is relevant since the analysis of potential genes and signaling pathways related to PD will be performed. The information presented in this work will help to accelerate the discovery of the genetic aspects of the disease to improve diagnosis and prognosis.

1.2 Huntington's disease

Huntington's disease is a fatal genetic disorder that causes the progressive breakdown of nerve cells in the brain. It deteriorates a person's physical and mental abilities during their prime working years and has no cure. HD is known as the quintessential family disease because every child of a parent with HD has a 50/50 chance of carrying the faulty gene. Many describe the symptoms of HD as having PD and AD – simultaneously [11]. Physical symptoms of HD can begin at any age from infancy to old age, but usually begin between 35 and 44 years of age. About 6% of cases start before the age of 21 years; they progress faster and vary slightly. The symptoms include personality changes, mood swings, depression, forgetfulness, impaired judgment, unsteady gait, involuntary movements (chorea), slurred speech, difficulty in swallowing, and significant weight loss [12]. Everyone has the gene that causes HD, but only those that inherit the expansion of the gene will develop HD and perhaps pass it on to each of their children. Every person who inherits the expanded HD gene will eventually develop the disease [12].

The progression of the disease can be roughly divided into three stages. The early stage HD usually includes subtle changes in coordination, perhaps some involuntary movements (chorea), difficulty thinking through problems and often a depressed or irritable mood. The effects of the disease may make the person less able to work at their customary level and less functional in their regular activities at home. In the middle stage, the movement disorder may become more of a problem. Occupational and physical therapists may be needed to help maintain control of voluntary movements and to deal with changes in thinking and reasoning abilities. Diminished speech and difficulty swallowing may require help from a speech language pathologist. Ordinary activities will become harder to do. In the late stage, the person with HD is totally dependent on others for

their care. Choking becomes a major concern. Chorea (neurological disorder characterized by jerky involuntary movements affecting especially the shoulders, hips, and face) may be severe or it may cease. At this stage, the person with HD can no longer walk and will be unable to speak. However, he or she is generally still able to comprehend language and retains an awareness of family and friends. When a person with HD dies, it is typically from complications of the disease, such as choking or infection and not from the disease itself. In all stages of HD, weight loss can be an important complication that can correspond with worsening symptoms and should be countered by adjusting the diet and maintaining appetite [11].

This work will propose a set of important genes in the presence of the disease that could help the acceleration of HD characterization. The work will also evaluate important pathways where the set of genes has already been found relevant in the literature.

1.3 Alzheimer's disease

Alzheimer's disease is a progressive brain ailment that causes memory loss, and eventually destroys the common functions such as thinking, the ability to perform simple tasks, and can – therefore- disrupt human behavior [13][14]. AD is usually diagnosed in people 65 years and older. However, 5 to 10 percent of cases are detected before this age [13]. AD is not easy to detect in people before several stages have already passed. The National Alzheimer's Association and the National Institutes of Health identify seven stages of this disease [14]. The first stage is known as "no impairment", when people do not present any sign of having the disease. The second stage is described as "very mild decline", when people start to forget words or the location of everyday objects, but the disease cannot be detected. 'Mild cognitive decline" (early stage) is the third stage,

when close people and doctors can detect the person's difficulties, for example performing tasks, remembering names, losing objects, among others. The fourth stage is "moderate cognitive decline", when people do not have the ability to perform complex tasks, mental challenges or even paying bills. The next stage is "moderately severe cognitive decline". In this stage people have gaps in memory and thinking, they cannot remember their own phone number or address, they also may need help to choose clothes. The sixth stage is known as "severe cognitive decline", when the person experiments personality changes, can lose awareness of recent experiences, and tend to wander or become lost, among others. The last stage is known as "very severe cognitive decline", in which the person loses the ability to carry a conversation, needs help with most of their daily personal care and the muscles grow rigid.

The brain is comprised of a complex network of nerve cells that controls the body; movement, thinking, learning, memory, senses, as well as all critical involuntary body functions that sustain life. Scientists are not sure how it happens, but somehow AD causes an inability of these cells to perform their individual jobs and eventually die [14].

This work presents a set of relevant information for AD. A set of important genes previously identified by our group was analyzed and the actual pathways were discussed. Also, the genetic information obtained was compared with PD and HD analysis and commonalities are discussed.

2. Background

2.1 Signaling Pathways

A signaling pathway describes a group of molecules in a cell that work together to control one or more cell functions, such as cell division or cell death [15]. It is a series of actions among molecules in a cell that leads to a certain product or change in a cell [16]. After the first molecule in a pathway receives a signal, it activates another molecule. This process is repeated until the last molecule is activated and the cell function is carried out [15]. Signaling pathways have a key role in various functions of a cell, for this reason they are of interest in diseases research. Studying the pathways that were disrupted by the genetic mutations could possibly narrow the search for improving treatments designed to combat neurological diseases development. As described in the following sections of this document, several methods in the literature identify genes of interest and possible pathways that match these genes. According to Rosas [17], 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. In our research group we have advocated the use of network optimization representations to approach the elucidation of proxies for genetic signaling paths by finding the most correlated structure among a series of candidate genes.

2.2 Microarrays

The main goal of a microarray analysis is finding a set of array probes related to genes that have an expression profile with an unusual level of under-expression or over-expression [18]. The DNA microarray is used to determine whether the DNA from a particular individual contains a mutation in genes. The chip consists of a small glass plate encased in plastic. On the surface, each chip contains thousands of short, synthetic, single-stranded DNA sequences, which together add up to the normal gene in question, and to variants (mutations) of that gene that have been found in the human population [18].

The amount of information generated by microarray analysis is particularly suited to certain specialized tasks such as biomarker discovery [19].

2.3 Differential Expression Analysis

Differentially expressed genes (DEG) may provide valuable information regarding the insights of certain diseases. For this reason differential expression analysis is used in this research to analyze if certain genes that significantly change their expression in control and illness-ridden samples affect certain cell functions or are related to pathways that could lead to the progression of the illnesses in study. Differential expression is a gene expression that responds to signals or triggers, for example: gene regulation and effects of certain hormones on protein biosynthesis [21].

The three postulates of differential gene expression, as presented in [22], are as follows:

1. Every cell nucleus contains the complete genome established in the fertilized egg. In molecular terms, the DNAs of all differentiated cells are identical.
2. The unused genes in differentiated cells are not destroyed or mutated, and they retain the potential for being expressed.

3. Only a small percentage of the genome is expressed in each cell, and a portion of the RNA synthesized in the cell is specific for that cell type.

There are multiple differentially expressed genes identified in the existing literature. Eleven methods for differential expression analysis of RNA-seq data were evaluated and compared in the publication [23]. Nine of them work on the count data directly: DESeq, edgeR, NBPSeq , TSPM, baySeq, EBSeq, NOISeq, SAMseq and ShrinkSeq. The remaining two combine a data transformation with limma for differential expression analysis. Limma is a package for differential expression analysis of data arising from microarray experiments. The package is designed to analyze complex experiments involving comparisons between many RNA targets simultaneously while remaining reasonably easy to use for simple experiments. For the interested reader, other techniques can be found in [24].

2.4 Literature Review

This work uses the data bases from the study "Polyamine pathway contributes to the pathogenesis of Parkinson disease", from Lewandowski, et al., published in 2010 [24]. These databases came from functional MRI, and were used to identify brainstem regions differentially affected and resistant to PD. CBV (Cerebral Blood Volumes) maps of the brainstem were generated with fMRI (Functional Magnetic Resonance Imaging) in five PD patients (mean age = 56.4 y) and five healthy age matched controls (mean age = 56.2 y). The samples were taken from autopsies.

The work mentioned previously [24] presented gene expression-profiling techniques such as microarray to identify molecular pathways contributing to the pathogenesis of PD. In this study their microarray data was re-analyzed to find correlations between the changes for those genes that

changed their expression the most. The study described in [24], pinpointed regions within the same brain structure that are differentially targeted by and resistant to a disease to determine genetic pathways. The methodology proposed in this thesis, based on network optimization methods, is capable of determining an optimal solution, which is novel and expected to differ from otherwise obtained structures. Another important point is that the work analyzed stipulated that one region of the brain was being affected (DMNV) in PD presence and the other region remained unaffected (ION). This is challenged in this work with our results that point to both regions being affected with the presence of the illness.

The study on HD ("Transcriptional modulator H2A histone family, member Y (H2AFY) marks Huntington disease activity in man and mouse") [26] provided data bases of human blood samples on people associated to the disease and control, used as input for HD analysis in this work. Differently from the work discussed previously [24] that stated the PD affects one region of the brain and other regions remain unaffected, this work stated that although HD symptoms reflected preferential neuronal death in specific brain regions, Huntingtin is expressed in almost all tissues and may cause detectable but clinically silent changes in gene expression and biochemistry in blood cells [26]. In this work [26], ninety-nine genes were classified as significantly differentially expressed in patients with HD, with a fold change ≥ 1.5 or ≤ 0.66 and a false discovery rate (FDR) <0.00002 based on 50,000 permutations of the dataset. These include the transcriptional modulator H2AFY, which was 1.6-fold overexpressed in cellular blood of patients with HD. A FDR is one way of conceptualizing the rate of type I errors in null hypothesis when conducting multiple comparisons. A type 1 error occurs

when a null hypothesis is rejected while being true. The error accepts the alternative hypothesis, despite it being attributed to chance [24].

The work described in [26] stated that gene expression data can be used to rank individuals according to molecular characteristics to generate hypotheses about disease mechanisms. Such data may be particularly useful for identifying prototype biomarkers for quantitatively and longitudinally tracking dynamic disease traits that cannot readily be explained by static variation in DNA sequence. This thesis will use gene expression data to find the most correlated path of differentially expressed genes (with potential bio marking characteristics) in the presence of HD.

The study described above [26] was not able to evaluate all genes from the samples, eliminating some genes from the study. This could imply that important information of differentially expressed genes was not analyzed. The limitation was due to technical variation and it the limitation was higher for genes with low average expression intensities (on Affymetrix Human GeneChip U133A arrays). Only genes with intensities ≥100 in at least one sample were considered for further analysis, the remaining genes were excluded from the analysis. This conservative statistical analysis keeps the number of false-positive results at a minimum, although the number of false-negative results is likely to remain high. In this thesis work, more comprehensive numbers of genes will be analyzed simultaneously in their relative expression and will provide a demonstrably optimal proxy for a signaling path.

The study "Gene Expression Profiling in Human Neurodegenerative Disease" [27] presented microarray human GEP (gene expression profiling) studies in the common neurodegenerative

diseases amyotrophic lateral sclerosis (ALS), PD and AD. Studies using samples from disease and controls included: postmortem spinal cord, substantia nigra (PD) and blood mononuclear cells and peripheral leukocytes (AD). Results from [27] included that for ALS and PD, gene expression related to RNA splicing and protein turnover is disrupted, and several studies in ALS support involvement of the cytoskeleton. GEP studies have implicated the ubiquitin–proteasome system in PD pathogenesis, and have provided evidence of mitochondrial dysfunction in PD and AD. In AD, a possible role for dysregulation of intracellular signaling pathways, including calcium signaling was highlighted. Similar to this thesis, [27] presented commonalities and prevalent pathways in three neurodegenerative diseases, although the combination of diseases was not the same. In this paper, the most significant changes were found using P-value for differentially expressed clusters that exhibited better functional enrichment than a similar number of the most differentially expressed individual genes. This implies that differentially expressed clusters contained genes that were more functionally similar to each other.

The use of different microarrays platforms and analysis techniques was identified in the study as one the limitations. In this thesis the results from the proposed method is not affected by the microarray platform, also the methodology and analysis will be the same for the three diseases.

In the paper "Integrative Gene Expression Analysis of lung cancer based on a technology-merging approach" [28], different data from several independent studies were integrated to tackle with the scarce number of samples that can be collected by only one study. They created a bigger database of lung cancer studies in order to obtain more significantly expressed genes in the differential gene expression. The scope of this paper is the comparison

between the final DEGs in a lung cancer study found by the meta-analysis approaches. Most identified genes were associated with the immune response. In comparison with the paper discussed, this thesis analyzes microarrays to find a set of array probes related to genes that have an expression profile with an unusual level of under-expression or over-expression. The work in [29] scanned through the literature to identify the genes, mostly associated with the immune response. Literature search aiming for biological evidence is also performed in this thesis. Both, the work described previously [29] and this thesis demonstrate that merging data from different studies is easily extensible to any other disease or differential analysis, as a methodology to integrate microarray data regardless of the experiment and origin of such data.

To reduce the false-positive discovery rate, in [29] the authors essayed different configurations varying several parameters: threshold of P-value, fold change, collapsing probes method, normalization of dataset, among others. Additionally, they obtained the intersection of the outstanding genes which are more likely to encode relevant information for the differential analysis. This step was done with meta-analysis. The normalization was needed because gene expression values must be comparable to each other before combining datasets. Consequently, those genes providing an expression profile which is significantly different between healthy and tumor cohorts according to these parameters were selected. The criteria for significance was an adjusted p-value<0.05 as well as a log fold change (FC) higher than 2. Consequently, those genes providing an expression profile significantly different between healthy and tumor cohorts according to these parameters were selected. By the contrary, this thesis proposes the use of analysis strategies based on mathematical optimization that facilitate the simultaneous analysis of

multiple experiments without the manipulation of parameters by the user while not requiring normalization of data.

The literature review presented here provides evidence of the novelty of using 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, as advocated by our group in [30].

2.5 Methodology Background

2.5.1 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 [30][31][32]. Because of the presence of conflict, an MCO problem does not find a single best solution but rather a set of best compromising solutions for the performance measures under analysis [35]. An example of a MCO problem is shown below.

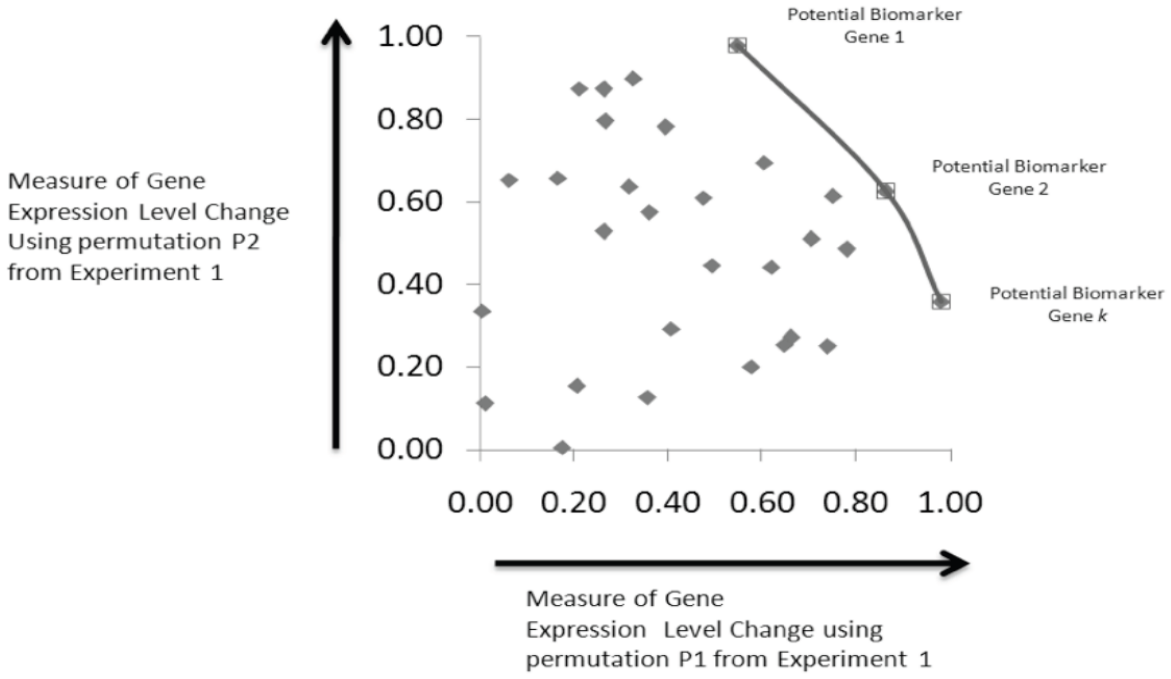


Figure 2: Example of an efficient frontier

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

Find x to

$$\textit{Minimize } f_j(x) \quad j=1,2,\ldots,J \tag{1}$$

The MCO problem in (1) can be discretized onto a set $\boldsymbol{K}$ with $|\boldsymbol{K}|$ points in the space of the decision variables so as to define particular solutions $\boldsymbol{x^k}$, $(\boldsymbol{k}=1,2,\ldots,|\boldsymbol{K}|)$ which can, in turn, be evaluated in the $\boldsymbol{J}$ performance measures to result in values $\boldsymbol{f_j}$ $(\boldsymbol{x^k})$.That is, the $\boldsymbol{k^{th}}$ combination of values for the decision variables evaluated in the $\boldsymbol{j^{th}}$ objective function [30].

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

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

$$\textbf{Minimize } f_j(x^k) \quad j=1,2,\ldots,J \tag{2}$$

The solutions to (2) are, then, the Pareto-efficient solutions of the discretized MCO problem. Considering formulation (2), a particular combination x^0 with evaluations $f_j(x^0)$ will yield a Pareto-Efficient solution to (2) if and only if no other solution x^Ψ exists that meets two conditions, from this point on called Pareto-optimality conditions:

$$f_j(x^\Psi) \leq f_j(x^0) \; \forall j \qquad \text{(Condition 1)}$$

$$f_j(x^\Psi) \leq f_j(x^0) \textbf{ in at least one } j \qquad \text{(Condition 2)}$$

Conditions (1) and (2) imply that no other solution x^Ψ dominates the solution under evaluation, x^0, in all performance measures simultaneously.

2.5.2 Traveling Salesman Problem

The Traveler Salesman Problem (TSP) is one of the most famous combinatorial optimization problems. TSP tries to construct the shortest tour through *n* cities for a salesperson to visit, usually going back to a preselected base city [33]. The object of TSP is to "Find the shortest tour that visits each city in a given list exactly once and then returns to the starting city" [33].

An illustration of how the resulting graph would look like for a five genes problem is shown in Figure 3. As an example, take the network presented in Figure 4. 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 with minimum total distance, there would be total of (5-1)! =24 possible cycles [17].

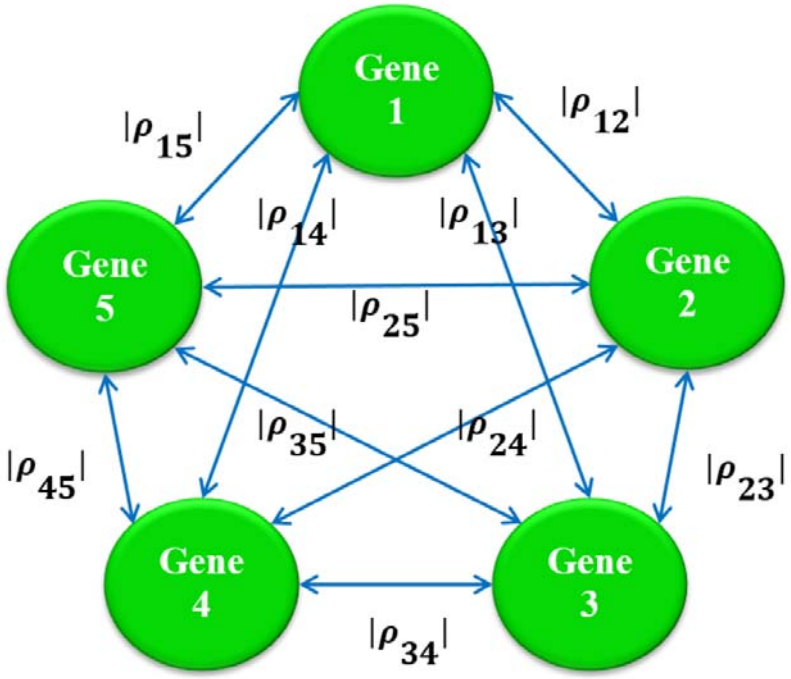


Figure 3: Representation of a Potential Sequence of a Signaling Pathway

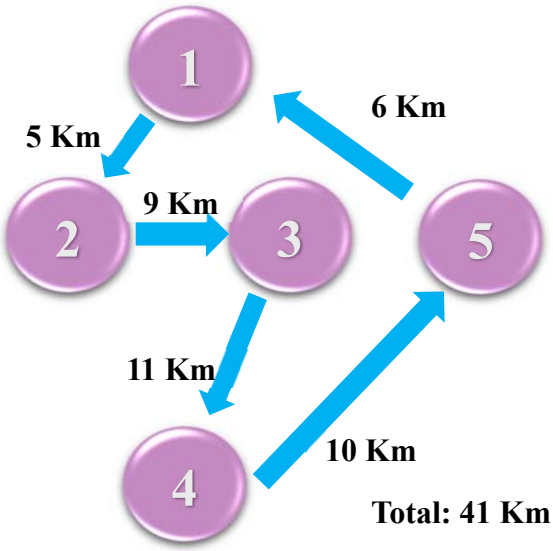


Figure 4: Representation of a TSP optimal solution example

The TSP optimization model is as follows:

$$\textbf{Minimize} \sum_{(i,j)\in A} c_{ij}y_{ij} \quad (3)$$

$$\sum_{1\le j\le n} y_{ij} = 1 \quad \forall i = 1, 2, \dots, n \quad (4)$$

$$\sum_{1\le i\le n} y_{ij} = 1 \quad \forall\, j = 1, 2, \dots, n \quad (5)$$

$$\mathcal{N}x = b \quad (6)$$

$$x_{ij} \le (n-1)y_{ij} \quad \forall (i,j) \in A \quad (7)$$

$$x_{ij} \ge 0 \quad \forall (i,j) \in A \quad (8)$$

$$y_{ij} = 0 \text{ or } 1 \quad \forall (i,j) \in A \quad (9)$$

Let **A' = {(i,j): y_{ij} =1}** and let **A'' ={(i,j): x_{ij} >0}**. The constraints (4) and (5) 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 (4) and (5) will be a union of disjoint cycles; if any such solution contains more than one cycle; they are referred to as sub tours, since they pass through only a subset of nodes.

In constraint (6) **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 xij. The column **Nij** has a +1 in the $\mathbf{i^{th}}$ row, a −1 in the $\mathbf{j^{th}}$ row; the rest of its entries are zero. Constraint (6) 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 (7) imply that A" is a subset A'. These conditions imply that the arc set A' is connected and thus cannot contain sub tours [33].

2.5.3 Minimum Spanning Tree

The Minimum Spanning Tree considers an undirected and connected network, a measure of the positive length (e.g., distance, cost, time, etc.) associated to each link [35] The MST methodology consist in 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 between each pair of nodes [38].

An example, explained in [17] of the MST is presented in Figure 8 with five nodes of a network with their potential links and the positive length for each if it is inserted into the network. It is important to ensure enough links are inserted to satisfy the requirement that there is a path between every pair of nodes. The objective of this method is to satisfy this requirement while at the same time minimizing the total length of the links inserted into the network. As an example the Figure 8 presents the solution, highlighted by the darker lines. 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 5 the solution is highlighted by the darker and thicker lines. In the case of a spanning tree the total number of possible solutions could be calculated with Cayley's formula n^{n-2} where n is the number of edges or arcs in the graph [17]. In this particular example there are a total of 5^{5-2} =125 possible solutions

.

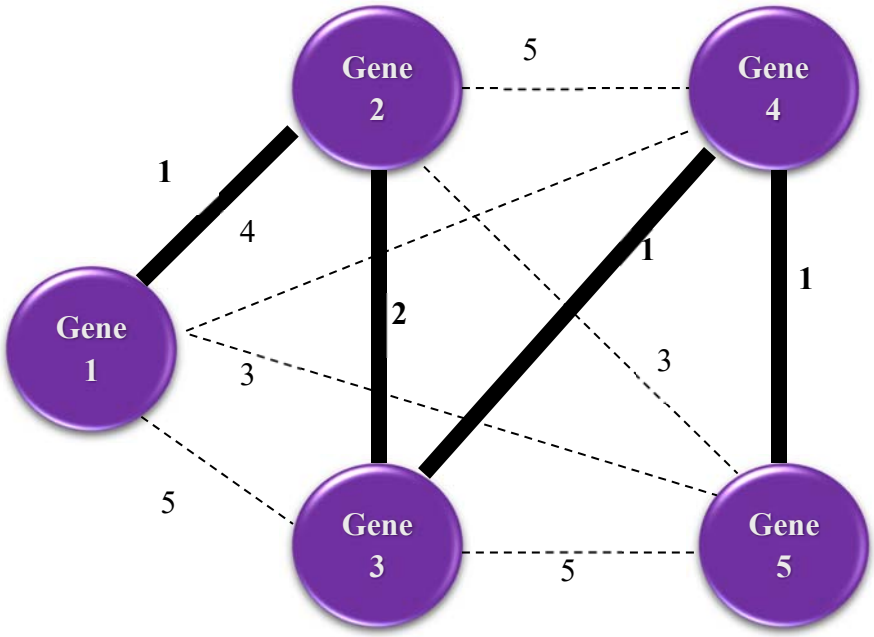


Figure 5: Example of a Minimum Spanning Tree (MST)

3. Methodology

PD is the second most common neurodegenerative disease. Although many of the pathogenic molecules underlying the rare autosomal-dominant forms of PD have been identified, the full complement of pathogenic pathways involved in the common "sporadic" form of PD remains unknown [25]. For this reason PD was the first neurological disorder analyzed and then, the methodology was replicated to HD and AD respectively. First, two microarray databases [25] were selected for analysis, each containing 22,277 probes in the microarray and their respective lectures in six samples in brain tissues for PD and five samples for control. Both databases were obtained from [23]. The first database came from a high-resolution variant of functional MRI (fMRI) that maps basal cerebral blood volume (CBV) with submillimeter resolution to show that

the DMNV (Dorsal Motor Nucleus of the Vagus) is dysfunctional in PD according with the study mentioned above. The second database came from a neighboring region relatively resistant to the disease, ION (Inferior Olivary Nucleus), taken from the same study. The PD samples were taken from autopsies from people with a mean age 56.4 years and control samples were taken from autopsies of people with mean age of 56.2 years [25].

The first step was to calculate the difference of PD and Control expressions in both databases simultaneously modeling the analysis as a Multiple Criteria Optimization (MCO) Problem as described in [34][35]. In brief, MCO deals with making decisions in the presence of multiple performance measures in conflict. 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 [34]. The general MCO 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 [36]. The performance measures used in this case, were the absolute median differences, calculated for PD and control samples in DMNV and also PD and control samples in ION databases as shown in Figure 6. The performance measures used were absolute median differences between control and PD tissues in each data base. Then both performance measure were graphed and analyzed using MCO.

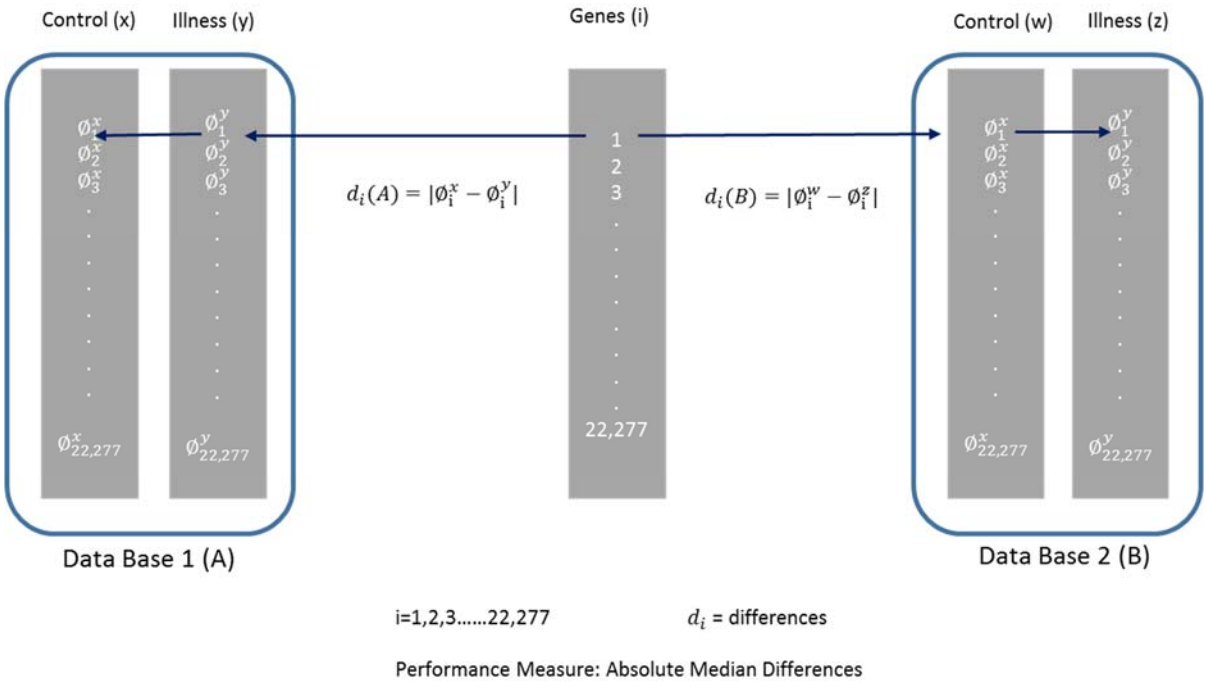


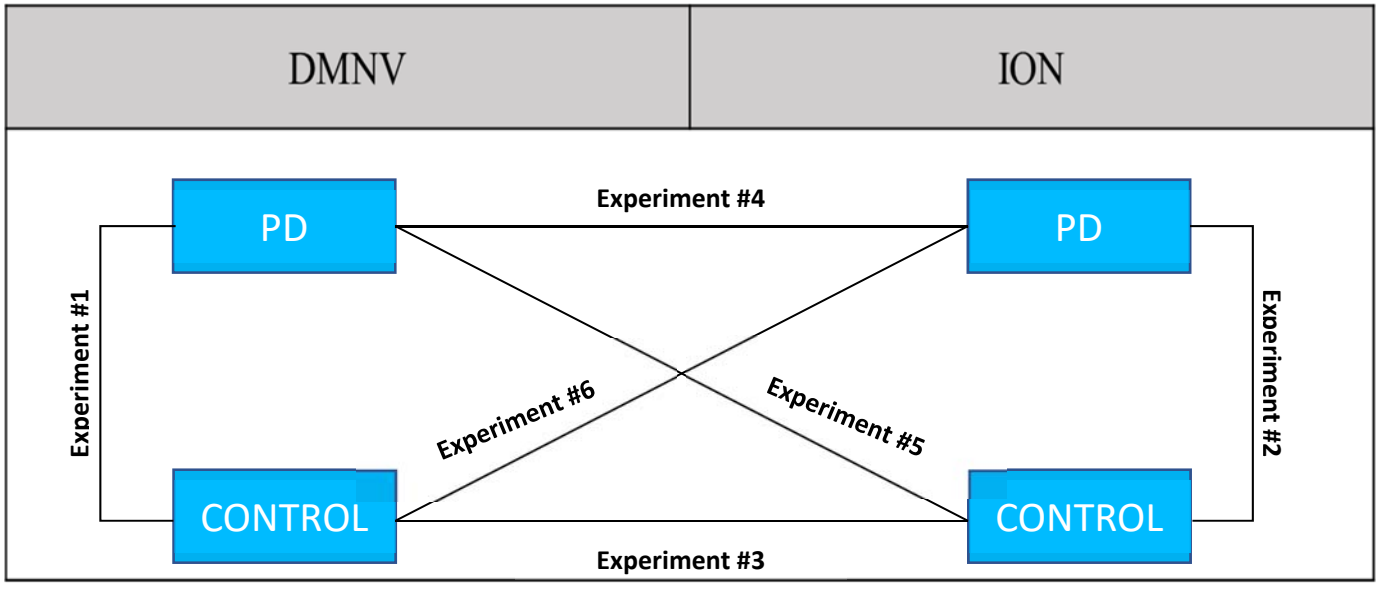
Figure 6: Gene expression differences between PD and control in DMNV and ION tissue

In this study, the set of solutions that provide the best compromise between the conflicting performance measures was found using the Pareto-Efficient Frontier of the candidate set of genes. The genes selected by this Frontier are those that change their expression the most between the controls and the PD samples using all performance measures at hand. For this study the first ten efficient frontiers were considered to avoid missing genes that also should be considered solutions, as explained in [35]. Due to computer memory constrains the databases were divided in 4 groups of 5,600 genes or less and iterated to converge to the final efficient frontier. An example of an efficient frontier analysis is shown in Figure 2, under a case with maximization of two performance

measures. Do notice that any performance measure to be maximized can be transformed linearly to become a perfectly equivalent performance measure to be minimized individually.

The original work from where the databases were obtained [2] stipulated that one region of the brain is being affected (DMNV) in PD presence and the other region remained unaffected (ION). This is challenged in our results as it will be shown later in this document. In addition, it is known that gene expressions works as a network and neighboring brain regions could be affected. For this reason our approach was to perform six analyses (presented in Table 1) to determine if ION was unaffected; or if the main variable driven the expression differences was the presence/absence of PD in the samples. All analyses were treated as MCO problems.

Table 1: Analysis with DMNV and ION databases



The first analysis was on the DMNV region samples (PD vs Control) where two performance measures (mean and median) were used, see Figure 7 as an example. The difference absolute value between each performance measure related to PD and control was calculated. In figure 7 it is shown the case when only one data base was used (in this example DMNV). For this specific case the performance measures used were the absolute mean difference and absolute median difference between PD and control gene expression.

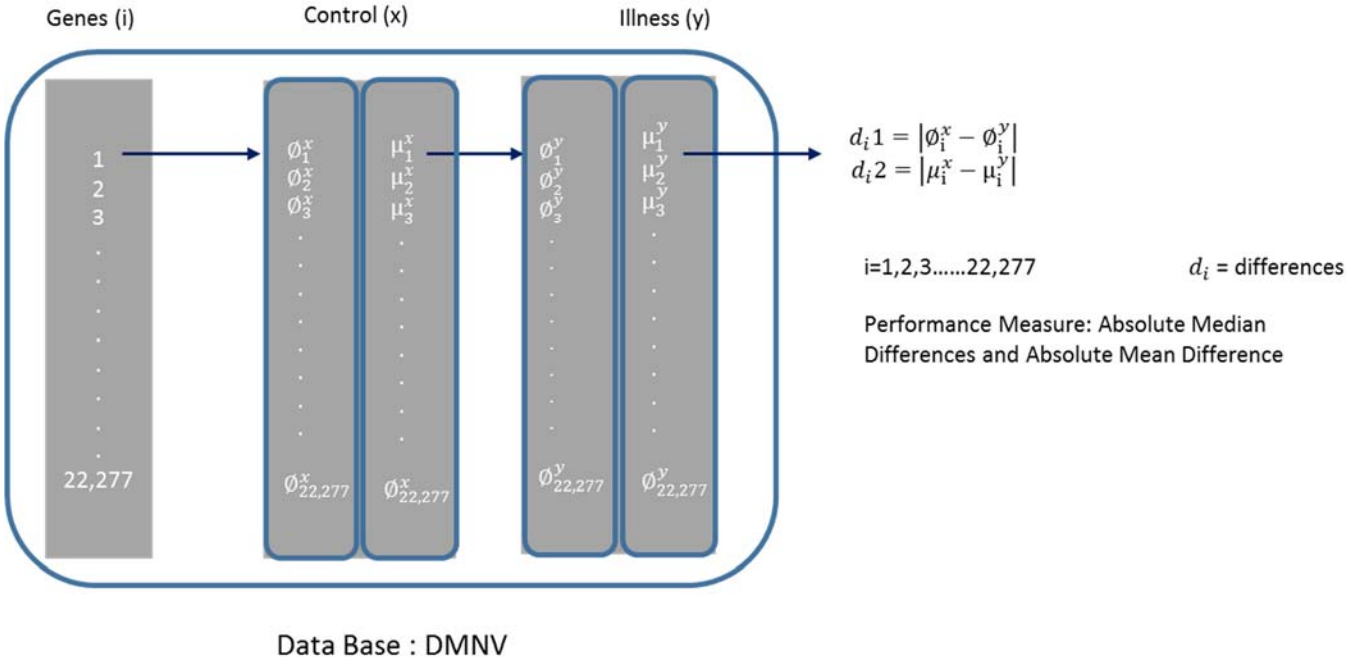


Figure 7: DMNV (PD vs control differences) Analysis

Ten series of Pareto Efficient Frontiers were identified on all analyses resulting in distinct sets of differentially expressed genes. As stated before, there were 22,777 genes initially, from which ten Pareto Efficient Frontiers were analyzed in each analysis. Accumulating the Pareto optimal

solutions (differentially expressed genes) from every analysis, there were a total of 44 differentially expressed genes. From those 44 differentially expressed genes, eight genes that were in common in every single analysis were further analyzed. Table 2 shows the 44 differentially expressed genes from every analysis. The genes that were found in a specific case have a number one in the table.

Table 2: Selected Genes each of the six comparative analyses

	DMNV	ION - DMNV-PD	ION	DMVN vs ION	DMNV PD-ION CONTROL	DMNN CONTROL -ION PD	ION-DMNV-Control
ARF3	1	1	1	1	1	1	1
CCL5	1	1	1	1	1	1	1
DDX39B	1	1	1	1	1	1	1
GDI2	1	1	1	1	1	1	1
PTPN21	1	1	1	1	1	1	1
RPL11	1	1	1	1	1	1	1
RPL21	1	1	1	1	1	1	1
THRA	1	1	1	1	1	1	1
UBA7	1	1	1	1	1	1	0
AHSA1	1	1	1	1	0	0	1
MXD4	1	1	1	1	0	0	1
SRP14	1	1	1	1	0	1	0
PAX8	1	1	1	0	0	1	0
PTGDS	0	1	0	1	0	1	1
PVT1	1	1	0	1	1	0	0
RABGGTB	0	1	0	1	0	1	1
ATG12	1	0	0	1	0	0	1
BCAT1	0	1	1	0	0	0	1
CD3D	1	0	0	1	0	0	1
GUCA1A	1	1	1	0	0	0	0
NENF	0	1	1	0	0	1	0
RAB11A	1	0	0	1	0	0	1
RAB3GAP1	0	1	0	1	0	0	1
RPS6	0	1	0	1	0	0	1
ATP2C1	0	0	0	1	0	1	0
CIRBP	1	0	0	1	0	0	0
ECHDC3	1	0	0	0	0	1	0
EIF4H	1	0	0	0	0	1	0
TFF2	1	0	0	1	0	0	0
VAMP4	1	0	0	0	0	1	0
RFC2	0	1	1	0	0	0	0
SLC5A12	0	0	1	0	0	0	1
IGKV4-1	1	0	0	0	0	0	0
ZBTB17	1	0	0	0	0	0	0
EIF3D	0	0	1	0	0	0	0
GYG1	0	0	1	0	0	0	0
ENOSF1	0	0	0	0	0	0	1

Continuation	DMNV	ION - DMNV-PD	ION	DMVN vs ION	DMNV PD-ION CONTROL	DMNN CONTROL -ION PD	ION-DMNV-Control
DCTN5	0	1	0	0	0	0	0
HSPA6	0	1	0	0	0	0	0
EIF4G2	0	1	0	0	0	0	0
HNRNPL	0	0	0	0	0	1	0
NCBP1	0	0	0	0	0	1	0
ESR2	0	0	0	0	0	1	0
LILRB3	0	0	0	0	0	1	0

The genes that obtained a one in every analysis were selected for future analysis. Table 3 shows the eight differentially expressed genes selected.

Table 3: Optimal Genes in every database analysis

Genes	DMNV	ION	ION-DMNV-Control	ION - DMNV-PD	DMNV PD-ION CONTROL	DMNN CONTROL -ION PD	DMVN vs ION
CCL5	1	1	1	1	1	1	1
DDX39B	1	1	1	1	1	1	1
GDI2	1	1	1	1	1	1	1
PTPN21	1	1	1	1	1	1	1
RPL11	1	1	1	1	1	1	1
RPL21	1	1	1	1	1	1	1
THRA	1	1	1	1	1	1	1

Using the selected differentially expressed genes from the last process, individually for each analyzes and for which behavior was measured as a statistical correlation, the next step was to find coordinated behavior among the expression changes of the selected genes. As explained in [17], 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, for example 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 is 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:

$$\boldsymbol{r_{xy}} = \frac{\sum_{i=1}^{n}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})}{(\boldsymbol{n}-\boldsymbol{1})\boldsymbol{s_x}\boldsymbol{s_y}} = \frac{\sum_{i=1}^{n}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}})}{\sqrt{\sum_{i-1}^{n}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^2 \sum_{i-1}^{n}(\boldsymbol{y_i} - \bar{\boldsymbol{y}})^2}} \quad (10)$$

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 [33].

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. 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[17].

The statistical correlation was computed as linear as a first approach and carried out in a pairwise manner. The linear correlation values found are proxies for suppressed or stimulated behavior in the expression level changes of the two genes under analysis. Because these values range from -1 to 1, their absolute values are computed to handle them as quantities to be maximized. Thus two genes will be strongly correlated if the absolute of their correlation value is close to 1. The correlations calculated for each pair of gene were further arranged in a correlation matrix (Figure

9). To construct this matrix, the differences taken in consideration in each analysis (see Table 1) 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 correlation of genes in each analysis was the combined to converge on a final solution as illustrated in Figure 8.

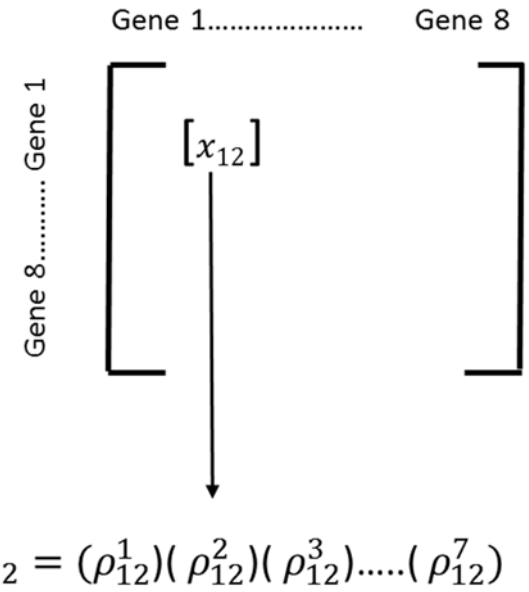


Figure 8: Combined correlations solution matrix example

	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8
Gen 1	0.000	0.967	0.089	0.839	0.933	0.910	0.980	0.902
Gen 2	0.967	0.000	0.066	0.911	0.970	0.970	0.983	0.970
Gen 3	0.089	0.066	0.000	0.035	0.065	0.045	0.069	0.040
Gen 4	0.839	0.911	0.035	0.000	0.944	0.943	0.912	0.957
Gen 5	0.933	0.970	0.065	0.944	0.000	0.969	0.985	0.964
Gen 6	0.910	0.970	0.045	0.943	0.969	0.000	0.961	0.991
Gen 7	0.980	0.983	0.069	0.912	0.985	0.961	0.000	0.953
Gen 8	0.902	0.970	0.040	0.957	0.964	0.991	0.953	0.000

Figure 9: Correlations solution matrix for PD

This method consisted in multiplying the absolute correlation values of each gene combination from all analyses on the same position. This ensured that changes in correlation values be reflected in a combined correlation factor. This combined correlation factor represented each pair of genes.

The methodology was explained using PD analysis for reference. For HD and AD the methodology was replicated, with the only difference that only one analysis was performed for each disease, and the correlation matrix used for optimization purposes was the correlation matrix from that particular analysis, not a combined matrix was needed. The absolute value correlation matrix for HD and AD disease are shown below in respective figures.

	Gen 1	**Gen 2**	**Gen 3**	**Gen 4**	**Gen 5**	**Gen 6**
Gen 1	0	0.043	0.03	0.08	0.187	0.001
Gen 2	0.043	0	0.351	0.355	0.383	0.42
Gen 3	0.03	0.351	0	0.889	0.637	0.398
Gen 4	0.08	0.355	0.889	0	0.768	0.599
Gen 5	0.187	0.383	0.637	0.768	0	0.567
Gen 6	0.001	0.42	0.398	0.599	0.567	0

Figure 10: Correlations solution matrix for HD

	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8	Gen 9	Gen 10
Gen 1	0	0.989	0.959	0.971	0.690	0.529	0.167	0.618	0.996	0.745
Gen 2	0.989	0	0.942	0.972	0.704	0.558	0.200	0.626	0.992	0.774
Gen 3	0.959	0.942	0	0.951	0.654	0.601	0.117	0.651	0.958	0.660
Gen 4	0.971	0.972	0.951	0	0.680	0.548	0.131	0.603	0.977	0.715
Gen 5	0.690	0.704	0.654	0.680	0	0.859	0.670	0.917	0.706	0.888
Gen 6	0.529	0.558	0.601	0.548	0.859	0	0.565	0.952	0.551	0.655
Gen 7	0.167	0.200	0.117	0.131	0.670	0.565	0	0.639	0.184	0.579
Gen 8	0.618	0.626	0.651	0.603	0.917	0.952	0.639	0	0.628	0.713
Gen 9	0.996	0.992	0.958	0.977	0.706	0.551	0.184	0.628	0	0.756
Gen 10	0.745	0.774	0.660	0.715	0.888	0.655	0.579	0.713	0.756	0

Figure 11: Correlations solution matrix for AD

If each gene is represented through a node in a graph, then the undirected arc joining a pair of genes can hold their absolute correlation value. This led to the Traveling Salesman Problem (TSP) formulation, where the idea is to find the most correlated complete tour [17].

3.1. Traveling Salesman Problem

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 [37]. TSP tries to construct the shortest tour through n cities for a salesperson to visit, usually going back to a preselected base city [39]. The object of TSP is to "Find the shortest tour that visits each city in a given list exactly once and then returns to the starting city" [41]. If one rephrases the quoted sentence as "Find the most correlated tour that visits each potential differentially expressed gene in a given list exactly", then it is clear that such solution might shed light on how PD works. In this particular research, the TSP is used to obtain an optimal sequence maximizing linear correlations among the eight genes considered as potential differentially expressed genes.

3.2. Minimum Spanning Tree

Due to the fact that the Traveling Salesman 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 is the minimum spanning tree (MST).

The methodology mentioned in Section 2.6.3 is applied as an alternative to the TSP. The purpose of this additional methodology to develop a signaling pathway from a list of genes identified as potentially expressed genes (see Table 4), same used in TSP methodology . Utilizing the same matrix developed from the linear correlations of the genes of interest (Figure 9). The logic of the MST is applied in order to find a minimal tree with the largest correlation among the nodes of interest.

3.3 GeneMANIA

GeneMANIA is a web-based tool. It is a web interface for generating hypothesis about gene function, analyzing gene lists and prioritizing for functional analysis [43]. This method was discussed and explained in [17]. Data sets utilized by GeneMania are collected from publicly available databases [41]. 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 (43).

The researcher must enter a list of predetermined genes of interest, optionally also selected 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. It displays an interactive association network, which illustrates the relations among the genes and data sets [43]. 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 [43].

4. Results for each disorder

4.1 Results for Parkinson's disease

In order to demonstrate the proposed analysis pipeline, Table 4 presents the resulting differentially expressed genes. The potential differentially expressed genes were taken from two microarray databases [25] and were selected for analysis, each containing 22,277 probes in the microarray and their respective lectures in six samples in brain tissues for PD and five samples for control.

Table 4: Set of potential differentially expressed genes

Gene Number	Gene Name
1	ARF3
2	CCL5
3	DDX39B
4	GDI2
5	PTPN21
6	RPL11
7	RPL21
8	THRA

4.2 Signaling Pathways for PD utilizing TSP

The optimal solution to this particular TSP is the tour among the genes of interest with the largest possible correlation, shown in Figure 12 (the signs in the figure represent the original correlation coefficient value). For this particular case there are a total of (8-1)! ≈5,040 ways in which a cyclic path can be drawn among the 8 genes, the optimal path was found in this research.

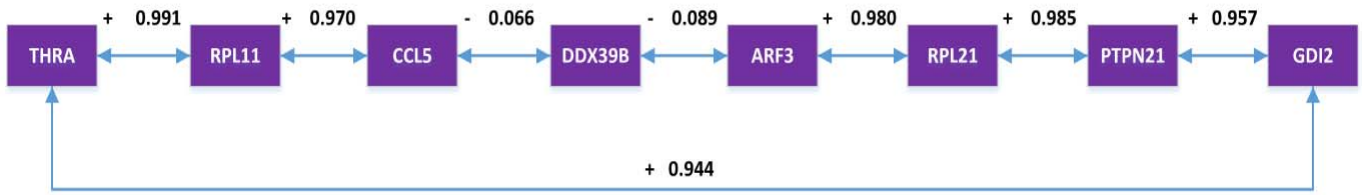


Figure 12: Optimal solution path for potential expressed genes

4.3 Signaling Pathways for PD utilizing MST

As an alternative method to the TSP, the MST was applied to obtain a tree that maximizes the linear correlations of the eight genes that were identified previously as potential differentially expressed genes. Parting from the list of genes of interest, the MST method was applied and an optimal structure that maximizes these correlation values was obtained MST is applied as an alternative to the TSP as a means to develop a signaling pathway from a list of genes identified as potential differentially expressed genes (see Table 4). Starting from the same matrix developed from the linear correlations of the genes of interest the logic of the MST is applied in order to find a minimal tree with the largest correlation among the nodes of interest. The MST model is shown in Figure 13.

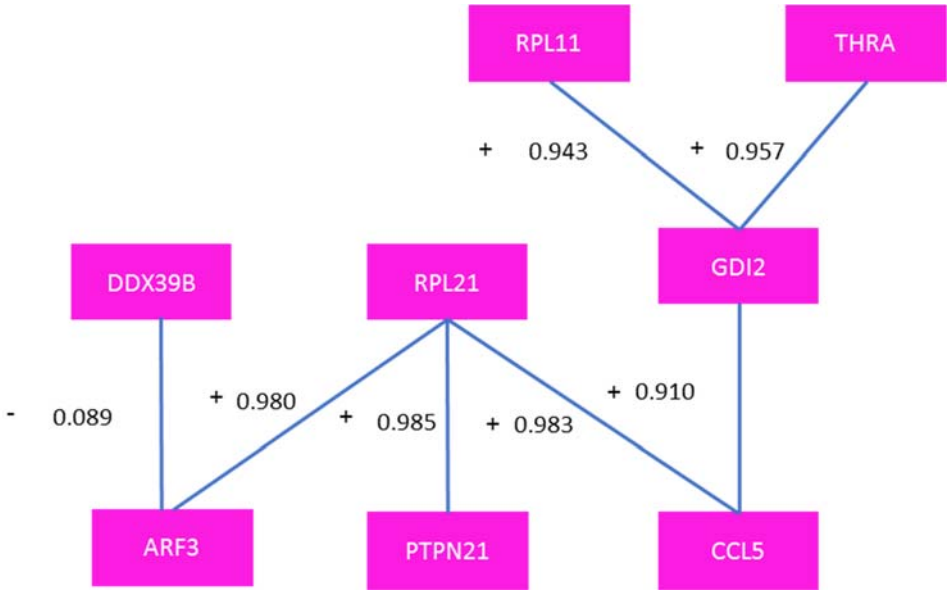


Figure 13: Optimal solution network for potential expressed genes

4.4 Signaling Pathways utilizing GeneMANIA

In order to have a comparison with an already existing tool, geneMANIA was used. As mentioned before the program requires the list to be entered in its query list located at its website. Figure 14 shows the resulting diagram.

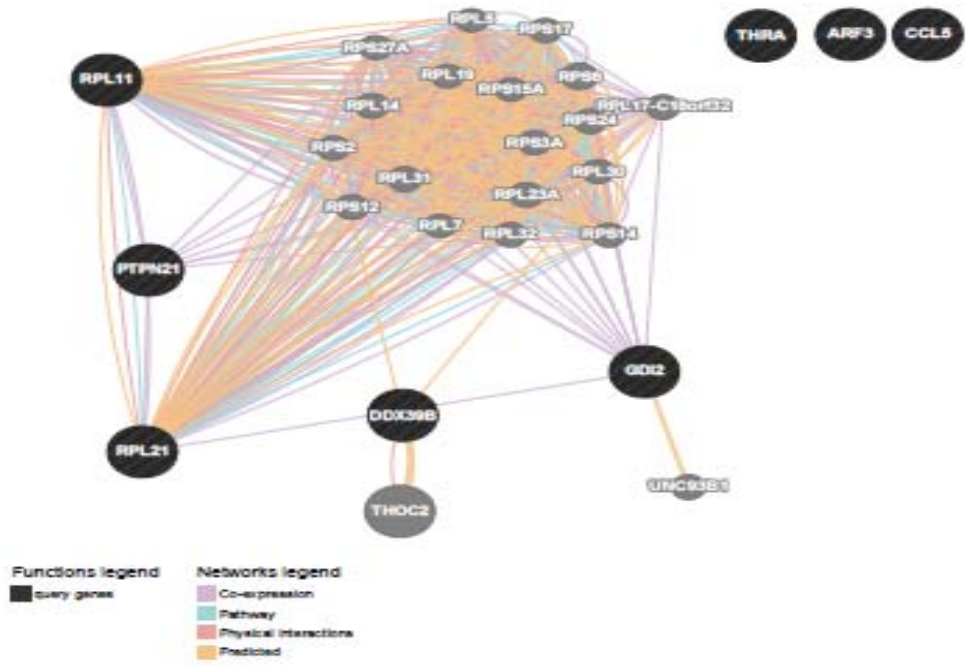


Figure 14: Solution diagram for most correlated changes in genes expressions using geneMANIA tool

It must be noted that GeneMANIA was unsuccessful in including all our original genes in the network. The genes THRA, APF3 and CCL5 were not connected to the rest of the genes even though they significantly changed expression with the method presented in this work.

4.5 Biological Discussion for PD genes

The biological expertise in this section was importantly contributed by Clara Isaza and Janice Garcia. The differentially expressed genes and relations among them, found with this methodology, are not necessarily associated with PD at this moment. A search for biological pathways and characteristics for each genes in KEGG [49] databases and Pubmed publications [47] was conducted. It was found that ARF3 is localized in the Flemming body and play important

roles in cytokinesis. For CCL5; recent advances in discovering the role in metastatic breast cancer, cell gene regulation, autoimmunity inmunoregulatory and inflammatory processes. It was also related to polarization of the immune response in ovalbumin-induced airway inflammation, herpes simplex infections, chaga disease and influenza A. Diseases associated with CCL5 include angina pectoris and ulcer of lower limbs. Gene DDX39B has association with clinical outcomes of patients with asmodium vivax malaria, the risk of late-onset Alzheimer's Disease in Iranian Population, RNA transport and Influenza A. The gene GDI2 regulates ciliogenesis after mitosis, drivers of metastatic dissemination in sonic hedgehog medulloblastoma. The following gene, PTPN21, exerts pro-neuronal survival and neuritic elongation via ErbB4/NRG3 signaling, identified by genome-wide association study data-mining and replication, are associated with schizophrenia, frameshift mutations in coding repeats of protein tyrosine phosphatase genes in colorectal tumors with microsatellite instability. In other hands, the RPL11 analysis of psoriasis reveals discordant and concordant changes in mRNA and protein abundance, molecular mechanisms underlying the pathology of Diamond-Blackfan anemia, contributes to the biogenesis of 60S ribosomal subunits and influences nucleolar morphology. Studies on RPL21, assemble characteristics of large subunit ribosomal proteins in S. cerevisiae, mutation in ribosomal protein L21 underlies hereditary hypotrichosis simplex. Last, THRA is related to a neuroactive ligand-receptor interaction, thyroid hormone signaling pathway, transcriptional network associated with the contractile phenotype of smooth muscle cells in human carotid atherosclerosis.

The eight differentially expressed genes were investigated in PUBMED to understand where those genes were reported in published literature. PubMed comprises more than 25 million citations for biomedical literature from MEDLINE, life science journals, and online books [46].

In the table below are shown the differentially expressed genes and where they have been reported in PUBMED. Mostly of the genes has been reported in apoptosis, cell proliferation, inflammation, mitochondria and neurodegenerative. The gene RPL21 was already related to Parkinson's disease.

Table 5: Genes reported in PUBMED in respective literature

	DDX39B	ARF3	CCL5	THRA	RPL11	GDI2	PTPN21	RPL21	Total
Apoptosis	0	1	1	1	1	1	1	0	**6**
Cell Proliferation	1	1	1	1	1	1	0	0	**6**
Inflammation	1	1	1	1	0	1	0	0	**5**
Mitochondria	1	1	1	1	1	0	1	0	**6**
Neuro-Degenerative	1	1	1	0	0	0	0	0	**3**
Parkinson's Disease	0	0	1	0	0	0	0	0	**1**
Immuno-regulatory and Inflammatory processes	0	0	1	0	0	0	0	0	**1**
Cholera toxin	0	1	1	0	0	0	0	0	**1**
N/A	0	0	0	0	0	0	0	1	**1**
Total	**4**	**6**	**8**	**4**	**3**	**3**	**2**	**1**	

The differentially expressed genes pathways were also investigated in KEGG (Kyoto Encyclopedia of Genes and Genomes). KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (46). In the table below it is shown the pathways related with each gene. The pathways ribosome and influence were related

with more than one differentially expressed gene. Also the gene CCL5 was related with twelve pathways and the gene DDX39B was related with four pathways. The genes ARF3, GDI2 and PTPNB1 were not related with any pathway in KEGG.

Table 6: Genes reported in KEGG in respective literature

Parkinson's Disease Differentially Expressed Genes									
	ARF3	CCL5	DDX39B	GDI 2	PTPN2 1	RPL11	RPL21	THRA	Total
Alzheimer's Disease	0	0	1	0	0	0	0	0	**1**
Squizophrenia	0	0	0	0	1	0	0	0	**1**
RNA transport	0	0	1	0	0	0	0	0	**1**
Spliceosome	0	0	1	0	0	0	0	0	**1**
Influenza A	0	1	1	0	0	0	0	0	**2**
mRNA surveillance pathway	0	0	1	0	0	0	0	0	**1**
Ribosome	0	0	0	0	0	1	1	0	**2**
Chemokine signalling pathway	0	1	0	0	0	0	0	0	**1**
Herpes simplex infection	0	1	0	0	0	0	0	0	**1**
NOD-like receptor signalling pathway	0	1	0	0	0	0	0	0	**1**
TNF signalling pathway	0	1	0	0	0	0	0	0	**1**
Toll-like receptor signalling pathway	0	1	0	0	0	0	0	0	**1**

Parkinson's Disease Differentially Expressed Genes									
	ARF3	CCL5	DDX39 B	GDI2	PTPN2 1	RPL1 1	RPL2 1	THRA	Total
Cytokine-cytokine receptor interaction	0	1	0	0	0	0	0	0	**1**
Chagas disease (American trypanosomiasis)	0	1	0	0	0	0	0	0	**1**
Cytosolic DNA-sensing pathway	0	1	0	0	0	0	0	0	**1**
Rheumatoid arthritis	0	1	0	0	0	0	0	0	**1**
Epithelial cell signalling in Helicobacter pylori infection	0	1	0	0	0	0	0	0	**1**
Prion diseases	0	1	0	0	0	0	0	0	**1**
Neuroactive ligand-receptor interaction	0	0	0	0	0	0	0	1	**1**
Thyroid hormone signalling pathway	0	0	0	0	0	0	0	1	**1**
n/a	1			1	1				**3**
Total	**1**	**12**	**4**	**1**	**1**	**1**	**1**	**2**	

4.6 Results for Huntington's disease

The genes that changed the expression the most using the HD data base are presented in Table 6. These genes were selected using Pareto Efficient Frontier, discussed in methodology section, using data bases of human blood samples on people associated to the disease and control [26]. This particular databases presented data from patients diagnosed with HD and patient without a HD diagnostic, mostly spouses of the affected people. Each expression was presented in probes as shown in table 7. The data base analyzed contained 22,284 probes, each probe with eleven sample

from HD patients and eight control samples. Each probe was investigated in gene cards and the genes that were part of each probes were compiled in table 7.Some probes contained the same genes, in this particular case they were grouped together and the sample data was analyzed together to have more samples for each case. The probes that were grouped since they contained the same genes are color coded in the table below (probes with the same genes share the same color) . Some probes that were also optimal, were controls from the previous experiment, from where the databases were used initially. Some control probes came from mouse or tomatoes tissues, they were excluded for further analysis.

Table 7: Set of potential differentially expressed genes and probes where the genes were found

Probes	Genes
204018_x_at	HAB2
209116_x_at	HBD
211696_x_at	HBD
211699_x_at	HBA2, HBA1
211745_x_at	HAB2
AFFX-r2-P1-cre-5_at	Control
209458_x_at	HAB2
AFFX-r2-P1-cre-3_at	Control
AFFX-CreX-3_at	Control
AFFX-hum_alu_at	Control
AFFX-CreX-5_at	Control
AFFX-r2-Ec-bioD-3_at	Control
AFFX-BioDn-3_at	Control
214414_x_at	HAB2
217414_x_at	HAB2
201315_x_at	IFITM3, IFITM1, LOC100101246, LOC391020, IFITM9P, IFITM8P, LOC144383
217232_x_at	HBD, RPS2P5, RPS2P8, RPS2P40, RPS2P7, RPS2P11, RPS2P29, RPS2P17
203107_x_at	RPS2P46, RPS2P5, RPS2P8, RPS2P40, RPS2P7, RPS2P11, RPS2P29, RPS2P17, RPS2P6, RPS2P12, RPS2P35, RPS2P12, RPS2P4, RPS2P55, RPS2P20, RPS2P52, RPS2P31, RPS2P32, RPS2P50, RPS2P1, RPS2P18, RPS2P48, RPS2P21
204848_x_at	HBG2
213828_x_at	H3F3A

For optimization analysis only one gene of each probe was used as reference, mostly because they are grouped together in probes because they are part of the same family. The genes used for TSP and MST are presented in table 7.

Table 8: Genes selected from each probe for further analysis

Gen Number	Gen Name
1	HAB2
2	HBD
3	IFITM3
4	RPS2P46
5	HBG2
6	H3F3A

4.7 Signaling Pathways for HD utilizing TSP

The optimal solution to this particular TSP is the tour among the genes of interest from HD database, with the largest possible correlation is shown in Figure 15. For this particular case there are a total of (6-1)! ≈120 ways in which a cyclic path can be drawn among the 6 genes, the optimal path was found in this research.

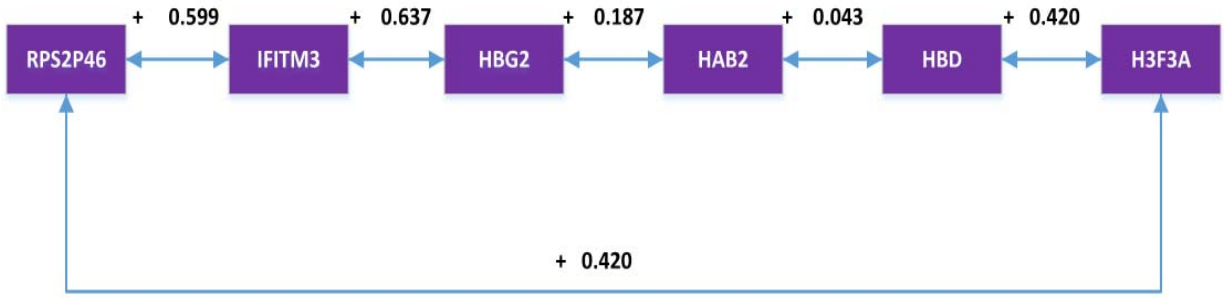


Figure 15: Optimal solution path for potential expressed genes

4.8 Signaling Pathways for HD utilizing MST

The MST was applied to obtain a tree that maximizes the linear correlations of the six genes that were identified previously. Starting from the same matrix developed from the linear correlations of the genes of interest the logic of the MST is applied in order to find a minimal tree with the largest correlation among the nodes of interest.

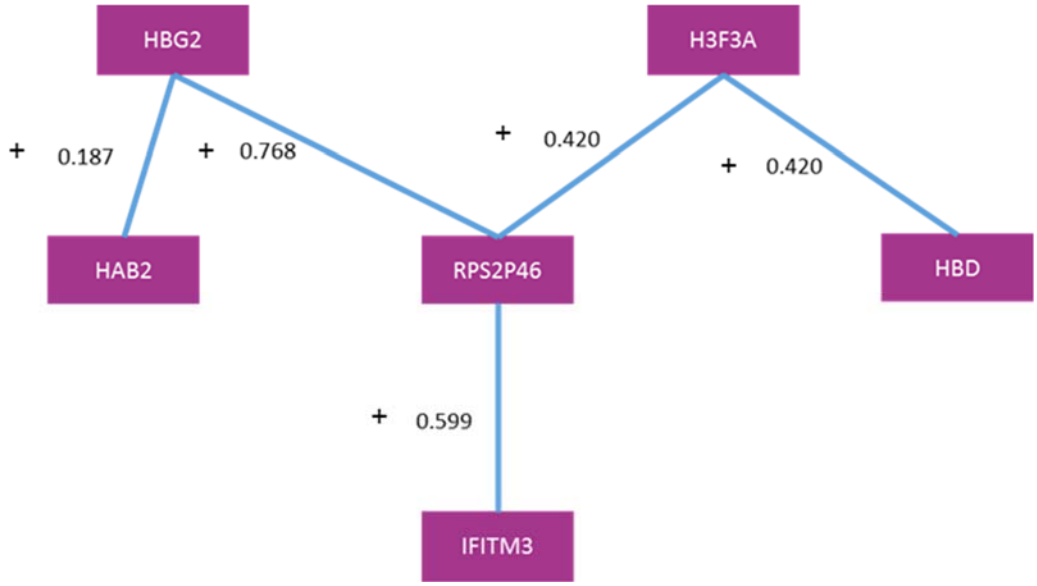


Figure 16: Optimal solution network for potential expressed genes

4.9 Biological Discussion for HD genes

Some of the gene reported in HD were related to viral infections. Also immune system and cytokine (messenger of immune system) were pathways of genes that changed their expression the most in HD. HBG2 has been reported already in patients with Alzheimer's disease. Several of the genes that change their expression in HD are related to hemoglobin. It was interesting that in a study that our research group is conduction on autism using similar techniques, IFITM2 and HBB (found as part of the optimal probes in HD) were found optimal for Autism analysis.

Analyzing the genes' information available in the literature the delta (HBD) and beta (HBB) genes are normally expressed in the adult: two alpha chains plus two beta chains constitute HbA, which in normal adult life comprises about 97% of the total hemoglobin. Two alpha chains plus two delta chains constitute HbA-2, which with HbF comprises the remaining 3% of adult hemoglobin. Five beta-like globin genes are found within a 45 kb cluster on chromosome 11 in the following order: 5'-epsilon--Ggamma--Agamma--delta--beta-3'. Mutations in the delta-globin gene are associated

with beta-thalassemia. Some diseases associated with HBD include hemoglobin lepore-beta-thalassemia syndrome and fetal hemoglobin quantitative trait locus 1. Among its related pathways are factors involved in megakaryocyte development and platelet production and Platelet activation, signaling and aggregation [46].

The HBA2 gene provides instructions for making a protein called alpha-globin. This protein is also produced from a nearly identical gene called HBA1. These two alpha-globin genes are located close together in a region of chromosome 16 known as the alpha-globin locus. Alpha-globin is a subunit of a larger protein called hemoglobin, which is the protein in red blood cells that carries oxygen to cells and tissues throughout the body. Hemoglobin is made up of four subunits: two subunits of alpha-globin and two subunits of another type of globin. Alpha-globin is a component of both fetal hemoglobin, which is active only before birth and in the newborn period, and adult hemoglobin, which is active throughout the rest of life [45] Diseases associated with HBA2 include thalassemias, alpha- and hemoglobin h disease, nondeletional [46].

For the gene IFITM3, the protein encoded by this gene is an interferon-induced membrane protein that helps confer immunity to influenza A H1N1 virus, West Nile virus, and dengue virus. Two transcript variants, only one of them protein-coding, have been found for this gene. Another variant encoding an N-terminally truncated isoform has been reported, but the full-length nature of this variant has not been determined [46]. Diseases associated with IFITM3 include influenza, severe and west nile virus [46].

Ribosomal protein S2 pseudogene 46, gene RPS2P46, is expressed in organism parts: temporal lobe, lung, gallbladder, spleen, gastroesophageal junction, subcutaneous adipose tissue, breast, liver, uterus, arm muscle, etc. [50].

The gamma globin genes (HBG1 and HBG2) are normally expressed in the fetal liver, spleen and bone marrow. Two gamma chains together with two alpha chains constitute fetal hemoglobin (HbF) which is normally replaced by adult hemoglobin (HbA) at birth. In some beta-thalassemias and related conditions, gamma chain production continues into adulthood. The two types of gamma chains differ at residue 136 where glycine is found in the G-gamma product (HBG2) and alanine is found in the A-gamma product (HBG1). The former is predominant at birth. Diseases associated with HBG2 include cyanosis, transient neonatal and fetal hemoglobin quantitative trait locus 1[46].

H3F3A gene codes for a replication-independent member of the histone H3 protein family. Diseases associated with H3F3A include blepharophimosis-intellectual disability syndrome, sbbys type and ohdo syndrome. Among its related pathways are activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3 and Transcriptional misregulation in cancer [46].

Searches for the selected genes were performed in PUBMED, GeneCards, and KEGG to find pathways for the genes found in HD that changed their expression the most. This information was summarized in the tables below.

Table 9: Genes reported in PUBMED & Gene Cards in respective literature

HD Pathways: PUBMED & Gene Cards							
	HBD	HBA2	IFITM3	RPS2P46	HBG2	H3F3A	Total
Hemostasis	1	0	0	0	1	0	**2**
Megakaryocyte	1	0	0	0	1	0	**2**
Immune System	0	0	1	0	0	0	**1**
Interferon Signalling	0	0	1	0	0	0	**1**
Cytokine Signalling in Immune system	0	0	1	0	0	0	**1**

Continuation: HD Pathways: PUBMED & Gene Cards							
	HBD	HBA2	IFITM3	RPS2P46	HBG2	H3F3A	Total
Interferon alpha/beta signalling	0	0	1	0	0	0	**1**
p70S6K Signalling	0	0	0	0	1	0	**1**
IL-4 Pathway	0	0	0	0	1	0	**1**
Activated PKNI stimulates transcription of AR (androgen receptor) regulated genes LK2 and KLK3	0	0	0	0	0	1	**1**
Cellular Senescence	0	0	0	0	0	1	**1**
Mitotic Prophase	0	0	0	0	0	1	**1**

Continuation: HD Pathways: PUBMED & Gene Cards							
	HBD	HBA2	IFITM3	RPS2P46	HBG2	H3F3A	Total
Cell Cycle, Mitotic	0	0	0	0	0	1	**1**
Development NOTCHI-mediated pathway for NF-KB activity modulation	0	0	0	0	0	1	**1**
Infection disease	0	0	1	0	0	1	**2**
Alzheimer's Disease	0	0		0	1		**1**
N/A	0	1		1			**2**
Total	**2**	**1**	**5**	**1**	**5**	**6**	

Table 10: Genes reported in KEGG in respective literature

HD Pathways: KEGG							
	HBD	HBA2	IFITM3	RPS2P46	HBG2	H3F3A	Total
Diabetic complications	1	0	0	0	0	0	**1**
Metabolic pathways	1	0	0	0	0	0	**1**
Complement and coagulation cascades	1	0	0	0	0	0	**1**
Carbon metabolism	1	0	0	0	0	0	**1**
Benzoate degradation	1	0	0	0	0	0	**1**
Amino benzoate degradation	1	0	0	0	0	0	**1**
Butanoate metabolism	1	0	0	0	0	0	**1**
Microbial metabolism in diverse environments	1	0	0	0	0	0	**1**
B cell receptor signalling pathway	0	0	1	0	0	0	**1**
n/a	0	0		1	1	1	**3**
Total	**8**	**0**	**1**	**1**	**1**	**1**	

4.10 Results for Alzheimer's disease

The differently expressed genes using AD database are presented in Table 11. The analysis was done with the use of a microarray database GDS2795 [47] related to AD disease that is focused on neurofibrillary tangles. The samples consisted of Entorhinal cortex of 19 cases and 14 control samples supplied from the brain banks of the AD Center (ADC) program. The database has a total of 54,675 genes. This data base also gave the probe, the respective genes are presented in table 11.

Table 11: Genes selected from each probe for further analysis

Probe	Gen Number	Gen Name
1553551_s_at	1	ND2
1553538_s_at	2	COX1
224373_s_at	3	ND4, Hnrnpm, DCAF6
1555653_at	4	HNRNPA3
1553588_at	5	ND3, SH3KBP1
201492_s_at	6	RPL41
212788_x_at	7	FTL
203540_at	8	GFAP
200095_x_at	9	RPS10
229353_s_at	10	NUCKS1

4.11 Signaling Pathways for AD utilizing TSP

The optimal solution to this particular TSP is the tour among the genes of interest in AD with the largest possible correlation is shown in Figure 17 .For this particular case there are a total of (10-1)! ≈ 362,880 ways in which a cyclic path can be drawn among the 6 genes, the optimal path was found in this research.

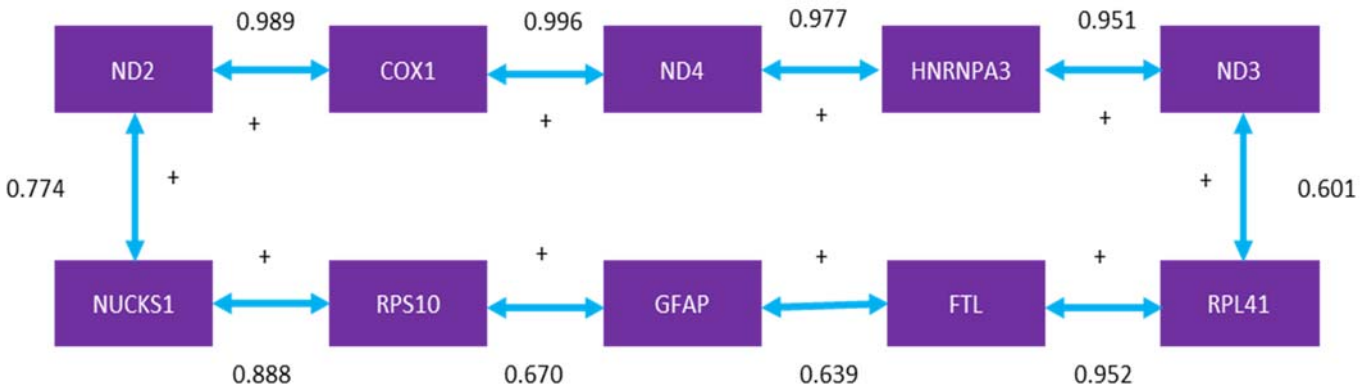


Figure 17: Optimal solution path for potential expressed genes

4.12 Signaling Pathways for AD utilizing MST

MST was applied to obtain a tree that maximizes the linear correlations of the ten genes that were identified previously as potential differentially expressed genes. The MST graphical representation is shown in Figure 18.

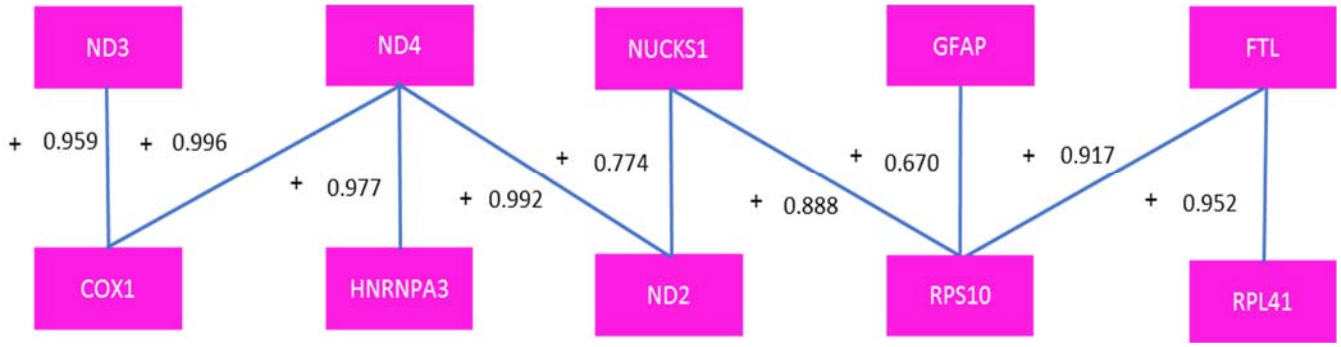


Figure 18: Optimal solution network for potential expressed genes

4.10 Biological Discussion for AD genes

Based on previous work by our research group "Characterization of AD: An Operations Research Approach"[54] the biological interpretation of the genes that changed their expression the most was performed. The interpretation of that work is reviewed in this thesis. From the literature review performed in that work [48], some of the probes recognize more than one gene for example probe 1553588_at recognizes ND3 and SH3KBP1. Probe 224373_s_at recognizes ND4, Hnrnpm and DCAF6. The protein encoded by the MT-DN4, or mitochondrially encoded NADH Dehydrogenase 4, forms part of the core subunit of the mitochondrial NADH dehydrogenase (Complex I). The mRNA expression for this gene has been reported to decrease in the hippocampus and inferior parietal lobule of AD patients DCAF6 is related to androgen receptors which are a path way to AD, this is proof that the method can move past the intermediary gene, as well as in the temporal cortex of AD patients. Mutations of this gene have been also link to maternally inherited schizophrenia. Drosophila ND2 mutants show progressive neurodegeneration. A study for the expression of mitochondrial ND2 and ND4 genes in

amyotrophic lateral sclerosis (ALS), found that the anterior neurons in the cervical spinal cord had reduced mtDNA gene levels and an increment in the amount of mtDNA deletions [51][52]. The next gene, MT-ND3, codes for another member of Complex I. The product of this gene was shown to bind to a peptide corresponding to 25 amino acids of the C-terminal of amyloid-beta [51][92]. The authors of the report proposed that the ND3 Amyloid- beta interaction could explain in part the lower activity of Complex I in astrocytes and neurons.

Next gene, COX3, is also a mitochondrial gene. It encodes a protein that forms part of the cytochorome C Oxidase enzyme complex. In an experiment that was testing the protective effect of melatonin, COX3 was found to have an increment in its expression [54][55]. PTPRO codes for the protein tyrosine phosphatase receptor type O, one of the proteins involved in the development of growth and branch morphology of axonal branches of sensory neurons under development is not directly linked to AD but it has been reported that it controls negative PH receptors which have been coined recently as to being one of the causes of Alzheimer's Disease [56].

NUCKS1 codes for nuclear casein kinase and cyclin-dependent kinase substrate 1. Its product has been shown very recently to participate in homologous recombination DNA repair[57][58]. NUCKS1 gene product is also used by the HIV-1 for the viral transcription of its genetic material and has been reported as a biomarker for some cancer [59][60].The expression the NUCKS1 gene has a link with mood disorders, It is related to Parkinson's disease [60][63].

RPS10 gene codes for one of the proteins of the 40S ribosomal subunit. The expression of this gene was found to be lower in Schizophrenia patients than in controls [65]. Changes in expression

of this gene has been observed in colorectal cancer [66]. Mutations in the RPS10 are link to diamond-blackfan anemia [68]. RPS10 is part of the ribosomal protein family that has been reported to change its expression in neurodegenerative diseases such as AD [67].

FTL gene codes for the ferritin light polypeptide protein. Ferritin is the main intracellular iron storage protein. It has been reported that levels of ferritin are lower in the peripheral blood mononuclear cells from AD patients, and it has been proposed that this change is one of the factors responsible for the dysregulation of iron found in AD patients .This gene is associated with neurodegenerative disorder associated with iron accumulation in the brain, primarily in the basal ganglia and enhances oxidative damage [71][72].

The Ribosomal Protein L41, encoded by RPL41, has been suggested to play roles in cell proliferation and differentiation during neurogenesis..This protein was found also to help with virus replication in some avian viruses: infectious bursal disease virus[47] and Sindbis virus [73]. It also promotes the expression of the c-myb proto-oncogene [74]. The mutation in this gene is responsible for Alexander disease (a rare disorder of the central nervous system), to leukodystrophy and AD. The disease causes the destruction of myelin [76].

MEG3 is a long non-coding RNA tumor suppressor [77]. Experiments in mice suggest that MEG3 could participate in the early development of the central nervous system and problems with it may be related with cortical malfunctions. MEG3 [80] has been found to be downregulated in Huntington´s disease [81].

Grouping the genes that code for proteins that form part of a bigger complex, there are ND4, ND2, and ND3, mitochondrial encoded NADH dehydrogenase (Complex I) genes. Another mitochondrial encoded gene is COX3 and its product also forms part of one of the electron transport complexes. The correlations between the expression changes for these genes is high, more than 0.9. These results coincide with the reports of mitochondrial genes expression change in AD [83][85]and other neurodegenerative diseases [82][95].There are two genes that code for different ribosomal proteins: RPS10 and RPL4, with a correlation coefficient of 0.86. in the TSP they are separated by FTL. Some of the selected genes have been reported to change their expression in different cancers and some are known to help in viral infections.

Similarly on what was done for PD and HD, a search for published pathways was performed in PUBMED, Gene Cards and KEG. The information is detailed in the tables below.

Table 12: Genes reported in PUBMED & Gene Cards in respective literature

AD Pathways: PUBMED & Gene Cards											
	COX1	ND2	ND3	HNRNPA 3	RPS10	RPL41	GFAP	FTL	ND4	NUCKS1	**Total**
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	1	1	1	0	0	0	0	0	1	0	**4**
Alzheimer's Disease	1	0	0	0	1	0	1	0	1	0	**4**
Metabolic Pathways	1	1	1	0	0	0	0	0	1	0	**4**
Apoptosis	1	0	0	0	0	0	0	0	0	0	**1**
Metabolism	1	1	1	0	0	0	0	0	1	0	**4**
Parkinson's Disease	1	1	1	0	0	0	0	0	1	1	**5**
Huntinton's Disease	1	0	0	0	0	0	0	0	0	0	**1**
Gene expression	1	0	0	1	0	1	0	0	0	0	**3**
Effects of nitric oxide	1	0	0		0		0	0	0	0	**1**
Generic Transcription Pathway	1	0	0	0	0	0	0	0	0	0	**1**
mRNA Splicing - Major Pathway	0	0	0	1	0	0	0	0	0	0	**1**
Spliceosome	0	0	0	1	0	0	0	0	0	0	**1**
Viral mRNA Trans lation	0	0	0	0	1	0	0	0	0	0	**1**

Continuation: AD Pathways: PUBMED & Gene Cards											
	COX1	ND2	ND3	HNRNPA 3	RPS10	RPL41	GFAP	FTL	ND4	NUCKS1	**Total**
Metabolism ofamino acids and derivatives	0	0	0	0	1	0	0	0	0	0	**1**
Activation ofthe mRNA upon binding of the cap-binding complex and elFs, and subsequent binding to 43S	0	0	0	0	1	0	0	0	0	0	**1**
Influenza Viral RNA Transcriptio n and Replication	0	0	0	0	1	1	0	0	0	0	**2**
Influenza	0	0	0	0	1	1	0	0	0	0	**2**
Infectious disease	0	0	0	0	1	1	0	0	0	0	**2**
rRNA processing	0	0	0	0	1	1	0	0	0	0	**2**
Viral mRNA Translation	0	0	0	0	0	1	0	0	0	0	**1**
Signaling by GPCR	0	0	0	0	0	0	1	0	0	0	**1**
Neural Stem Cell Differentiati on Pathways and Lineage-specific Markers	0	0	0	0	0	0	1	0	0	0	**1**
Spinal Cord Injury	0	0	0	0	0	0	1	0	0	0	**1**
ERK Signaling	0	0	0	0	0	0	1	0	0	0	**1**

Continuation: AD Pathways: PUBMED & Gene Cards											
	COX1	ND2	ND3	HNRNPA 3	RPS10	RPL41	GFAP	FTL	ND4	NUCKS1	**Total**
Clathrin derived vesicle budding	0	0	0	0	0	0	0	1	0	0	**1**
Vesicle-mediated transport	0	0	0	0	0	0	0	1	0	0	**1**
Binding and Update of Ligands by Scavenger Receptors	0	0	0	0	0	0	0	1	0	0	**1**
Integrated pancreatic cancer pathways	0	0	0	0	0	0	0	0	0	1	**1**
Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds	0	0	0	0	0	0	0	0	0	1	**1**
Cell cycle pathway	0	0	0	0	0	0	0	0	0	1	**1**
Cell proliferation	0	0	0	0	0	1	0	0	0	1	**2**
Schizophrenia	0	0	0	0	1	0	0	0	0	1	**2**
Total	**10**	**4**	**4**	**3**	**9**	**7**	**5**	**3**	**5**	**6**	

Table 13: Genes reported in PUBMED & Gene Cards in respective literature

AD Pathways: KEGG											
	COX1	ND2	ND3	HNRNPA 3	RPS10	RPL41	GFAP	FTL	ND4	NUCKS1	**Total**
Metabolic pathways	1	1	1	0	0	0	0	0	1	0	**4**
Oxidative phosphory lation	1	0	1	0	0	0	0	0	1	0	**3**
Porphyin and chlorophyll metabolism	1	0	0	0	0	0	0	0	0	0	**1**
Biosynthesis of secondary metabolites	1	0	0	0	0	0	0	0	0	0	**1**
Two-component system	1	0	0	0	0	0	0	0	0	0	**1**
Cardiac muscle contraction	1	0	0	0	0	0	0	0	0	0	**1**

Continuation: AD Pathways: KEGG											
	COX1	ND2	ND3	HNRNPA 3	RPS10	RPL41	GFAP	FTL	ND4	NUCKS1	Total
Non-alcoholic fatty liver disease (NAFLD)	1	0	0	0	0	0	0	0	0	0	**1**
Alzheimer's disease	1	0	0	0	0	0	0	0	0	0	**1**
Parkinson's disease	1	1	1	0	0	0	0	0	1	0	**4**
Huntington's disease	1	0	0	0	0	0	0	0	0	0	**1**
Transcriptional misregulation in cancer	0	1	0	0	0	0	0	0	0	0	**1**
Hedgehog signaling pathway	0	1	0	0	0	0	0	0	0	0	**1**
Cell proliferation	0	1	0	0	0	0	0	0	0	0	**1**
HTLV-I infection	0		1	0	0	0	0	0	0	0	**1**
Viral carcinogenesis	0	1	1	0	0	0	0	0	0	0	**2**

Continuation: AD Pathways: KEGG											
	COX1	ND2	ND3	HNRNPA 3	RPS10	RPL41	GFAP	FTL	ND4	NUCKS1	Total
Neomycin, kanamycin and gentamicin biosynthesis	0	1	0	0	0	0	0	0	0	0	**1**
Meiosis - yeast	0	1	0	0	0	0	0	0	0	0	**1**
Prolactin signalling pathway	0	1	0	0	0	0	0	0	0	0	**1**
Polycyclic aromatic hydrocarbon degradation	0	1	0	0	0	0	0	0	0	0	**1**
Microbial metabolism in diverse environments	0	1	0	0	0	0	0	0	0	0	**1**
FoxO signalling pathway	0	1	0	0	0	0	0	0	0	0	**1**
Cell cycle	0	1	1	0	0	0	0	0	0	0	**2**
Wnt signalling pathway	0	1	1	0	0	0	0	0	0	0	**2**
Hippo signalling pathway	0	1	1	0	0	0	0	0	0	0	**2**
Jak-STAT signaling pathway	0	1		0	0	0	0	0	0	0	**1**

Continuation: AD Pathways: KEGG											
	COX1	ND2	ND3	HNRNPA 3	RPS10	RPL41	GFAP	FTL	ND4	NUCKS1	Total
Serotonergic synapse	0	1	0	0	0	0	0	0	0	0	**1**
Measles	0	1	1	0	0	0	0	0	0	0	**2**
MicroRNAs in cancer	0	1	0	0	0	0	0	0	0	0	**1**
Jak-STAT signalling pathway	0	0	0	0	0	0	1	0	0	0	**1**
Mineral absorption	0	0	0	0	0	0	0	1	0	0	**1**
p53 signalling pathway	0	0	1	0	0	0	0	0	0	0	**1**
P13K-Akt signalling pathway	0	0	1	0	0	0	0	0	0	0	**1**
n/a	0	0	0	1	1	1	0	0	0	1	**4**
Total	**10**	**19**	**11**	**1**	**1**	**1**	**1**	**1**	**3**	**1**	

5. Commonalities Analysis

Analyzing the three diseases together, it was interesting to easily notice there are a lot of commonalities between PD, AD and HD. This support the initial idea of this thesis: the genetic behavior of all three diseases has commonalities. Even though the same genes were not repeated among the diseases, it was found that the genes shares common pathways that are already reported in the literature and also, some genes found in a particular disorder were already reported in one of the others disorder under analysis.

The gene COX1 was found optimal in AD and it has been related with PD, HD and AD pathways in the literature. The genes COX1, ND2, ND3, ND4 and NUCKS1 were found in AD and all has been related with PD pathways already. Also COX1, RPS10, ND4 and GFAP were found in AD and they have been reported in AD pathways. In other hands, COX1 was found in AD and have been reported with pathways in HD in literature. RPL21 was found in PD and reported in PD pathways. HBG2 was found in HD and reported in AD pathways and DDX39B was found in PD and reported in AD pathways. From this analysis it can be deduced that COX1 is optimal for all disorders. Four genes (COX1, RPS10, ND4 and GFAP) were found optimal in AD analysis with the proposed methodology in this work and are already reported in AD in the literature. The same case with RPL21 found in PD with this methodology and reported in PD pathways in literature. This shows that the proposed methodology is powerful and successfully identified genes that were reported already using other mythologies. Also some genes were optimal in a particular disease with the methodology proposed in this work and published in one of the other two diseases under discussion. This is presented in Table 21.

Table 15: Common pathways between PD, HD and AD

Common Pathways				
Pathway	**Parkinson's Disease**	**Huntington's Disease**	**Alzheimer's Disease**	**Total Genes**
Metabolic Pathways		HBD	COX1, ND2, ND3, ND4	**5**
Parkinson's Disease	CCL5		NUCKS 1, COXI, ND2, ND3,ND4	**6**
Cell proliferation	DDX39B, ARF3, CCL5, GDI2, THRA, RPL11		RPS10, NUCKS 1	**8**
Apoptosis	ARF3, PTPN21, CCL5,GDI2, THRA, RPL11		COX1	**7**
Metabolism		HBD	COX1,ND2,ND3,ND4	**5**
Influenza	CCL5, DDX39B		RPS10, RPL41	**4**
Infectious Disease		HBA2, RPS2P46, IFTM3	RPS10, RPL41	**5**
Cell Cycle pathway		H3F3A	ND2, ND3	**3**
Cytokine-Cytokine receptor interactor	CCL5	IFITM3		**2**
Alzheimer's disease	DDX39B	HBG2	RPS10, GFAP,ND4, COX1	**5**
Schizophrenia	PTPN21		RPS10, NUCKS 1	**2**
Huntington's Disease			COX1	1

Also it is important to mention that the disorders also share common pathways. This means that some optimal genes in a particular disorder are related in the literature with a pathway and some other genes that were optimal in another disease was also related to the same pathway in the literature. PD and AD share the following pathways: Parkinson's disease, cell proliferation, apoptosis, Alzheimer's disease, schizophrenia and influenza. PD and HD shares the pathways: Cytokine-cytokine receptor interaction and Alzheimer's disease. HD and AD shares: metabolic pathway, cell cycle pathway, Alzheimer's disease, metabolism and infectious disease. It was noted that all three diseases has pathways reported already related to PD.

6. Cost Model

6.1. Costs related in PD, AD and HD available in literature

A cost model focused on patients with a neurological disease like PD, AD and HD was developed in this work. The purpose is to give an idea of the cost of living with a disease like one these. Also the model could be used for planning purposes. A research in the literature was performed to understand what cost were considered in previous studies. According to the Alzheimer Foundation of America [101] in 2010, the average cost care per dementia case was between $41,000 and $56,000.

For PD, incidence increases with age, but an estimated four percent of people with PD are diagnosed before the age of 50 [96]. In previous research it was estimated that the number of people living with PD will double by 2040. The average age when people is diagnose with PD is 60 years old [117]. The average age at death is 81 years old [117].

Medication costs for an individual person with PD average $2,500 a year, and therapeutic surgery can cost up to $100,000 dollars per patient [94]. A model performed by Kowal SL [98] estimates disease prevalence, excess healthcare use and medical costs, and nonmedical costs for each demographic group defined by age and sex. The national economic burden of PD exceeds $14.4 billion in 2010 (approximately $22,800 per patient). In this particular study they stated that he population with PD incurred medical expenses of approximately $14 billion in 2010, $8.1 billion higher ($12,800 per capita) than expected for a similar population without PD. Indirect costs (e.g., reduced employment) are conservatively estimated at $6.3 billion (or close to $10,000 per person with PD). According to APDA (American Parkinson's Disease Association) the average yearly cost associated to a person with PD is $22,800 [99]. This cost includes $2500 medication cost per year. People with PD are more likely to live in nursing homes [97], nursing home care is accounted for people with PD were more than ten times as likely as people who did not have PD [97].

According to the Alzheimer's Association [100], since AD is a progressive disease, the type and level of care needed will change over time. The common care cost for AD provided by the Alzheimer's Association [100] are:

- Ongoing medical treatment for Alzheimer's-related symptoms, diagnosis and follow-up visits
- Treatment or medical equipment for other medical conditions
- Safety-related expenses, such as home safety modifications or safety services for a person who wanders

- Prescription drugs
- Personal care supplies
- Adult day care services
- In-home care services
- Full-time residential care service

The average cost for long-term care services in USA provided by the Alzheimer's Association [98] are:

- $220 per day or $80,300 per year for a semi-private room in a nursing home
- $250 per day or $91,250 per year for a private room in a nursing home
- $3,600 per month or $43,200 per year for basic services in an assisted living facility
- $20 per hour for a home health aide
- $59 per day for adult day services

Among people age 70, 61 percent of those with AD are expected to die before the age of 80 compared with 30 percent of people without AD — a rate twice as high. Of the 5.4 million Americans with AD, an estimated 5.2 million people are age 65 and older, and approximately 200,000 individuals are under age 65 (younger-onset AD).One in nine people age 65 and older has Alzheimer's disease [102] They are 28 percent more likely than other adults to eat less or go hungry

because they cannot afford to pay for food. At the same time, many survey respondents had misconceptions about what expenses Medicare and Medicaid cover, leaving them unprepared to handle the tremendous costs associated with the disease [102]. AD is the most expensive disease in America, costing more than cancer and heart disease [10].

Nearly 800 people with mild Alzheimer's disease, about half of whom went to the hospital over the course of the study period due to falls, infections or other problems [109]. Being hospitalized was associated with nearly twice the likelihood of having a poor outcome, including mental decline and death, and being delirious while hospitalized increased the risk by about 12 percent. "Delirium can be quite a problem for patients even with mild Alzheimer's disease, and preventing it may be a more effective treatment strategy than the current medications," said Dr. Tamara Fong, an assistant professor of neurology at Harvard Medical School and lead author of the study [109].

The Alzheimer Association [109], stated that people with Alzheimer's disease are three times as likely to spend time in the hospital. Between 20 percent and 40 percent of AD patients are hospitalized each year for an average of about four days, the study authors noted.

For HD, Divino [104] presented patients were classified by disease stage (Early/Middle/Late) by a hierarchical assessment of markers of disease severity, confirmed by literature review and key opinion leader input. Costs were measured over the follow-up time of each patient with total costs per patient per stage annualized using a patient-year cost approach. According to Parkinson Association [104], the mean total annualized cost per patient increased by stage (average of commercial $4,947- $22,582 and Medicaid: $3,257- 37,495). Outpatient costs were the primary

healthcare cost component. The vast majority (73.8%) of Medicaid Late stage patients received nursing home care and the majority (54.6%) of Medicaid Late stage costs were associated with nursing home care. In comparison, only 40.6% of commercial late stage patients received nursing home care, which contributed to only 4.6% of commercial late stage costs [104].According to this study [104], the annual direct economic burden of HD is substantial and increased with disease progression. More late stage Medicaid HD patients were in nursing homes and for a longer time than their commercial counterparts, reflected by their higher costs (suggesting greater disease severity). The average cost of HD could be decreased with Medicaid savings as much as 20 percent [106].

Physical therapy is popular in PD treatment. It cannot cure PD, because neurological damage cannot be reversed. But therapy can help the person compensate for the changes brought about by the condition. These "compensatory treatments," include learning about new movement techniques, strategies, and equipment. A physical therapist teach exercises to strengthen and loosen muscles. Many of these exercises can be performed at home. The goal of physical therapy is to improve independence and quality of life by improving movement and function and relieving pain. For patients covered by health insurance, out-of-pocket costs typically consist of a co-pay of $10-$75 per session or coinsurance of 10%-50% or more. Physical therapy typically is covered by health insurance when medically necessary [115].

Physical therapy can help with: balance problems, lack of coordination, fatigue, pain, gait, immobility and weakness [114]. Treatments in physical therapy often can be completed in one to

three office visits. The first appointment includes an evaluation and recommendations for exercises. The following appointments check your progress and review and expand your home program. Most hospitals can provide additional sessions of outpatient therapy if needed [114].People with Parkinson's disease (PD) typically visit their neurologist two to four times a year [114] and according with Alzheimer's Association [109] the average fee to visit a neurologist is $15 with a medical insurance and around $250 without medical insurance.

In Puerto in specific, one of the most popular health insurance is Triple S [107], using the information from the Triple S official website [109] the monthly rate of a health insurance coverage varies from $330 to $480. Selecting one of the available medical insurance plans in Triple S, the specialist fee goes from $15-$20, per visit and the hospitalization fee is in average $75 per night. Beneficiaries with dementia are in the hospital 3.4 times more often than other elderly beneficiaries, totalizing around seven times per year [110].

According to "Tu Cubierta" web page [105] the average fee of an emergency room with a medical insurance coverage is $15 per visit. If laboratory or radiology is needed then the fee will be the 30% of the cost [109]. In 2013, the emergency room visits per patient was 1.01, and people with dementia usually goes 3.4 times more [110], this gives an average of 3.43 visits to emergency room per year. Also in 2003, 17% of people went one time per year to emergency room, 4%, two times and 6% three times or more [112]. Blood work pricing at a laboratory can range anywhere from $100 for one simple test, to $3,000 for several complex tests. On average, to get blood work done at a lab when the patient is uninsured will cost around $1,500 [113]. Some medical insurances cover 70 % of this cost and the insured person only pays 30% of the cost [109].

6.2. Costs variables and model definition

With all the information related of the high yearly cost of the diseases discussed in this work, and taking in consideration that the cost varies with the disease's stage, this work focuses in a cost model. This cost model could be used by any patient or patient's relative for planning and budget management purposes. The projection of the number of affected patients are increasing, so it is important the study of the diseases' cost. The cost also depends on the stage of the disease, for convenience the cost model will be adapted for three stages using severity as reference (Low, Medium, and High) similar to the severity hierarchy explained by Divino [104]. The cost model is compounded by individual cost, shown in a cost diagram in Figure19.

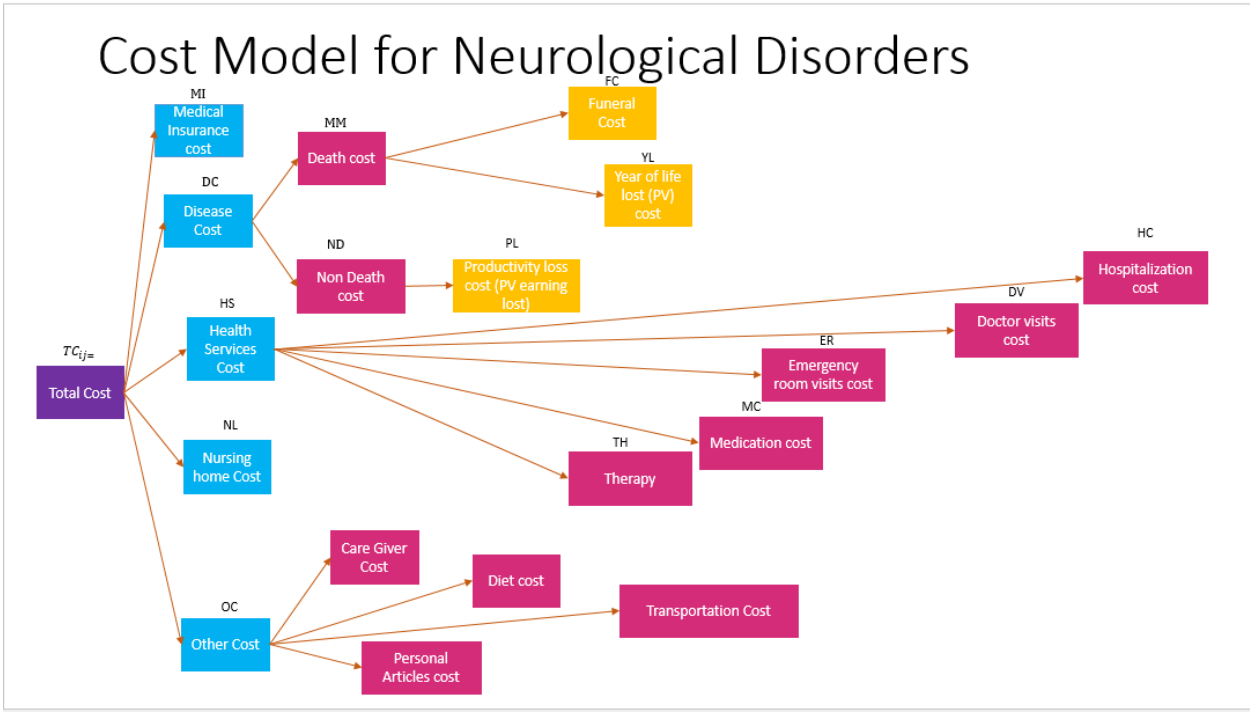


Figure 19: Cost Model for Neurological Disorders Representation

Cost Variable definitions:

i=1,2,3, type of disease (1=PD, 2=HD, 3=AD)

j=1,2,3 Stage of disease I (1=Low, 2=Medium, 2=High)

$TC_{ij=}$ Total Cost of disease i in stage j

MI= Cost of medical insurance

DC= Cost of a patient with the disease

MM= Cost of a patient death

FC= Cost of Funeral Services

YL= Annual Cost calculated from Present value (PV) of Year of life lost (opportunity cost). Years of life lost caused by the disease

- AE=Average earning
- YL-Year lost

ND=Cost of living with the disease

PL=Productivity Loss, Present value (PV) of earning lost due to productivity reduction

- AE=Average earnings
- AD=Day non Productive (Absence days in work)

HS=Health services cost

HC=Hospitalization Cost

HD=Days spend in Hospital

DV=Doctor Visit cost

TD=Times visited doctor office

ER=Emergency Room Visit

TE=Times visited Emergency Room

MC=Medication and prescription cost

TH=Therapy cost

TT=Times in therapy

NL=Nursing Home Cost

OC=Other Costs including diet, transportation, caregiver, personal articles, walkers, x-rays, laboratories, electronic beds, etc.

The cost model is then the sum of all cost of the disease i in stage j.

$$\boldsymbol{TC_{ij=}MI+DC+HS+NL+OC} \quad (11)$$

Where:

$$\boldsymbol{DC=MM+ND} \quad (12)$$

$$\boldsymbol{MM=FC+(YL)} \quad (13)$$

$$\boldsymbol{YL} = \left(\frac{\boldsymbol{CF}}{\boldsymbol{(1+r)^n}}\right)\left(\mathbf{1} - \frac{\frac{\mathbf{1}}{\boldsymbol{(1+r)^n}}}{\boldsymbol{r}}\right) \quad (14)$$

- Where:

  YL=Present Value

  CF= cash flow in future period

  r = the periodic rate of return or interest

  n = number of periods

$$\boldsymbol{PL = \frac{CF}{(1+r)^n}} \quad (15)$$

$$\boldsymbol{AC = 1 - \frac{\frac{1}{(1+r)^n}}{r}} \quad (16)$$

$$\boldsymbol{HS=HC(HD)+DV(TD)+ER(TE)+MC+TC(TT)} \quad (17)$$

As it was stated at the beginning of this document each disorder has several stages. For PD, there are five stages [10]. For convenience in this cost model the stages will be grouped as Low (stage 1 and 2), Medium (stage 3) and High (stage 4 and 5).This stages were similar as explained in [104] by Divino, where patients were classified by disease stage (Early/Middle/Late) by a hierarchical assessment of markers of disease severity. In the Low stage some symptoms appear [106], for example tremor, movement and walking problems, poor posture, etc. This stage would mainly require several medical appointments. For the Medium stage, [106] the loss of balance is one of the main symptoms, resulting in falls that could end in several Emergency Room visits and some medication cost. The patients have problems eating, getting dressed and performing daily activities. This represent a higher cost in emergency room visit and medical appointments, since the frequency of visits increases. Also this stage includes cost of care giver, other costs (diet, transportations, and personal items) and therapy costs. For the stage High [106], the person has hallucinations and dilutions, it is impossible to stand and walk , they need help with daily living

and they can't live alone, eventually the person could end confined to bed. In this stage the cost include nursing cost as a major cost. Also all the other cost are present in this stage. In all stages the medical insurance cost may be added if the patient has it, this would decrease cost of Emergency room visits, medical appointments, hospitalization, medication, therapy, etc. Also the opportunity cost, loss of productivity is present in all stages and it would increase proportional to the stage. The death cost (funeral and year of live lost) could be present at any stage.

For HD there are 3 stages, each stage will be assigned to Low, Medium and High as explained before. In Low stage [106], the major symptoms are depression, involuntary movements and changes in coordination, the person starts to be less able to work. In this stage medical appointments and productivity loss cost are easily present. In Medium stage [104], all the ordinary occupational and psychical activities are harder to do. The productivity loss cost increases, therapy (occupational and physical) and care giver cost are evident as well as other costs. Medication cost is also present in this stage. In High stage [106], they are not able to talk or walk. They are totally dependent on others so the nursing cost is the main cost in most cases in combination with all other cost discussed. For all stages medical insurance could be present, decreasing another costs mentioned. Opportunity cost, loss of productivity is present in all stages and it would increase proportional to the stage. The death cost (funeral and year of live lost) could be present at any stage.

For AD, The National Alzheimer's Association and the National Institutes of Health identified seven stages of this disease [110]. For convenience and for this cost model this seven stages will be reduced to Low, Medium and High. In the first stage [110], people do not present any sign of having the disease and in the second stage people start to forget words or the location of everyday objects, but the disease cannot be detected. It is in the third stage when people and doctors can

detect the person's difficulties, for example performing tasks, remembering names, losing objects, among others. These first three stages will be grouped as Low. The mainly cost present in this Low stage are: Medical Appointments (Doctor Visits), medical insurance, productivity loss and medication cost. In the fourth stage, people do not have the ability to perform complex tasks, mental challenges or even paying bills. In the fifth stage, people have gaps in memory and thinking, they cannot remember their own phone number or address, and they also may need help to choose clothes. This two stages will be grouped as Medium Stage, the main costs that are present in this Medium stage are: Other Costs (care giver mainly), increased Productivity Loss, Medical Insurance, Doctor Visits , Hospitalizations and some Therapy Cost.

The sixth stage [110] is when the person experiments personality changes, can lose awareness of recent experiences, and tend to wander or become lost, among others. The last stage when the person loses the ability to carry a conversation, needs help with most of their daily personal care and the muscles grow rigid. This two stages will be grouped together in the High Stage. The main costs related with the high stage are: Nursing Home Cost, Medication Cost, Therapy Cost, Other Costs, etc. Depending on each particular case the costs mentioned may or may not be included as well as all the other cost presented in this model.

Also, the opportunity cost, loss of productivity is present in all stages and it would increase proportional to the stage. The death cost (funeral and year of live lost) could be present at any stage. The Medical Insurance Cost may or may not be present in all stages (depending if the person have a government or private medical plan).

6.3. Model Verification

In order to verify the cost model, the literature available estimates will be used to calculate the cost of a patient in the Medium Stage of PD, as a case study. The total cost calculated with the proposed model will be compared with yearly estimates provided in other similar studies. For validation processes, it will be assumed that the person has daily visits of a care giver, since in this stage they cannot perform every task alone, the person has a medical insurance and the age of the person is 60 years old (average age provided by Parkinson's National Foundation [117]).

The cost used to verify the model will be discussed and compared with recent studies estimates. The Medical Insurance Cost [109] used was the average between the monthly payments (average between $330 and $480= $405). The medication cost per year used is $2,500 as stated in [96]. The Hospitalization Cost used was $2,100 per year, using 4 days per hospitalization [109], 7 times per year [109] with the average hospitalization fee of $75 per night [110]. The cost of emergency room was calculated using the average fee of an emergency room with a medical insurance coverage- $15 per visit and 3.43 visits per year in average [110], totalizing $50.10. The doctor visits cost for insured people is $15 per visit [113], in average 4 visits per year [116], with a yearly cost of $60.00.The radiology cost (OC) is between $138-$265 (excluding esophagus radiology ($595) and Usi w/ air w/sm bovel ($1139))[113]. Supposing a patient needs x-rays in at least one of the Emergency room visits, then the average radiology cost would be $201.5. Supposing that at least one blood laboratory (OC) is needed then the cost will be the average laboratory cost of $1500 per 0.30 [113], that is percentage that the insured person is responsible for according to [109]. Personal

articles cost discussed in [119] includes: equipment costs to make a home comfortable and safe - electric beds ($3,000 to $5,000), walkers ($100 to $450), bath lift ($1,200) and ramps ($200 to $8,000) If we used the cheaper estimates the personal articles cost are $4,500 plus cost of adult dippers that the estimated cost is $2,160 per year. The equipment cost could be used and bought in any stage, and in different years during the life of the person, so it was not included in this specific case. Only the diapers cost were included. In this case, using a daily caregiver minimizes transportation cost, so it was not included in this specific example. A specific diet may be or may not be required, $20 per day is the average cost of diet of any patient with a dementia disorder [121], totalizing $7,300 per year.

Physical therapy cost for insured people goes from $10 to $75 [113]. In average [115], two visits to the office to get therapy is enough for PD patients. The total cost is the average fee per 2 visits per year, $127.50. To calculate nursing home cost per year, $59 per day was used [100], excluding weekends, this gives a total cost of $15,340 per year.

The Productivity Loss (YL) was calculated using an annual interest pay per period of 0.75%, this is the USA rate in 2016 presented by Trading Economics website [123] and presented in equation (14). The first approach was to use Present Value (PV) using the PMT (payment) as the average monthly salary in PR in 2016 ($2,244.58) [119]. The periods used was 36 months, since the average age of PD patients is 60 years old [117] and the average retirement age in all USA states and PR is 63 years old [118]. This means that the person will loss at least 3 years of working/productive years (36 months). The yearly interest pay per period was converted to

monthly (0.00625%). The PV cost is $72,158.68, the equation was presented in equation (16). Then in order to annualize the PV the Annual Cost (AC) calculation was used, this was presented in equation (15).

These costs were used as examples, the total cost could change depending each specific patient and the stage of the disorder. The medical insurance could also have influence, the use of nursing homes or care givers and surgeries interventions could also affect the cost.

Comparing this cost with other studies, in [101] the average cost for people with dementia disorders were between $41,000 and $56,000 per year. The study performed in [99], only including nursing home and medication cost estimates $22,800 per year. In [98] the cost of AD was $43,200 per year for basic services in an assisted living facility including medical treatment and equipment, prescription drugs and personal care supplies. The model proposed by this work considers the cost in table 15, and total cost, including an opportunity cost- productivity loss (PL) is $62,898.13. This cost is higher than the others but this consider Productivity Loss cost that was neglected in the other studies. If we ignore this additional cost to compare with the studies mentioned then the model TC would be $35,150.45.

Table 15: Costs considered in Model Validation

Type	Cost Calculation Details	Formula	Yearly Cost
MI	MI=(Average Monthly Fee for Medical Insurance)x 2	MI=12*((330+480)/2)	$ 4,860.00
MC	MC=Average Medication Cost per year	MC=2500	$ 2,500.00
HC	HC=Average Hospitalization Fee/night x(Average days per hospitalization)x Hospitalization times/year	HC=75x7x4	$ 2,100.00
DV	DV= (Doctor Visit Average Fee) x (Average Visits/year)	DV=4x15	$ 60.00
ER	ER= (Emergency Room Average Fee)x Average visits/year	ER=3.43x15	$ 51.45
TH	TH= (Average therapy cost)x Average therapist visits	TH=((10+75)/2)x3	$ 127.50
OC (Care Giving)	OC=(Care giver cost/day)x days/year	OC=59x52x5	$ 15,340.00
OC(Labs)	OC=Laboratory Cost x Percentage not covered by Insurance	OC=1500x0.3	$ 450.00
OC(X-Rays)	OC=Average radiology costs	OC=(13+265)/2	$ 201.50
OC (Personal Articles)	OC=Diapers cost	OC=2,160	$ 2,160.00
OC(Diet)	OC=Average Diet cost/day x (days/year)	OC=20 x 365	$ 7,300.00
YL	YL(Present Value) =PV(interest rate, number of payments, payments amount) YL (Annual Cost=PMT(interest rate, number of periods, amount)	PV(0.00625,36,-2244.58)= -72,158.56 PMT(0.075,3,-72,158.56)	$ 27,747.68
TC	**TC=MI+MC+HC+DV+ER+TH+OC+YL**	**TC=4860+2500+2100+60+ 51.45+127.5+15340+450+ 201.5+2160+7300+27747. 68**	**$ 62,898.13**

7. Conclusions

This work proposes the study of commonalities among three important neurological disorders: PD, AD and HD. The interrelations of the genes that change their expression the most can be enhanced using optimization techniques. The purpose of the proposed method is the detection of a biological pathway, as a combinatorial problem similar to the Traveling Salesman Problem and Minimum Spanning Tree. This implies an optimal solution exists. It 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 characterization of PD, AD and HD.

The analysis has been applied for PD, HD and AD. For PD results of 8 genes that change their expression the most were obtained, some of them related with inflammation, apoptosis, neurological disorders, Parkinson's disease, mitochondria and proliferation. The same analysis was replicated to HD obtaining results of 6 genes. Most of them related to viral infections, metabolism, hemoglobin and Alzheimer's disease. Some genes were also found in Autism using similar analysis in a work from my research group. In AD the analysis was performed and 10 genes were found. Most of them related with metabolic pathways, cell peliferation, AD, PD and HD. This analysis proposes genes that has not been explored in PD, AD and for HD yet and could suggest new experiments.

It was concluded that the three disorders have genetic communalities between them. Some genes from each disorder shares pathways with one or two of the other disorders. Also, some genes found in a particular disorder were already reported in one of the others disorder under analysis.

A cost model was performed to represent the actual average cost that is incurred in patient with a neurological disease. Depending on the stage of the disorder this cost may vary. The main costs include: medical insurance, disease cost, health services costs, nursing and other costs. Using the proposed formulation and costs already available in the literature the yearly average cost of a neurological disorder is $62,863.21.

As future work a biological comparison between MST and TSP results will be performed. Also the genetics over expression or under expression biological's effects between a pair of genes will be analyzed by our research group. The analysis of using next generation sequencing instead of microarrays as gene expressions measure could be performed by our research group to improve the availability of the data. Medical effects of this results are also intended to discuss and characterize them in the future.

References

[1] Enfermedad de Huntington [Internet]. Estadísticas. [cited 2016Apr26]. Retrieved from: http://enfermedad-huntington-biologiacr.blogspot.com/2010/12/estadisticas.html

[2] Statistics on Parkinson's [Internet]. Parkinson's Disease Foundation (PDF). [cited 2016Apr26]. Retrieved from: http://www.pdf.org/en/parkinson_statistics

[3] Alzheimer's Statistics [Internet]. Alzheimers.net. [cited 2016Apr26]. Retrieved from: http://www.alzheimers.net/resources/alzheimers-statistics/

[4] Instituto de Estadísticas de Puerto Rico [Internet]. Instituto de Estadísticas de Puerto Rico. [cited 2016Apr26]. Retrieved from: http://www.estadisticas.pr/iepr/

[5] Help End Alzheimer's [Internet]. Alzheimer's Association. [cited 2016Apr26]. Retrieved from: http://www.alz.org/

[6] Scherzer CR, Jensen RV, Gullans SR. Gene expression changes presage neurodegeneration in a Drosophila model of Parkinson's disease. PubMed. 2003Aug:2457–66.

[7] National Parkinson Foundation: Believe in Better [Internet]. National Parkinson Foundation. [cited 2016Apr26]. Retrieved from: http://www.parkinson.org/understanding-parkinsons/what-is-parkinsons

[8] National Parkinson Foundation: Believe in Better [Internet]. National Parkinson Foundation. [cited 2016Apr26]. Retrieved from: http://www.parkinson.org/understanding-parkinsons/what-is-parkinsons

[9] What is Parkinson's Disease? Causes, Symptoms, Treatments [Internet]. WebMD. WebMD; [cited 2016Apr26]. Retrieved from: http://www.webmd.com/parkinsons-disease/tc/parkinsons-disease-topic-overview

[10] National Parkinson Foundation: Believe in Better [Internet]. National Parkinson Foundation. [cited 2016Apr26]. Retrieved from: http://www.parkinson.org/

[11] What Is Huntington's Disease? [Internet]. Huntingtons Desease Society of America What is HD Comments. [cited 2016Apr26]. Retrieved from: http://hdsa.org/what-is-hd/

[12] Huntington's Disease [Internet]. Huntington's Disease. [cited 2016Apr26]. Retrieved from: http://www.huntingtons.org.za/index.php/about/huntington-s-disease

[13] Brookmeyer, R., Gray, S., & Kawas, C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. American Journal of Public Health, 1998: 1337–42.

[14] Clinical Criteria for Alzheimer's Diagnosis | Research Center | Alzheimer's Association [Internet]. Alzheimer's Association. [cited 2016Apr28]. Retrieved from: http://www.alz.org/research/diagnostic_criteria

[15] NCI Dictionary of Cancer Terms [Internet]. National Cancer Institute. [cited 2016Apr28]. Retrieved from: http://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=561720

[16] Biological Pathways Fact Sheet [Internet]. Biological Pathways Fact Sheet. [cited 2016Apr28]. Retrieved from: http://www.genome.gov/27530687

[17] Rosas J., Cabrera M., Isaza C. Biological Signaling Pathways and Potential Mathematical Network Representations: Biological Discovery through Optimization [dissertation]. 2015.

[18] DNA Microarray Technology [Internet]. DNA Microarray Technology. [cited 2016Apr28]. Retrieved from: https://www.genome.gov/10000533/dna-microarray-technology/

[19] J C-K, J K, L F. Gene expression profiling in human neurodegenerative disease. PubMed. 2012Aug14: 518–30.

[20] What is meta-analysis [Internet]. Medicine. [cited 2016Apr28]. Retrieved from: http://www.medicine.ox.ac.uk/bandolier

[21] Definition: 'Differential Gene Expression' [Internet]. Differential Gene Expression. [cited 2016Apr28]. Retrieved from: http://www.medilexicon.com/medicaldictionary.php?t=31078

[22] Differential Gene Expression [Internet]. National Center for Biotechnology Information. U.S. National Library of Medicine; [cited 2016Apr28]. Retrieved from: http://www.ncbi.nlm.nih.gov/books

[23] Soneson C., Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013Mar9:1471–2105.

[24] Smith G. BioC2005 Conference [Internet]. Bioconductor. 2013 [cited 2016Jul16]. Retrieved from: https://www.bioconductor.org/help/course-materials/2005/bioc2005

[25] Lewandowski, N. M., Ju, S., Verbitsky, M., Ross, B., Geddie, M. L., Rockenstein, E. Polyamine pathway contributes to the pathogenesis of Parkinson disease. National Academy of Sciences of the United States of America.2010, 16970–16975.

[26] Hu Y., Chopra V., Chopra R. Transcriptional modulator H2A histone family, member Y (H2AFY) marks Huntington disease activity in man and mouse. Proceedings of the National Academy of Sciences. 2011Mar;108(41):17141–6.

[27] Burg JMVD, Björkqvist M., Brundin P. Beyond the brain: widespread pathology in Huntington's disease. The Lancet Neurology. 2009;8(8):765–74.

[28] Cooper-Knock J., Kirby J., Ferraiuolo L. Gene expression profiling in human neurodegenerative disease. Nature Reviews Neurology Nat Rev Neurol. 2012;8(9):518–30

[29] Soto J.M., Ortuno F.M., Rojas I. Integrative gene expression analysis of lung cancer based on a technology-merging approach. IEEE EUROCON 2015 - International Conference on Computer as a Tool (EUROCON). 2015

[30] Lorenzo E., Camacho-Caceres K.I., Ropelewski A., Rosas J., Ortiz-Mojer M., Perez-Marty L., et al. An Optimization-Driven Analysis Pipeline to Uncover Biomarkers and Signaling Paths: Cervix Cancer. Microarrays. 2015; 4(2):287–310.

[31] Sánchez-Peña M.L., Isaza C.E., Pérez-Morales J., Rodríguez-Padilla C, Castro J.M., Cabrera-Ríos M. Identification of potential biomarkers from microarray experiments using multiple criteria optimization. [Internet]. Cancer medicine. U.S. National Library of Medicine; 2013 [cited 2017Apr23].

[32] Camacho-Cáceres K.I., Acevedo-Díaz J.C., Pérez-Marty L.M., Ortiz M., Irizarry J., Cabrera-Ríos M., et al. Multiple criteria optimization joint analyses of microarray experiments in lung cancer: from existing microarray data to new knowledge. Cancer Medicine. 2015;4(12):1884–900.

[33] Laporte G. A Concise Guide to the Traveling Salesman Problem. The Journal of the Operational Research Society.2010: 35-40

[34] Sanchez-Peña M., Watts-Oquendo E., Isaza C., Cabrera-Ríos M. Potential Colon Cancer Biomarker Search Using More Than two Performance Measures in a Multiple Criteria Optimization Approach. Puerto Rico Health Sciences Journal. 2013: 31(2): 59-63

[35] Camacho-Cáceres, K. I., Acevedo-Díaz, J. C., Pérez-Marty, L. M., Ortiz, M., Irizarry, J., Cabrera-Ríos, M., & Isaza, C. E. (2015, December). Multiple criteria optimization joint

analyses of microarray experiments in lung cancer: from existing microarray data to new knowledge.

[36] Sánchez-Peña, Matilde L., et al. “Identification of potential biomarkers from microarray experiments using multiple criteria optimization.” Cancer Medicine, vol. 2, no. 2, 2013, pp. 253–265., doi:10.1002/cam4.69.

[37] Cruz-Rivera Y., Lorenzo E., and Ortiz N.J. Models to Lead Genetic Signaling Path Discovery: Preliminary Ideas and Results. Proceedings of the 2013 Biomedical Engineering Society Annual Meeting, Seattle, WA

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

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

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

[40] Mougeot J.L., Bahrani-Mougeot F.K., Lockhart P.B., Brennan M.T. 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: 446-52

[41] Hashler, Hornik.TSP-Insfraestructure for the traveling salesman problem.Journal of Statistical Software (in press)

[42] Zhai Y., Kuick R., Nan B., Ota I., Weiss S.J., Trimble C.L., Fearon, Cho K.R. Gene Expression Analysis of Preinvasive and Invasive Cervical Squamous Cell Carcinomas Identifies HOXC10 as a Key Mediator of Invasion. Cancer Res. 2007, 67:10163-72

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

[44] 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

[45] Grice D.M., Vetter I., Faddy H.M., Kenny P.A., Roberts-Thomson S.J., 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.

[46] Wilting S.M., de Wilde J., Meijer C.J., Berkhof J., Yi Y., van Wieringen W.N., Braakhuis B.J., Meijer G.A., Ylstra B., Snijders P.J., Steenbergen R.D. Integrated genomic and transcriptional profiling identifies chromosomal loci with altered gene expression in cervical cancer. Genes Chromosomes Cancer 2008, 47:890-8905

[47] Using PubMed [Internet]. National Center for Biotechnology Information. U.S. National Library of Medicine; [cited 2016Apr29]. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed

[48] GENECARDS [Internet]. GENECARDS. [cited 2016Sep9]. Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=hbd

[49] KEGG: Kyoto Encyclopedia of Genes and Genomes [Internet]. KEGG: Kyoto Encyclopedia of Genes and Genomes. [cited 2016Apr29]. Retrieved from: http://www.genome.jp/kegg/

[50] HBA2 - Genetics Home Reference [Internet]. U.S National Library of Medicine. U.S. National Library of Medicine; [cited 2016Sep10]. Available from: https://ghr.nlm.nih.gov/gene/hba2

[51] Genes and mapped phenotypes [Internet]. National Center for Biotechnology Information. U.S. National Library of Medicine; [cited 2016Sep15]. Available from: http://www.ncbi.nlm.nih.gov/gene/10410

[52] Dunckley T., Beach T. G., K. Ramsey E., A. Grover, D. Mastroeni, Walker D. G., LaFleur B. J., Coon K. D., Brown K. M., Caselli R.. Gene expression correlates of neurofibrillary tangles in Alzheimer's disease. Neurobiol. 2016 Aging, vol. 27, no. 10, pp. 1359–71

[53] EBI SEARCH [Internet]. EBI SEARCH. [cited 2016Sep9]. Available from: http://www.ebi.ac.uk/ebisearch/search.ebi?db=atlas-genes&t=rps2p46

[54] Cruz Y., Santiago Y., Gonzales V., Isaza C., Cabrera-Rios M. Characterization of Alzheimer's disease: An Operations Research Approach. Proceedings of the 2015 International Conference on Operations Excellence and Service Engineering Orlando, Florida, USA, September 10-11, 2015

[55] Keeney P. M., Bennett J. P. ALS spinal neurons show varied and reduced mtDNA gene copy numbers and increased mtDNA gene deletions. Mol. Neurodegener 2010 vol. 5, p. 21

[56] Munguia M. E., Govezensky T., Martinez R., Manoutcharian K., Gevorkian G.. Identification of amyloid-beta 1-42 binding protein fragments by screening of a human brain cDNA library. Neurosci. Lett., vol. 397, no. 1–2, pp. 79–82

[57] Swerdlow R.H. Mitochondria and cell bioenergetics: increasingly recognized components and a possible etiologic cause of Alzheimer's disease.Antioxid. Redox Signa; 2012 vol. 16, no. 12, pp. 1434–55

[58] Anisimov V. N., Popovich I. G., Zabezhinski M. A., Anisimov S. V, Vesnushkin G. M.. Melatonin as antioxidant, geroprotector and anticarcinogen.. Biochim. Biophys. Acta, vol. 1757, no. 5–6, pp. 573–89

[59] Cissé M., Checler F. Eph receptors: new players in Alzheimer's disease pathogenesis. Neurobiol. Dis 2015 vol. 73, pp. 137–49

[60] Gatto G., Dudanova, P. Suetterlin I., Davies A. M., Drescher U., Bixby J. L., Klein.Protein R. Tyrosine phosphatase receptor type O inhibits trigeminal axon growth and branching by repressing TrkB and Ret signaling. J. Neurosci 2013 vol. 33, no. 12, pp. 5399–410

[61] Parplys A. C., Zhao W., Sharma N., Groesser T., Liang F., Maranon D. G., Leung S. G, Grundt K, Dray E, Idate R., Østvold A. C., Schild D. NUCKS1 is a novel RAD51AP1 paralog important for homologous recombination and genome stability.Nucleic Acids Res 2015 vol. 43, no. 20, pp. 9817–34

[62] Gu L., Xia B., Zhong L., Ma Y., Liu L., Yang L., Lou G. NUCKS1 overexpression is a novel biomarker for recurrence-free survival in cervical squamous cell carcinoma.Tumour Biol 2014 vol. 35, no. 8, pp. 7831–6

[63] Kikuchi A., Ishikawa T., Mogushi K., Ishiguro M., Iida S., Mizushima H., Uetake H., Tanaka K. Sugihara K. Identification of NUCKS1 as a colorectal cancer prognostic marker through integrated expression and copy number analysis.Int. J. Cancer 2013 vol. 132, no. 10, pp. 2295–302,

[64] Liu X., Cheng R., Verbitsky M., Kisselev S., Browne A., Mejia-Sanatana H., Louis E. D., Cote L. J, Andrews H., Waters C, Ford B, Frucht S, Fahn S, Marder K, Clark L.N. Genome-wide association study identifies candidate genes for Parkinson's disease in an Ashkenazi Jewish population. BMC Med. Genet 2011 vol. 12, p. 104

[65] Savitz J., Frank M. B., T Victor T., Bebak M., Marino J. H.,. Bellgowan P. S. F, McKinney B. A., Bodurka J., Kent Teague T., Drevets W. C.Inflammation and neurological disease-related genes are differentially expressed in depressed patients with mood disorders and

correlate with morphometric and functional imaging abnormalities. Brain. Behav. Immun. 2013 vol. 31, pp. 161–71.

[66] Martins-de-Souza M., Gattaz W. F., Schmitt A, Rewerts C., Marangoni S,. Novello J. C, Maccarrone G., Turck C. W., Dias-Neto E. Alterations in oligodendrocyte proteins, calcium homeostasis and new potential markers in schizophrenia anterior temporal lobe are revealed by shotgun proteome analysis. J. Neural Transm 2019 vol. 116, no. 3, pp. 275–89,

[67] Frigerio J. M., Dagorn J. C., Iovanna J. L. Cloning, sequencing and expression of the L5, L21, L27a, L28, S5, S9, S10 and S29 human ribosomal protein mRNAs. Biochim. Biophys 1995 Acta, vol. 1262, no. 1, pp. 64–8

[68] Abbas W., Dichamp I., Herbein G.The HIV-1 Nef protein interacts with two components of the 40S small ribosomal subunit, the RPS10 protein and the 18S rRNA. Virol. J. 2012 vol. 9, p. 103, Jan. 2012.

[69] Doherty L., Sheen M. R, Vlachos A., Choesmel V., O'Donohue M.-F., Clinton C., Schneider H. E.,. Sieff C. A., Newburger P. E., Ball S. E., Niewiadomska E., Matysiak M., Glader B., Arceci R. J., Farrar J. E., Atsidaftos E., Lipton J. M.,. Gleizes. P.-E. Ribosomal protein genes RPS10 and RPS26 are commonly mutated in Diamond-Blackfan anemia. Am. J. Hum. Genet 2010., vol. 86, no. 2, pp. 222–8

[70] Fittschen M., Lastres-Becker I., Halbach M. V, Damrath E., Gispert S., Azizov M., Walter, M., Müller S., Auburger G..Genetic ablation of ataxin-2 increases several global translation factors in their transcript abundance but decreases translation rate. Neurogenetics 2015 vol. 16, no. 3, pp. 181–92

[71] Crespo Â. C. Silva , B., Marques L., Marcelino E., vMaruta C., Costa, A. Timóteo, A. Vilares, F. S. Couto, P. Faustino, A. P. Correia, A. Verdelho, G. Porto, M. Guerreiro, A.

Herrero, C. Costa, A. de Mendonça, L. Costa, and M. Martins, “Genetic and biochemical markers in patients with Alzheimer’s disease support a concerted systemic iron homeostasis dysregulation.,” Neurobiol. Aging, vol. 35, no. 4, pp. 777–85, Apr. 2014.

[72] Li X., Liu Y., Zheng Q., Yao G., Cheng P., Bu G., Xu H., Zhang Y., Ferritin light chain interacts with PEN-2 and affects γ-secretase activity. Neurosci. Lett 2013 vol. 548, pp. 90–4

[73] Maciel P., Cruz V. T., Constante M., I. Iniesta, Costa M. C., Gallati S., Sousa N., Sequeiros J., Coutinho P. Neuroferritinopathy: missense mutation in FTL causing early-onset bilateral pallidal involvement.Neurology 2015 vol. 65, no. 4, pp. 603–5

[74] Barbeito A. G., Garringer H. J., Baraibar M. A., Gao X., Arredondo M., Núñez M. T., Smith M. A., Ghetti B., Vidal R. Abnormal iron metabolism and oxidative stress in mice expressing a mutant form of the ferritin light polypeptide gene..J. Neurochem 2019 vol. 109, no. 4, pp. 1067–78

[75] Vidal R., Miravalle L., Gao X., Barbeito A. G., Baraibar M. A., Hekmatyar S. K., Widel M., Bansal N., Delisle M. B. Expression of a mutant form of the ferritin light chain gene induces neurodegeneration and iron overload in transgenic mice.J. Neurosci. 2018 vol. 28, no. 1, pp. 60–7

[76] Baraibar M. A., Barbeito A. G., Muhoberac B. B., Vidal R. A mutant light-chain ferritin that causes neurodegeneration has enhanced propensity toward oxidative damage. Free Radic. Biol. Med 2012 vol. 52, no. 9, pp. 1692–7

[77] Ueno M., Nakayama H., Kajikawa S, Katayama K., Suzuki K., Doi K. Expression of ribosomal protein L4 (rpL4) during neurogenesis and 5-azacytidine (5AzC)-induced apoptotic process in the rat. Histopathol. 2012 vol. 17, no. 3, pp. 789–98

[78] Chen Y., Lu Z., Zhang L., Gao L., Wang N., Gao X., Wang Y., Li K., Gao Y., Cui H., Gao H., Liu C., Zhang Y., Qi X., Wang X. "Ribosomal protein L4 interacts with viral protein VP3 and regulates the replication of infectious bursal disease virus. Virus Res., vol. 211, pp. 73–8

[79] Green L., Houck-Loomis B., Yueh A., Goff S. P. Large ribosomal protein 4 increases efficiency of viral recoding sequences. J. Virol 2012 vol. 86, no. 17, pp. 8949–58, Sep. 2012.

[80] Egoh A., Nosuke Kanesashi S., Kanei-Ishii C., Nomura T., Ishii.Ribosomal S. protein L4 positively regulates activity of a c-myb proto-oncogene product..Genes Cells, 2010 vol. 15, no. 8, pp. 829–41

[81] Wang A., Xu S., Zhang X., He J., Yan D., Yang Z., Xiao S. Ribosomal protein RPL41 induces rapid degradation of ATF4, a transcription factor critical for tumour cell survival in stress.J. Pathol 2011 vol. 225, no. 2, pp. 285–92

[82] Fayaz S.M., RajanikantmG.K. ATF4: the perpetrator in axonal-mediated neurodegeneration in Alzheimer's disease. CNS Neurol. Disord. Drug Targets 2014 vol. 13, no. 9, pp. 1483–4

[83] Baleriola J., Walker C. A., Jean Y. Y., Crary J. F., Troy C. M., Nagy P. L. Axonally synthesized ATF4 transmits a neurodegenerative signal across brain regions. Cell 2014 vol. 158, no. 5, pp. 1159–72

[84] Heyne H. O., Lautenschläger S., Nelson R., Besnier F., Rotival M., Cagan A., Kozhemyakina R., Plyusnina I. Z., Trut L., Carlborg Ö. Genetic influences on brain gene expression in rats selected for tameness and aggression. Genetics 2014 vol. 198, no. 3, pp. 1277–90

[85] Kim Y.-J., Cho Y.-E., Kim Y.-W., Kim J.-Y., Lee S., Park J.-H. Suppression of putative tumour suppressor gene GLTSCR2 expression in human glioblastomas. J. Pathol 2008 vol. 216, no. 2, pp. 218–24.

[86] Kim J.-Y., Cho Y.-E., Park J.-H. "The Nucleolar Protein GLTSCR2 Is an Upstream Negative Regulator of the Oncogenic Nucleophosmin-MYC Axis" Am. J. Pathol 2015 vol. 185, no. 7, pp. 2061–8

[87] Kim J.-Y., Cho Y.-E., An Y.-M., Kim S.-H., Lee Y.-G., Park J.-H., and Lee S., "GLTSCR2 is an upstream negative regulator of nucleophosmin in cervical cancer.," J. Cell. Mol. Med., vol. 19, no. 6, pp. 1245–52, Jun. 2015.

[88] Yoon J. C., Ling A. J. Y., Isik M., Lee D.-Y. D., Steinbaugh M. J., Sack L. M., Boduch A. N., Blackwell T. K., Sinclair D. A., Elledge S. J. "GLTSCR2/PICT1 links mitochondrial stress and Myc signaling. Proc. Natl. Acad. Sci. U. S. A 2014., vol. 111, no. 10, pp. 3781–6

[89] Ben Haim L., Carrillo-de Sauvage M.-A., Ceyzériat K., Escartin C. Elusive roles for reactive astrocytes in neurodegenerative diseases" Front. Cell. Neurosci 2015, vol. 9, p. 278

[90] Chen Y.-S., Lim S.-C., Chen M.-H., Quinlan R. A., Perng M.-D. Alexander disease causing mutations in the C-terminal domain of GFAP are deleterious both to assembly and network formation with the potential to both activate caspase 3 and decrease cell viability. Exp. Cell Res. 2011 vol. 317, no. 16, pp. 2252–66

[91] Liu Q., Sun S., Yu W., Jiang J., Zhuo F., Qiu G., Xu S., Jiang X. Altered expression of long non-coding RNAs during genotoxic stress-induced cell death in human glioma cells. J. Neurooncol 2015 vol. 122, no. 2, pp. 283–92

[92] Kagami M., O'Sullivan M. J., Green A. J., Watabe Y., Arisaka O., Masawa N., Matsuoka K., Fukami M., Matsubara K., Kato F., Ferguson-Smith A. C., Ogata T. "The IG-DMR and the MEG3-DMR at human chromosome 14q32.2: hierarchical interaction and distinct functional properties as imprinting control centers. PLoS Genet. 2010 vol. 6, no. 6, p. e1000992.

[93] Qu C., Jiang T., Li Y., Wang X., Cao H., Xu H., Qu J., Chen J.-G. Gene expression and IG-DMR hypomethylation of maternally expressed gene 3 in developing corticospinal neurons. Gene Expr. Patterns vol. 13, no. 1–2, pp. 51–6

[94] Johnson R. "Long non-coding RNAs in Huntington's disease neurodegeneration.," Neurobiol. Dis., vol. 46, no. 2, pp. 245–54, May 2012.

[95] Francis B. M., Yang J., Song B. J., Gupta S., Maj M., Bazinet R. P., Robinson B., Mount H. T. J. Reduced levels of mitochondrial complex I subunit NDUFB8 and linked complex I + III oxidoreductase activity in the TgCRND8 mouse model of Alzheimer's disease. J. Alzheimers. Dis 2014 vol. 39, no. 2, pp. 347–55, Jan. 2014

[96] Bhat A. H., Dar K. B., Anees S., Zargar M. A., Masood A., Sofi M. A., Ganie S. A. "Oxidative stress, mitochondrial dysfunction and neurodegenerative diseases; a mechanistic insight. Biomed. Pharmacother. = Biomédecine pharmacothérapie 2015 vol. 74, pp. 101–10

[97] Brockington A., Heath P. R., Holden H., Kasher P., Bender F. L. P., Claes F., Lambrechts D., Sendtner M., Carmeliet P., Shaw P., J. Downregulation of genes with a function in axon outgrowth and synapse formation in motor neurones of the VEGFdelta/delta mouse model of amyotrophic lateral sclerosis.BMC Genomics 2010 vol. 11, p. 203

[98] Understanding Parkinsons [Internet]. Parkinson's Disease Foundation. [cited 2017Jan15]. Available from: http://www.pdf.org/

[99] Dollar Cost of Parkinson's Underscores Need for Research [Internet]. Parkinson's Disease Foundation (PDF) - Hope through Research, Education and Advocacy. [cited 2017Jan15]. Available from: http://www.pdf.org/en/science_news/release/pr_1363095060

[100] S.L. K., T.M. D. The current and projected economic burden of Parkinson's disease in the United States. PubMed [Internet]. 2013Mar [cited 2017Jan15];:311–28. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23436720

[101] American Parkinson Disease Association. N.p., n.d. Web. 15 Jan. 2017.

[102] "Planning for Care Costs | Caregiver Center." Alzheimer's Association. N.p., n.d. Web. 15 Jan. 2017.

[103] "Care Costs." Alzheimer's Foundation of America - Alzheimer's Disease and Caregiving Support. N.p., n.d. Web. 15 Jan. 2017.

[104] "Care Costs." Alzheimer's Foundation of America - Alzheimer's Disease and Caregiving Support. N.p., n.d. Web. 15 Jan. 2017.

[105] "Latest Alzheimer's Facts and Figures." Latest Facts & Figures Report | Alzheimer's Association. N.p., 29 Mar. 2016. Web. 15 Jan. 2017.

[106] "Costs of Alzheimer's to Medicare and Medicaid." Alzheimer's Association - Alzheimer's Association. N.p., n.d. Web. 15 Jan. 2017.

[107] Divino V., Dekoven N., and Warner J.H. "The direct medical costs of Huntington's disease by stage. A retrospective commercial and Medicaid claims data analysis." PUBMED (2013): 1043-050. Web. 15 Jan. 2017. "The costs associated with the Huntington's Disease parity act" CQ Roll Call. N.p., n.d. Web. 15 Jan. 2017.

[108] Divino, Victoria, Mitch Dekoven, and John Howard. "See all › 5 Citations See all › 24 ReferencesShare Request full-text The Direct Medical Costs of Huntington's Disease by Stage. A Retrospective Commercial and Medicaid Claims Data Analysis." ResearchGate - Share and discover research. N.p., June 2013. Web. 15 Jan. 2017.

[109] "Compara los principales Planes Médicos de Puerto Rico." TuCubierta. N.p., n.d. Web. 15 Jan. 2017.

[110] "Elige un Plan Médico." Triple-S Salud. N.p., n.d. Web. 15 Jan. 2017.

[111] Alzheimer's Disease and Chronic Health Conditions: The Real Challenge for 21st Century Medicare [Internet]. Alzheimer's Association. [cited 2017Jan18]. Available from: https://www.alz.org/

[112] Storrs C. Hospitalization May Be 'Tipping Point' for Alzheimer's Decline [Internet]. Consumer HealthDay. 2012 [cited 2017Jan18]. Available from: https://consumer.healthday.com/cognitive-health-information-26/alzheimer-s-news-20/hospitalization-may-be-tipping-point-for-alzheimer-s-decline-665866.html

[113] "Emergency department visit rate (per 1,000 beneficiaries)." Healthindicators. N.p., n.d. Web. 15 Jan. 2017.

[114] Gallup, Inc. "One in Four Americans Visited Hospital Emergency Room in Past Year." Gallup.com. N.p., 09 July 2003. Web. 15 Jan. 2017.

[115] "Patient Price List." HonorHealth. N.p., n.d. Web. 15 Jan. 2017. [111] http://www.walkinlab.com

[116] "Parkinson's Disease: Physical and Occupational Therapy." WebMD. N.p., n.d. Web. 15 Jan. 2017.

[117] "How Much Does Physical Therapy Cost? - CostHelper.com." CostHelper. N.p., n.d. Web. 15 Jan. 2017.

[118] "Making the Most Out of Your Visit to the Neurologist: A Check-up Checklist." Making the Most Out of Your Visit to the Neurologist: A Check-up Checklist - Parkinson's Disease Foundation (PDF). N.p., n.d. Web. 15 Jan. 2017. <http://www.pdf.org/en/yy_doctor_checklist>.

[119] "Believe in Better." National Parkinson Foundation. N.p., n.d. Web. 15 Jan. 2017.

[120] Wallace, Nick. "The Average Retirement Age in Every State in 2015." SmartAsset. N.p., 19 Oct. 2016. Web. 15 Jan. 2017.

[121] "Estadisticas de Puerto Rico." Trabajo PR. N.p., n.d. Web. 15 Jan. 2017.

[122] Leong, Melissa. "The unexpected costs of caring for your elderly parents." Financial Post. N.p., n.d. Web. 15 Jan. 2017.

[123] "How Much Should I Budget for Adult Diapers?" How Much Should I Budget For Adult Diapers? | iDiaper.com. N.p., n.d. Web. 15 Jan. 2017.

[124] "Demystifying Psychiatry." Psychology Today. N.p., n.d. Web. 15 Jan. 2017.

[125] United States Fed Funds Rate Forecast 2016-2020 [Internet]. United States Fed Funds Rate Forecast 2016-2020. [cited 2017Jan25]. Available from: http://www.tradingeconomics.com/united-states/interest-rate/forecast

[126] Kumari S., Nie J. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. PLoS One. 2012

Appendix A

A1- Pareto Efficient Frontier Program used to calculate Pareto Efficient Frontier

```
%pareto2criteriosBio.m
%Analisis de frontera Pareto de dos criterios
dataT=load('Group_1_It2.txt'); %Cargar la data
[x,y]=size(dataT); % data completa x=num filas, y=num columnas
data=dataT(:,2:end); %Cargar la data de los criterios
[n,m]=size(data); %n=num filas (k), m=num columnas (j)
c1 = 1000*ones(n,n,m);   % matriz primera condicion
for j=1:m %recorrido para cada criterio
    for a=1:n %recorrido fila
        for b=1:n %recorrido columna
            if data(a,j)==data(b,j) %condicion 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 dos criterios
c2=zeros(n,n);  %matriz segunda condicion
for a=1:n %recorrido para filas
    for b=1:n %recorrido de columnas
        if c1(a,b,1)+c1(a,b,2)==0
            c2(a,b)=1000;
        elseif c1(a,b,1)+c1(a,b,2)==1000
            c2(a,b)=1000;
        elseif (c1(a,b,1)+c1(a,b,2))>=2000;
            c2(a,b)=2000;
        end
    end
end

% Procedimiento para encontrar el conjunto dominado cd, y el 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,:)); %comando para sumar la fila a, : de todas las
columnas
    if sumfila>=2000; % condicion para sacar el 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; %contadores de los datos
disp([round(index') cd]);
disp([round(index') cnd]);


%Graficar el conjuno dominado y el conjunto no dominado
scatter(cd(1:i,2),cd(1:i,3),'bo','linewidth',2); % grafica del conjunto
dominado
%scatter(cd(1:i,2),cd(1:i,3),'m*','linewidth',2); % grafica del conjunto
dominado
hold on  % mantiene el grafico anterior
scatter(cnd(1:j,2),cnd(1:j,3),'r*','linewidth',8); % grafica del conjunto no
dominado
axis([-1 5 -1 6]); % tamanho de los ejes
%xlabel('F1 - Objective 1','fontsize',20,'fontweight','b','color','k');
%ylabel('F2 - Objective 2','fontsize',20,'fontweight','b','color','k');
xlabel('Transformed(|Mean(Case)-
Mean(Control)|)','fontsize',15,'fontweight','b','color','k'); %nombre del eje
x
ylabel('Transformed(|Med(Case)-
Med(Control)|)','fontsize',15,'fontweight','b','color','k'); %nombre del eje
y
title('Pareto Efficient
Frontier','fontsize',18,'fontweight','b','color','k');
%title('Minimizacion case','fontsize',20,'fontweight','b','color','k');
%titulo

%Mostrar el Conjunto no dominado en un notepad, con los datos de
%x1,x2,f1,f2
disp('   Conjunto no dominado    ');
%disp('Gen Accesion       F1(H)       F2(C)');
cnd=cnd(1:j,:);
filecnd = fopen('cndcase1_It2.txt','w');
%cnd=cnd(1:j,:);
fprintf(filecnd,'%6s   %12s  %12s\r\n','Posicion Genes','F1','F2');
fprintf(filecnd,'%6.4f     %12.4f   %12.4f\r\n',cnd');
fclose(filecnd);
```

A2-TSP Matlab Program used

```
%tsp.m
function [costo secuencia matriz_solucion]=tsp(correlaciones)

% Función para encontrar la secuencia con la máxima la suma de las
correlaciones
% Se utiliza programación binaria con partición de subtours

% Argumentos y datos de salida del programa
% correlaciones: matriz de correlaciones de tamaño nxn
% costo: varlor óptimo de la solución
% secuencia: secuencia óptima (tamaño 1xn)
% matriz_solucion: representaci?n de la solución óptima mediante matriz
% binaria

% Los informes de fallos y las sugerencias para mejorar el código pueden ser
% enviadas a rodriguez@ingenieros.com
% Código elaborado por Jesus Andres Rodriguez Sarasty


nnodos=length(correlaciones);

correlaciones=-correlaciones+eye(nnodos)*nnodos*100;
[Aeq beq f]=generador(correlaciones);
A=[];
b=[];
solucion_inicial=bintprog(f,A,b,Aeq,beq);
matriz_solucion=reshape(solucion_inicial,nnodos,nnodos)';
[menor_tour min_tour contador_tour]=identificacion_subtour(matriz_solucion);

secuencia=menor_tour;
matriz=matriz_solucion';
x=matriz(:);
fprintf('Numero subtours \t  Tour m?nimo \n');

if min_tour<nnodos
    while min_tour<nnodos
        disp(contador_tour);
        disp(menor_tour);
        menor_tour2=[menor_tour(2:end),menor_tour(1)];
        matriz_tour=full(sparse(menor_tour,menor_tour2,1,nnodos,nnodos));
        matriz_tour=matriz_tour';
        vector_restriccion=matriz_tour(:);
        A=[A;vector_restriccion'];
        b=[b;min_tour-1];
        x = bintprog(f,A,b,Aeq,beq);
        matriz_solucion=reshape(x,nnodos,nnodos)';
        [menor_tour min_tour
contador_tour]=identificacion_subtour(matriz_solucion);
    end
end
[respuesta longitud]=identificacion_subtour(matriz_solucion);
if longitud==nnodos
    secuencia=respuesta;
else
    display('No se obtuvo solucion')
```

```
end
costo=f'*x;
end


function [Aeq beq f]=generador(correlaciones)

n=length(correlaciones);
nvariables=n*n;
matriz=zeros(n,nvariables);
vector=ones(1,n);
matriz1=zeros(n,nvariables);
for i=1:n
    j=(i-1)*n+1;
    matriz1(i,j:(j+n-1))=vector;
end

for i=1:n
  vector=i:n:nvariables;
  coeficientes=sparse(1,vector,1,1,nvariables);
  matriz(i,:)=full(coeficientes);
end
Aeq=[matriz1;matriz];
beq=ones(2*n,1);
f=correlaciones';
f=f(:);

end


function [menor_tour min_tour
contador_tour]=identificacion_subtour(distancias)

A=distancias;
nnodos=length(A);
completo=false;
min_tour=inf;
restantes=1:nnodos;
contador_tour=0;
while completo==false
 contador_tour=contador_tour+1;
 nnodos=length(A);
 indices=1:nnodos;
[fila columna]=find(A,1);
tour=zeros(1,size(A,1));
tour(1)=fila;
tour(2)=columna;
siguiente=find(A(tour(2),:));
contador_secuencia=3;
while (any(siguiente==tour)==0)
    tour(contador_secuencia)=siguiente;
    siguiente=find(A(tour(contador_secuencia),:));
    contador_secuencia=contador_secuencia+1;
end
```

```
longitud=find(tour,1,'last');

if longitud<min_tour
    min_tour=longitud;
    tour(tour==0)=[];
    menor_tour=restantes(tour);
end

tour(tour==0)=[];
tour_real=restantes(tour);
restantes=setdiff(restantes,tour_real);
indices=setdiff(indices,tour);

if isempty(restantes)
    completo=true;
else
    A=A(indices,indices);
end

end

end
```

Appendix B

I-Minimum spanning tree and TSP method applied to this work example and steps

1) Extract data of relative expression for control and disorder tissues- Table B1

2) Pareto Efficient Frontier to obtain the genes that changes their expression the most- Figure B1

3) Calculate pairwise difference between disorder tissues and control tissues- Table B2

4) Calculate correlation of differences between genes - Table B3

5) Select the most correlated genes lineal/network path

- Method Comparison (Adjacent genes communalities between TSP and MST in each disorder-Tables B4-B6

6) Establish commonalities between PD, AD and HD- Table B7

An example for Parkinson's disease is shown below:

Table B1- Relative Expression of control and Alzheimer's disease tissues (Extract of samples)- This is only a sample to show the process followed

	Parkinson's Tissues								Control Tissues						
ID_REF	GSM 488119	GSM 488121	GSM 488123	GSM 488125	GSM 488127	GSM 488129	Mean Parkinsons	Median Parkinsons	GSM 488111	GSM 488113	GSM 488115	GSM488117	GSM488131	Mean Control	Median Control
AFFX-BioB-5_at	2162.99	11103.4	270.707	1143.74	3271.21	24768.3	7120.057833	3271.21	455.15	1404.95	1983.89	4598.15	792.953	1847.0186	1404.95
AFFX-BioB-M_at	3905.63	21632.9	285.144	1892.79	5048.94	49440.9	13701.05067	5048.94	706.512	2110.51	3325.02	8718.43	1092.8	3190.6544	2110.51
AFFX-BioB-3_at	2261.9	13273.4	157.89	987.834	2653.62	31552	8481.107333	2653.62	348.136	1052.73	1920.3	6283.5	546.263	2030.1858	1052.73
AFFX-BioC-5_at	6682.78	38601	556.883	3054.54	9212.57	83009.2	23519.4955	9212.57	1233.56	3569.3	5694.18	13654.5	1733.73	5177.054	3569.3
AFFX-BioC-3_at	5791.09	29114.7	527.853	3302.69	9182.67	60231.5	18025.08383	9182.67	1157.29	3795.85	5564.71	11331.3	2008.19	4771.468	3795.85
AFFX-BioDn-5_at	13940.3	70544.3	1147.63	7979.72	22040.6	154317	44994.925	22040.6	2163.94	7353.54	12903.1	28337.9	4144.98	10980.692	7353.54
AFFX-BioDn-3_at	32559.6	117643	2671.7	13429.2	33598.6	190331	65038.85	33598.6	9916.3	24809.9	21534.7	44414.6	8138.11	21762.722	21534.7
AFFX-CreX-5_at	55227.8	171971	3409.59	20070.4	45417.2	257086	92196.99833	55227.8	9710.49	31033.5	29978.1	59864.2	12239.5	28565.158	29978.1
AFFX-CreX-3_at	60401	161689	3959.19	20140.5	46911.6	264270	92895.215	60401	14883.3	42507.9	31232.6	61641.2	12709.4	32594.88	31232.6
AFFX-DapX-5_at	38.905	162.18	18.5607	51.4297	105.683	105.111	80.31156667	80.31156667	40.8911	51.3019	94.9411	27.6157	22.9417	47.5383	40.8911
AFFX-DapX-M_at	18.8934	106.697	18.4864	23.1113	90.8538	40.8585	49.81673333	40.8585	6.54427	137.517	31.9783	14.3447	7.47238	39.57133	14.3447
AFFX-DapX-3_at	46.2818	129.005	47.9228	9.30178	140.303	96.8995	78.28564667	78.28564667	115.21	109.422	9.41568	34.0378	12.5667	56.130436	34.0378
AFFX-LysX-5_at	46.9195	212.861	18.8189	19.7122	71.4243	38.8989	68.1058	46.9195	20.3508	7.84472	13.1848	37.1045	15.3913	18.775224	15.3913
AFFX-LysX-M_at	63.3079	42.4143	18.8521	27.1947	127.076	33.2498	52.0158	42.4143	5.72838	37.8953	65.704	55.2062	7.64183	34.435142	37.8953
AFFX-LysX-3_at	13.7668	58.5993	13.4824	110.23	162.007	277.855	105.9900833	105.9900833	45.9831	77.5259	58.1438	9.58607	37.7022	45.788214	45.9831
AFFX-PheX-5_at	2.15659	35.8335	18.494	54.4808	26.3082	62.1438	33.23614833	33.23614833	4.41727	47.296	17.0992	5.15612	17.0649	18.206698	17.0649
AFFX-PheX-M_at	7.10183	36.4759	3.1783	10.0365	36.2602	6.87408	16.65446833	10.0365	31.2316	6.62176	16.5458	2.19077	3.38339	11.994664	6.62176
AFFX-PheX-3_at	22.1622	77.6666	36.0164	25.396	99.1528	188.17	74.76066667	74.76066667	18.1593	31.4929	79.0156	22.5544	63.9842	43.04128	31.4929
AFFX-ThrX-5_at	7.11719	151.039	7.29883	12.5215	13.9975	23.3252	35.88320333	13.9975	49.9272	8.35776	105.722	21.3704	17.2875	40.532972	21.3704
AFFX-ThrX-M_at	4.37134	109.634	1.97002	17.2376	41.609	27.5096	33.72192667	27.5096	15.1678	6.20277	8.77102	36.0744	5.32198	14.307594	8.77102
AFFX-ThrX-3_at	16.3692	16.8281	6.861	10.9011	33.4407	85.199	28.26651667	16.8281	40.0469	34.9572	9.22921	13.4269	8.62801	21.257644	13.4269
AFFX-TrpnX-5_at	11.4245	30.6689	27.1027	49.9285	26.2969	11.6823	26.18396667	26.2969	88.0414	36.6183	32.2498	27.773	9.89596	38.915692	32.2498
AFFX-TrpnX-M_at	12.3962	320.438	6.0268	26.5029	32.1306	23.694	70.19808333	26.5029	9.08208	30.9958	48.9993	7.10897	7.14293	20.665816	9.08208
AFFX-TrpnX-3_at	1.50849	22.1743	19.5209	6.99452	119.049	227.302	66.091535	22.1743	7.50269	5.86959	33.7388	13.2527	26.47	17.366756	13.2527
AFFX-r2-Ec-bioB-5_at	3245.55	16747.4	357.625	1723.74	4046.48	31505.5	9604.3825	4046.48	751.598	1610.23	3280.94	6413.08	991.623	2609.4942	1610.23

AFFX-r2-Ec-bioB-M_at	3601.68	22163.9	214.912	2028.59	5232.21	49252.4	13748.94867	5232.21	781.156	1191.83	3519.9	7829.61	905.391	2845.5774	1191.83
AFFX-r2-Ec-bioB-3_at	2532.19	16779	265.252	1530.49	4464.98	37191.9	10460.63533	4464.98	404.252	1505.14	2686.23	8380.71	1150.28	2825.3224	1505.14

Figure B1- One of the ten Pareto efficient frontier from program "Pareto2CriteriosBio"- This was one of efficient frontiers obtained.

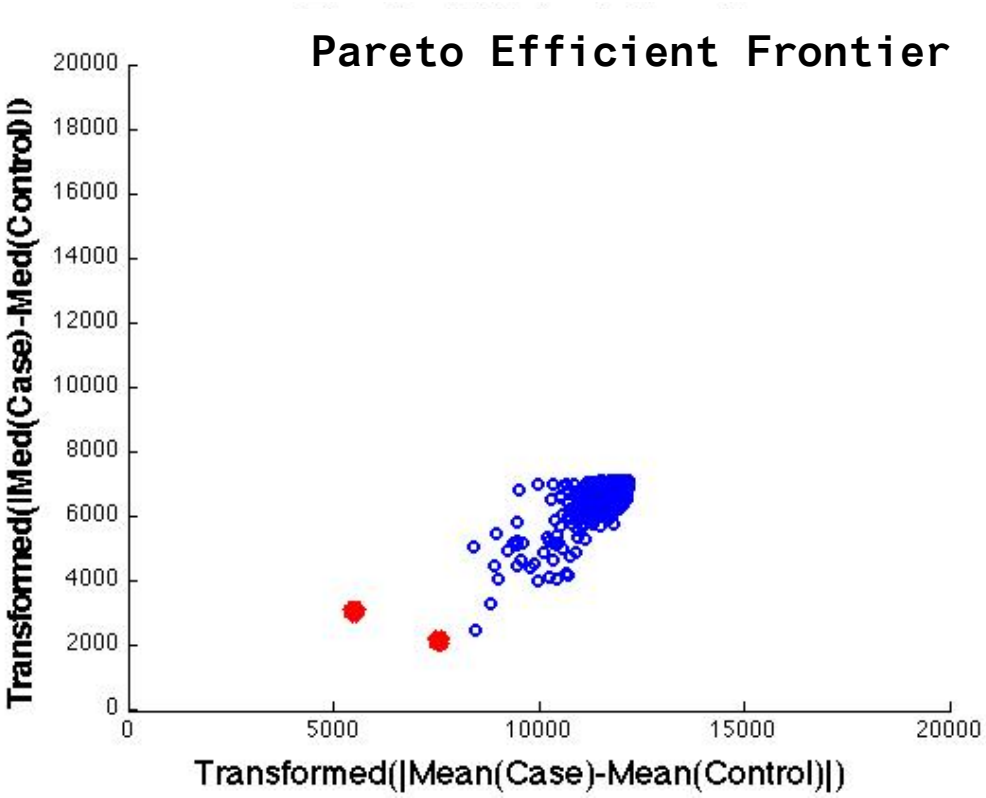


Table B2- Example of Pairwise difference one of the six analysis performed in PD (Used in the sixth Analysis)

Pairwise Differences							
diff1	**diff2**	**diff3**	**diff4**	**diff5**	**diff6**	**diff7**	**diff8**
-175010	-129039	2116.167	-76165.5	-150648	-65250.6	-176592	-57296.5
7794.8	125.2	-83.6	-4337.98	-4356.41	1193.62	6157.7	-1244.8
-49216.4	-30877	431.97	-33612.6	-28706.8	-16342.2	-44917.4	-19760.6
-584334	-576439	1531.19	-514709	-560757	-429636	-664274	-473862
-287287	-276151	598.98	-263269	-307689	-240294	-353640	-237894
-14435.4	-17716	-396.9	-19082.5	-25341.6	-11829.1	-17379.8	-17151.2
-144831	-101414	1634.767	-61109.4	-129325	-52522.9	-150272	-42402.9
37973.5	27749.8	-565	10718.1	16966.6	13921.32	32477.5	13648.8
-19037.7	-3252.4	-49.43	-18556.5	-7383.8	-3614.5	-18597.6	-4867
-554155	-548814	1049.79	-499653	-539434	-416908	-637954	-458968
-257108	-248526	117.58	-248213	-286366	-227566	-327320	-223000
15743.3	9908.6	-878.3	-4026.4	-4018.6	898.6	8940	-2257.6
-159640	-112689	1722.997	-56325.2	-130380	-53513.2	-157024	-45678.1
23165.1	16474.5	-476.77	15502.3	15911.2	12931.02	25725.1	10373.6
-33846.1	-14527.7	38.8	-13772.3	-8439.2	-4604.8	-25350	-8142.2
-568964	-560089	1138.02	-494868	-540489	-417898	-644706	-462243
-271917	-259801	205.81	-243428	-287421	-228556	-334072	-226275
934.9	-1366.7	-790.07	757.8	-5074	-91.7	2187.6	-5532.8
-130351	-82280.8	2085.787	-31289.2	-100494	-35398	-128842	-22798.2
52453.5	46883.1	-113.98	40538.3	45797.3	31046.22	53907.7	33253.5
-4557.7	15880.9	401.59	11263.7	21446.9	13510.4	2832.6	14737.7
-539675	-529681	1500.81	-469832	-510603	-399783	-616524	-439363
-242628	-229393	568.6	-218392	-257535	-210441	-305890	-203395
30223.3	29041.9	-427.28	25793.8	24812.1	18023.5	30370.2	17347.1
-185971	-131213	2037.777	-74939.3	-148119	-68682	-185434	-59074.7
-3166.2	-2048.7	-161.99	-3111.8	-1827.4	-2237.82	-2684.7	-3022.99
-60177.4	-33050.9	353.58	-32386.4	-26177.8	-19773.6	-53759.8	-21538.8
-595295	-578613	1452.8	-513482	-558228	-433067	-673116	-475640
-298248	-278325	520.59	-262042	-305160	-243725	-362482	-239672

-25396.4	-19889.9	-475.29	-17856.3	-22812.6	-15260.5	-26222.2	-18929.4

Table B3- Correlation Matrix for the sixth analysis in PD in absolute values

	Gen 1	Gen 2	Gen 3	Gen 4	Gen 5	Gen 6	Gen 7	Gen 8
Gen 1	0	0.996	0.618	0.985	0.996	0.984	0.998	0.981
Gen 2	0.996	0	0.554	0.995	0.997	0.993	0.998	0.994
Gen 3	0.618	0.554	0	0.479	0.566	0.481	0.58	0.461
Gen 4	0.985	0.995	0.479	0	0.992	0.998	0.991	0.999
Gen 5	0.996	0.997	0.566	0.992	0	0.994	0.999	0.99
Gen 6	0.984	0.993	0.481	0.998	0.994	0	0.992	0.998
Gen 7	0.998	0.998	0.58	0.991	0.999	0.992	0	0.989
Gen 8	0.981	0.994	0.461	0.999	0.99	0.998	0.989	0

Table B4- TSP and MST comparison- Table to show adjacent correlated genes results in common in both methods in PD

Genes Relation	Number of correlated genes	Adjacent TSP	Adjacent MST
DDX39B-ARF3	2	✓	✓
ARF3-RPL21-PTPN21	3	✓	✓
GD12-THRA	2	□	□

Table B5- TSP and MST comparison- Table to show adjacent correlated genes results in common in both methods in HD

Genes Relation	Number of correlated genes	Adjacent TSP	Adjacent MST
HAB2-HBG2	2	✓	✓
HBD-H3F3A-RPS2P46-IFITM3	4	✓	✓

Table B6- TSP and MST comparison- Table to show adjacent correlated genes results in common in both methods in AD

Genes Relation	Number of correlated genes	Adjacent TSP	Adjacent MST
ND2-NUCKS1	2	✓	✓
COX1-ND4-HNRNPA3	3	✓	✓
NUCKS1-RPS10-GFAP	3	✓	✓
FTL-RPL41	2	✓	✓

Table B7- Table of pathways found in each disorder

Pathways	Parkinson's Disease	Huntington's Disease	Alzheimer's Disease	Total
Metabolic pathways	0	1	1	**2**
Oxidative phosphorylation	0	0	1	**1**
Pomhyrin and chlorophyll metabolism	0	0	1	**1**
Biosynthesis of secondary metabolites	0	0	1	**1**
Two-component system	0	0	1	**1**
Cardiac muscle contraction	0	0	1	**1**
Non-alcoholic fatty liver disease (NAFLD)	0	0	1	**1**
Alzheimer's disease	1	1	1	**3**
Parkinson's disease	1	0	1	**2**
Huntington's disease	0	0	1	**1**
Transcriptional misregulation in cancer	0	0	1	**1**
Hedgehog signaling pathway	0	0	1	**1**

Cell proliferation	1	0	1	**2**
Continuation: Pathways	Parkinson's Disease	Huntington's Disease	Alzheimer's Disease	Total
HTLV-I infection	0	0	1	**1**
Viral carcinogenesis	0	0	1	**1**
Neomycin, kanamycin and gentamicin biosynthesis	0	0	1	**1**
Meiosis -yeast	0	0	1	**1**
Prolactin signaling pathway	0	0	1	**1**
Polycyclic aromatic hydrocarbon degradation	0	0	1	**1**
Microbial metabolism in diverse environments	0	0	1	**1**
FoxO signaling pathway	0	0	1	**1**
Wnt signaling pathway	0	0	1	**1**
Hippo signaling pathway	0	0	1	**1**
Jak-STAT signaling pathway	0	0	1	**1**
Serotonergic synapse	0	0	1	**1**
Measles	0	0	1	**1**
MicroRNAs in cancer	0	0	1	**1**
Mineral absorption	0	0	1	**1**
p53 signaling pathway	0	0	1	**1**

P13K-Akt signaling pathway	0	0	1	**1**
Continuation: Pathways	Parkinson's Disease	Huntington's Disease	Alzheimer's Disease	Total
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins	0	0	1	**1**
Apoptosis	1	0	1	**2**
Metabolism	0	1	1	**2**
Gene expression	0	0	1	**1**
Effects of nitric oxide	0	0	1	**1**
Generic Transcription Pathway	0	0	1	**1**
mRNA Splicing - Major Pathway	0	0	1	**1**
Spliceosome	0	0	1	**1**
Viral mRNA Translation	0	0	1	**1**
Metabolism of amino acids and derivatives	0	0	1	**1**
Activation of the mRNA upon binding of the cap-binding convlex and elFs, and subsequent binding to 43S	0	0	1	**1**

Influenza Viral RNA Transcription and Replication	0	0	1	**1**
Influenza	1	0	1	**2**
Continuation: Pathways	Parkinson's Disease	Huntington's Disease	Alzheimer's Disease	Total
Infectious disease	0	1	1	**2**
rRNA processing	0	0	1	**1**
Signaling by GPCR	0	0	1	**1**
Neural Stem Cell Differentiation Pathways and Lineage-specific Markers	0	0	1	**1**
Spinal Cord Injury	0	0	1	**1**
ERK Signaling	0	0	1	**1**
Clathrin derived vesicle budding	0	0	1	**1**
Vesicle-mediated transport	0	0	1	**1**
Binding and Uptake of Ligands by Scavenger Receptors	0	0	1	**1**
Integrated pancreatic cancer pathways	0	0	1	**1**
Transport of glucose and other sugars, bile salts and organic acids, metal ions and amine compounds	0	0	1	**1**
Cell cycle pathway	0	1	1	**2**
Schizoprenia	0	0	1	**1**

RNA transport	1	0	0	**1**
mRNA surveillance pathway	1	0	0	**1**
Ribosome	1	0	0	**1**
Continuation: Pathways	Parkinson's Disease	Huntington's Disease	Alzheimer's Disease	Total
Chemokine signaling pathway	1	0	0	**1**
Herpes simplex infection	1	0	0	**1**
NOD-like receptor signaling pathway	1	0	0	**1**
TNF signaling pathway	1	0	0	**1**
Toll-like receptor signaling pathway	1	0	0	**1**
Cytokine-cytokine receptor interaction	1	1	0	**2**
Chagas disease (American trypanosomiasis)	1	0	0	**1**
Cytosolic DNA-sensing pathway	1	0	0	**1**
Rheumatoid arthritis	1	0	0	**1**
Epithelial cell signaling in Helicobacter pylori infection	1	0	0	**1**
Prion diseases	1	0	0	**1**
Neuroactive ligand-receptor interaction	1	0	0	**1**
Thyroid hormone signaling pathway	1	0	0	**1**
Inflammation	1	0	0	**1**

Neurodegenerative	1	0	0	**1**
Hemostasis	0	1	0	**1**
Megakaryocyte	0	1	0	**1**
Inmune System	0	1	0	**1**
Continuation: Pathways	Parkinson's Disease	Huntington's Disease	Alzheimer's Disease	Total
Interferon Signaling	0	1	0	**1**
Interferon alpha/beta signaling	0	1	0	**1**
p70S6K Signaling	0	1	0	**1**
IL-4 Pathway	0	1	0	**1**
Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3	0	1	0	**1**
Cellular Senescence	0	1	0	**1**
Mitotic Prophase	0	1	0	**1**
Developnent NOTCHI-mediated pathway for NF-KB activity modulation	0	1	0	**1**
Diabetic complications	0	1	0	**1**
Complement and coagulation cascades	0	1	0	**1**
Carbon metabolism	0	1	0	**1**

Benzoate degradation	0	1	0	**1**
Aminobenzoate degradation	0	1	0	**1**
Butanoate metabolism	0	1	0	**1**
Continuation: Pathways	Parkinson's Disease	Huntington's Disease	Alzheimer's Disease	Total
Microbial metabolism in diverse environments	0	1	0	**1**
B cell receptor signaling pathway	0	1	0	**1**
Schizoprenia	1	0	1	
Total	**23**	**25**	**56**	

B8-Accronyms

- PD- Parkinson’s disease
- AD- Alzheimer’s disease
- HD- Huntington’s disease
- DEG- differentially expressed genes
- CBV - Cerebral Blood Volumes
- FMRI- Functional Magnetic Resonance Imaging
- DMVN- Dorsal Motor Nucleus of the Vagus
- ION- Inferior Olivary Nucleus
- ALS- amyotrophic lateral sclerosis
- MCO- Multiple Criteria Optimization
- PV- Present Value
- PMT- Payment
- AC- Annual Cost
  MST-

II- Paper published in ISERC 2016 Conference

Commonalities in Genetic Signatures and Signaling Pathways in Neurological Disorders

Nicole J Ortiz
University of Puerto Rico
Mayagüez, PR

Mauricio Cabrera
University of Puerto Rico
Mayagüez, PR

Clara Isaza
Ponce School of Medicine
Ponce, PR

Abstract

In this research, mathematical optimization is used to detect potential genetic differentially expressed genes and proposes probable signaling structures in Alzheimer's disease, Parkinson's disease and Autism. The characterizations of these three affections have been elusive in the literature, although their impact in society is projected to dramatically increase in the next decades worldwide. This project proposes the study of commonalities in the genetic signatures and genetic pathways, through optimization analysis, among these three neurological disorders. The search for the most correlated path is carried out using very popular integer optimization formulations: Travelling Salesman Problem (TSP) and Minimum Spanning Tree (MST). For both, a set of differentially expressed genes is identified previously through multiple criteria optimization; a

correlation coefficient is used to link every pair of genes. This information is used to create a mathematical graph, where genes are nodes and the absolute value of the correlation coefficient is applied to each directed arc. A branch-and-bound approach is used to find the optimal path, that is, the most correlated one. The analysis has been partially applied for Parkinson's disease, obtaining preliminary results of 8 differentially expressed genes, some of them related with inflammation, apoptosis, neurological disorders, mitochondria and proliferation.

Keywords

Traveling Salesman Problem; Signaling Pathways; Multiple Criteria Optimization; Parkinson's disease

1. Introduction

This project proposes the study of commonalities among three important neurological conditions: Alzheimer's disease (ALS), Parkinson's disease (PAR) and Autism (AUT). The characterizations of these three conditions have been elusive in the literature, although their impact in society is projected to dramatically increase in the next decades worldwide. In Puerto Rico alone, current estimates for people affected with these disorders add up to 48,000 for ALS, 14,000 for PAR, and 7,800 for AUT [1]. In addition, it must also be noted that Caribbean-Americans are 1.5 times as likely to suffer dementia as White Americans [1]. The high incidence of these conditions in Puerto Rico call for accelerating their understanding and characterization.

There is a large amount of publicly-available data associated to mRNA microarrays and microRNA arrays, as well as several other types of high throughput biological experiments related to ALS, PAR and AUT with the potential to be analyzed in a coordinated manner [2, 3]. Our

research group has specialized in designing analysis strategies based on mathematical optimization that facilitate the simultaneous analysis of multiple experiments without the manipulation of parameters by the user. In this sense, our strategies offer consistent convergence to a manageable number of key pieces of information for each disorder that can be correlated with convenience in search for commonalities. Figure 1 schematically shows the proposed research design for this pilot project. Moving horizontally from left to right, the first stage involves the simultaneous analysis of multiple microarrays (or microRNA arrays) to detect potentially important genes or regulatory molecules. This first stage will be approached through our originally-devised multiple criteria optimization approach. The second stage will use the lists of potentially important genes from the previous stage to determine the optimally correlated circular path among them as a proxy for a biological signaling path. Moving vertically downwards in Figure 1, these two stages will be carried out for all three conditions (ALS, PAR and AUT). In Stage 3, these analyses will then allow establishing commonalities in terms of differentially expressed genes and potential signaling paths. Biological and medical literature will be used to marshal evidence of similar mechanisms among illnesses and to propose mechanisms that have eluded discovery to date.

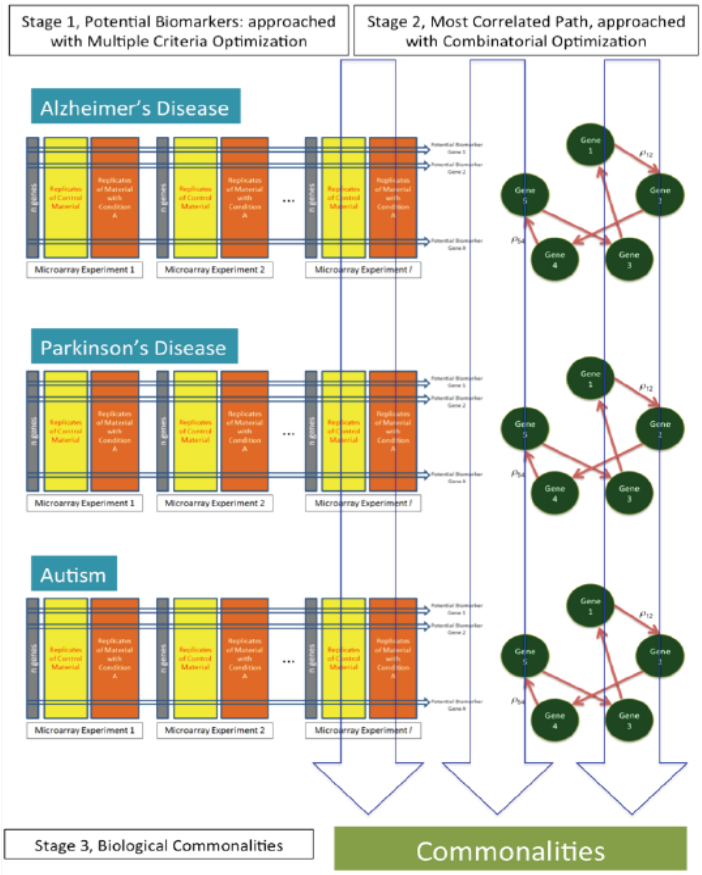


Figure 1: Research Design for Project

Even though this work goal is to characterize commonalities between the three neurological disorders, at the moment only PD has been analyzed and will be discussed in this research paper.

2. Methodology

Parkinson's disease (PD) is the second most common neurodegenerative disease. Although many of the pathogenic molecules underlying the rare autosomal-dominant forms of PD have been identified, the full complement of pathogenic pathways involved in the common "sporadic" form

of PD remains unknown [4]. For this reason PD was the first neurological disorder analyzed at this time and it is the one discussed in this research paper. First, two microarray databases [4] were selected for analysis, each containing 22,277 probes in the microarray and their respective lectures in six samples in brain tissues for PD and five samples for control. Both databases were obtained from [4]. The first database came from a high-resolution variant of functional MRI (fMRI) that maps basal cerebral blood volume (CBV) with submillimeter resolution to show that the DMNV (Dorsal Motor Nucleus of the Vagus) is dysfunctional in PD according with the study mentioned above. The second database came from a neighboring region relatively resistant to the disease, ION (Inferior Olivary Nucleus), taken from the same study. The PD samples were taken from autopsies from people with a mean age 56.4 years and control samples were taken from autopsies of people with mean age of 56.2 years [4].

2.1 Multiple Criteria Optimization

The first step was to calculate the difference of PD and control expressions in both databases simultaneously modeling the analysis as a Multiple Criteria Optimization (MCO) Problem as described in [5, 6], both works by our research group. In brief, MCO deals with making decisions in the presence of multiple performance measures in conflict. 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 [7]. The general MCO 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 [7]. The performance measure used in this case, was the median, calculated for PD and control samples in DMNV and also PD and control samples in ION databases as shown in Figure 2.

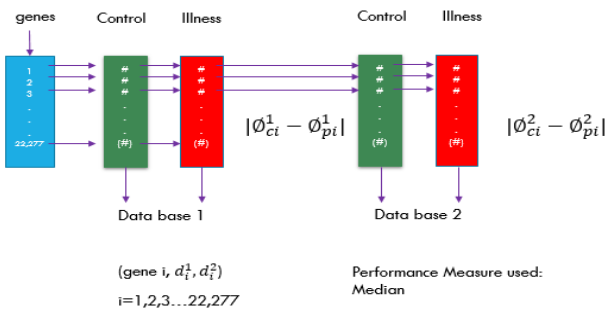


Figure 2: Gene expression differences between PD and control in DMNV and ION tissues

In this study, the set of solution that provide the best compromise between the conflicting performance measures were found using a Pareto-Efficient Frontier. The genes selected by this Frontier are those that change their expression the most between the control and the PD samples. For this study the first ten efficient frontiers were considered to avoid missing genes that also should be considered solutions. Due to computer memory constrains the databases were divided in 4 groups of 5,600 genes or less and iterated to converge to the final efficient frontier. An example of an efficient frontier analysis is shown in Figure 3, under a case with maximization of two performance measures. Do notice that any performance measure to be maximized can be transformed linearly to become a perfectly equivalent performance measure to be minimized individually.

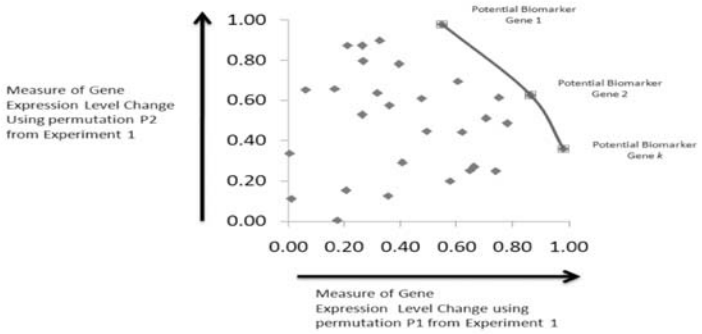


Figure 3: Example of an efficient frontier

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

Find x to

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

The MCO problem in (1) 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,…, |K|) which can, in turn, be evaluated in the J performance measures to result in values $\boldsymbol{f_j\ x^k}$.That is, the $\boldsymbol{k^{th}}$ combination of values for the decision variables evaluated in the $\boldsymbol{j^{th}}$ objective function.The MCO formulation under such discretization is, then as follows:

Find $\boldsymbol{x^k}$ (k∈K) to

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

The solutions to (2) are, then, the Pareto-efficient solutions of the discretized MCO problem. Considering formulation (2), a particular combination $\boldsymbol{x^0}$ with evaluations $\boldsymbol{f_j\ x^0}$ will yield a

Pareto-Efficient solution to (2) if and only if no other solution $\boldsymbol{x^{\psi}}$ exists that meets two conditions, from this point on called Pareto-optimality conditions

$$\boldsymbol{f_j(x^{\psi}) \leq f_j\ x^0}\ \ \forall \text{j} \qquad \text{(Condition 1)}$$

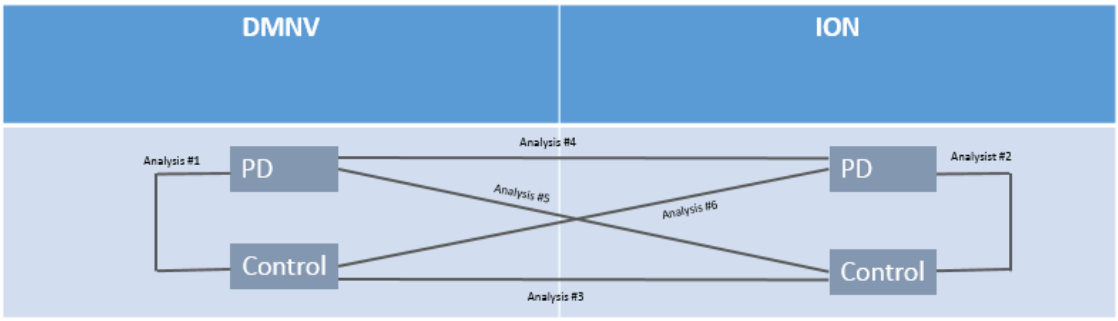
$$\boldsymbol{f_j(x^{\psi}) \leq f_j\ x^0}\ \text{ in at least one j}$$

(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.

The original work from where the databases were obtained [2] stipulated that one region of the brain is being affected (DMNV) in PD presence and the other region remained unaffected (ION). This was challenged in this study since it is known that gene expressions works as a network and neighboring brain regions could be affected. For this reason it was decided to perform six additional analyses (presented in Table 1) to validate if ION region was unaffected; or if the main variable driven the expression differences was the presence/absence of PD in the samples. All analyses were also treated as MCO problems.

Table 1: Analyses with DMNV and ION database



Series of Pareto Efficient Frontiers were identified on all analyses resulting in distinct sets of differentially expressed genes. As stated before, there were 22,777 genes initially, from which ten Pareto Efficient Frontiers were analyzed in each analysis. Accumulating the Pareto optimal solutions (differentially expressed genes) from every analysis, there were a total of 44 differentially expressed genes. From those 44 differentially expressed genes, eight genes that were in common in every single analysis were further analyzed. Using the selected differentially expressed genes from the last process, individually for each analyzes and for which behavior was measured as a statistical correlation, the next step was to find coordinated behavior among the expression changes of the selected genes. The statistical correlation was computed as linear as a first approach and carried out in a pairwise manner. The linear correlation values found are proxies for suppressed or stimulated behavior in the expression level changes of the two genes under analysis. Because these values range from -1 to 1, their absolute values are computed to handle them as quantities to be maximized. Thus two genes will be strongly correlated if the absolute of their correlation value is close to 1. The correlations calculated for each pair of gene were further arranged in a correlation matrix. To construct this matrix, the differences taken in consideration in each analysis (see Table 1) 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 correlation of genes in each analyses was the combined to converge on a final solution as illustrated in Figure 4.

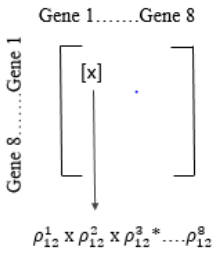


Figure 4: Combined correlations solution matrix example

This method consisted in multiplying the absolute correlation values of each gene combination from all analyses on the same position. This ensured that changes in correlation values be reflected in a combined correlation factor. This combined correlation factor represented each pair of genes.

2.2 Traveler Salesman Problem

If each gene is represented through a node in a graph, then the undirected arc joining a pair of genes can hold their absolute correlation value. This led to the Traveling Salesman Problem (TSP) formulation, where the idea is to find the most correlated complete tour.[8] The TSP is one of the most famous combinatorial optimization problems. TSP tries to construct the shortest tour through n cities for a salesperson to visit, usually going back to a preselected base city [9]. The object of TSP is to “Find the shortest tour that visits each city in a given list exactly once and then returns to the starting city” [10]. If one rephrases the quoted sentence as “Find the most correlated tour that visits each potential differentially expressed gene in a given list exactly”, then it is clear that such solution might shed light on how PD works. In this particular research, the TSP is used to

obtain an optimal sequence maximizing linear correlations among the eight genes considered as potential differentially expressed genes. An illustration of how the resulting graph would look like for a five genes problem is shown in Figure 5.

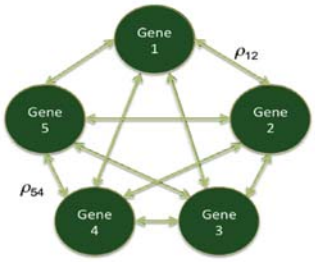


Figure 5: TSP illustration example

The TSP optimization model is as follows:

$$\boldsymbol{Minimize}\ \sum_{(i,j)\in A} c_{ij}y_{ij} \tag{3}$$

$$\sum_{1\le j\le n}(y_{ij}) = 1 \quad \forall i \tag{4}$$

$$\sum_{1\le i\le n}(y_{ij}) = 1 \quad \forall j \tag{5}$$

$$\aleph x = b \tag{6}$$

$$x_{ij} \le (n-1)y_{ij} \quad \forall (i,j) \in A \tag{7}$$

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

(8)

$$y_{ij} = 0 \; or \; 1 \qquad \forall (i,j) \in A \tag{9}$$

Let A' = {(i,j): yij =1} and let A'' ={(i,j): xij >0}. The constraints (4) and (5) 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 (4) and (5) will be a union of disjoint cycles; if any such solution contains more than one cycle; they are referred to as sub tours, since they pass through only a subset of nodes. In constraint (6) N is an nxm matrix, called the node-arc incidence matrix of the minimum cost flow problem. Each column Nij in the matrix corresponds to the variable xij. The column Nij has a +1 in the i^{th} row, a −1 in the j^{th} row; the rest of its entries are zero. Constraint (6) 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 (7) imply that A" is a subset A'. These conditions imply that the arc set A' is connected and thus cannot contain sub tours [11].

3. Results

The optimal solution to this particular TSP is the tour among the genes of interest with the largest possible correlation, shown in Figure 6. For this particular case there are a total of 8! ≈40,320 ways in which a cyclic path can be drawn among the 8 genes, the optimal path was found in this research.

Figure 6: Optimal solution path for potential expressed genes

The differentially expressed genes and relations among them, found with this methodology, are not necessarily associated with PD at this moment. A search for biological pathways and characteristics for each genes in KEGG [12] databases and PubMed publications [13] was conducted relating some of the genes with inflammation, apoptosis, neurological disorders, mitochondria and proliferation. All this information will be further analyzed and discussed as future work to validate the mathematical methods and biological structure presented in this work Also, the same method discussed in this work would be applied to ALZ, AUT, and it is also under consideration to include Huntington's disease.

4. Conclusions

This work proposes the study of commonalities among three important neurological disorders: Alzheimer's disease (ALS), Parkinson's disease (PAR) and Autism (AUT). As an initial state the characterization of PD was performed to search for potential differentially expressed genes. Their interrelations can be enhanced using optimization techniques. The purpose of the proposed method is the detection of a biological pathway, as a combinatorial problem similar to the Traveling Salesman Problem. This implies an optimal solution exists. It 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 characterization of PD, as well as ALS and AUT eventually.

Acknowledgements

This work was possible thanks to NIH MARC Assisting Bioinformatics Efforts at Minority Schools project

2T36GM008789.

References

1. Instituto de Estadisticas de Puerto Rico. (2015, January). Retrieved form http://www.estadisticas.pr/iepr
2. Alzheimer Association. (2015, January). Retrieved from http://www.alz.org
3. Scherzer C.R., Jensen R.V. and Gullans S.R., 2003 “Gene expression changes presage neurodegeneration in a Drosophila model of Parkinson’s disease”, Human Molecular Genetics 2003; 12(19); 2457-2466vv
4. Lewandowski N, Ju S., 2003 “Polyamine pathway contributes to the pathogenesis of Parkinson disease” PUBMED; 12(19):2457-66
5. Sanchez-Peña M., Isaza C. and Cabrera-Ríos M., 2013 “Potential Colon Cancer Biomarker Search Using MoreThan two Performance Measures in a Multiple Criteria Optimization Approach”, Puerto Rico Health Sciences Journal 2013: 31:2 59-63
6. Cruz-Rivera Y.E., Lorenzo E.L. and Ortiz N.J., 2013 “Models to Lead Genetic Signaling Path Discovery: Preliminary Ideas and Results”, Biomedical Engineering Society Annual Meeting, Seattle, WA
7. Lorenzo E., Camacho-Caceres K., 2013 “An Optimization-Driven Analysis Pipeline to Uncover Biomarkers and Signaling Paths: Cervix Cancer” Microarrays 2015, 4(2), 287-310
8. Hashler and Hornik, 2007, “TSP-Insfraestructure for the traveling salesman problem” Journal of Statistical
Software (in press)

9. Laporte G. “A Concise Guide to the Traveling Salesman Problem”, The Journal of the Operational Research Society; 2010,Vol. 61, No. 1, Journal: Challenges for OR (Jan., 2010), pp. 35-40
10. Sánchez-Peña M. L., Isaza, C., 2013 “Identification of potential biomarkers from microarray experiments using multiple criteria optimization”, CANCER MEDICINE 2013, 2: 253–265. doi: 10.1002/cam4.69
11. Kumari, S.; Nie, J. “Evaluation of gene association methods for coexpression network construction and biological knowledge discovery”, PLoS ONE 2012, 7, e50411. [Google Scholar] [CrossRef] [PubMed]
12. PubMed. (2016, January). Retrieved form http://www.ncbi.nlm.nih.gov/pubmed
13. KEGG. (2015, December). Retrieved from http://www.genome.jp/kegg/