Title
Computational Methods for Processing and Analyzing Large Scale Genomics Datasets

Author
Grujic, Olivera

Publication Date
2016

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Computational Methods for Processing and Analyzing Large Scale Genomics Datasets

A dissertation submitted in partial satisfaction of the requirements

for the degree Doctor of Philosophy in Computer Science

by

Olivera Grujic

2016

ABSTRACT OF THE DISSERTATION

Computational Methods for Processing and Analyzing Large Scale Genomics Datasets

by

Olivera Grujic

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2016

Professor Eleazar Eskin, Chair

Abstract

This dissertation develops computational methods for analyzing large-scale genomic and epigenomic datasets. We developed a supervised machine learning approach to predict non-exonic evolutionarily conserved regions in the human genome based on vast amount of functional genomics data. The resulting probabilistic predictions provide a resource for prioritizing functionally important regulatory regions in the human genome. We also developed a method for identifying from large-scale gene expression datasets genes that are differentially expressed in both blood and brain from 12 vervet monkeys, which we used to identify 29 transcripts whose expression is variable between individuals and heritable. Additionally, we developed a method using a global search optimization algorithm to successfully improve a model of human thyroid hormone regulation dynamics leading to a better fit of data for thyrotoxicosis. Together, these three approaches have the potential to impact the understanding and eventual treatment of disease.

The dissertation of Olivera Grujic is approved.

Milos D. Ercegovac

Jason Ernst

Matteo Pellegrini

Eleazar Eskin, Committee Chair

University of California, Los Angeles,

2016

TABLE OF CONTENTS

List of Tables

List of Figures

Acknowledgements

I am very grateful for support I received since I started grad school nine years ago, but am also thankful to wonderful people from other chapters of my life who taught me life-long skills and helped shape me to be the person and the scholar I am today. The list is mostly in chronological order, and contains only names. The impact all these individuals (and many more) had on this exciting and yet arduous process are locked in my heart. Many thanks to:

UCLA

Dr. Eleazar Eskin, Dr. Jason Ernst, Dr. Milos Ercegovac, Dr. Matteo Pellegrini

Dr. Anna Jasinska, Dr. Nelson Freimer, Dr. Jennifer Pike

David Smallberg, Vera Morgan, Steve Arbuckle, staff at UCLA CS Department

Eckerd College

Dr. Kelly Debure, Dean Lloyd Chapin, Dr. Ed Gallizzi, Dr. Eduardo Fernandez, Mark Fishman

High School "J J Zmaj"

Dr. Duska Pesic, Dr. Dragan Masulovic, Dr. Radivoje Stojkovic, Petnica Science Center

Elementary School "Vuk Karadzic"

Jasmina Dajevic, Stanko Mandic, Ramadan Dervisevic, Zlata Lalic

Friends who are like Family

Dr. Rada Smith and Dr. John Smith

Jane Colson and Dr. Gerald Grunska

Ken and Debra Sulewski

Sybil and Dick Israel

Nessie

Parents

Dr. Olgica Grujic, Vladimir Grujic, PE

Friends, former classmates, and colleagues around the world who kept reminding me to NEVER GIVE UP!

CURRICULUM VITAE
OLIVERA GRUJIC

EDUCATION

University of California, Los Angeles
MS, Computer Science, June 2009 GPA: 3.58 / 4.0 scale
Eckerd College
BS, Computer Science and German, January 2004 GPA: 3.96 / 4.0 scale

LANGUAGES

Fluent in Serbo-Croatian, Advanced in Spanish, Intermediate in German

RESEARCH EXPERIENCE

Principal Investigator
University of California, Los Angeles
March 2015 – June 2016
Project: Predictor and Identifier of Conserved Elements (PICEL)

Developed a method that integrates vast amount of various functional genomics datasets into a single track that prioritizes likely regulatory important bases in the human genome using supervised machine learning. Model incorporated over 10,000 distinct features and was trained on 10 million samples genome-wide using parallel computing. Obtained relatively effective predictions of the conserved regions (AUC = ~0.82), which have the potential to be an important resource for interpreting and prioritizing disease associated variants.
Chair: Jason Ernst, PhD

Research Assistant
University of California, Los Angeles
September 2007 – June 2009
Project: Analysis of Gene Expression Datasets in Vervet Monkey

Objectives were to determine differences in individual gene expressions of vervet monkeys related to stress response and identify candidate genes expressed both in blood and brain tissues. My part was managing and statistically analyzing large gene expression datasets.
Principal Investigator: Nelson Fraimer, M.D., Anna Jasinska, PhD

TEACHING EXPERIENCE

Teaching Assistant – Undergraduate Intro to Computer Science Courses I, II, III
University of California, Los Angeles
September 2008 – December 2008 (Fall Quarter)
September 2009 – March 2010 (Fall & Winter Quarters)
April 2011 – June 2011 (Spring Quarter)
January 2012 – June 2012 (Winter & Spring Quarters)
April 2013 – June 2014 (Spring Quarter)
June - August 2013 (Summer Session)
April 2013 – June 2014 (Spring Quarter)

June - August 2014 (Summer Session)
January 2015 – March 2015 (Winter Quarter)
April 2015 – June 2015 (Spring Quarter)
September 2015 – March 2016 (Fall & Winter Quarters)

Course integrated approximately ten assignments and programming projects to be implemented in C and C++. Duties included planning and implementation of weekly discussion sections, weekly office hours, assisting students with implementing and debugging programming assignments, and evaluation of student progress.

INDUSTRY EXPERIENCE

Graduate Technical Intern
Intel
Hillsboro, OR (June 2012- September 2012)
Designed and implemented web application, which automatically determines quality of current software release based on given test metrics and real-time data, used by mobile communications quality group.

Software Designer
Hewlett Packard
San Diego, CA (June 2011-December 2011); Boise, ID (April 2010 – March 2011)
Worked as part of R&D team to deliver NextGen SW applications to enable laser and inkjet printers. Lead a team of test engineers and distributed testing among 5 supplier companies located on and off shore, which lead to higher quality test reports and reduced testing cost. Adopted tools and processes to deliver high quality and robust SW that exceeded user expectations resulting in lower warranty cost.

PUBLICATIONS

Jasinska, A. J., Lin M. K., Service S., Choi O., DeYoung, J., **Grujic, O.**, et al. (August 2012). A non-human primate system for large-scale genetic studies of complex traits. *Human Molecular Genetics*, **21** (15):3307-3316.

Jasinska, A. J., Service, S., Choi, O., DeYoung, J., **Grujic, O.**, et al. (November 2009). Identification of brain transcriptional variation reproduced in peripheral blood: an approach for mapping brain expression traits. *Human Molecular Genetics*, **18** (22), 4415-4427.

Grujic, O., Fishman, M. (May 2004). Prototype of Automatic Intelligent Translator. *The Eckerd Scholar,* 2004, 64-70.

CONFERENCE PRESENTATIONS

Grujic, O*., Ernst, J. Poster: Systematic Prediction of Conserved Non-Exonic Bases from Epigenomic and Transcription Factor Binding Data. RECOMB, Santa Monica, CA April 2016.

Grujic, O*., DiStefano, J. Poster: Augmenting Thyrosim for Thyrotoxic Conditions. Tech Forum, University of California, Los Angeles, CA February 2015.

Grujic, O*., DiStefano, J. Poster: Global Sensitivity Analysis of Biomolecular Network Model NF-kB. CWIC Southern California, Santa Ana, CA April 2012.

Chapter 1

Introduction

1.1 Motivation

Human genome contains 3 billion base-pairs. Only 1.5% of it is coding for proteins. At the time of writing this dissertation, it is estimated that only 4% of the human genome (including the aforementioned 1.5%) is exonic, while 96% of the human genome is non-exonic. Therefore, any dataset obtained by measuring any kind of activity related to human genome (e.g. gene expression, GWAS, sequencing, ChIP-Seq, DNase-Seq, etc.) is considered large scale data. These datasets require statistical and computational methodology, in addition to incredible amount of computational resources (such as time, memory, and processing power) in order to be analyzed and produce scientific findings. The main objective of this dissertation is to develop and/or apply existing computational, statistical, numerical, and machine learning methods to analyze large-scale biomolecular and genomic data. The overarching motivation for this dissertation is to use computers and algorithms to help solve problems that would potentially lead to novel medical discoveries, such as insights, therapies, and medication for finding cures for some of currently major diseases such as cancer and various neuro-psychological disorders.

1.1.1 Epigenetic Marks for Genome Annotation

Epigenetic modifications in the human genome play diverse roles in gene regulations and are likely to have a causal role in human disease. The human genome comprises of 96% non-exonic regions. Genome-wide association studies (GWAS) have discovered thousands of loci associated with complex traits and diseases, and have shown that most variants are located in non-exonic regions. At this time, there are thousands of data tracks obtained by using various vet

lab techniques (assays) followed by next generation sequencing measuring various epigenetic processes, and yet the challenge remains to identify potentially important locations in 96% of non-exonic genome. Efforts are being made to functionally annotate human genome using functional genomics and comparative data.

We developed Predictor and Identifier of Conserved Elements (PICEL) method that uses supervised machine learning technique (logistic regression) and produces a single data track containing probabilities that predict conserved regions genome-wide. It integrates almost 11,000 data tracks of functional genomic datasets in order to prioritize important regulatory bases and provide a resource to the scientific community. This method also has for the objective to understand to what extent conservation could be predicted from functional genomics data. The resulting predicted peaks overlap known conserved elements genome-wide. False negatives provide potentially novel discoveries in non-exonic DNA and candidates for further human genetics studies.

1.1.2 Gene Expression Studies for eQTL mapping

We processed and analyzed large-scale dataset obtained by gene expression studies of data collected from vervet monkeys (Chlorocebus aethiopssabaeus). Vervets are excellent non-human primate model for this type of investigation because there is high degree of conservation of tissue expression profiles between vervets and humans. The data consisted of blood samples drawn from 347 vervet monkeys and eight brain regions from 12 vervet individuals. The goal was to find out how genes expressed in brain correlated with gene expression variation in blood from the same individuals and search for genes expressed in both brain and blood so that blood tissue could serve as surrogate for brain tissue in further experiments. By applying stringent

method and statistical criteria, we identified 29 transcripts whose expression is measurable, stable, replicable, variable between individuals, relevant to brain function and heritable. These findings provided means to investigate variation in gene expression relevant to human brain traits and neuropsychiatric diseases.

1.1.3 Global Parameter Search and Sensitivity Analysis

Dynamic system models are used for simulating intracellular functions in search of novel therapies for clinical disorders. Techniques involved in quantification cycle of nonlinear dynamic models, specifically methods for global sensitivity analysis and parameter search, have as an objective to efficiently improve selection of parameter subsets by reducing the search space and computational times. We carried out numerical experiments on NF-κβ biomolecular dynamic system model and have successfully augmented model of human thyroid hormone regulation dynamics to better fit the available data for thyrotoxicosis. Our study provides insight into model mechanics and identifies limitations of current methods, as well as serves educational purpose of demonstrating how human body processes toxic dose of thyroid hormone replacement over time.

1.2 Contribution of this Dissertation

The contribution of the dissertation can be split into three parts.

The first part contributes novel computational methodology for analyzing large-scale genomic and epigenomic datasets and offers a useful tool for future scientific discoveries.

The second part contributes to statistical methodology for analyzing and interpreting large-scale gene expression datasets and leads to novel finding significant in human genetics.

The third part analyzes numerical computational methods for parameter search and global sensitivity analysis for dynamic system models, and makes a contribution towards improving one of such models.

Together, the three computational and quantitative methods developed specifically for processing large-scale data and searching vast parameter space may lead to discoveries in biological chemistry, human genetics, and medicine.

1.3 Organization of Dissertation

The dissertation is organized in three chapters, each one pertaining to a different problem and methodology used to solve it.

The first chapter discusses methodology for dealing with large-scale genomic and epigenomic datasets and efforts to functionally annotate the human genome, which comprises of 96% non-exonic regions. At this time, there are thousands of data tracks obtained by using various vet lab techniques (assays) followed by next generation sequencing measuring various epigenetic processes, and yet identifying potentially important locations in the vast amount of data remains a challenge. We present Predictor and Identifier of Conserved Elements (PICEL) method that integrates more than 10 thousand of tracks into a single track and accurately predicts conserved regions in the human genome. This computational method is joint work with Dr. Jason Ernst.

The second chapter discusses methodology for interpreting large-scale gene expression datasets. The dataset contains gene expression values from brain and blood samples from vervet monkeys (non-human primate model). Result of data analysis identified 29 transcripts that are differentially expressed in both blood and brain whose expression is variable between individuals

and heritable, which lead to a further genetic studies and discoveries. Chapter 2 is joint work with Dr. Anna Jasinska and Dr. Nelson Freimer.

The third chapter addresses numerical computational methods for parameter search and global sensitivity analysis for dynamic system models. This type of modeling existed before large-scale datasets became available, but it does rely on clinically measured data (the data needs to fit the model, not vice versa). The models typically consist of systems ordinary differential equations, which contains several scores (or more) of parameters that are unknown (e.g. impossible to measure). The challenge remains to estimate the parameter values in order to get the model to fit the clinical data. The chapter describes application of global sensitivity analysis techniques on more complex version of NF-kB model dynamics and global parameter search or model of thyroid hormone dynamics, addressing in particular fitting the model to a thyrotoxic data. Chapter 3 was joint work with Dr. Joe DiStefano.

Finally, we conclude by summarizing results obtained from the three parts and discuss potential extensions to the methodology presented and its application to different areas in computer science.

Chapter 2

Predictor and Identifier of Conserved Elements (PICEL)

Abstract

A large majority of genome-wide association study (GWAS) hits fall into non-exonic regions of the genome. Determining likely causal variants among multiple ones in linkage disequilibrium (LD) remains a challenge. Large scale consortium projects such as NIH Roadmap Epigenomics and ENCODE as well as the collective effort of many individual labs have produced thousands of genome-wide experiments on regions of open chromatin, locations of transcription factor binding and histone modifications which can be a resource to prioritize important regulatory locations in the genome. However with the vast number of data sets available, large fractions of the genome showing signal, and in many applications the relevant cell types uncertain, it is often unclear how to prioritize genomic locations based on the data.

Here we present Predictor and Identifier of Conserved Elements (PICEL) method that integrates vast amount of various functional annotation datasets into a single track that prioritizes likely regulatory important bases in the human genome. PICEL uses supervised machine learning technique (logistic regression) to predict conserved regions in the human genome. Training set incorporates histone modification, transcription factor, DNaseI and DNaseI footprints data as well as chromatin states. The resulting training set is very large, it contains 10,836 distinct features. PICEL was trained on 10 million samples genomewide. The response labels were chosen from conserved region datasets: phastCons, GERP++, SiPhy-Pi and SiPhy-Omega elements. We applied PICEL genome-wide to make a probabilistic prediction for each

individual nucleotide as to whether it would fall into a non-exonic conserved region. We were able to obtain relatively effective predictions of the conserved regions (AUC = ~0.82).

Analysis of the false negative predictions of our method and the genes they are proximal to can be used to identify cell types or classes of genes, such as olfactory genes, that are not adequately represented in current functional genomic data sets and can suggest potentially addition cell and tissue types for experimental mapping. Our false positive predictions can suggest potential important recently evolved regulatory locations in the genome. Heritability analysis of complex traits such as BMI show that locations of greater probability of being conserved strongly tracked with locations that explained an increasing fraction of disease heritability suggesting our predictions have the potential to be an important resource for interpreting and prioritizing disease associated variants.

2.1 Introduction

Genome-wide association studies (GWAS) have identified loci associated with complex traits and diseases. However, detecting the causal variants remains a challenge due to loci containing multiple single nucleotide polymorphisms (SNPs) in linkage disequilibrium (LD). More recently, large scale consortium projects such as NIH Roadmap Epigenomics and ENCODE as well as the collective effort of many individual labs have produced thousands of genome-wide experiments on regions of open chromatin, locations of transcription factor binding and histone modifications which can be a resource to prioritize important regulatory locations in the genome.

Epigenetics is the study of change in gene expression or phenotype that occurs without changes in DNA (Bird, 2007; Goldberg *et al.*, 2007). Therefore, epigenetic modifications in the human genome can modulate the interpretation of the primary DNA sequence, without alternating the sequence. Functionally, epigenetic modifications can serve as markers to represent gene activity, expression, and chromatin state (Berger, 2007; Bernstein *et al.*, 2007; Kouzarides, 2007). Epigenetic research can help explain how cells carrying identical DNA differentiate into different cell types (Weinhold, 2006; Jaenisch e Young, 2008). Epigenetic markers across the genome are called epigenome.

Epigenetic mechanisms (Weinhold, 2006; Gal-Yam *et al.*, 2008; Consortium, 2012; Encode, 2012) include histone proteins associated with DNA (histone modifications) (Jenuwein e Allis, 2001; Berger, 2007; Bernstein *et al.*, 2007; Goldberg *et al.*, 2007; Kouzarides, 2007; Li *et al.*, 2007), chemical modifications to the cytosine residues of DNA (DNA methylation) (Laird, 2003; Feinberg e Tycko, 2004; Jones e Baylin, 2007; Esteller, 2008), small and non-coding RNAs (Mattick e Makunin, 2006; Pathways | SABiosciences, 2016), and chromatin architecture

(Li *et al.*, 2007; Encode, 2012). The nucleosome is the fundamental subunit of chromatin, composed of approximately two turns of DNA wrapped around histone octamer, which contains two copies of each of H2A, H2B, H3, and H4 histone proteins. Perhaps the most useful epigenetic information for detecting regulatory elements is post-translational modifications in the tails of histone proteins that package DNA into chromatin.

Epigenetic marks are used to annotate potential locations of functional elements of the human genome, including in non-coding regions, and therefore aid in understanding of the human genome and gene regulations. However, the functional roles of combinations of epigenetic modifications are still being discovered. Functional annotations are important because they could predict the functional effect of a variant and subsequently its likelihood to have a causal role in a disease. Similarly, open chromatin and transcription factor binding sites (TFBS) provide vast amount of epigenetic information and are likely to be predictive to conservation in the genome.

Next generation sequencing enabled studying of the genome-wide occupancy maps of transcription factors (TF's) by chromatin immunoprecipitation technique (assay) followed by sequencing (ChIP-seq). Many loci are occupied by multiple TF's in various cell types, indicating the existence of combinatorial regulation in cells. The encyclopedia of DNA Elements (ENCODE) (Encode, 2012) project has systematically mapped regions of TFBS.

Another commonly used experimental technique (assay) combined with next generation sequencing is Deoxylribonuclease I Hypersensitivity (DNase-seq) (Wang *et al.*, 2012). Deoxylribonuclease I (DNase I) is an enzyme that cleaves (cuts) links in the DNA backbone usually at sites that are 'hypersensitive' to DNase I. These are the sites where the chromatin is

open and accessible (where there are no histone proteins) (Encode, 2012) and therefore it is likely TFBS.

One more, among many experimental techniques that derive functional genomic datasets, is formaldehyde assisted isolation of regulatory elements (FAIRE-seq). This assay isolates genomic regions that are depleted of nucleosomes (Encode, 2012). Similarly to DNase-seq, this assay is also identifying open chromatin regions that are likely TFBS.

Data from genomes of many species including some that have recently been sequenced provided an opportunity to identify and interpret encoded functional elements by looking for sequences that are conserved across species (Siepel *et al.*, 2005; Davydov *et al.*, 2010; Pollard *et al.*, 2010; Rosenbloom *et al.*, 2015). The reason for sequence conservation across species is believed to be negative (purifying) selection. Sequences that are significantly more similar than would be expected if they were evolving under neutral revolution are considered to be conserved and therefore are likely to have critical functions. In other words, nucleotides that are functionally important tend to remain unchanged by evolution because mutations at those sites would reduce fitness and are therefore eliminated by natural selection (Cooper *et al.*, 2005; Siepel *et al.*, 2005). It is estimated that only ~1.5% of the human genome encodes proteins (Lander *et al.*, 2001), and yet comparison studies with genomes of other species (e.g. mouse (Waterston *et al.*, 2002), rat (Gibbs *et al.*, 2004), and dog comparison (Lindblad-Toh *et al.*, 2005)) showed that at least 5% of the genome is probably functional.

In this study, we integrated the vast amount of functional genomic datasets to predict conserved regions in the human genome, as indicated in the workflow schema in figure 1. The result from applying the trained model is a single track with a score for each base in the genome

that predicts the probability the base is conserved. We show that our method Predictor and Identifier of Conserved Elements (PICEL) accurately predicts known conserved regions in non-exonic DNA, and also discuss the known conserved regions that our method is not predicting. Our false positive predictions can suggest potential important recently evolved regulatory locations in the genome.

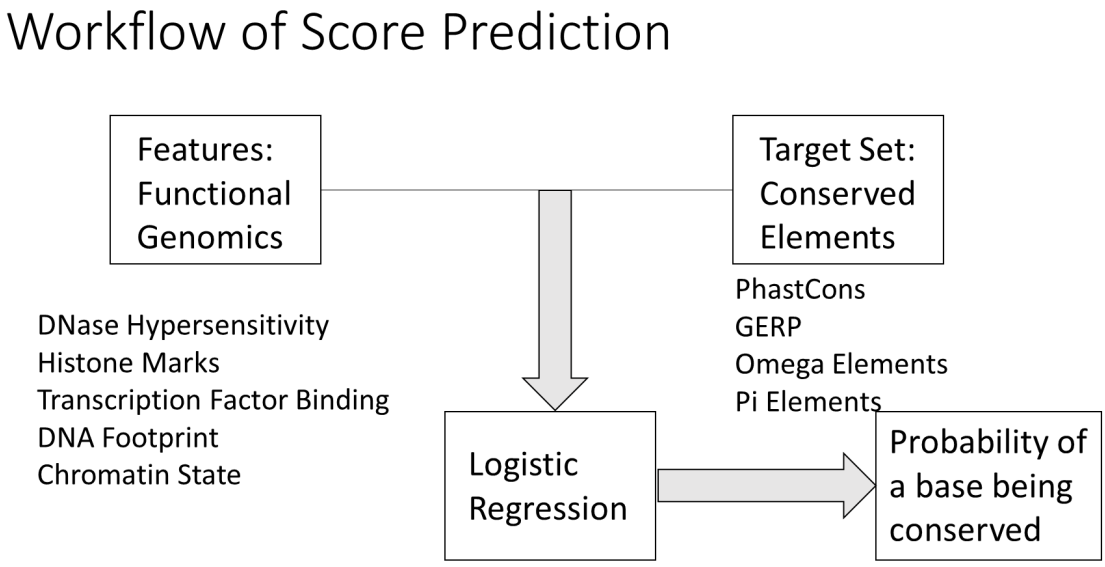

Figure 2.1: General Workflow of the Project

2.2 Background

In this section, we present in detail some currently available functional genomics and their limitations. In general, most of the conservation methods do not make use of functional information. Furthermore, methods designed to define conservation metrics across the genome are based on the assumption that genomic elements are present at orthologous locations and maintain similar function over long evolutionary time, but evolutional turnover may cause

inconsistencies between sequence orthology and functional homology (Gulko *et al.*, 2015). The shorter evolutionary time scales can be associated with information about genetic polymorphism in order to make this approach more robust to evolutionary turnover and less sensitive to errors in alignment, but polymorphic sites tend to be sparse along the genome (Fay *et al.*, 2001).

Combined Annotation-Dependent Depletion (CADD) method (Kircher *et al.*, 2014) is one of the first methods trying to close the gap and attempt to integrate functional and comparative data. CADD identifies differences between human genomes and the inferred human-chimpanzee ancestral genome in order to predict variants that are pathogenic. It uses samples from simulated and observed datasets, and supervised machine learning technique support vector machine (SVM) with linear kernel to accomplish this task (the latest version uses logistic regression). It was trained on 166 features derived from 63 annotations from functional and comparative data, with total of 949 features. Ten models were trained independently on observed and simulated data, and an average of those models was applied to score 8.6 billion possible single-nucleotide variants (SNVs) of the human reference genome (GRCh37). The model predictions are compiled into a single-value metrics named C score. Therefore, the advantage of CADD over other methods is that integrates diverse annotations of genetic variation into a single score.

Fitness consequence (fitCons) (Gulko *et al.*, 2015) is another method that attempts to integrate cell type specific functional genomic and comparative data. It estimates the probability that a point mutation at each position genome-wide will influence fitness. Therefore, fitCons scores are designed as evolution-based measure of genomic function. They are derived from ENCODE (Consortium, 2012) data for three human cell types and are inferred from patterns of genetic variation. In particular, genomic positions are clustered by their joint functional genomic

fingerprints from three data types: DNase-seq, RNA-seq, and ChIP-seq, and also using ChromHMM (Ernst e Kellis, 2010) with 25 states. fitCons method is reported to perform better than other discussed methods: phyloP (Rosenbloom *et al.*, 2015), phastCons (Siepel *et al.*, 2005), GERP++ (Cooper *et al.*, 2005; Davydov *et al.*, 2010), and CADD (Kircher *et al.*, 2014). However, this could be an artifact of discrepancy among the definitions of genomic function. fitCons represent the fractions of positions at which point mutations will have fitness consequences, but they do not account if sequences that would have fitness were deleted. fitCons scores in different cell types were concordant.

Another method that makes an attempt to integrate functional genomic data is Eigen (Ionita-Laza *et al.*, 2016). Similarly to CADD, Eigen proposes single meta-score that differentiates among disease-associated and benign variants in both coding and non-coding regions. Eigen method uses unsupervised spectral approach for scoring variants, which does not make use of labeled training data. This approach has an advantage when labeled data is limited, which is the case with annotation data, as different annotations measure different properties of a variant (e.g. degree of conservation, the effect of regulatory element in non-coding region, etc.). Like CADD's C-score, Eigen score needs to be interpreted with caution, because different annotations can measure different properties of a variant.

Among the most recent methods, GenoCanyon (Lu *et al.*, 2015) uses unsupervised statistical learning to accomplish whole-genome annotation. It uses 22 computational and experimental annotations (regions near significant loci from GWAS data) to predict many of the known functional regions and attempt to predict potentially new ones. The unsupervised learning is accomplished by estimating 49 model parameters using EM algorithm. The resulting score is prediction of functionality for each position in the human genome.

(Schrider e Kern, 2015) used supervised learning approach to predict conserved regions in the human genome. In particular, they trained support vector machine (SVM) classifier using known functional and non-functional portions of the genome and combined it with allele frequencies from low-coverage 1000 Genomes Project dataset. Some of the datasets they used as features are: disease-associated SNPs from GWAS compiled by (Hindorff *et al.*, 2009), and phastCons elements (Siepel *et al.*, 2005) from an alignment of 29 mammalian genomes, but ignoring human. The method recognizes previously known constrained portions of the genome (identified by phylogenetic methods), and uncovers new regions where gains and losses of function might have occurred, specifically in central nervous system and near neurotransmitter receptor genes.

2.3 Materials and Methods

2.3.1 Functional Genomics Datasets

Encyclopedia of DNA elements ENCODE (Consortium, 2012) project has a goal to map all functional elements in the human genome. So far ENCODE project has systematically mapped regions of transcription, TF association, chromatin structure, and histone modification, among others. The elements mapped and approaches used include RNA transcribed regions (RNA-seq, CAGE, RNA-PET and manual annotation), TFBS (ChIP-seq), chromatin structure (DNase-seq, FAIRE-seq), histone marks (ChIP-seq and MNase-seq), and DNA methylation sites (RRBS assay). Out of those, for our model training, we have only used ENCODE datasets with TFBS, histone marks, DNase peaks, and chromatin state core and imputed marks (Ernst e Kellis, 2010; Ernst *et al.*, 2011; Roadmap Epigenomics *et al.*, 2015) derived by ChIP-seq, DNase-seq, and FAIRE-seq assays. We used hg19 assembly of human genome and exclude exons as they are

already well-annotated and focusing on regulatory regions. We excluded RNA-seq datasets as RNA-type features are associated with exons.

DNaseI footprinting enables discovery of the sequence-specific TFs at a higher resolution. Footprints are quantitative markers of TF occupancy at nucleotide resolution because DNaseI cleavage is not uniform within DNaseI hypersensitive sites. (Neph *et al.*, 2012) provided an extensive map of the footprints, resulting with millions of TF sites (short sequence elements) in 41 cell lines (~1.1 million high-confidence footprints per cell type and collectively 45,096,726 6-40 base-pair footprint events across all cell types). The study found strong correlations between footprint occupancy and phylogenetic conservation for diverse TFs. For that reason, the footprint dataset was incorporated in training of our model.

The NIH Roadmap Epigenomics Consortium (Roadmap Epigenomics *et al.*, 2015) is, like ENCODE, another major project that attempts to map epigenomic data. It generated so far the largest collection of human epigenomes for different cells and tissues, profiled for histone modifications, DNA accessibility and methylation, and RNA expression. Regulatory elements are catalogued in global maps, based on their activators and repressors. Some of the resulting datasets from ChIP-seq and DNase-seq assays are selected to train our model.

In addition to ENCODE and Roadmap projects, there is a lot of data available publicly from other ChIP-seq experiments. (Griffon *et al.*, 2015) integrated 395 available human public (non-ENCODE) datasets from ChIP-seq experiments (downloaded from Gene Expression Omnibus and ArrayExpress databases) and created a map of regulatory elements by compiling the genomic localization of 132 different TF's across 83 different cell lines and tissue types. They compared resulting TF occupancy map with the one from ENCODE TF catalogue and

realized that public regulatory elements catalogue was complementary to ENCODE region. They identified 8.9 million ChIP-seq peaks bound by TFs in the human genome, out of those 5.4 million they classified as non-redundant. For our model training we're using original 395 non-ENCODE datasets as features.

Chromatin states data (Ernst e Kellis, 2010; Ernst *et al.*, 2011; Roadmap Epigenomics *et al.*, 2015) provides an annotation of the human genome based on diverse combinations of epigenetic marks. Two datasets we used for this work are a result of chromatin model applied to 9 (Ernst *et al.*, 2011) and 127 epigenomes (Roadmap Epigenomics *et al.*, 2015) respectively of different cell types using ChromHMM algorithm. ChromHMM uses a multivariate Hidden Markov Model to reveal chromatin states in many cell types, based on combinations of chromatin marks. Each chromatin state shows specific enrichments in functional annotations, sequence motifs and other characteristics related to biological roles. ChromHMM analysis revealed 15 distinct states for both of these datasets. One of the 15-state chromatin models consists of 8 active and 7 repressed states that were recurrently recovered (such as promoter, enhancer, insulator, transcribed regions, repressed regions and inactive state) and show evolutionary conserved non-exonic regions. The second chromatin state data set we used for model training also consisted of 15-states, but different ones due to being derived across 127 epigenomes. The third chromatin state data set we used for model training is 25-state model also based across 127 epigenomes (Roadmap Epigenomics *et al.*, 2015), but also contains imputed epigenomic signal. It is used to complete missing histone marks and therefore complement observed data, but also imputed data can be used even when observed data is available.

2.3.2 Conserved Regions Datasets

Several methods to detect conservation already exist and resulting datasets of identified conserved elements are available. For this study, we used four different datasets delineating conserved regions in the human genome: phastCons (Siepel *et al.*, 2005; Rosenbloom *et al.*, 2015), GERP++(Cooper *et al.*, 2005; Davydov *et al.*, 2010), SiPhy-Pi (Pi) and SiPhy-Omega (Omega) Elements (Garber *et al.*, 2009; Lindblad-Toh *et al.*, 2011), named after methods that were used to derive them. The four datasets contain differences not only because they are derived by different methods, but also because there are discrepancies in quality and granularity among them in terms of density and length of conserved regions they define. Each dataset is used separately to derive a response vector (labels) to train our method as described in methods section. The names of the files we used and sources where we obtained them are given in the table in data section, and all of them pertain to hg19 assembly of the human genome. The four sets of resulting predictions resulted with similar predictions, but there are slight differences that are later compared and contrasted in the discussion section. In this section, we give brief description of each of the four datasets used for model training.

PhastCons dataset (Siepel *et al.*, 2005) was generated by phastCons program, which is part of software package called PHAST (Phylogenetic Analysis with Space/Time models). PhastCons model uses shortened phylogenetic tree for comparison of genomes of various species. PhastCons method aims to identify regions that show slower substitution rates than sequences evolving neutrally and is therefore classified as rate-based method. The original study using phastCons method compared total of 18 genomes (five vertebrate, four insect, two C. elegans and seven yeast genomes, but not compared to each other). For labels in our method, we used a newer dataset that aligns 100 vertebrate genomes (Rosenbloom *et al.*, 2015).

GERP++ dataset (Cooper *et al.*, 2005; Davydov *et al.*, 2010) is an improvement to GERP (Genomic Evolutionary Rate Profiling) program (Cooper *et al.*, 2005). GERP method is also rate-based and uses the concept of rejected substitutions to discover sequences with fewer mutations than would be expected for sequences evolving neutrally. Rate-based methods are successful and widely-used, but do not capture all aspects of the evolution of functional sequences and the notion that the functional constraint can also act as a biased substitution pattern. GERP++ provides both nucleotide and element-level constraint scores. It uses maximum likelihood evolutionary rate estimation for position-specific scoring and dynamic programming to define constraint elements. It ranks the candidate elements based on statistical significance. GERP++ identified over 1.3 million constrained elements spanning over 7% of the human genome. We have used GERP++ dataset for labels in order to train our model (newer version than original GERP method).

SiPhy-Pi and SiPhy-Omega elements datasets (Garber *et al.*, 2009; Lindblad-Toh *et al.*, 2011) were generated by SiPhy (SIte-specific PHYlogenetic analysis) program. SiPhy-Pi method takes into account biases in the substitution patterns (selection acting on the pattern of mutations) in addition to the substitution rate, such as positions free to mutate between A and G as the last nucleotide in lysine codons AAA and AAG. SiPhy-Omega method only takes into account substitution rate. The study using SiPhy-Pi method, comparing 29 eutherian (placental) genomes across the four major mammalian clades, has greater power to detect evolutionary constraint than human-mouse-rat-dog comparison (Lander *et al.*, 2001; Waterston *et al.*, 2002; Gibbs *et al.*, 2004; Lindblad-Toh *et al.*, 2005) due to greater effective branch length (~4.5 substitutions per site) and high sequence accuracy (estimated error rates of only 1-3 miscalled bases per kilobase, that is ~50 fold lower than between the species nucleotide sequence difference).

The study concluded that 4.2% of the human genome is detectable (confidently identifiable) and at least 5.5% is constrained. Specifically, 3.6 million conserved elements were identified at a finer resolution of 12 basepairs. Most of the newly detected elements were present in non-coding regions. Further, the study interpreted ~60% of the identified bases as functional, playing protein-coding, RNA, regulatory and chromatin roles, while ~40% were unclassified. Roles of elements residing in intronic and intergenic regions were characterized by their overlap with evolutionary signatures (Kellis *et al.*, 2003; Stark *et al.*, 2007) of specific types of features.

When compared to PhastCons and GERP methods at false discovery rate (FDR) at 5%, SiPhy method identified significantly more bases as constrained in comparison to PhastCons and GERP, indicating that significant number of SiPhy-Pi elements do not overlap with PhastCons and GERP. Specifically, only 56% of the elements are shared by all three methods. SiPhy-Pi detects more degenerate sequences than the other two methods. Specifically, SiPhy elements showed enrichment in unique exonic and promoter bases in regions that are rich with degenerate regulatory motifs. The SiPhi-Pi set that we used for labels to train our model has conservation score of approximately 5%.

2.3.3 Data Access

Datasets for training features were obtained from ENCODE and Roadmap projects, and publicly available database compiled by (Griffon *et al.*, 2015). Table 1 illustrates datasets used to train our model and number of features obtained from each set of files. Table 2 contains file names of datasets used as labels in our model. Further information and links to each specific datasets are given in tables 3 and 4 in the supplementary material.

Source	Feature Types	Features
ENCODE	TFBS Uniform	690
ENCODE	DNase Uniform	125
ENCODE	Histone Peaks	274
ENCODE	Open Chrom Dnase	100
ENCODE	DNase	236
ENCODE	Dgf	64
ENCODE	TFBS – Sydh	359
ENCODE	TFBS – Univ. of Chicago	6
ENCODE	TFBS – Univ. of Washington	114
ENCODE	Open Chrom Chip	55
ENCODE	Histone	29
ENCODE	Histone	207
ENCODE	Open Chrom Faire	38
ENCODE	TFBS	692
ENCODE	Histone	280
ROADMAP	Peaks (narrow)	1915
ROADMAP	DNA Footprints	42
Public Data	Peaks	395
ROADMAP	Chromatin State Hmm	135
ROADMAP	Chromatin State Core Marks	1905
ROADMAP	Chromatin State Imputed Marks	3175

Table 2.1: Datasets used to train our model and number of features obtained from each set of files

Name	File Name
GENCODE v19	gencode.v19.annotation.gtf.gz
phastCons	phastconselements_100_hg19.txt
GERP	hg19.GERP_scores.tar.gz
Omega Elements	hg19_29way_omega_lods_elements_12mers.chr_specific.fdr_0.1_with_scores.txt
Pi Elements	hg19_29way_pi_lods_elements_12mers.chr_specific.fdr_0.1_with_scores.txt

Table 2.2: Datasets used as labels in our model

2.3.4 Data Access Supplementary

Datasets for training features were downloaded from ENCODE and Roadmap projects, and other publicly available databases. Links to each specific datasets are given in the table 3. The single-feature files contain histone mark peak calls in the first three columns (chromosome number, start and end coordinates of the peak). The coordinates in these files are 0-based with the first coordinate inclusive and the second coordinate exclusive. Each file is treated as one feature (total of 5226 files and therefore 5226 features).

Source	Files	Features	Link
ENCODE	690	690	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/*.narrowPeak.gz
ENCODE	125	125	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgDnaseUniform/*.narrowPeak.gz
ENCODE	274	274	http://www.broadinstitute.org/~anshul/projects/encode/rawdata/peaks_histone/mar2012/narrow/combrep_and_ppr/
ENCODE	100	100	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeOpenChromDnase/*.narrowPeak.gz
ENCODE	236	236	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/*.narrowPeak.gz
ENCODE	64	64	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDgf/*.narrowPeak.gz
ENCODE	359	359	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/*.narrowPeak.gz
ENCODE	6	6	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUchicagoTfbs/*.narrowPeak.gz
ENCODE	114	114	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwTfbs/*.narrowPeak.gz
ENCODE	55	55	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeOpenChromChip/*.narrowPeak.gz
ENCODE	29	29	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhHistone/*.narrowPeak.gz
ENCODE	207	207	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwHistone/*.narrowPeak.gz
ENCODE	38	38	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeOpenChromFaire/*.narrowPeak.gz
ENCODE	692	692	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEnco

			deHaibTfbs/*.broadPeak.gz
ENCODE	280	280	ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone/*.broadPeak.gz
ROADMAP	1915	1915	http://egg2.wustl.edu/roadmap/data/byFileType/peaks/unconsolidated/narrowPeak/*.narrowPeak.gz
ROADMAP DNA Footprints	42	42	http://egg2.wustl.edu/roadmap/data/byDataType/dgfootprints/*.narrowPeak.gz
Publicly Curated Data	1	395	http://tagc.univ-mrs.fr/remap/download/All/nrPeaks_public.bed.gz
ROADMAP Chromatin State	9	135	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHmm/*
ROADMAP Chromatin State	127	1905	http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/*mnemonics.bed.gz
ROADMAP Chromatin State	127	3175	http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imputed12marks/jointModel/final/*mnemonics.bed.gz

Table 2.3: Datasets used to train our model and access links for each feature file

The multi-feature files (public currated data and chromatin states) contain feature names in the fourth column, in addition to chromosome number, start and end coordinates of the peak in the first three columns. The coordinates of the peaks in these files are 0-based with the first coordinate inclusive and the second coordinate exclusive. Each unique entry in the fourth column is a separate feature so the combination of the file and the fourth column uniquely identifies the feature. There is a total of 264 files and 5610 features, depending on how many features each file contains. Publicly curated data contains 1 file, which contains 395 distinct features, which are corresponding to different datasets. For chromatin model, every genome position should be in one chromatin state in each cell type by the Roadmap ChromHMM. First Chromatin State dataset contains 9 cell types (there is one file per cell type) and 15 states per file and therefore 135 features, the second dataset contains 127 files and 15 states per file totaling 1905 features, and the third dataset contains 127 files and 25 states per file, totaling 3175 features.

For conserved regions, we used GENCODE v19 (hg19) as reference, in order to remove exons from our model. In this file, exon positions are those lines for which the third column has the entry exon, the first column is chromosome, the fourth column is the start position and the fifth column is the end position. These coordinates are 1-based with start and end inclusive. Phastcons file is a single file. The coordinates are 0-based, with start inclusive and end exclusive. GERP file is a single zipped file that contains separate files for each chromosome named hg19_chr#_elems.txt (where # stands for chromosome number). These coordinates are 1-based and start and end inclusive. Omega and Pi elements files are a single file each. The coordinates are 0-based, start and end inclusive. Links to each specific datasets are given in the table 4.

Name	File Name	Link
GENCODE v19	gencode.v19.annotation.gtf.gz	ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_19/
phastCons	phastconselements_100_hg19.txt	http://hgdownload.cse.ucsc.edu/
GERP	hg19.GERP_scores.tar.gz	http://mendel.stanford.edu/SidowLab/downloads/gerp/
Omega Elements	hg19_29way_omega_lods_elements_12mers.chr_specific.fdr_0.1_with_scores.txt	http://www.broadinstitute.org/scientific-community/science/projects/mammals-models/29-mammals-project-supplementary-info
Pi Elements	hg19_29way_pi_lods_elements_12mers.chr_specific.fdr_0.1_with_scores.txt	http://www.broadinstitute.org/scientific-community/science/projects/mammals-models/29-mammals-project-supplementary-info

Table 2.4: Datasets used as labels in our model and access links for each feature file

2.5 Method

PICEL was trained using logistic regression to predict non-exonic in the human genome. In order to accomplish this task, we have chosen 10,836 functional genomics features to be directly related to gene regulation. This number of features supersedes by far the amount of features used by other similar methods described earlier in this manuscript such as CADD (Kircher *et al.*, 2014), fitCons (Gulko *et al.*, 2015), Eigen (Ionita-Laza *et al.*, 2016), GenoCanyon (Lu *et al.*, 2015), and method by (Schrider e Kern, 2015).

Majority of the features chosen to train PICEL (5579 features, which equals to 51.5% of all features) are peak or domain calls from DNase-seq and FAIRE-seq assays, and histone modification TFBS derived by ChIP-seq experiments from various sources. In addition, a small subset of features for training (~.5%) also included DNase I footprint data. As described earlier, footprints are quantitative markers of TF occupancy at nucleotide level and demonstrate protein-DNA interaction. TFBS are likely to be regulatory elements and therefore provide useful epigenetic information and are therefore likely to be predictive to conservation in the genome.

Approximately another half of the features used to train PICEL (5215 features, which equals to 48% of all features) are chromatin state datasets (with core and imputed marks). As discussed earlier, chromatin states (Ernst e Kellis, 2010; Ernst *et al.*, 2011; Roadmap Epigenomics *et al.*, 2015) provide diverse classes of epigenetic function in the human genome. Two datasets we used for this work are a result of chromatin model applied to 9 and 127

epigenomes respectively of different cell types using ChromHMM algorithm. The third chromatin state dataset we used for model training is imputation based 25-state model across 127 epigenomes, which contains imputed epigenomic signal. It is used to complete missing histone marks and therefore complement observed data, but also imputed data can be used even when observed data is available. Combined patterns of chromatin marks show evolutionary conserved non-exonic regions and are therefore good candidates for our training set.

Each feature subset and data source is discussed in more detail elsewhere in this manuscript and the information about the number of individual features in each subset and source from where they were obtained is provided in the tables.

The response vector for training PICEL contains binary labels indicating whether the sampled location in the human genome is conserved or not. The labels are derived from a dataset that already contains known conserved regions, except that locations within exons are excluded because exploring non-exonic regions of the human genome is of primary interest in this project. In particular, if randomly sampled location happens to be an exon, it was labeled as non-conserved region. Also, we have discarded any conserved regions located at random fragments that were present in GERP++ file, as those weren't pertaining to a specific chromosome (or at least their location hasn't been confirmed). Also, we have excluded chromosome Y from our model training in all cases. We repeated model training with four different sets of labels, each coming from a different dataset: phastCons, GERP, Pi, and Omega elements.

2.6 Model Training

The model training was done by randomly sampling positions from genome-wide locations on all chromosomes, except for one chromosome that was designed as a test chromosome. The same procedure was repeated 23 times, as each time a different chromosome was designed as the test chromosome. Sampling of the positions on the chromosomes was done proportional to chromosome size. 10 sets of samples of 1 million each were trained individually and predictions were averaged across the 10 resulting models, totaling to 10 million samples. The same procedure has been repeated 23 times, where each time a different chromosome was designed as a test chromosome, and the model was trained on samples drawn from all but the test chromosome. Furthermore, the entirely described procedure was repeated four more times, for four different sets of response labels (phastCons, GERP, Pi and Omega elements) as mentioned earlier.

Given the large amount of samples, the training matrix was logged in sparse format, indicating only matrix coordinates of the locations at which features were present (non-zero entries), and not listing locations where the features were not present (zero entries). However, even in sparse format, the files are very large. The number of non-zero entries in datasets was approximately 50 million.

The matrix was created in programming language C and the source code was unified to do the sampling, feature extraction from various data sources, and label identification all at once. The code was executed on the cluster computer with 32GB of memory (RAM). The code is also able to handle two million samples at once (twice as many), when 64GB of memory (RAM) is available (twice as much).

We were able to reduce the time to create training matrix down to several hours from expected several days or longer (the matrix size was very large: 1,000,000 rows x 10,836 columns). In order to do that, we made several algorithmic improvements.

One of the major improvements was to create training vectors separately for each feature and write them to intermediate files. This way we didn't need to wait for 10,836 features to be processed serially one by one, which was not only computationally intensive task (for each feature). Instead, feature vectors were created in parallel, which significantly reduced computational time, but it also reduced any additional I/O access to original datasets (features files) because all the information about training features was already processed. In addition, this added flexibility to the code, as more features (datasets or tracks) could be added without previously collected features having to be recalculated for the same training set.

However, about 50% of the features were contained in files that consisted of multiple features (not in single file), as mentioned in data section. Not all training vectors were able to be created simultaneously in parallel as repeated passes through multi-feature files were needed. Therefore, to compromise between the two (single vs. multiple feature files) and to accommodate for flexibility for adding more dataset tracks later, our code was split into sections, so that given the argument passed when running the program, a certain section of features gets processed. All sections were executed in parallel.

Having intermediate vector files also helped with redundancy (back up),, and thanks to this improvement the process could be continued from the breaking point instead of all the way from the beginning (e.g. if 80% of the features had already been processed, then the run could get restarted to only complete processing the remaining 20% of the features). Requiring extra

space for the intermediate vector files was a minor issue in comparison to the gain of overall performance improvement, flexibility and reliability. Besides, if disk space were an issue, those files could've been written to the "scratch" partition that was designed to occasionally automatically get purged.

Once training vectors were written to separate files, they were collected into a matrix, and the matrix was transposed. Transposing the training matrix was memory and computationally intensive task, but was fortunately needed to be done only once per training set. Since row vector files were containing only binary values, we took advantage of Boolean type in C that required less memory (at least 4 times less memory than integer value). Given that intermediate row vector files were identical in format (unlike feature files) and smaller than originally feature files (datasets containing more information), we did not experience any significant issues loading them after figuring out the memory requirement for the size of the matrix. The time to transpose the matrix and create a file in the format required by Liblinear package to perform logistic regression was measured in hours (e.g. around 12 hours, depending on the power of the cluster node assigned for the job), which is still significantly short given the size of the problem (which could take days).

One more improvement included the code for determining labels (whether sampled genome position was considered a conserved region or not). All four datasets used for labels were processed in parallel, and separate label vectors were created in parallel (again, using Boolean type in C that uses at least 4 times less memory than integer). Therefore, when training matrix was transposed, the labels were appended for four different datasets and four different files (in the file format required by Liblinear) were created right away. Therefore, there was no

need to transpose the matrix four times for the same training set, just because it needed different labels.

Another major improvement to feature extraction algorithm was addition of temporary 2D array (matrix) to help speed up creation of the training vector for each feature. First, we sorted all feature files using sort() utility in Unix so that tuples (chromosome number, start position, end position) were sorted by chromosome number in ascending order, then by start position also in ascending order. Given the number and size of the datasets (feature files), this was time consuming task (it took approximately half a day), but was only done once.

The training samples were sorted as well (after being randomly sampled genome-wide), in a manner that all positions pertaining to a particular chromosome were logged in a single row (and sorted in ascending order), and each row of the matrix pertained to one chromosome (again in ascending order, starting from chromosome 1 up to chromosome 23, which was chromosome X, as we ignored the information from chromosome Y).

The temporary 2D array (matrix) was initialized to all zeros, and if the feature was present at the particular sampled location (if the entry containing the interval of the sampled position was found in the feature file), the matrix value for that sample was changed to one. Because the feature files were sorted, logging whether a feature was present at a certain sampled location on the chromosome was speeded up because as the program was parsing through a feature file, it would log features present at a particular chromosome in a single row of a matrix, then when the chromosome number changed in the feature file, it would hop to the next row of the matrix. This way access to 2D array was serialized, and given that the access was row by row (which is how matrices are stored in memory) and not column by column, this provided

significant speedup in processing single feature file. Feature vector was created also by passing through temporary 2D array row-wise.

Finally, we took advantage of job parallelization by simultaneously scheduling array jobs that could run independently in parallel on the cluster computer.

2.7 Logistic Regression

Logistic regression was performed using Liblinear package (Fan *et al.*, 2008) on the training matrix created by PICEL method. Here we give a brief overview of logistic regression.

Logistic regression is traditionally one of the tools used for discrete data analysis, possibly because it often works well as a classifier. It models the conditional probability P(Y=1|X=x) as a function of x (where X is a feature vector and x is an individual feature) for a binary output variable Y. The hypothesis function p ($\theta^T x$), where θ is vector of parameters, is chosen to be sigmoid function (termed "logistic function"):

$$p\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

The output values of the sigmoid function could be any real value in the range [0..1] (including zero and one, which are boundaries). Any unknown parameters in the function can be estimated using maximum likelihood.

Formally, the model for logistic regression is (Shalizi, 2015):

$$\log \frac{p(x)}{1 - p(x)} = \beta_0 + x \cdot \beta$$

Where solving for p gives:

$$p(x;b,w)=\frac{e^{\beta_0+x\cdot\beta}}{1+e^{\beta_0+x\cdot\beta}}=\frac{1}{1+e^{-(\beta_0+x\cdot\beta)}}$$

The decision boundary separating two classes is the solution of the linear term:

$$\beta_0+x\cdot\beta$$

This makes logistic regression a linear classifier. The classifier not only shows where the boundary between the two classes is located, but that the class probabilities depend on the distance from the boundary in a particular way, which go towards the extremes (zero or one) more rapidly depending on the value of β. Most importantly, logistic regression predicts probabilities and is also well calibrated, for example, for classes C and "not C" p(C|X)=[0..1] and p("not C"|X) > .5

The performance of a model trained using logistic regression is measured by applying it to test samples and calculating area under the curve (AUC) value, which ideally would be an estimate of the probability that the classifier will assign a higher score to a randomly chosen positive example than to a randomly chosen negative example. Statistically, AUC is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one(Fawcett, 2006) and is based on calculating true positive rates and false positive rates. Receiver operating curve (ROC) plots true positive vs. false positive rate for all possible cutoffs, and AUC is simply the area (integral) under ROC curve. Therefore, the purpose of AUC is to tell us something about family of tests, not an individual test, but one for each possible cutoff value.

The pseudo code for the simple way of calculating AUC is the following:

First, rank scores obtained for test examples in decreasing order, then:

auc = 0.0

height = 0.0

for each training example x_i, y_i

    if y_i = 1.0

        height = height + tpr

    else

        auc = auc + height * fpr

return auc

2.8 Model Testing, and Validation

Given that the model was trained using large number of samples, over-fitting ended up not being a problem. We have run many experiments in this area in order to make sure that is the case. The model was trained multiple times, including times when regularization was attempted. For regularization, various values of regularization parameter lambda were used and in the range from 10^{-5} to 100, with step 10 in terms of power in order to understand when the AUC value reaches saturation point on both ends (lower and upper bound). Also, two types of logistic regression regularization were attempted: l1 and l2. L1 regularization is based on Laplacian distribution which contains a term with absolute value (e.g. |x|) and therefore favors zero weights, while l2 regularization is based on Gaussian distribution which contains square term (e.g. x^2) and favors weights that are close to zero, but not exactly zero. The reason is that zero weights penalize some features and favor others, which could be good approach in some cases. It is noteworthy, though, that if there are two feature with the same weight and one of them gets favored randomly the other one might get ignored. In our training, almost all of the resulting

weights were non-zero whether using l1 weights for most values of the regularization parameter or l2 regularization. Table 5 illustrates some of the average weights of 10 classifiers for some of the features (rows) used for model training. Each column indicates a model trained with a different set of labels (conserved element sets: Omega, Pi, GERP++, and phastCons). The only notable difference between l1 and l2 regularization was speed, because l1 took much less time to run than l2 (approximately 10-15 minutes vs. 10-15 hours).

Average Weights			
Omega	Pi	GERP++	phastCons
-0.03	-0.03	-0.01	0.00
-0.02	0.01	-0.03	-0.02
-0.02	0.00	-0.03	-0.03
0.02	0.04	0.02	0.02
0.03	0.00	0.02	0.03
-0.03	-0.02	-0.03	-0.03
0.01	0.02	0.01	0.00
0.02	0.02	0.04	0.02
-0.01	0.00	-0.02	-0.02
-0.04	-0.04	-0.05	-0.05
-0.05	-0.04	-0.04	-0.03
-0.03	-0.01	-0.02	-0.02
-0.02	-0.02	-0.03	-0.02
-0.01	0.00	-0.01	-0.01
0.05	0.06	0.06	0.05
0.12	0.07	0.15	0.12
-0.01	0.00	-0.01	0.00
-0.09	-0.07	-0.09	-0.08

Table 2.5: Average weight of 10 classifiers for some of the features (rows) used for model training. Each column indicates a model trained with a different set of labels (conserved element sets: Omega, Pi, GERP++, and phastCons).

The area under the curve (AUC) value reached its saturation point when the value of regularization parameter lambda was not small (as mentioned earlier, values of regularization parameter lambda were used in the range from 10^{-5} to 100, with step 10 in terms of power), as indicated in table 6. This showed that the amount of samples was large enough to avoid overfitting and therefore any non-small value of parameter lambda would suffice (regularization was not needed).

Type	l1		l2	
Samples	100 million	10 million	100 million	10 million
lambda	AUC			
0.00001	0.5823	0.574	0.7381	0.7343
0.0001	0.6218	0.6227	0.8056	0.8009
0.001	0.7511	0.7493	0.823	0.8181
0.01	0.8141	0.8097	0.829	0.8236
0.1	0.8284	0.823	0.8303	0.8242
1	0.8309	0.8241	0.8306	0.8242
10	0.8309	0.8237	0.8306	0.8243
100	0.8309	0.8236	0.8306	0.8244

Table 2.6: Resulting AUC values after regularization (varying the value of parameter lambda). Model trained using l1and l2 logistic regression on 100 million and 10 million data points representing genome-wide chromosome positions on all chromosomes except chromosome 1.

To push the limits even further, we have successfully trained the model on 100 million samples in total by training 100 models of one million samples each and averaging resulting predictions. This particular run was done only once, where training set were all chromosomes except chromosome 1, and chromosome 1 was designed as a test chromosome. As shown in table 6, AUC values were only slightly higher, indicating that adding more samples was reaching diminishing returns. Likewise, averaging two models of one million samples each or training a

single model of two million samples didn't show significant difference (similar AUC values in all three cases, as shown in table 7). Having more than two million samples in a single training set was not feasible with current technical abilities of the cluster computer we were using and model implementation. Alternatively, it might've been possible to use stochastic gradient descent regression, that doesn't require all samples to be provided at once, but the number of samples our implementation could process was satisfactory.

Samples	10x1 million	1x2 million	10x2 million
Chromosome	AUC		
chr1	0.828552	0.816697	0.828579
chr2	0.821355	0.807488	
chr3	0.822076	0.810995	
chr4	0.822461	0.811989	
chr5	0.821542	0.810629	
chr6	0.822243	0.811814	
chr7	0.822673		
chr8	0.822734		
chr9	0.821938		
chr10	0.821636		
chr11	0.82241		
chr12	0.822625		
chr13	0.82271		
chr14	0.822914		
chr15	0.822295		
chr16	0.822132		
chr17	0.822155		
chr18	0.823042		
chr19	0.822897		
chr20	0.822805		
chr21	0.822756		
chr22	0.823071		
chr23	0.821613		

Table 2.7: AUC values received when testing the model. Training set contained genome-wide positions from all chromosomes, except single chromosome, test set *contained the* excluded chromosome. Model trained using l1 logistic regression with regularization parameter lambda = 1. In the first column, predictions were averaged from 10 models,

each one containing 1 million samples (total of 10 million samples). In the second column, predictions were based on a single model containing 2 million samples. In the third column, predictions were averaged from 10 models of 2 million samples each (total of 20 million samples). AUC values are similar enough to indicate that the optimal amount of samples for model training is 10 x 1 million samples (when computational time and memory are taken into account).

Therefore, based on experimental results and taking into account technical availabilities (e.g. computational time and memory), we have chosen the amount of samples for model training to be 10 x 1 million samples. Thus, for final model training we have settled on using l1 logistic regression with regularization parameter lambda =1 (default value), and 10 million samples (10 models trained on one million samples each and then averaged predictions) per training set for each test chromosome.

Finally, we have applied our trained model parameters genome-wide and obtained a score for each position in the entire human genome. This was computationally intensive task and required some clever algorithmic design and parallelization, which reduced the memory requirement 4 times (from 32 GB to 8 GB needed for creating a training set), and computational time 2-fold (from initial 1 to 5 days to only 12 to 60 hours) depending on the size of the chromosome (taking into consideration that chromosome 1 is five times larger than chromosome 21).

First of all, the genome-wide process was parallelized by working on each chromosome separately (23 array jobs were running in parallel). Even though this was major time-saving step, it would've still taken (more) days to complete the computations. The second major improvement was splitting each of the jobs operating on a single chromosome into 10 parts. This ended up being 230 jobs times four for four different models (due to four different datasets for labels), which totaled 920 jobs running in parallel (theoretically). Practically, the scheduler did

not allow this many jobs to run in parallel, and certainly wouldn't prioritize scheduling any of the jobs if they required a lot of memory, so the second major algorithmic improvement was to reduce memory requirement.

The significant memory reduction was achieved by storying as little information as possible in order to calculate probabilities (scores). While the parameters of all training vectors were stored in the matrix (in order to keep calculations for multiple models efficient and parallel), whether the feature was present or not at each particular position on the chromosome was not stored in a matrix. Instead, we only kept a single vector that was keeping the running sum. To be more specific, the vector was initialized to zero. Once the first feature was processed, all the entries in the vector pertaining to positions that contained that particular feature were changed to the weight of that feature (specifically, the new weight was added to the originally initialized zero weight). Then, once the next feature was processed, also the weight for that particular feature was added to the previous value in that position in the vector. Once the last feature was processed, the vector contained the running sum for each position on the chromosome. Then the value of the intercept was added and the running sum was plugged into a logistic function, in order to calculate probabilities (scores). Since the final scores were averaged of 10 (or more) models, we introduced an extra vector to keep track of accumulated probabilities at each position for all models, which were averaged in the end. None of these steps required 2D arrays (except for the array keeping track of parameter vectors for each model), and therefore didn't have as large memory requirement.

As mentioned earlier, each run per chromosome was split to be processed as 10 runs in parallel, each one pertaining to the different part of the chromosome. Therefore, the above described vectors were 10 times shorter than the chromosome size, which resulted in memory

requirement per job reduced 8 times from the original version. This was possible to do, because the scores for each genome position were independent.

However, this did require extra coding effort to make sure that each part of the chromosome was processed correctly (e.g. that each position where the feature is located is logged in the right section). Even more specifically, start and end position of a particular feature location could span multiple chromosome parts and it required extra caution that they those features get logged in all places (parts). It is worth noting that some feature files contained positions of the peaks starting before the beginning of the chromosome and end positions ending beyond the ends of the chromosome, probably due to read errors during sequencing.

Some of the algorithmic improvements were similar to already described in model training section. The processing of feature files (datasets or tracks) was split into sections, so that given the argument passed when running the program a certain section of features would get processed. That way I/O access to feature files (datasets) was minimized. All sections were executed in parallel. Similarly to also previously described, given that feature files (datasets) were sorted, only sections of the files containing lines pertaining to a certain chromosome needed to be considered, while lines for other chromosomes were skipped.

Finally, concatenated 10 resulting files from each part of the chromosome into a single .wig file with scores for the entire chromosome that can be loaded into genome browser. Similarly, we concatenated .wig files for all chromosomes into a single .wig file for more comprehensive viewing.

2.9 Results

Receiver operating characteristic (ROC) curve for each element set along with Area under the Curve (AUC) value is given in figures 2, 3, 4, and 5 respectively. The highest AUC value for GERP element set was 0.83 and 0.82 for both SiPhy-Pi and SiPhy-Omega elements, indicating good model performance. Only the AUC value for phastCons elements was slightly lower: 0.76.

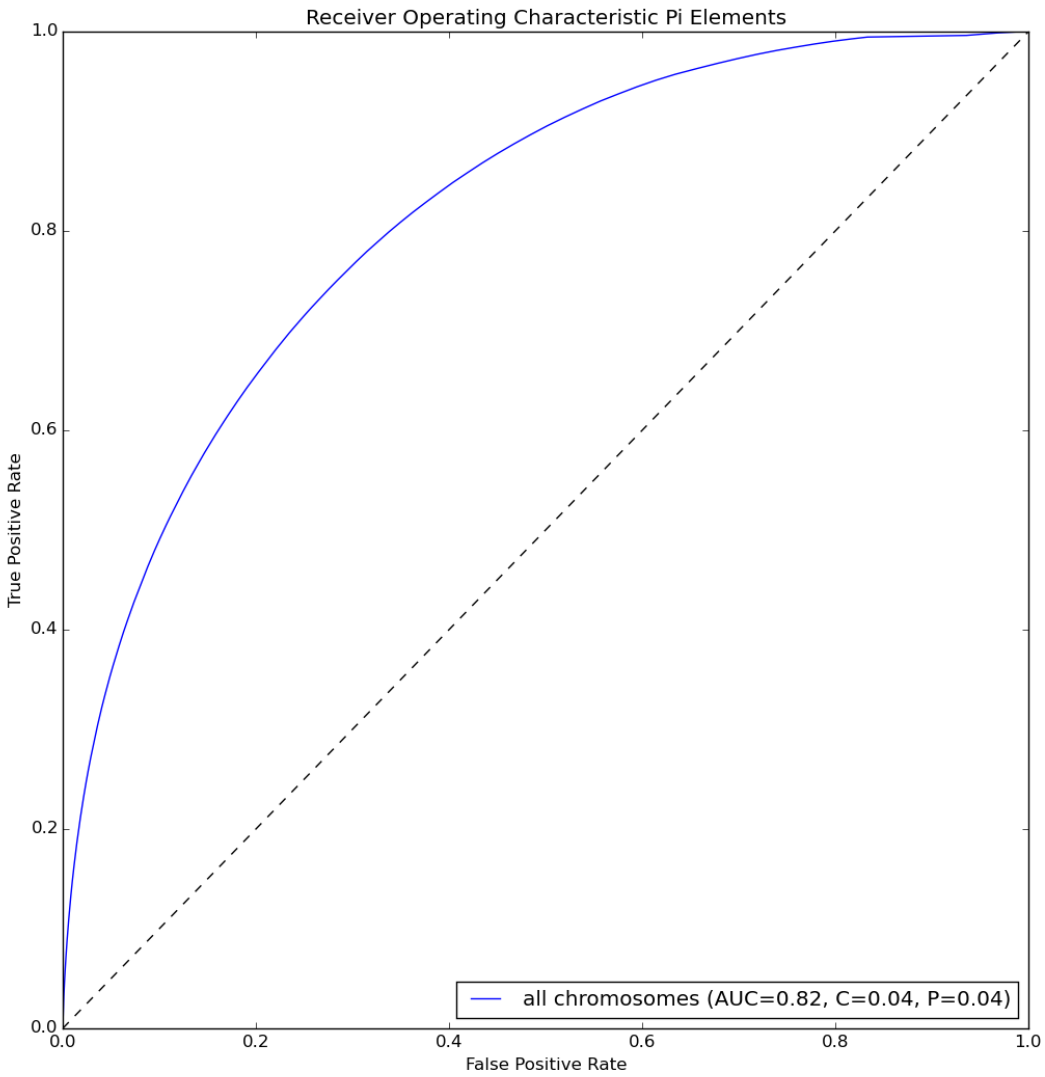

Figure 2.2: Single ROC curve for all models trained using10 million sampled points from all chromosomes genome-wide except the chromosome numbered on the plot and labels based on Pi elements dataset.

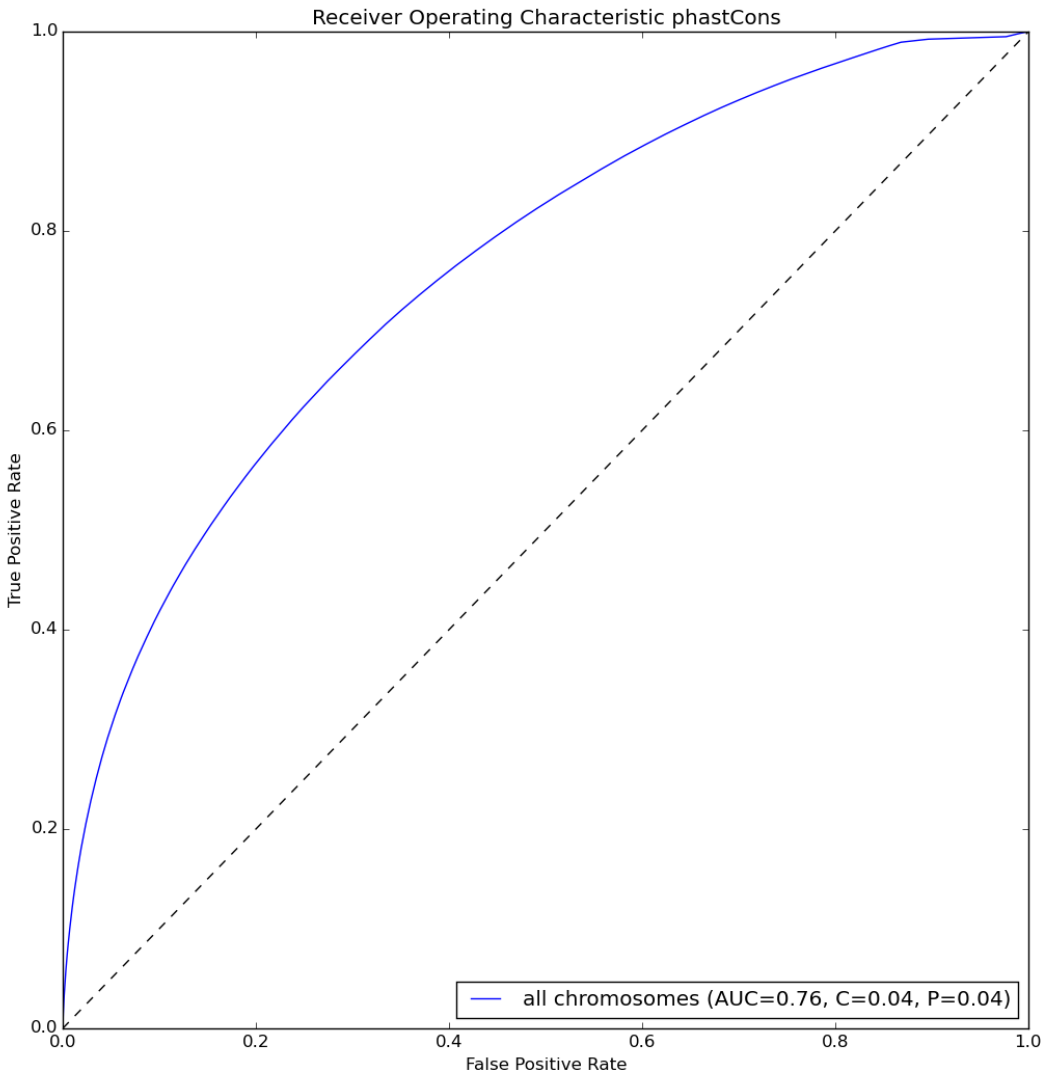

Figure 2.3: Single ROC curve for all models trained using10 million sampled points from all chromosomes genome-wide except the chromosome numbered on the plot and labels based on phastCons dataset.

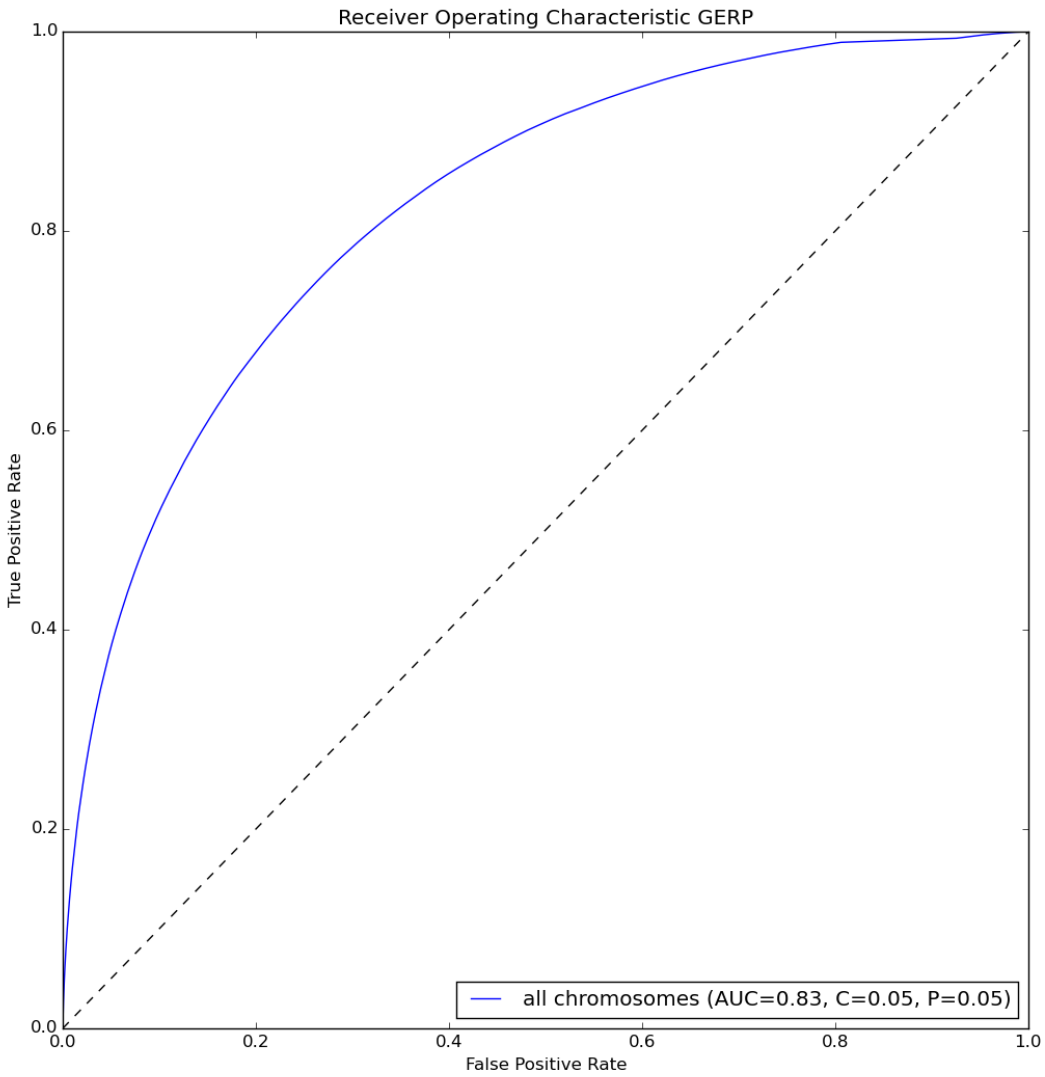

Figure 2.4: Single ROC curve for all models trained using10 million sampled points from all chromosomes genome-wide except the chromosome numbered on the plot and labels based on GERP dataset.

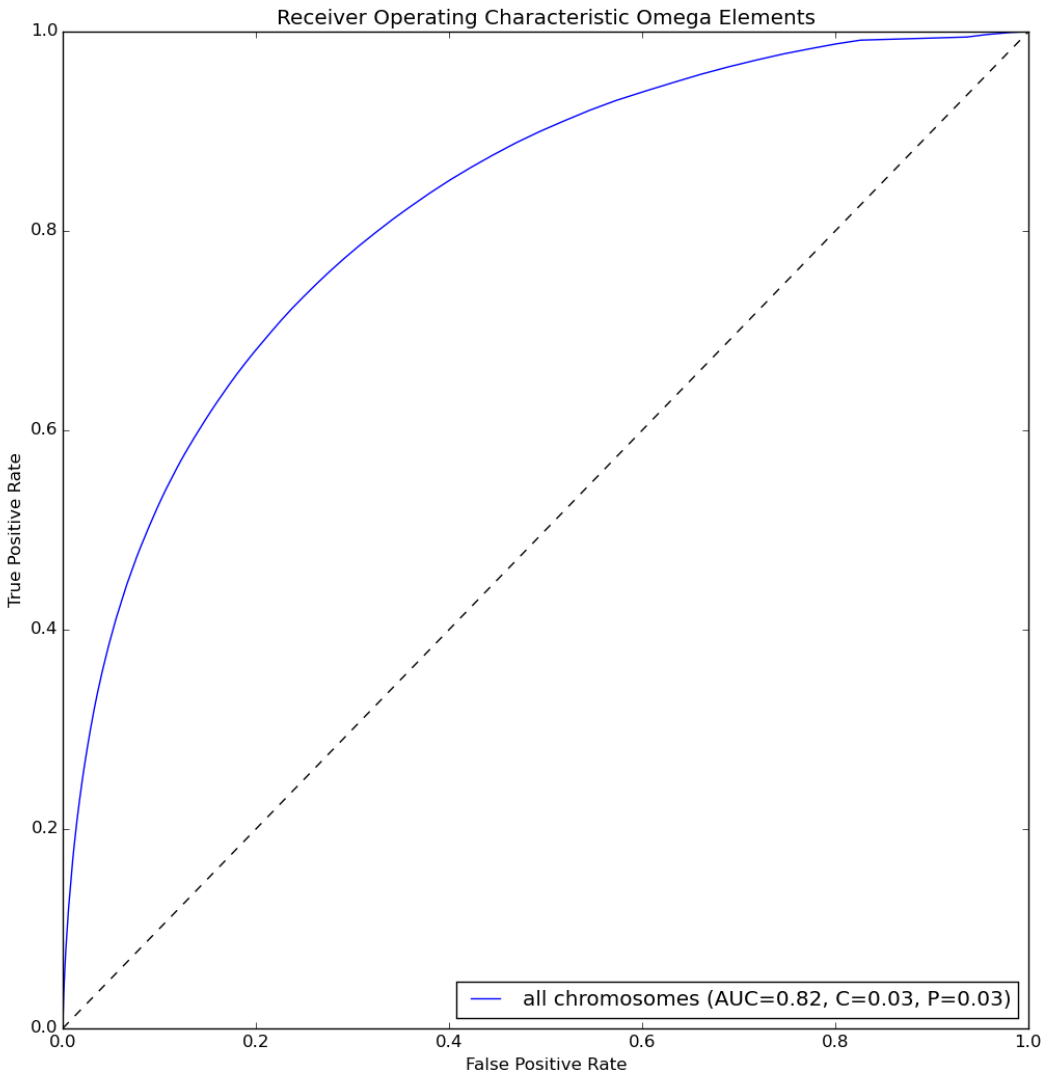

Figure 2.5: Single ROC curve for all models trained using10 million sampled points from all chromosomes genome-wide except the chromosome numbered on the plot and labels based on Omega elements dataset.

The AUC values for each test per chromosome are given in table 8.

10 milion	Area Under the Curve (AUC) Value			
lambda = 1	Pi	phastCons	GERP	Omega
chr1	0.77	0.76	0.83	0.82
chr2	0.80	0.76	0.82	0.82
chr3	0.79	0.75	0.81	0.81
chr4	0.80	0.75	0.81	0.81
chr5	0.80	0.75	0.81	0.81
chr6	0.80	0.75	0.81	0.81
chr7	0.82	0.75	0.82	0.82
chr8	0.80	0.76	0.82	0.81
chr9	0.85	0.78	0.84	0.85
chr10	0.80	0.75	0.82	0.81
chr11	0.81	0.75	0.81	0.80
chr12	0.80	0.74	0.82	0.81
chr13	0.84	0.79	0.84	0.85
chr14	0.84	0.80	0.85	0.85
chr15	0.85	0.79	0.85	0.85
chr16	0.83	0.78	0.84	0.82
chr17	0.80	0.74	0.81	0.80
chr18	0.81	0.76	0.82	0.83
chr19	0.85	0.73	0.84	0.83
chr20	0.82	0.75	0.82	0.82
chr21	0.86	0.80	0.86	0.86
chr22	0.87	0.81	0.86	0.86
chr23	0.81	0.71	0.78	0.82

Table 2.8: Comparison of AUC values on per single chromosome basis for the four conserved element sets: Pi, Phastcons, GERP, and Omega. Lambda is a regularization parameter (non-small value of lambda indicates very little regularization).

ROC curves for each set of conserved regions and each test per chromosome are given in figures 6, 7, 8, and 9 respectively.

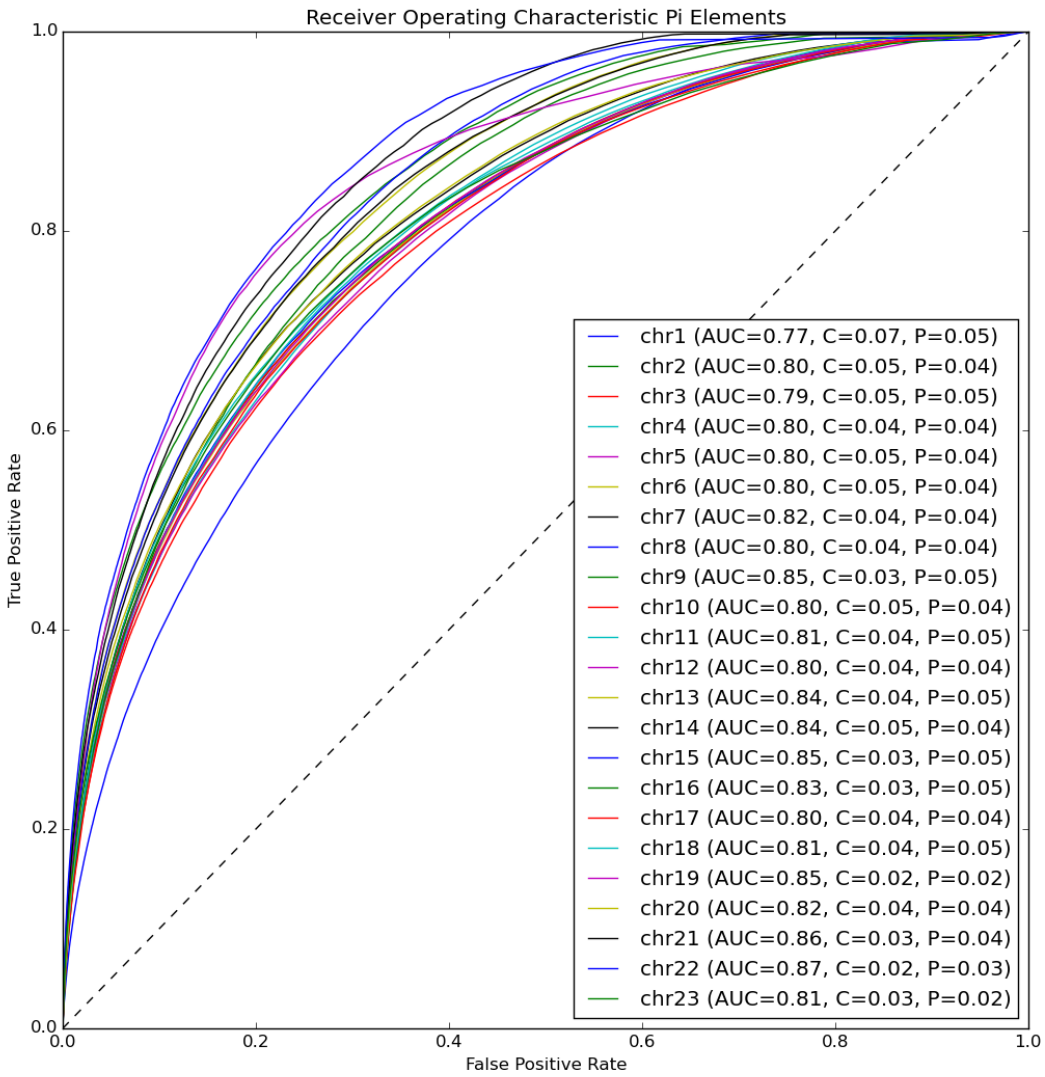

Figure 2.6: ROC curves for model trained using10 million sampled points from all chromosomes genome-wide except the chromosome numbered on the plot and labels based on Pi elements dataset.

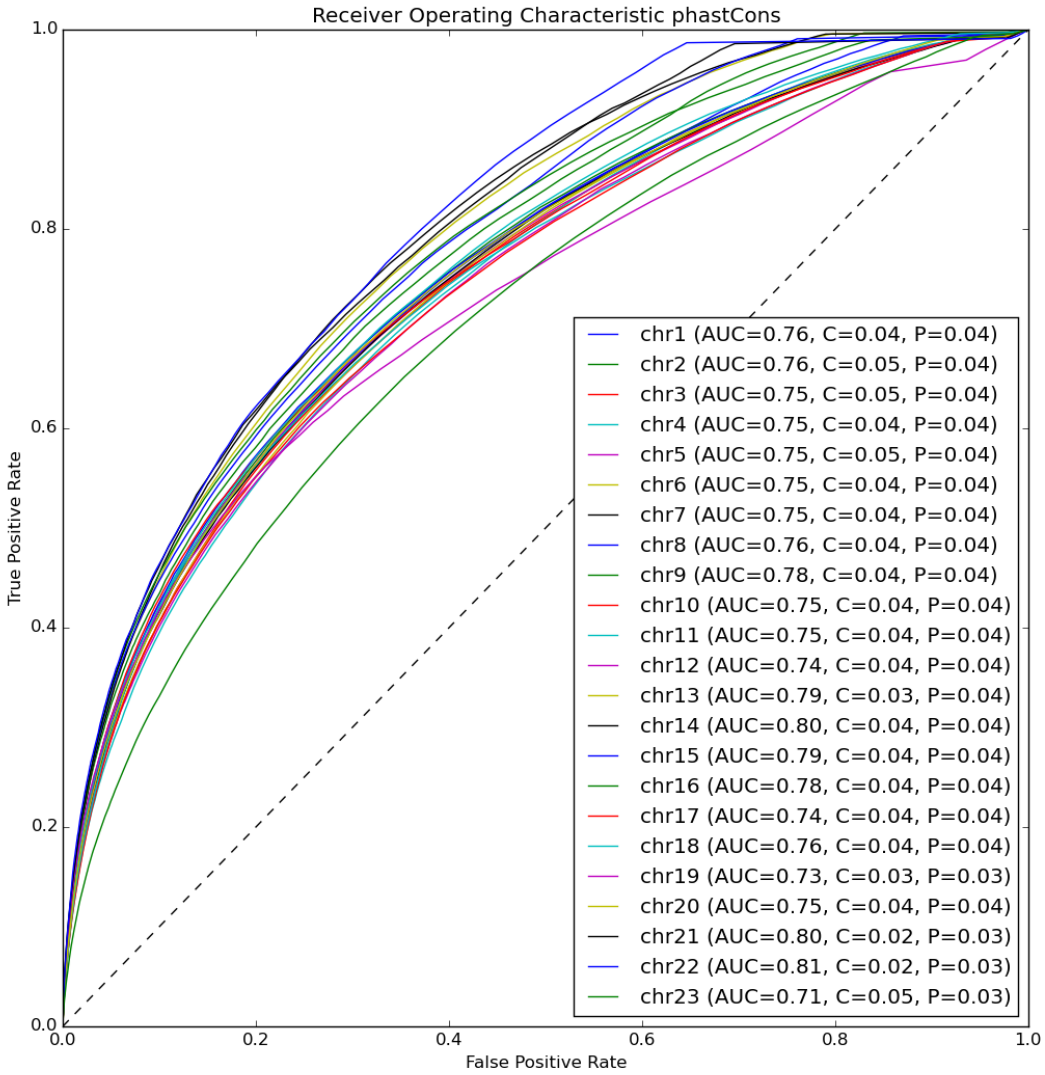

Figure 2.7: ROC curves for model trained using10 million sampled points from all chromosomes genome-wide except the chromosome numbered on the plot and labels based on phastCons elements dataset.

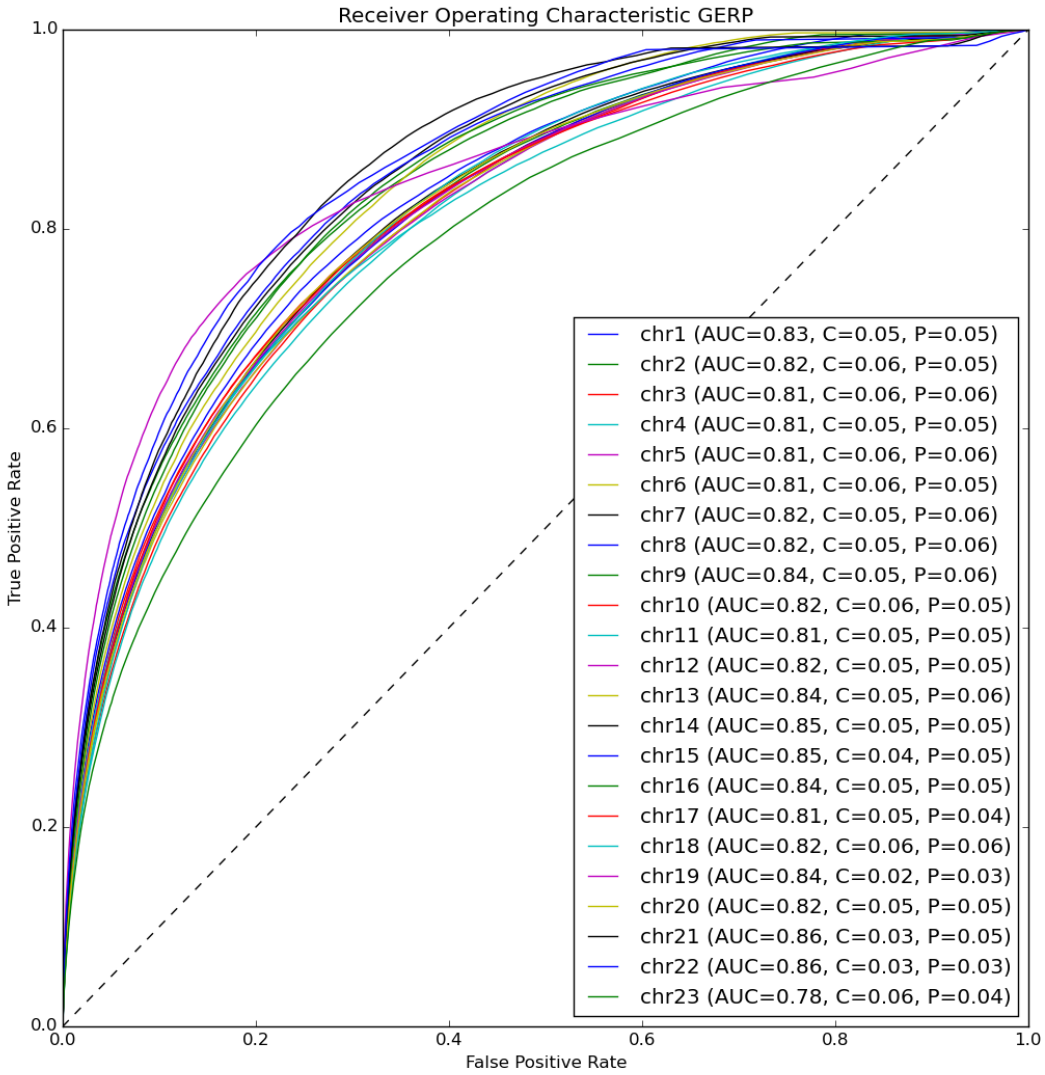

Figure 2.8: ROC curves for model trained using10 million sampled points from all chromosomes genome-wide except the chromosome numbered on the plot and labels based on GERP elements dataset.

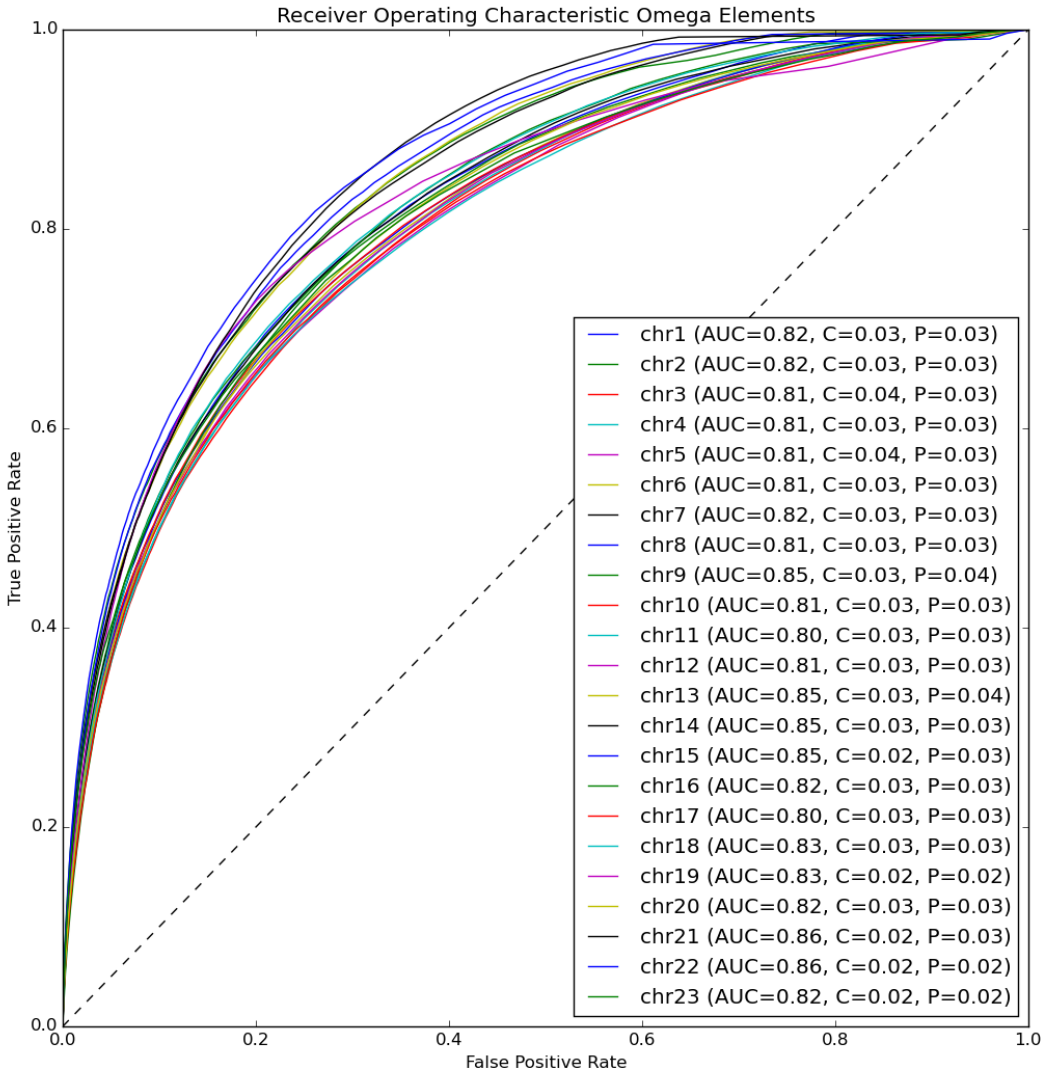

Figure 2.9: ROC curves for model trained using10 million sampled points from all chromosomes genome-wide except the chromosome numbered on the plot and labels based on Pi elements dataset.

In addition, AUC curves for each test chromosome are plotted separately (not overlaid for better clarity) for Pi elements in figure 10.

Figure 2.10: Individual ROC curves for model trained using10 million sampled points from all chromosomes genome-wide except the chromosome numbered on the plot and labels based on Pi elements dataset.

The resulting genome-wide scores are stored in the .wig file format and each score indicates how likely the particular base in the human genome is a conserved region. Scores are rounded to three decimal digits in order to make the file smaller, as scores are available for each of the three billion bases in the human genome. The .wig file can be imported in genome browser and all genome-wide predictions get plotted in the form of predicted peaks where the conserved regions are located. The genome-wide cumulative distribution of all scores is given in figure 11. The histogram of scores for all bases is given in figure 12 also for Pi element set. Similarly, cumulative distribution of the genome-wide scores for all four element sets (phastCons, GERP, Pi and Omega) is given in figure 13.

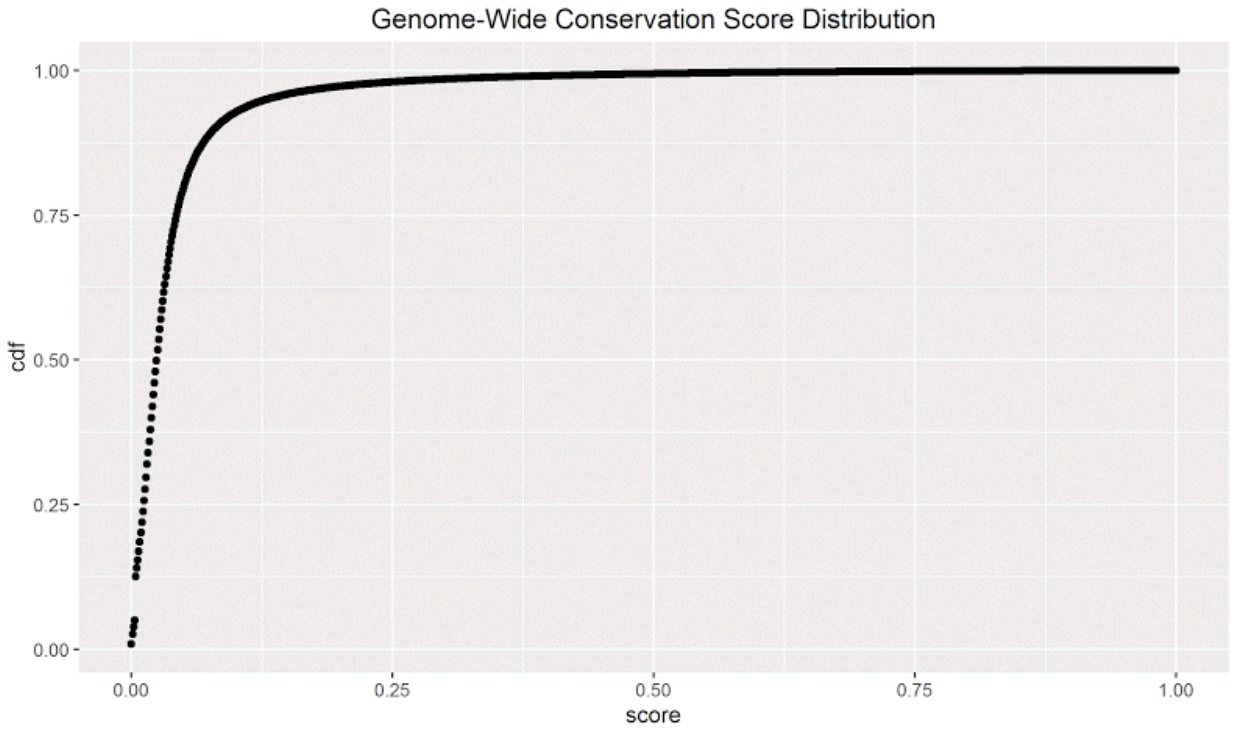

Figure2. 11: Cumulative distribution of scores for all positions genome-wide based on Pi conserved elements dataset

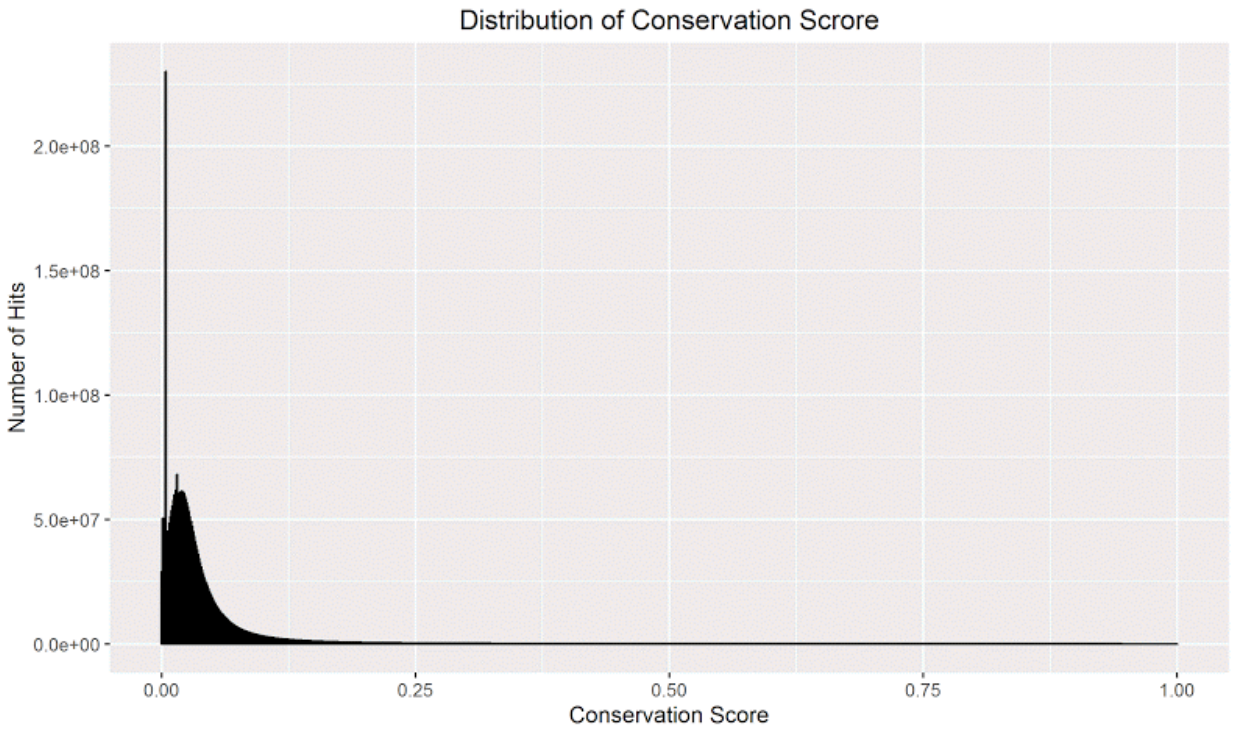

Figure 2.12: Histogram of scores for all positions genome-wide based on Pi elements

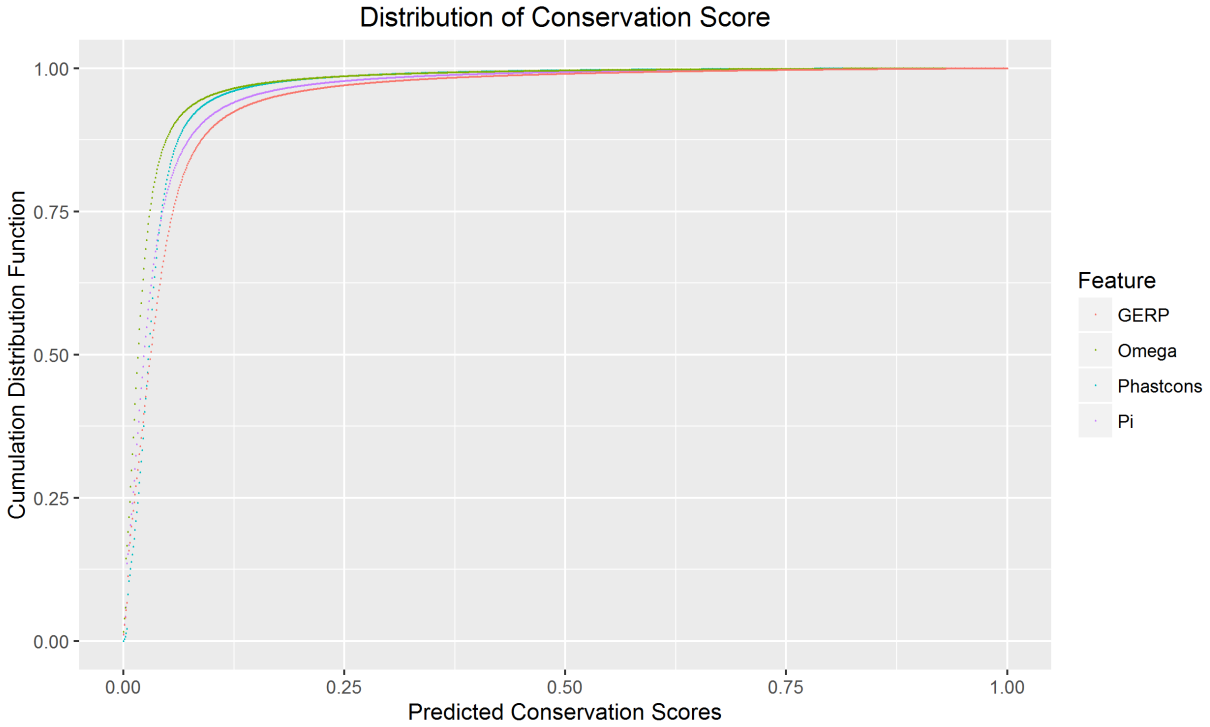

Figure 2.13: Cumulative distribution of conservation scores of all positions genome-wide based on all four conserved element datasets (phastCons, GERP, Pi, and Omega)

In order to show robustness of the scores, we have repeated sampling and model training on another set of genome-wide samples and calculated correlation among predicted scored genome-wide from two models with different samples to be 0.97. This is a very high correlation and indicates that the model predictions are robust and independent of the randomly chosen sample of positions genome-wide. Furthermore, we calculated pairwise correlations of scores obtained by applying the models that were trained based on different set of labels. The resulting correlations among each of the two sets of lables are given in table 9. Scores are highly correlated, pairwise correlation among any of the two sets was greater than 0.9.

Element Set 1	Element Set 2	Pairwise Correlation
GERP++	Pi	0.9259
GERP++	Omega	0.9549
GERP++	phastCons	0.9496
Omega	Pi	0.9497
Omega	phastCons	0.9553
Pi	phastCons	0.9095

Table 2.9: Pairwise correlations of scores obtained by applying the models that were trained based on different sets of labels. The resulting correlations among each of the two sets of lables (phastCons, GERP, Pi, and Omega).

Resulting predicted peaks overlap known conserved elements in non-exonic DNA. The example from genome browser is given in figure 14, where top track contains PICEL's predictions, and track on the bottom shows existing conserved regions as identified by Pi elements set. It is clear from the figure that PICEL accurately identifies known regions under constraint.

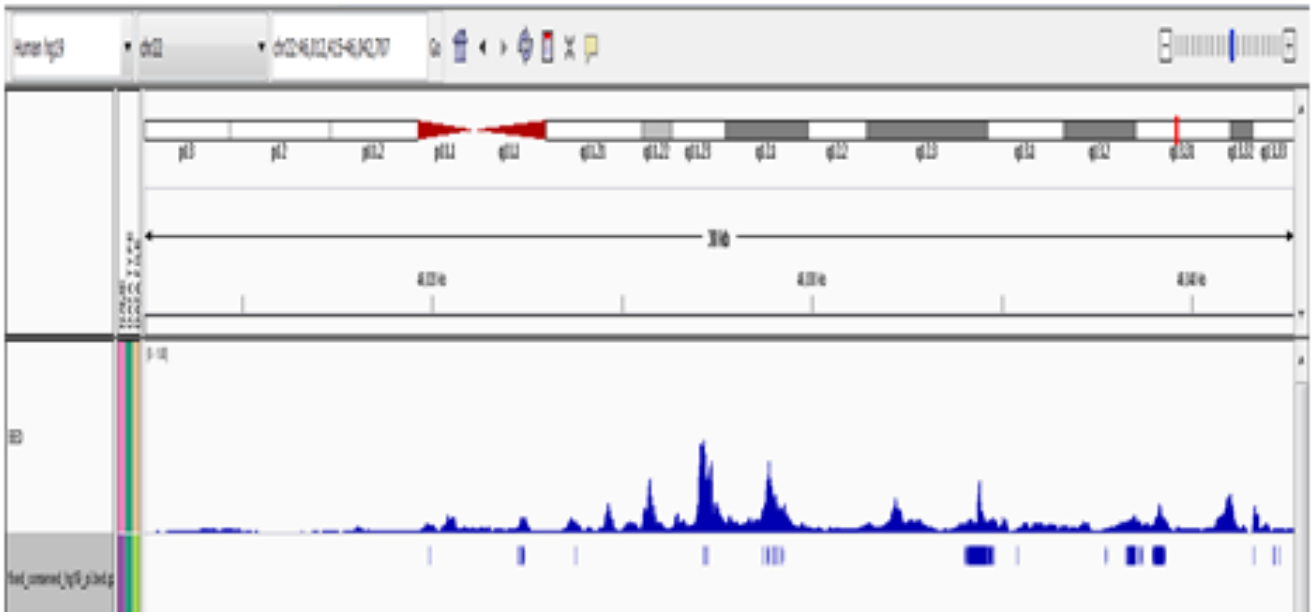

Figure 2.14: Resulting predictions for positions on chromosome 21 depicted in genome browser. Model trained using10 million sampled points from all chromosomes genome-wide except chromosome 21 and labels based on Pi elements dataset. Predicted peaks overlap known conserved elements.

As discussed earlier, Pi element set seems to show one of the best results overall and therefore we have continued to focus on results obtained using Pi element set. 56% of all bases are covered by at least one DNase peak. Cumulative distribution of scores under DNase sites in comparison to scores on all bases (figure 15) and histogram (figure 16) indicate that more bases under DNase peaks have higher scores on average. Compared to all bases, bases covered by DNase peaks showed decreased population with very low scores and increased mean value of scores (from 0.042 to 0.057). Most bases with a score of 0.25 or above are covered by at least one DNase peak as illustrated in subplot of figure 16.

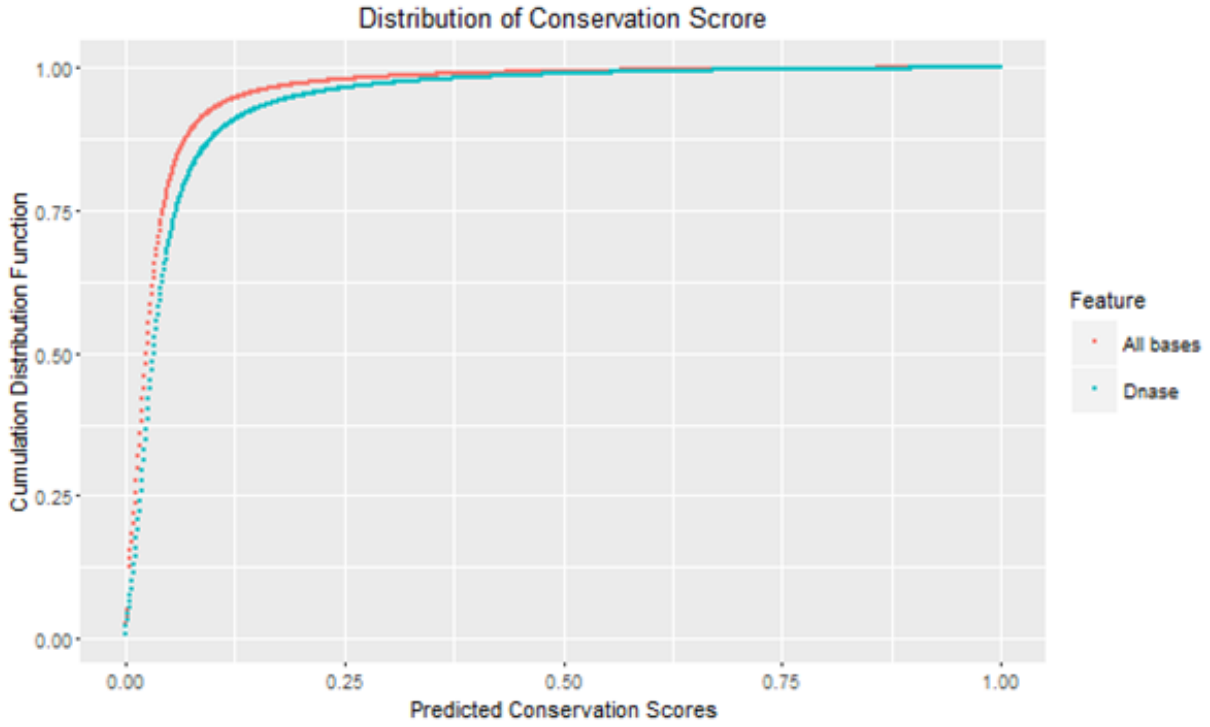

Figure 2.15: Comparison of cumulative distribution of scores at DNase sites genome-wide and all genome-wide positions based on Pi conserved elements dataset. More bases under DNase peaks have higher scores.

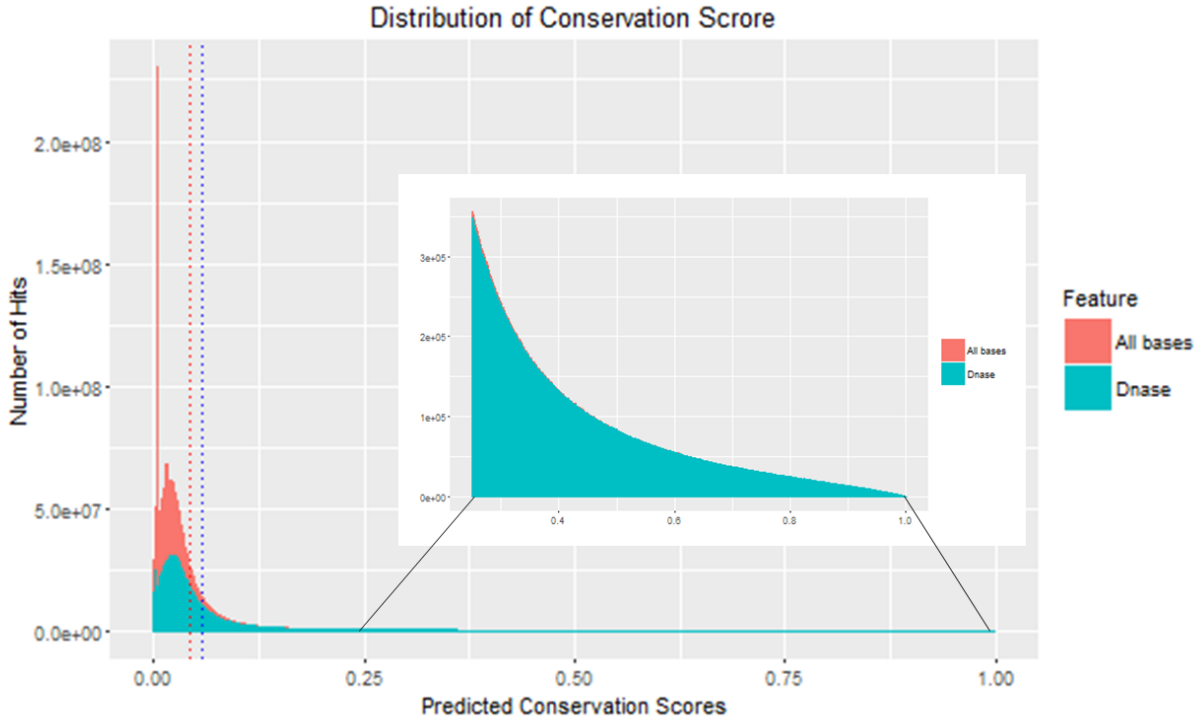

Figure 2.16: Histogram of conservation scores of all positions genome-wide based on Pi element dataset. Most bases with a score of 0.25 or above are covered by at least one DNase peak.

Even though PICEL accurately predicts most conserved regions, there are some known conserved regions that it misses. Prediction score of 0.004 is a threshold indicating that the particular genome position (base) is likely not within conserved region. For this reason, we identified all known conserved elements that PICEL is not predicting (with score less than 0.005) for each of the four conserved element sets and performed the analysis using Genomic Regions Enrichment of Annotations Tool (GREAT) (Mclean *et al.*, 2010). For example, GREAT analysis on Pi dataset returned a set of olfactory genes in various cell types to be located in that region (results and with p-values shown in table 10). It is presumed that olfactory genes are well-conserved across species as great variation exists in the number of genes among vertebrates and it is difficult to infer about individual receptors across species.

Recently, stratified LD score regression method has become available (Finucane *et al.*, 2015) that is using GWAS summary statistics and explicitly modelling for LD in order to partition heritability of complex traits. Since their findings indicate large enrichment of heritability in conserved regions across many traits, including heritability of body mass index (BMI), and since the method is flexible enough to allow for custom annotations, we were able to perform heritability analysis of our conservation scores using their method for BMI.

We partitioned our scores in 20 bins (each bin indicating certain interval, so that all intervals from 0 to 1 are included), and each bin contained approximately 5% of the SNPs associated with BMI that had the score that falls within the interval. This gave us approximately even spread of SNPs per bin, ~5% in each of the 20 bins. We created custom annotation track with binary indicator whether the SNP associated with BMI fell into that interval (bin label) or not. We then ran the method by comparing one bin against the remaining 19 bins, and repeated the procedure for all 20 bins. The results we obtained were proportion of heritability of complex trait (BMI in this case) explained by the SNP's in that interval, and the enrichment value, which was defined as proportion of heritability divided by proportion of SNPs in the same interval. The obtained proportion of heritability and the enrichment scores along with p-values for each test are listed in table 11, and also shown on plot in figure 19. The table is sorted by fold enrichment (highest to lowest), while the bar chart is sorted by intervals in each bin (again from highest to lowest). It is remarkable that results showed eight-fold enrichment in the regions that our method scored the highest (the most likely to be conserved). Also, other bins with high conservation scores also showed 4-6 times enrichment. These results indicate an increasing fraction of disease heritability.

Binom Rank	Binom Raw P-Value	Binom FDR Q-Val	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	Hyper FDR Q-Val	Hyper Fold Enrichment	Hyper Observe Gene Hit
1	1.5106e-263	5.5709e-260	2.7828	1,548	2.62%	3,346	1.0000	0.6201	170
2	5.9533e-208	1.0978e-204	21.2419	225	0.38%	1,352	1.0000	1.1147	4
3	3.3772e-121	4.1516e-118	23.3124	125	0.21%	3,303	1.0000	0.3483	1
4	1.1077e-117	1.0213e-114	21.7855	125	0.21%	2,888	1.0000	0.6967	1
9	2.5229e-73	1.0338e-70	5.7976	176	0.30%	2,760	1.0000	0.9016	11
11	3.1459e-71	1.0547e-68	60.9461	51	0.09%	1,654	1.0000	1.3933	1
13	3.0476e-66	8.6458e-64	8.4329	118	0.20%	381	1.0000	1.3933	5
14	1.2905e-58	3.3996e-56	40.1532	48	0.08%	1,030	1.0000	1.3933	2
16	9.1125e-53	2.1004e-50	22.6771	54	0.09%	1,030	1.0000	1.3933	2
18	2.3868e-51	4.8902e-49	15.9555	62	0.10%	1,654	1.0000	1.3933	1
19	8.1206e-51	1.5763e-48	4.5912	151	0.26%	1,002	1.0000	1.0837	7
20	1.1846e-50	2.1843e-48	19.0790	56	0.09%	1,030	1.0000	1.3933	2
21	1.6556e-50	2.9075e-48	23.3009	51	0.09%	1,654	1.0000	1.3933	1
22	2.1895e-50	3.6704e-48	18.1737	57	0.10%	1,030	1.0000	1.3933	2
23	1.9991e-48	3.2056e-46	3.9675	169	0.29%	2,773	1.0000	0.8708	5
24	2.8883e-47	4.4383e-45	3.8469	171	0.29%	600	1.0000	1.2192	7
25	8.7232e-47	1.2868e-44	2.1383	449	0.76%	1,427	1.0000	0.9952	30
26	2.6025e-46	3.6916e-44	15.7897	56	0.09%	2,625	1.0000	0.9289	2
27	1.4545e-45	1.9867e-43	10.3356	70	0.12%	1,030	1.0000	1.3933	2
28	4.9477e-45	6.5169e-43	2.0947	453	0.77%	1,491	1.0000	0.9817	31

Table 2.10: Results of Gene Ontology (GO) Analysis using GREAT tool applied to some known conserved regions that it misses. Top result shows a set of olfactory genes in various cell types to be located in that region.

2.10 Discussion

As indicated in results section, the resulting predicted peaks overlap known conserved elements in non-exonic DNA. However, the predictions vary slightly (correlation greater than 0.9) based on which set of conserved elements (phastCons, GERP, Pi, or Omega) was used for training. As discussed in earlier sections, the four conserved element sets are derived by methods that use different approaches to identify constrained regions. Also as mentioned in one of the sections earlier, the quality of data varies from dataset to dataset.

According to the AUC values presented in Results section, GERP++ elements have the largest AUC values (by 0.01 greater than the AUC of Pi and Omega elements), while phastCons have the lowest. The explanation for this discrepancy is that phastCons conserved region data set, even though widely used (and based on a newer alignment). Pi elements are covering more of the genome. On the other hand, GERP++ seems to be covering fewer genome locations, but it has longer conserved region segments, which would hypothetically be less likely to be missed by our method. In addition, phastCons method is only a substitution based method, and methods that incorporate other information in addition to substitution rates (such as Pi elements) are more likely to have higher accuracy. The reason is, as explained earlier in the manuscript, that SiPhy-Omega method doesn't include substitution patterns when identifying conserved regions. Due to these trade-offs, we decided to perform most of our analysis on model obtained by using Pi elements as labels.

Results of GREAT analysis discussed in Results section indicated that false negative predictions of our method and the genes they are proximal to can be used to identify cell types or classes of genes, which are not adequately represented in current functional genomic data sets.

Furthermore, this type of analysis could suggest additional cell and tissue types for experimental mapping.

GWAS studies have uncovered 1000's of loci associated with complex traits and diseases (Bush e Moore, 2012; Marbach *et al.*, 2016), but they can't tell whether the variant is causal due to most of them being in linkage disequilibrium (LD) and located in non-exonic regions. As indicated in figure 18, two lead SNPs at specific loci on chromosome 21 fall in regions where conservation is scored highly by our method. This type of analysis can potentially suggest locations of candidate causal SNPs.

Furthermore, heritability analysis of complex traits such as BMI shows that locations of greater probability of being conserved were strongly correlated with locations that explained an increasing fraction of disease heritability. This suggests our predictions have the potential to be an important resource for interpreting and prioritizing disease associated variants.

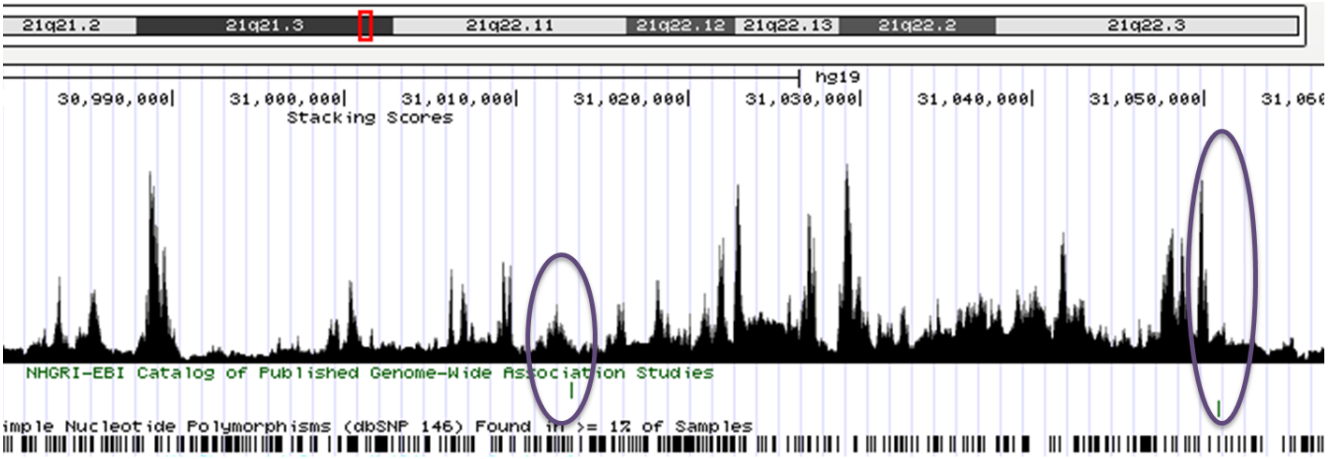

Figure 2.18: Resulting predictions for specific positions on chromosome 21 depicted in genome browser and two causal single nucleotide polymorphisms (SNP's) that fall in the regions where conservation is scored highly.

Begin Score	End Score	Prop._SNPs	std error	Enrichment	p-value
0.102	1.100	4.83%	4.75%	8.7	1.26E-13
0.068	0.102	5.07%	4.93%	6.7	2.44E-08
0.055	0.068	5.26%	5.77%	6.4	2.16E-06
0.048	0.055	5.01%	6.83%	5.3	2.03E-03
0.043	0.048	5.20%	6.50%	4.3	8.08E-03
0.039	0.043	5.55%	7.40%	2.2	3.48E-01
0.010	0.013	4.89%	6.38%	2.1	3.98E-01
0.013	0.016	5.53%	6.99%	2.1	3.93E-01
0.016	0.018	3.97%	7.61%	2.0	6.11E-01
0.036	0.039	5.15%	7.86%	0.6	7.72E-01
0.007	0.010	4.06%	5.78%	-0.1	4.19E-01
0.018	0.021	6.32%	7.98%	-0.6	1.89E-01
0.000	0.007	4.81%	2.49%	-2.4	1.14E-09
0.034	0.036	3.93%	7.45%	-2.7	5.23E-02
0.021	0.023	4.45%	7.47%	-4.2	1.80E-03
0.031	0.034	6.61%	6.97%	-4.2	5.53E-06
0.023	0.025	4.73%	8.61%	-6.2	8.19E-05
0.029	0.031	4.81%	6.46%	-6.7	8.04E-08
0.025	0.027	4.89%	6.20%	-6.9	2.14E-09
0.027	0.029	4.95%	5.80%	-7.6	4.29E-11

Table 2.11: Results of heritability analysis for body mass index (BMI): Enrichment scores sorted highest to lowest for 20 bins. Fold enrichment calculated when 1 bin is compared to the remaining 19 bins. Top score is associated with BMI Heritability (8 fold enrichment).

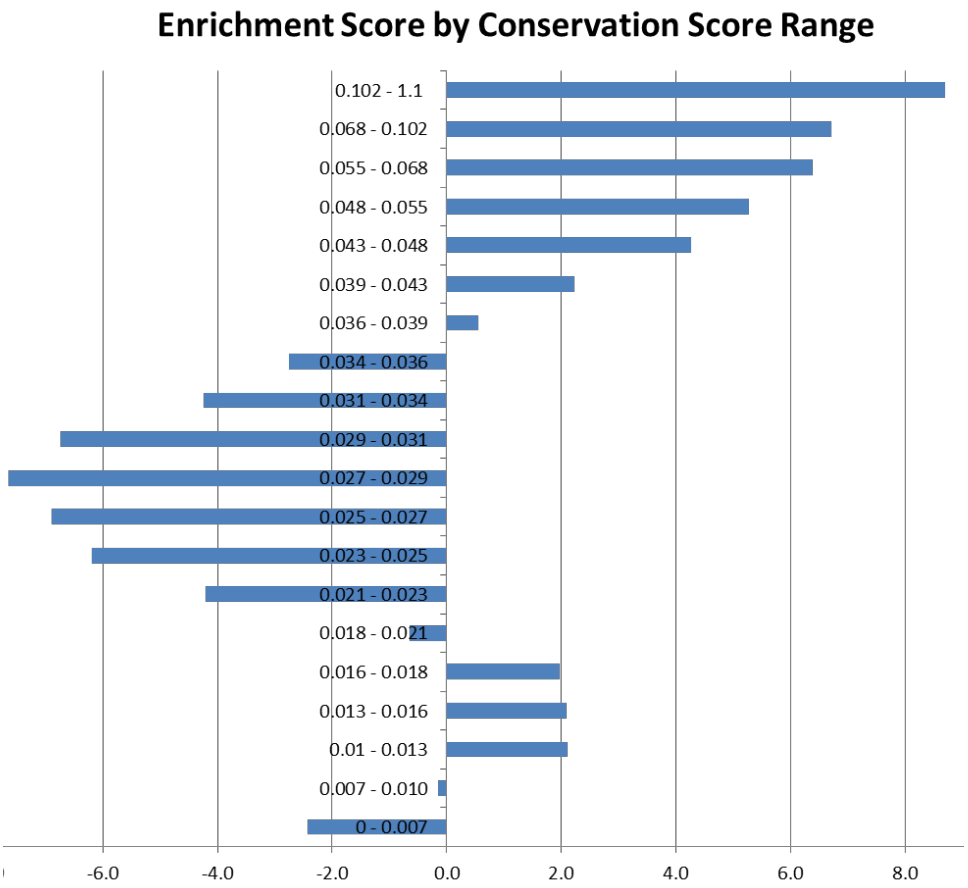

Figure 2.19: Results of heritability analysis for body mass index (BMI): Enrichment scores for 20 bins (sorted by bins containing highest to lowest scores). Fold enrichment calculated when 1 bin is compared to the remaining 19 bins. Top score is associated with BMI Heritability (8 fold enrichment).

2.11 Acknowledgment

The author would like to thank all her lab mates in the Ernst lab for support during this project, in addition to Dr. Ernst. Specifically, big thank you goes to Petko Fiziev for helpful discussion regarding the project, Adriana Sperlea for performing heritability analysis for BMI SNPs, and Xiaorui Fan for creating score plots.

Chapter 3

Gene Expression Data Analysis in Vervet Monkey

Abstract

This study was joint work with Dr. Nelson B. Freimer and Dr. Anna J. Jasinska (Jasinska *et al.*, 2009) from center of Neurobehavioral Genetics at University of California, Los Angeles. We analyzed gene expression profiles in tissues derived from blood samples from 347 vervet monkeys (Chlorocebus aethiopssabaeus) and eight brain regions from 12 vervet individuals. The goal was to find out how genes expressed in brain correlated with gene expression variation in blood from the same individuals. Because of high degree of conservation of tissue expression profiles between vervets and humans, our findings provide means to investigate variation in gene expression relevant to human brain traits and neuropsychiatric diseases. By processing large-scale datasets resulting from gene expression studies and applying stringent method and statistical criteria, we identified 29 transcripts whose expression is measurable, stable, replicable, variable between individuals, relevant to brain function and heritable. The follow up study (Jasinska *et al.*, 2012) localized one eQTL (at B3GALTL) to a region of <200kb by conducting SNP genotyping and association analysis.

3.1 Introduction

Vervets, also known as African green monkeys (chlorocebus aethiops sabaeus), are native to Africa. However, small number of vervet individuals had been moved on ships from Africa to the islands of the Caribbean, during colonial times in the 1600's. These individuals have formed their own habitat on the Caribbean islands, where they have lived for centuries (and still do). During the period from 1975 to 1989, UCLA researchers trapped 57 individual vervet monkeys from various regions of the Caribbean islands, St. Kitts and Barbados in particular, and brought them to UCLA in order to form a vervet research colony. Over time, the colony became an inbred pedigree as the monkeys could only mate with each other. In January 2008, when the colony contained approximately a thousand members, it's members have made the cross-country journey from UCLA on the West coast to Wake Forest University on the East coast of the United States. The colony has been permanently relocated at Wake Forest University in Winston-Salem, North Carolina.

Vervet monkeys are non-human primate (NHP) biomedical model that is more directly relevant to human biology and disease than rodents or other commonly used animal models. They permit invasive and longitudinal investigations that are not possible (or ethical) to be performed in humans. Therefore, prior to relocation of the vervet colony, UCLA researchers collected data in the form of blood and brain tissue samples. In particular, blood samples were drawn from 347 individuals as well as two blood replicates from 18 individuals, in order to correct for (or avoid) technical issues during sample collection. In addition, 12 individuals were sacrificed (and therefore did not make the cross-country move!) and their brain tissues were collected from eight brain regions. The regions of the vervet brain are: head of caudate,

cerebellar vermis, hippocampus, frontal pole, dorsolateral prefrontal cortex, orbital frontal cortex, pulvinar, and occipital pole.

The objective of this particular data collection was to perform gene expression studies on collected tissue samples. However, at the time, there were several challenges related to studying gene expression in vervet. First of all, vervet genome had not yet been sequenced and microarrays used for gene expression studies didn't contain vervet probes. The second challenge was that not much was known about gene expression in primate brain. The additional challenge was that brain tissue samples were scarce and the large percentage of data that was available for the analysis contained only blood samples. Therefore, the first goal was to evaluate probe-target sequence compatibility of the available gene expression microarrays. If the available microarray probes were compatible enough, the other two important goals were to characterize regional gene expression in brain and identify transcripts with low variability between brain and blood (also known as peripheral biomarker of brain expression). The idea here was to find genes expressed in both brain and blood and be able to use blood as a surrogate for brain tissue sample for further studies.

3.2 Methods

BLAST was used for evaluating probe-target sequence compatibility. Hierarchical clustering was used to characterize regional gene expression in vervet brain. Percent variation (PV) inter-individual and intra-individual and among tissues was used to address the third goal and identify peripheral biomarkers of brain expression (genes expressed in both brain and blood).

3.3 Probe Comparison Outcome

We compared 22,184 50-nucleotide long probe sequences located on the Illumina BeadStudio HumanRef-8 version 2 chip against 341, 172 available vervet sequences that had been sequenced, since the whole vervet genome didn't exist at the time. We used BLAST method on the cluster machine in order to execute this comparison. When parsing the output file of BLAST method, we were specifically looking for top hit for each probe and counted frequencies for each probe length match (e.g. how many probes had 48 nucleotides matched). The breakdown number of probes per number of nucleotide matcheed is illustrated in figure 1. The results showed that 46% of human probes matched the vervet probes, but that even 88% of vervet probes matched the human probes. Therefore, since approximately 90% of the vervet probes could bind to human probes on the Illumina HumanRef-8 version 2 chip, we were able to continue with gene-expression studies in vervet using blood and brain data.

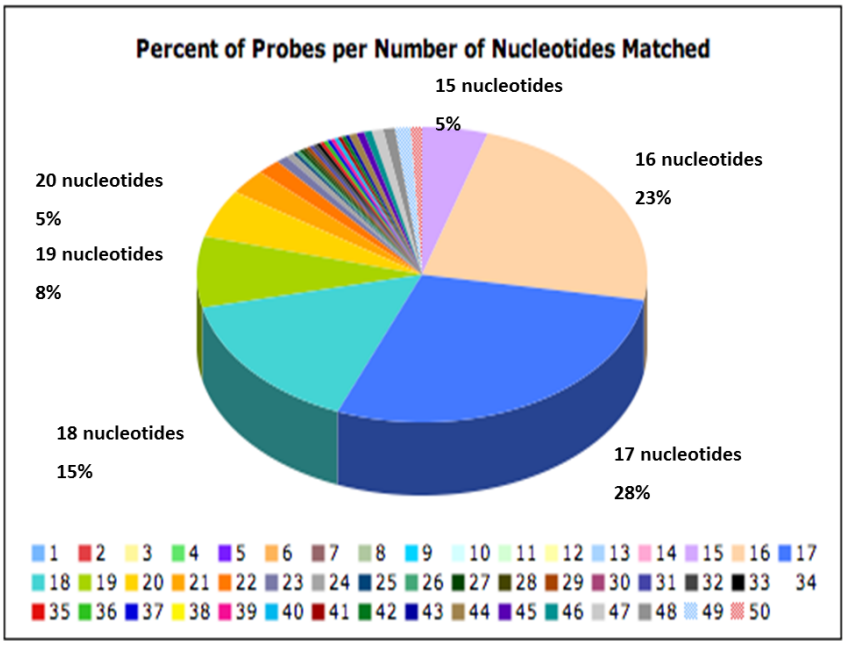

Figure 3.1. Percent of probes per number of nucleotides matched (number of nucleotides ranges from 1 to 50.

3.4 Workflow for Identifying Transcripts

Having data measured from the same tissues and from the same individuals allowed us to evaluate sources of transcript variation within and between individuals. The workflow in figure 2 shows that we focused on two classes of transcripts characterized by high variation of expression across brain regions or high variance between individuals. High inter-individual variation between brain and blood and between independent blood samples allowed us to investigate heritable brain gene expression traits in peripheral blood.

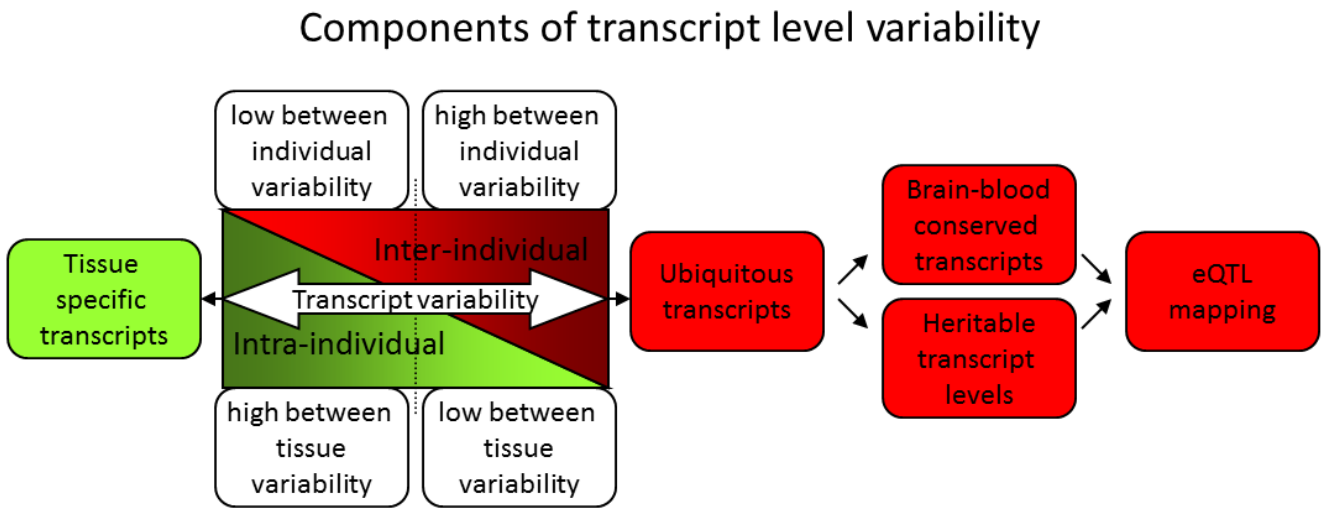

Figure 3.2. Components of transcript level variability. Inter-individual variation (green) and intra-individual or between tissue (red) variation. Transcripts characterized by much higher intra-individual than inter-individual variation provide insight into the functional relationships between different tissues and are of interest as candidates for mapping brain eQTL using blood as a surrogate tissue.

3.5 Results

3.5.1 Results for Gene Expression Differences between Brain Tissues

We used hierarchical clustering in order to measure gene expression differences between brain tissues. Distances between tissues were estimated based on probes that showed the most differences in terms of number of shared detections (either detected in single one or in most brain

tissues). Similarities between the pairs of tissues are illustrated in heat-map in figure 3(A). Cerebellum (cerebellar vermis) is the most distinct among brain regions. Interestingly, it could be inferred from the dendogram in figure 3(B) that expression of some genes was specific to a single cortical region, in particular the three cortical regions are clustering together: orbital frontal cortex, dorsolateral prefrontal cortex, and frontal pole.

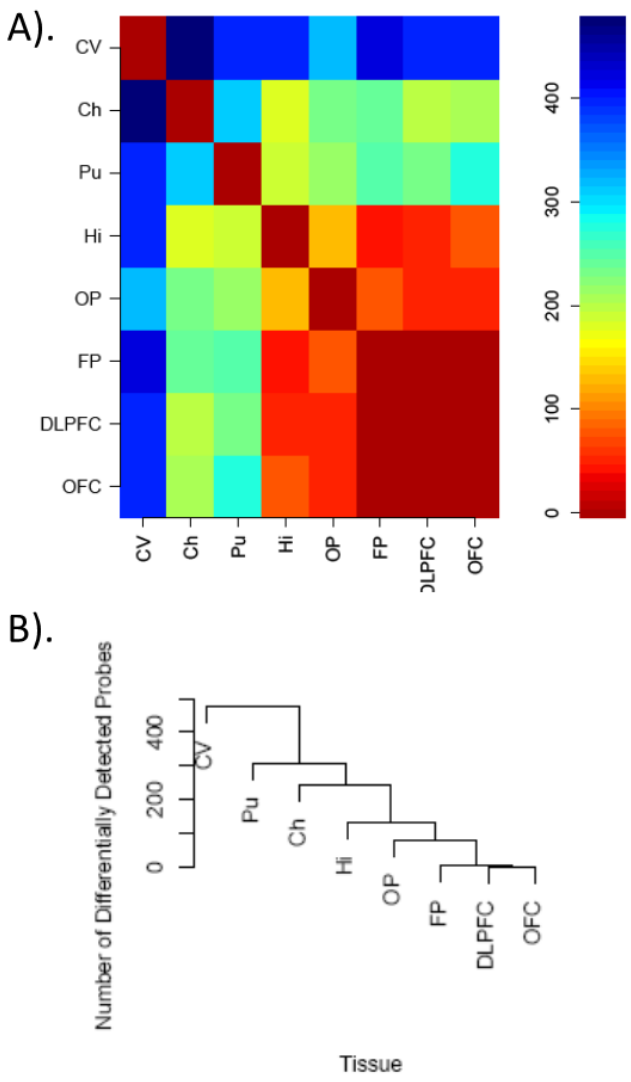

Figure 3.3. Gene expression differences between brain tissues. Pairwise comparison between all eight brain regions is presented in heat map (A). Corresponding hierarchical clustering of tissues is presented on a dendogram (B). Labels: CV-cerebellar vermis, Pu-pulvinar, Hi-hippocampus, OP-orbital pole, FP-frontal pole, DLPFC-dorsolateral prefrontal cortex, OFC-orbital frontal cortex.

3.5.2 Results for Candidate Transcripts for eQTL Mapping

There were 2481 probes (representing 2430 genes) where gene expression was detected in all 12 vervet brains and in all 8 brain regions and in blood. Out of those, 474 probes (31%) had percent variation between brain and blood (PV BB) greater than 0.55 and Spearman Rank Correlation (SRC) greater than 0.55, as illustrated in figure 4. The abundance of a gene transcript is directly modified by polymorphism in regulatory elements. Consequently, transcript abundance might be considered as a quantitative trait (eQTL) that can be mapped with considerable power. Expression Quantitative Trait Loci (eQTL) may act in cis (locally) or trans (at a distance) to a gene. By using aforementioned stringent criteria, we identified 29 transcripts whose expression is measureable, stable, replicable, variable between individuals, relevant to brain function and heritable.

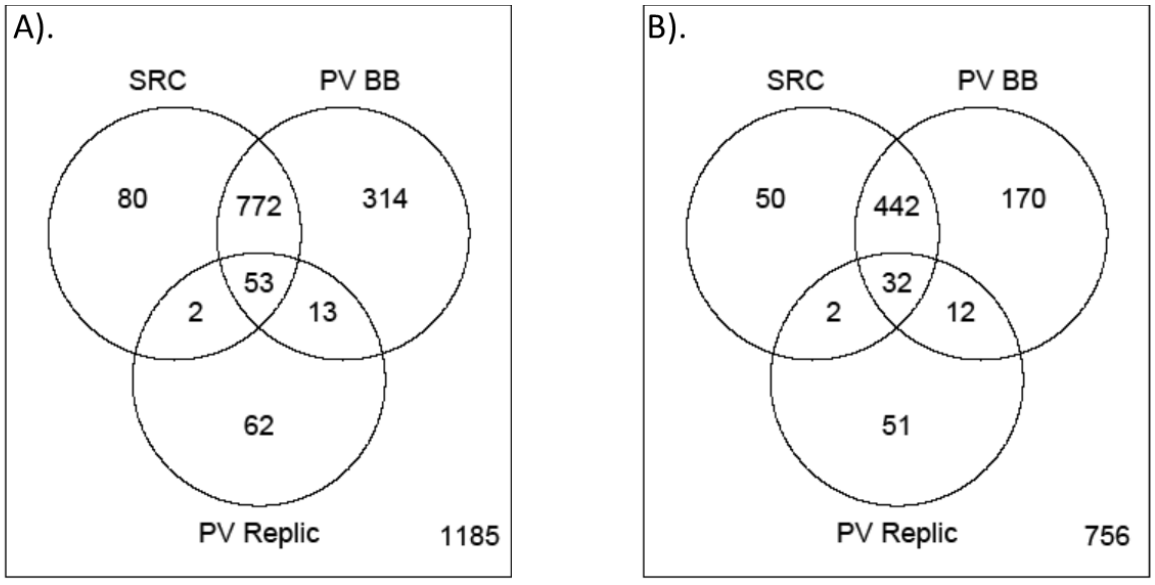

Figure 3.4. Selection of candidate transcripts for mapping brain eQTL in peripheral blood. The diagram represents the set of probes in the brain-blood gene expression comparison that passed the 55% threshold for PV (PV BB) and the 0.55 threshold for the SRC for any of the brain regions. (A) All 2481 probes that passed detection thresholds. (B) The subset of 1515 probes that also passed detection thresholds in in the replicate blood sample dataset.

3.6 Discussion

As discussed in goals, methods, and several results sections, by using stringent criteria, we were able to identify 29 transcripts whose expression is measureable, stable, replicable, variable between individuals, relevant to brain function and heritable. Many heritable transcript levels are specifically up or down regulated in several or only one tissues and therefore there could exist genetic variants regulating functions of a narrow set of brain tissues. This has been illustrated well and discussed in detail in the published manuscript (Jasinska *et al.*, 2009). A follow up study (Jasinska *et al.*, 2012) further investigated these findings by performing quantitative trait linkage analysis using 261 microsatellite markers that were identified as significant and suggested linkages for 12 of these transcripts, including both cis- and trans-eQTL. For one cis-eQTL (at B3GALTL, beta-1,3-glucosyltransferase), the study conducted follow-up single nucleotide polymorphism (SNP) genotyping and fine-scale association analysis in a sample of unrelated Caribbean vervets, and localized a single eQTL to a (small) region of less than 200kb.

Chapter 4

Global Sensitivity and Parameter Search for Biomolecular Dynamic Models

Abstract

Dynamic system models are often used for simulating intracellular functions in search of novel therapies for clinical disorders. Systematic and efficient algorithms are needed for discovering primary components and interactions in biomolecular network model dynamics. The methods are expected to work effectively for high-dimensional models. Complex theory and computation make it very challenging. A major problem is that models are difficult to quantify from experimental data. Current algorithms for accomplishing this suffer from high computational demands and convergence issues. Improvements are needed for large models containing scores to hundreds of parameters, which are commonly found in systems biology.

This study focuses on techniques involved in quantification cycle of biomolecular nonlinear dynamic models, specifically methods for global parameter search and sensitivity analysis, which are closely related to model reduction and optimal experiment design. The objective is to adequately and efficiently improve selection of parameter values by reducing the search space and computation times. We carried out numerical experiments for local and global sensitivity analysis on combination of synthetic and real biological datasets of TNF-α meditated NF-κβ model and performed detailed comparison of results. Our analysis provides insight into model mechanics and attempts to identify limitations of current methods. We discuss Sobol method in detail, as it proved to be the most comprehensive approach for calculating global parameter sensitivities. Its main disadvantage is high computational cost, which could be significantly reduced by using cluster computer. As a result, we augmented computer model of human thyroid hormone regulation dynamics to better fit the available data for thyrotoxicosis.

4.1 Introduction

Biomolecular processes are difficult to be measured in vivo, within the cell. It is especially difficult to predict oscillations, different levels of interactions, and other dynamic responses of the cellular system that are time-dependent and happen on different levels (among cell organelles, proteins, or genes) and across different scales (cytoplasm or nucleus). Dynamic system modeling helps fill-in the gaps.

The dynamics of biomolecular networks are typically described by nonlinear ordinary differential equations (ODEs) based on principles of control theory. They simulate cell dynamics or other biological processes in the organism. Kinetics of chemical reactions among molecules is incorporated into the model by applying rate laws based on Michaelis-Menten, mass-balance and flux-balance principles (stoichiometric analysis). Dynamic system models are also referred to as mechanistic models.

For model quantification, estimation of model variables and parameter values remains a bottleneck. Besides the usual modeling challenges related to transforming experimental data into structural models, more difficult challenges involve structural and numerical identifiability, i.e. knowing what model parameters can be quantified well. Even though most models are simplified and based on approximations and assumptions, they are still very complex, with tens or scores of state variables and even more parameters, and highly non-linear. The main question here is which of these state variables (nodes) and parametric couplings (edges) dominate the dynamical responses of the network to input perturbations, and which others can be neglected, eliminated or otherwise approximated.

Due to complexity, only subsets of parameters are usually estimated, but finding the best subsets is difficult. Sensitivity analysis is a technique intended to aid in accurate identification of

model network structure and underlying dynamics. Discovery should confirm and extend our mechanistic understanding of biological processes and metabolic disorders (such as vulnerability to infection, diabetes, and obesity), leading to the identification of novel targets and therapies for intervention.

4.2 Problem Statement

Current methods for sensitivity analysis and parameter set selection suffer from high computational demands and algorithmic convergence issues. This study focuses improving the application of global sensitivity analysis and parameter search techniques involved in quantification cycle of biomolecular nonlinear dynamic models to examine one biological dynamic system model (NF-kB) and improve upon another one (human thyroid hormone regulation dynamics).

4.2.1 Model Reduction

The complexity and thus number of parameters in dynamic system models is very high, which drastically increases computational demands (Chu e Hahn, 2007). Importantly, the way these models are developed typically yield overly complex, over-parameterized models. Sensitivity analysis of the entire model is an effective first step in model reduction, but it is computationally very intensive. Therefore, systematic and efficient algorithms are needed for discovering the primary components (state variables and parameters) that govern the dynamics (the real biology).

Selecting the structurally identifiable set of model parameters and estimating them from data is a well-established and important problem (Miao *et al.*, 2011). However, a structurally

identifiable parameter cannot be estimated reliably if the model output is insufficiently sensitive to that parameter. On the other hand, the parameter can be estimated with high precision from the output data when the model output is highly sensitive to that parameter (Distefano, 2014). Model parameter sensitivity analysis explores sensitivity of model variables to variations in parameter values, i.e. how much the model variables might change if a model parameter changes. The changes could be related to variation of a single parameter (local sensitivity) or combined variations of multiple parameters (global sensitivity), which includes interactions between parameters. Mathematically speaking, local (linear) sensitivity analysis approximates parameter variations using only first order terms in Taylor expansion of delta-X (for state X) with respect to parameter variations delta-p (for parameter p) about nominal point in parameter space. Global sensitivity analysis (GSA) includes nonlinear effects of higher-order terms, which characterize interactions/nonlinear couplings among parameters on deltaX ("impact").

Solution methods currently proposed in literature include analysis of sensitivity vectors via principal component analysis, singular value decomposition, and correlation analysis (Chu e Hahn, 2007). The most recent methods involve algorithms for clustering parameters into distinct groups based on the dynamic effect that changes in parameters have on the output of the dynamic system model (Chu e Hahn, 2008). Some algorithms take advantage of sampling methods to determine ranges of parameters (e.g. orthogonal array sampling, Latin hypercube sampling, and least squares estimation) and rely on statistical approaches (e.g. Kolmogorov-Smirnov statistics, Pearson correlation coefficient) to rank parameter sensitivities. Numerous other techniques for sensitivity analysis were applied in engineering and other disciplines (physics, economics, food safety, etc.).

4.2.2 Parameter Estimation

Estimation of model variables and parameter values remains a bottleneck for biological dynamic models, especially for larger systems. Besides the usual modeling challenges related to collecting experimental data (such as noisy or incomplete data, missing data points, time series, etc.), there are more difficult challenges regarding extracting implicit data. For example, kinetic constants are often derived from different organisms or under different experimental conditions. These nominal values are often just "guestimates". Consequently, simulated model output may not be consistent with biological observations.

In addition, high computational demands are associated to model simulation due to "combinatorial explosion" and issues related to algorithmic convergence and numerical round-off error are frequently encountered. Currently proposed algorithmic solutions focus on circumventing integration of differential equations, reducing the complexity of the task, smoothing noisy data, optimizing parameter values, and constraining parameter search space [Chou & Voit 2009]. Many of parameter estimation methods rely on Fisher information matrix (FIM), as it serves as a measure of how much information about the parameters is possible to extract from an experiment. FIM is dependent upon initial states, inputs and parameter values. If singular, it is impossible to estimate parameter values. The inverse of FIM is a lower bound of parameter covariance matrix as postulated by Cramer-Rao theorem.

4.2.3 Sensitivity Analysis

Sensitivity analysis is a technique intended to aid in parameter estimation, accurate identification of model network structure and underlying dynamics, and potentially lead to model reduction. It is also intended to reduce experimental and computational cost for future

experiments. When a model output is highly sensitive to a particular parameter, the parameter is considered a key-player in regulating that output and can be estimated with relatively high precision. Sparse, noisy experimental data pose additional challenges, and sensitivity analysis deals directly with these as well, with global sensitivity analysis addressing the overall problem more comprehensively than local sensitivity analysis [Saltelli A 2004].

Local sensitivity analysis explores sensitivity of model variables to local (linear) variations in parameters about single nominal parameter points in parameter space, i.e. how much model outputs (and thus state variables) change when model parameters vary a small degree from particular nominal values. Global sensitivity analysis goes two steps further than local analysis. It explores how higher-order, nonlinear parameter interactions affect model dynamics, and – in principle – it also considers the dependence of sensitivities on the assumed nominals, for the entire feasible parameter space.

4.2.4 Parameter Set Selection

Parameter estimation is challenging not only because of sparse and noisy experimental data, but also because models can contain hundreds or even thousands of parameters. Usually, it is impossible to accurately estimate values of all the parameters from experimental data. Recent research efforts have been geared toward using parameter set selection approaches for model reduction (Yue *et al.*, 2006; Chu e Hahn, 2007; Jaewook Joo, 2007; Chu e Hahn, 2008; Cintron-Arias, 2009). Due to complexity of parameter estimation, only a subset of parameters is usually estimated, while others are fixed to constant values.

Algorithmic solutions currently proposed in literature typically employ discretized sensitivity matrix (DSM) and look at directions of sensitivity vectors to identify a subset of parameter set to be estimated (see Methods). DSM is constructed by stacking sensitivity matrices for each of sample time points, and therefore could be very large. Proposed solutions usually identify collections of suboptimal parameter sets, as optimal set is difficult to distinguish in practice. This is combinatorial optimization problem and an active area of research.

4.3 Materials and Methods

4.3.1 Sensitivity Analysis Methods

In this section, we provide a brief overview of several recent and commonly used methods for Local (LSA) and Global Sensitivity Analysis (GSA)(Zi *et al.*, 2005; Zi *et al.*, 2008).

4.3.1.1 Finite Difference

Several software packages (e.g. Copasi(Institute *et al.*) and SBML-SAT (Zi *et al.*, 2008)) use central difference formula to approximate for calculations of partial derivatives for calculating local sensitivity functions. For a system with state variable x_i, and parameter θ_j, sensitivity s_{ij} is defined as:

$$s_{ij}(t) = \frac{\partial x_i(t)}{\partial \theta_j}$$

and approximated by final difference formula, where $\Delta\theta_j$ is defined as small change in parameter value (in reference to its nominal value):

$$s_{ij}(t) = \frac{\partial x_i(t)}{\partial \theta_j} = \frac{x_i(\theta_j + \Delta\theta_j, t) - x_i(\theta_j, t)}{\Delta\theta_j}$$

This method is used most often to calculate sensitivities(Daun *et al.*, 2008).

4.3.1.2 Weighted Average of Local Sensitivities (WALS)

WLSA is local sensitivity analysis calculated at multiple random points within parameter space. Therefore, distribution reveals global impact. Sensitivity is computed by weighted average of local sensitivities (normalized). Boltzmann distribution $exp\left(-\frac{E}{kbT}\right)$ is used for parameter ranking. In this formula, E represents error between model simulation and data, while kbT stands for customized scaling factor. In practice, least squares error (LSE) is used for the term in the denominator and minimum of least square errors is used for the term in the numerator of Boltzmann distribution: $w_k = exp\left(-\frac{LSE(x_k)}{min_{i=1\ldots N} LSE(x_i)}\right)$.

4.3.1.3 Multi-Parametric Sensitivity Analysis (MPSA)

Latin Hypercube Sampling (LHS) is used to randomly generate parameter values from probability distributions (rather than using fixed values). The method calculates objective function for each parameter set, which is defined as sum of square errors between random and reference parameter set. Range of parameter distributions is usually determined from literature or experience. Parameter sensitivity is based on cumulative frequency of parameter set, and

Kolmogorov-Smirnov statistics. Sensitivity values range on the interval between 0 and 1. Arbitrary threshold value is used to classify parameters from highly sensitive to non-sensitive.

4.3.1.4 Partial Rank Correlation Coefficient analysis (PRCC)

This method is based on rank transformed linear regression analysis. It seeks linear (monotonic) relationship between ranks of output function and input parameters and uses Pearson correlation coefficients to calculate parameter sensitivities, which map to sensitivity interval from -1 to 1. As by the previous method, LHS is used here to generate parameter values.

4.3.1.5 Sobol Method

Method proposed by (Sobol, 2001) belongs to variance-based class of methods, which have recently become the preferred approach for sensitivity analysis across many disciplines (Saltelli, 2004; Saltelli, 2007). It provides decomposition of output variance into factors of increasing dimensionality (variances of parameters), and is therefore able to account for interactions among parameters. Variance decomposition is defined by the following formula, where V(y) represents total output variance, V_i, V_{ij}, and $V_{12\ldots k}$ represent partial variances contributed by parameter combinations, and k represents total number of parameters:

$$V(y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + V_{12\ldots k}$$

$V_i = V(E(Y|x_i))$ (first order effect)

$V_{ij} = V(E(Y|x_i, x_j)) - V_i - V_j$ (second order effect)

$V_{12\ldots k}$ is the last term of order k effect. There are 2^k-1 terms.

First order S_j sensitivity is defined using statistical identity as ratio of first order effect to overall variance:

$$S_j = \frac{V(E(y|x_j))}{V(y)}$$

Total effect sensitivity is defined as total contribution to the variance of y due to non-x_j. For k=3 example, the formula is:

$$S_1^T = \frac{V(y) - V(E(y|\mathbf{x}_{-1}))}{V(y)} = \frac{E(V(y|\mathbf{x}_{-1}))}{V(y)} = S_1 + S_{12} + S_{13} + S_{123}$$

When the model is purely additive, sensitivity values sum to 1: $\sum_{i=1}^{k} S_i = 1$

As illustrated by equations above, this method enables computing sensitivities based on total effect on output (combinations of parameter subsets) in addition to contributions of individual parameters. However, due to its high computational complexity, only 1st order and Total Effect sensitivities are calculated in practice (e.g. S_1 and S_{123}, but not S_{12} and S_{13}), bringing the computational cost down to n(2k+2) from n(S^k-1) for all effects. Cost for calculating 1st order (local) sensitivity alone is n2^k (n is the number of samples, and 2 is due to needing two matrices to calculate mean value E). Monte Carlo estimates are used to generate parameter values, as large number of random samples is required for estimates to be stable (Rosolem *et al.*, 2012).

Sobol method (Sobol, 2001) is one of the most robust ones currently available for sensitivity analysis, with its main disadvantage being high computational cost. Improvements suggested by

several groups have only been partially successful and the method seems to fail on large datasets (Kiparissides, 2008; Kiparissides et al., 2009).

Given that it relies on Monte Carlo sampling, which generates large numbers of random samples (~1000), and that variance decomposition is based on some of many individual factors/terms, our approach of implementing it in C programming language and executing it on cluster computer resulted in improved running time and accuracy, as well as more-likely convergence for large models.

4.3.2 Parameter Set Selection Methods

Several groups have recently developed methods to reduce the number of parameters to be estimated by identifying a subset of parameters which are more likely to be estimated accurately. Approaches that appear to be most effective are listed below.

4.3.2.1 Cintron-Arias Approach

The method (Cintron-Arias, 2009) lists all possible subsets of parameters and performs parameter estimation on each possible subset. Parameters not in the subset are fixed to nominal values. Parameter selection score (length of the vector of parameter coefficient of variance) and condition number (ratio between the highest and lowest singular values of the DSM) are assigned to each subset. Subsets with the lowest PSS and condition number emerge as winners. The reasoning is that PSS would be high for parameters with high uncertainty making them less identifiable, while large condition number would indicate “ill-conditioned” DSM and therefore high-dependencies among parameters.

The limitation of this method is that in practice it only works for very small models, due to combinatorial explosion caused by large parameter sets. The effectiveness of the method was demonstrated on a model consisting of only 3 state variables and 11 parameters. The authors additionally constrained the parameter subset space to 8 parameters, as they decided that particular 3 parameters needed to be included in the model. Even parameter sets of moderate size would require exponentially large number of computations to be performed simultaneously.

4.3.2.2 Sequential Cintron-Arias Approach

Due to constraints of Cintron-Arias and similar approaches that evaluate parameter sets simultaneously, sequential methods are frequently used for parameter estimation of models with large number of parameters. These types of approaches look for ways to reduce the model by selecting and removing one parameter at a time and reevaluating the model after each step. Therefore, they are computationally feasible (linear time). Their main disadvantage is that the best combination of parameters might be missed, due to parameters selected at earlier steps (Chu e Hahn, 2007; 2008).

Recently, modification of the Cintron-Arias approach had been suggested, to make it applicable sequentially. Instead of evaluating all possible combinations, search space gets reduced by evaluating the combinations containing P-1 parameters, where P is the number of parameters in the unreduced model (Sin, 2012). The new model (P-1) gets accepted with the lowest PSS and condition number. The next step evaluates the models with (P-1)-1 parameters, and again the updated model with lowest PSS and condition number gets accepted. The sequence of steps continues until a suitable model has been reached. Suitable, in this context, means when

an upper limit for the magnitude of %CVs is reached, but the criteria could be adjusted for other models (e.g. lower limit for singular values).

4.3.2.3 Daun, Rubin Approach

Another example of sequential approach for reducing over-parameterized models is found in (Daun *et al.*, 2008), and similar method is found in (De Pauw *et al.*, 2008). The approach is based on analyzing individual parameter sensitivities and prioritizing parameters with strong pairwise correlations that have minimal effects on the output as candidates for reduction. The idea is that strong pairwise correlations suggest parameters that are difficult to distinguish, and parameters with low sensitivity values indicate lower effect of those parameters on model output.

In (Daun *et al.*, 2008),parameter correlations are computed first, starting from the most correlated pair of parameters. Sensitivity of each parameter is evaluated by finding the length of the sensitivity vector. If the model is not sensitive to at least one of the parameters, the one with lower sensitivity is fixed, otherwise the next parameter pair is considered. Every time a parameter is fixed, new correlation matrix is calculated and the same sequence of steps is repeated until the stopping criterion is reached. Possible choices for stopping criteria are: number of parameters to be reduced, lower bound on the length of the sensitivity vector, upper bound on the correlation between two parameters, etc.

4.3.2.4 Pairwise Clustering Approach

This technique uses slightly different approach by splitting parameter set into groups and choosing only one parameter from each group as representative for parameter estimation. The

idea is to form groups of pairwise indistinguishable parameters (their sensitivity vectors are parallel and therefore those parameters cannot be estimated individually). The similarity measure between parameter pairs is defined by (Chu e Hahn, 2008) as cosine of the angle formed between two sensitivity vectors s_i and s_k:

$$\cos\phi_{ik} = \frac{|s_i^T s_k|}{\|s_i\|_2 \|s_k\|_2}$$

Values close to 1 indicate that two parameters are highly dependent (small angle), while values close to 0 indicate parameters that could be easily discerned from each other (angle close to 90°).

Agglomerative hierarchical clustering is used for this purpose. Number of groups of parameters (clusters) is chosen by cutting hierarchical tree at certain level, according to an arbitrary (problem-specific) threshold. Parameters with largest sensitivity vectors are chosen as group representatives (candidates for estimation). Discrepancy (between the original model and the reduced model) value of 5% and corresponding smallest similarity measure are considered to be the threshold for determining number of clusters. The number of parameters to be estimated (length of a subset) is determined based on the number of singular values of sensitivity matrix, until there is a gap of an order of magnitude or more between the singular values.

To demonstrate the effectiveness of this method, the authors applied it to sample model of 115 parameters total. It turned out that the 9th through 115th singular values of sensitivity matrix were close to zero, and therefore only 8 parameters were needed to be estimated. In addition, authors suggested fixing (not estimating) all parameters with sensitivity vector lengths less than 5% of the largest sensitivity vector. There turned out to be 70 parameters satisfying the criterion and therefore only 45 parameters were candidates for estimation. Number of clusters needed was determined to be 11 by observing that the discrepancy value dropped below 0.05

(0.095 for 10 clusters). Conveniently, a problem boiled down to choosing parameter subset consisting of only 8 parameters out of 11 to satisfy D-optimality criterion (maximizes the logarithm of the determinant of FIM), drastically decreasing computational demands from $\sim 2 \times 108$ possible combinations (for 45 parameters) to 165.

The limitation of this method is that its effectiveness was demonstrated on a single case study with arbitrary set thresholds, and the idea of clustering contributed to only a portion of parameter set reduction (reducing parameter set from 45 to 11 parameters), whereas other criteria were used to prune the parameter set (from 115 to 45) and determine the length of subset (8).

4.3.2.5 Multiple-Criteria Screening Approach

(Rosolem *et al.*, 2012) recently proposed parameter grouping technique based on Pareto ranking concept (Goldberg, 1989). They obtained sensitivity indices for their Simple Biosphere 3 (geophysics) model using Sobol method for all three model outputs (heat flux, latent heat flux, and net ecosystem exchange of CO_2) with respect to 43 parameters.

In order to obtain a more objective parameter ranking given differing sensitivity values for each of the outputs, they proposed this screening approach, which assigns a "group rank" to each parameter based on simultaneous maximization of its individual total order sensitivities. Even though this only worked for models with multiple outputs, the general idea of having groups of factors that have strong or weak contributions to overall model performance is important. This approach is yet to be applied to biomolecular models.

4.3.3 Software

Software packages for implementing and analyzing dynamic system biomolecular models are available as standalone applications or as toolboxes (add-ons) for Matlab (Mathworks). Many are difficult to install, not at all easy to use, have high-level of system dependency (often incompatible), and cannot be modified for expansion or correction. Some of those, such as Cell Designer and SimBiology (Mathworks), do not provide simple interface for manually entering mathematical equations (ODE's) and do not use standard compartmental modeling nomenclature, which greatly limits their use for existing complex models. Furthermore, they don't handle models with time-delays, for example, common in many network models (e.g. gene networks). Delay-ODEs are often used to describe transport and translation delays in mRNA induction, but the method in which delayed synthesis gets calculated does not behave very well with repeated sampling of time steps by most ODE solvers. During our preliminary studies, the software was not fully functional for models of only moderate complexity. Most of these programs are very limited in their sensitivity analysis capabilities such as COPASI (Institute *et al.*; Kent *et al.*, 2012).

Recently, Systems Biology Markup Language (SBML) and Systems Biology Toolbox for Matlab were created that enabled more flexible software environment for simulation and analysis of bio-systems (Schmidt e Jirstrand, 2006). Subsequently, SBML-based sensitivity analysis tool, SBML-SAT was created by (Zi *et al.*, 2008) which has proven useful. Aside from providing method for performing Local Sensitivity Analysis (LSA), which uses central difference formula to approximate for calculations of partial derivatives, it enables users to run several methods for global sensitivity analysis (GSA). In particular, the four methods for performing GSA implemented in the SBML-SAT package are: Multi-Parametric Sensitivity Analysis (MPSA),

Partial Rank Correlation Coefficient analysis (PRCC), Weighted Average of Local Sensitivities (WASL), and Sobol method. COPASI software package also provides a feature for performing LSA based on similar technique.

Amigo (Balsa-Canto e Banga, 2010) is one of the most-comprehensive software packages to date for dynamic model quantification. It provides parameter rankings and correlation plots between pairs of parameters, but expects the user to decide what parameters to fix or what experiments to do. When user gets ranking of 128 different parameters (or more), it's difficult to determine what combinations of parameters should be fixed. For example, parameters with the lowest ranking could be recommended to be fixed or eliminated from the model, while the highly ranked parameters could be considered essential ("key players"). Regardless of its limitations, if used with caution, Amigo is of great use for experimenting with sensitivity analysis and parameter set selection for bio-system dynamic models.

4.3.4 Parallel Computing

There are two general parallel programming paradigms used to improve computational performance. The first one is to divide matrices in blocks that enable multiple cores (nodes) to perform computations simultaneously. This means that sum operations could run in parallel (e.g. sum), while others that are highly dependent on the rest of the matrix (e.g. multiplication) continue to run sequentially. This is similar to popular concept of Map Reduce widely used in large data processing – "Map" stage is when multiple nodes work simultaneously (e.g. each one is counting frequencies of words in subset of pages assigned to it) and "Reduce" stage is when results from all nodes get sorted and summarized together. The second paradigm is related to optimal scheduling of nodes/clusters in order to get the best performance possible. Due to system

constraints, adding more nodes would not necessarily achieve increased performance. Some of the ideas include feeding larger matrices to nodes with more memory, reversing the order of operations (that are associative) in order to keep all nodes busy at all times, minimizing interactions among nodes by incorporating more care in assigning tasks to each node, etc.

4.4 Data

4.4.1 p53 Model

The model proposed by (Sin, 2012) consists of four state variables and 23 parameters. It describes negative feedback loop of p53 regulation (activation and degradation) in the cell. p53 is essential tumor suppressor protein, its function being to stop cell cycle progression (cell arrest) and repair DNA of damaged cells, or initiate apoptosis (cell death) where repair is not possible. Experimental data was obtained from (Wang *et al.*, 2007) based on phosphorylation at the serine-15 epitope of p53 and mRNA response of tumorigenic mammary epithelial cells (MCF7) following exposure to Neocarzinostatin (NCS), which causes double stranded DNA breaks. This model appears to be over-parameterized and needs to be reduced in order to be fully quantified (Javier, 2009).

4.4.2 NF-kB Model

Nuclear factor κB (NF-κB) regulates genes that are involved with cellular signaling, stress response, cell growth, survival, and apoptosis (cell death). Elucidating inner workings of NF-κB would aid in development of new drugs for chronic inflammatory diseases. Notably first computational model describing NF-κB signal transduction pathway was presented by

(Hoffmann *et al.*, 2002) based on experimental data and computer simulations. It consisted of 15 state variables and 30 parameters. The model was improved by (Lipniacki *et al.*, 2004) to contain 24 state variables and 64 parameters, by (Werner *et al.*, 2008) to contain 33 state variables and 110 parameters, and by (Koh e Lee, 2011) to contain 68 state variables and 127 parameters. Since this is a very important model, many research groups have tried to improve (augment or reduce) it using both experimental and computer simulated data. In conclusion, new studies are needed to help elucidate complex biomolecular mechanics (Basak *et al.*, 2012).

4.4.3 Alga Model

Genome-scale metabolic network of Chlamydomonas reinhardtii (model alga) was reconstructed by integrating biological and optical data by (Chang *et al.*, 2011). This type of modeling encompasses existing knowledge about an organism's metabolism and genome annotation and is employed to study diverse biological processes (e.g. photosynthesis, light-driven metabolism). This particular model was reconstructed by analyzing photon absorption (biochemical activity spectra) in light-driven reactions (varying light wavelengths) using stoichiometric representation of the spectral composition of emitted light. The authors claim that they have experimentally validated their simulated model by applying it to measure efficiency of light utilization from diverse light sources.

This is an example of whole-cell modeling approach that is becoming more and more utilized due to due to recent technological advances. More studies are needed to uncover the dynamics of regulatory mechanisms. This network models 2190 chemical reactions and 1706 metabolites in total, divided into 18 compartments. Model simulations were performed using flux-balance analysis. Mechanistic (ODE) model for this system is also likely to have very large

number of state variables and parameters in the near future. However, it is important to note that this is not a dynamic system model as the derivatives in ODEs are set to zero (non-dynamic model).

4.4.4 Human Thyroid Hormone Regulation Dynamics Model

4.4.4.1 Human Thyroid Hormone Regulation Dynamics Model Overview

A computer simulation model of human thyroid hormone (TH) dynamics has previously been developed and validated against published clinical data sets (Eisenberg *et al.*, 2006; 2008; Eisenberg e Distefano, 2009; Eisenberg *et al.*, 2010; Ben-Shachar *et al.*, 2011). It models regulation of TH in blood, tissues and the hypothalamus-pituitary-thyroid axis. The model is represented by 25 ordinary differential equations (ODEs) and more than 50 parameters. It has been used to predict pharmacological and pathophysiological conditions in adults and children. It is useful for computer experiments for hypothetically dosing human patients with hormone supplements (e.g. pills, injections, intravenous infusion) with T3 and T4 content, and obtaining temporal dynamics of hormone levels (T3, T4, and TSH) in plasma. The limitation of the current model is that it seriously underestimates clinically observed T3 plasma levels for times greater than 24 hours following a thyrotoxic dose of 3000 ug of T4 (Leboff *et al.*, 1982).

4.4.4.2 Human Thyroid Hormone Regulation Dynamics Detailed Model Structure

The thyroid gland secretes thyroid hormones (TH) thyroxine (T4) and triiodothyronine (T3), which play a role in metabolism and growth of the human body (Ho, 2013). Thyroid stimulating hormone (TSH) regulates the secretion rate of T4 and T3. The pituitary gland is stimulated to release TSH, which in turns signals the thyroid gland to secrete more TH when its levels in the bloodstream are low. The process is reversed when TH levels in the bloodstream are high via negative feedback loop. In this section, we describe modeling process of human TH dynamics in detail (Eisenberg *et al.*, 2006; Eisenberg *et al.*, 2010; Ben-Shachar *et al.*, 2011).

The closed-loop human TH regulation model is primarily structured as a feedback control system (FBCS). It contains dual suppressor inputs: plasma concentrations of T3 and T4 ($T3_p$ and $T4_p$), two controlled outputs: $T3_p$ and $T4_p$, and six interconnected subsystems (blocks) representing source and sink components based on tissue types (organs). The three source blocks are: hypothalamus (HYP), anterior pituitary (ANT PIT), and thyroid gland (THYROID). The three sink, also called distribution and elimination (D&E), components are: thyrotropin-releasing hormone (TRH) D&E, thyroid stimulating hormone (TSH) D&E, and thyroid hormone (TH) D&E. Concentrations of TSH in plasma (TSH_p) are driven implicitly by TRH and embodied in the $TSH_p(t)$ data forcing function, since hypothalamic and pituitary submodels cannot be quantified experimentally. Therefore, four of the six blocks were subsumed, and only the remaining two, TH D&E and THYROID were needed explicitly. Subsequently, T3 and T4 feedback effects (secretion rates in response to TSH stimulation) were used as (TSH_p) input (output data as input forcing function) and fitted to closed-loop data as indicated in figure 1.

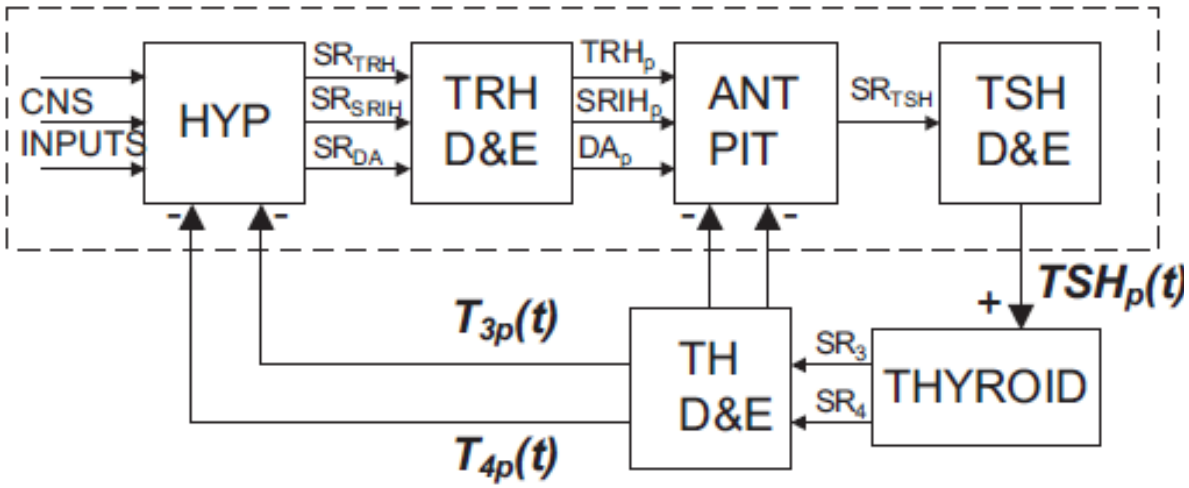

Figure 4.1: Adult thyroid hormone feedback control system (FBCS) composed of three source (HYP, ANT PIT) and three sink (TRH D&E, TSH D&E, and TH D&E) submodels. HYP= hypothalamus, TRH = thyroid releasing hormone, ANT PIT = anterior pituitary, TSH = thyroid stimulating hormone, TH = thyroid hormone. (Eisenberg *et al.*, 2006; 2008)

Originally, the group proposed and quantified the six-compartment TH D&E model structure, linear and uniquely identifiable (leaks k_{o3} and k_{06} set to equal zero). Compartments 1 and 4 represent the free hormone in plasma, T4 and T3 respectively. Compartments 2 and 5 represent total T4 and T3 in tissues with fast exchange, while compartments 3 and 6 represent total T4 and T3 in those with slow exchange. Extravascular enzymatic T4 to T3 conversions from compartment 3 to 6 and from 2 to 5 (k_{52} and k_{63}) are nonlinear. Michaelis-Menten (M-M) kinetic terms represent enzymatic conversions of D1 and D2. For this purpose, two known constants were used: K_{mD1} and K_{mD2} and unknown V_{maxS}, and it was assumed that all fast compartment conversion occurred via D1 and 80% of slow pool conversions was via D2.

The more complete model contains several additional submodels. The binding submodel models protein-bound hormone representing the reversible and competitive binding of both T3 and T4 to three plasma proteins: TBG, HSA, and TTR. It expresses plasma free hormone as functions of plasma total hormone concentrations. Total hormone uptake rate constants k_{21}, k_{31},

k_{54}, and k_{64} are converted into free hormone uptake rate constants k_{21free}, k_{31free}, k_{54free}, and k_{64free} using normal steady state free fraction values. The thyroid gland secretion submodel is a simple three-parameter time-delay model, with linear coefficients representing the secretory responses (stimulating T3 and T4) to plasma TSH concentrations. Parameter τ represents the approximate time delay for thyroidal secretion in response to TSH stimulation. The gut absorption submodel serves for pharmacokinetic (PK) simulation studies using oral dose inputs of TH hormone supplements. The TH absorption rates were estimated from data (e.g. 88% of T3 is absorbed). The nonlinear brain submodel represents the dynamics of brain components (hypothalamus and pituitary combined) as shown in figure 2.

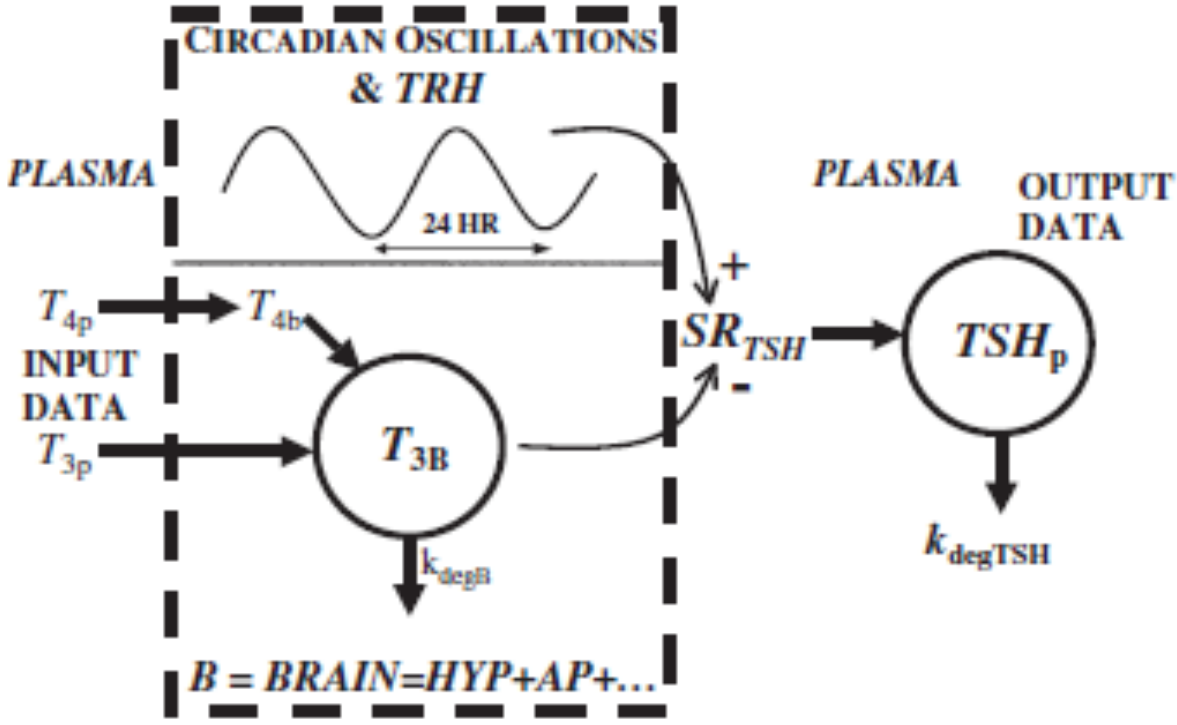

Figure 4.2: Adult lumped brain submodel for TSH, TRH D&E, and TSH secretion from Figure 10 (Eisenberg *et al.*, 2006; 2008). TSHp is driven implicitly by TRH, and dual suppressor inputs---plasma T3 and T4 concentrations, T3p(t) and T4p(t).

It simulates TRH and TSH secretion, and TRH D&E. It contains a single-output, TSH secretion rate, two (dual supressor) inputs: plasma T3 and T4 concentrations (T_{3p} and T_{4p}), and two new compartments T3 in brain (T_{3B}) and TSH D&E. In particular, T_{3B} models how T3 in brain directly or indirectly affects TSH secretion in brain. TSH secretion rate is represented as a harmonic oscillator, to model TSH secretion circadian oscillations. The FBCS Hypothalamo-Pituitary-Thyroid Axis (H-P-T Axis) model is shown in figure 3.

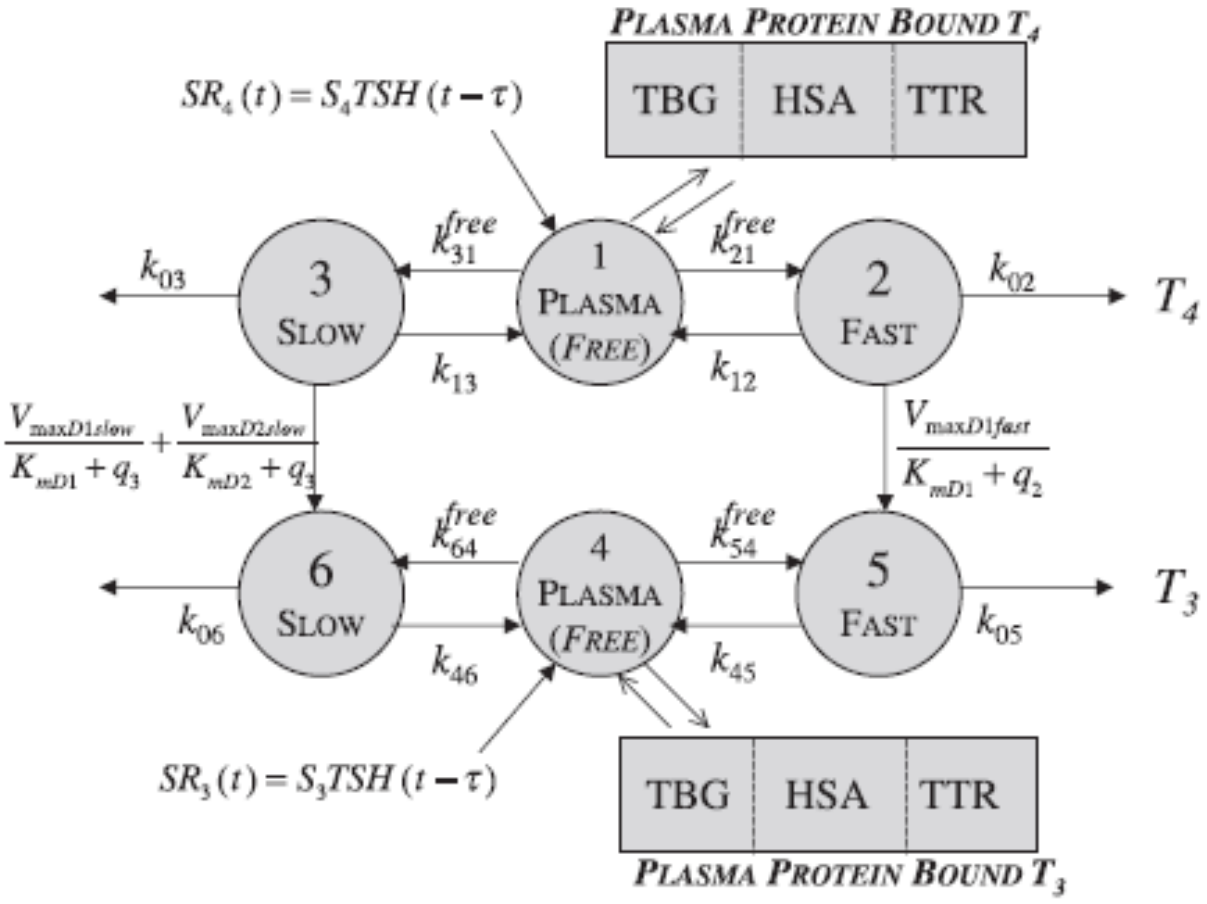

Figure 4.3: Adult T3 and T4 D&E submodel (Eisenberg *et al.*, 2006; 2008). Compartment 1: free plasma T4, compartment 4: free plasma T3; non-linear, extravascular enzymatic T4 to T3 conversions: from compartment 3 to 6 and from compartment 2 to 5. Rates based on Michaelis-Menten (M-M) kinetics. The secretion of T3: $SR_3(t)$, T4: $SR_4(t)$. Time-delay estimate τ for $SR_3(t)$ and $SR_4(t)$ responses to TSH stimulation that yields the best fit to the closed-loop data (Eisenberg *et al.*, 2008).

4.5 Results

4.5.1 Global Sensitivity Analysis of NF-kB Model

We focused on model reduction methodology for elucidating biomolecular pathway dynamics. Specifically, we addressed computational aspects of this problem by examining more efficient methods for sensitivity analysis and parameter set selection for NF-kB model with128 parameters in reference to 69 state variables.

One of the arguments for adding complexity to the NF-kB model was that apoptosis is a complex process and interactions between all pertinent molecular pathways needed to be taken into consideration. In order to identify molecular key-players, (Koh e Lee, 2011) performed sensitivity analysis. However, the two methods they used (LSA and MPSA) yielded very different parameter rankings, which we interpreted as unreliable. We reproduced their results and proceeded by applying other methods to the same dataset in an attempt to identify and isolate key components with higher confidence.

In particular, we applied four algorithms for GSA to TNF-α meditated NF-κB model by (Koh e Lee, 2011) using SBML_SAT software. MPSA had previously been applied to this model by (Koh e Lee, 2011), while PRCC, SOBOL's, and WASL are the original contributions of this study. Running time of Sobol method was 12 hours on a cluster computer, while the other three methods took only 1-2 hours to converge. Resulting 3D plots are included in figures 4-7, and results are summarized in tables 1 – 3.

Rank	Parameter	PRCC Sensitivity	Parameter	SOBOL's Sensitivity	Parameter	WALS Sensitivity
1	ki18	5.23E-01	ke11	-1.5704e+00	ki28	6.67E+03
2	ki22	4.29E-01	ki2	-1.8715e+00	ki18	2.83E+02
3	ki1	1.74E-01	kr4	-2.3107e+00	ke2	2.82E+02
4	ki2	1.32E-01	kn16	-2.4765e+00	ke4	2.82E+02
5	ke9	7.61E-02	ki22	-2.5107e+00	ki13	2.82E+02
6	ki4	6.77E-02	p	-2.7649e+00	kn1	2.82E+02
7	ki5	6.11E-02	ki5	-2.9830e+00	kn17	2.82E+02
8	kr4	5.75E-02	kn18	-3.1194e+00	kr13	2.82E+02
9	ke3	5.49E-02	ke1	-3.1733e+00	ke1	2.08E+02
10	kc5	5.22E-02	kc5	-3.2919e+00	ke5	1.57E-76

Table 4.1: Top 10 sensitive parameters in reference to state variable s57 obtained using PRCC, SOBOL's, and WALS sensitivity analysis methods.

Rank	Parameter	PRCC Sensitivity	Parameter	SOBOL's Sensitivity	Parameter	WALS Sensitivity
1	kr15	7.28E-01	kr1	-1.0262e+05	ke2	9.48E+03
2	kn18	2.35E-01	kn18	-1.0405e+05	ke4	9.48E+03
3	kn5	1.90E-01	kn13	-1.0469e+05	ki13	9.48E+03
4	kn15	1.58E-01	ke5	-1.0901e+05	ki18	9.48E+03
5	kr5	1.37E-01	kr5	-1.1075e+05	kn1	9.48E+03
6	kr6	1.32E-01	kn20	-1.1278e+05	kn17	9.48E+03
7	kn1	4.49E-02	ke3	-1.1357e+05	kr13	9.48E+03
8	kn12	4.41E-02	kr17	-1.1361e+05	ke2	3.99E-75
9	kn10	4.31E-02	kr4	-1.1364e+05	ke5	3.99E-75
10	ki23	3.66E-02	ke4	-1.1370e+05	ke11	3.99E-75

Table 4.2: Top 10 sensitive parameters in reference to state variable s194 obtained using PRCC, SOBOL's, and WALS sensitivity analysis methods.

Rank	Parameter	PRCC Sensitivity	Parameter	SOBOL's Sensitivity	Parameter	WALS Sensitivity
1	kr2	5.35E-01	ki25	-1.1809e-01	ki28	-8.62E+02
2	kr1	5.28E-01	kn15	-1.2499e-01	ke7	-8.16E+02
3	kr4	4.25E-01	kn14	-1.3377e-01	ki29	-8.16E+02
4	kn18	1.60E-01	kn16	-1.3455e-01	kn14	-3.24E-82
5	ke9	1.12E-01	kr2	-1.5075e-01	kn16	-3.01E-82
6	kn5	1.09E-01	ke4	-1.5324e-01	kr4	-2.69E-82
7	kn10	7.84E-02	kn3	-1.6955e-01	kr6	-2.11E-82
8	kn15	7.37E-02	kn6	-1.7660e-01	kn9	-1.83E-82
9	k_r17	6.91E-02	kn4	-1.8115e-01	kr5	-1.21E-82
10	kr19	4.30E-02	kn12	-1.8154e-01	ke3	-4.74E-83

Table 4.3: Top 10 sensitive parameters in reference to state variable s222 obtained using PRCC, SOBOL's, and WALS sensitivity analysis methods.

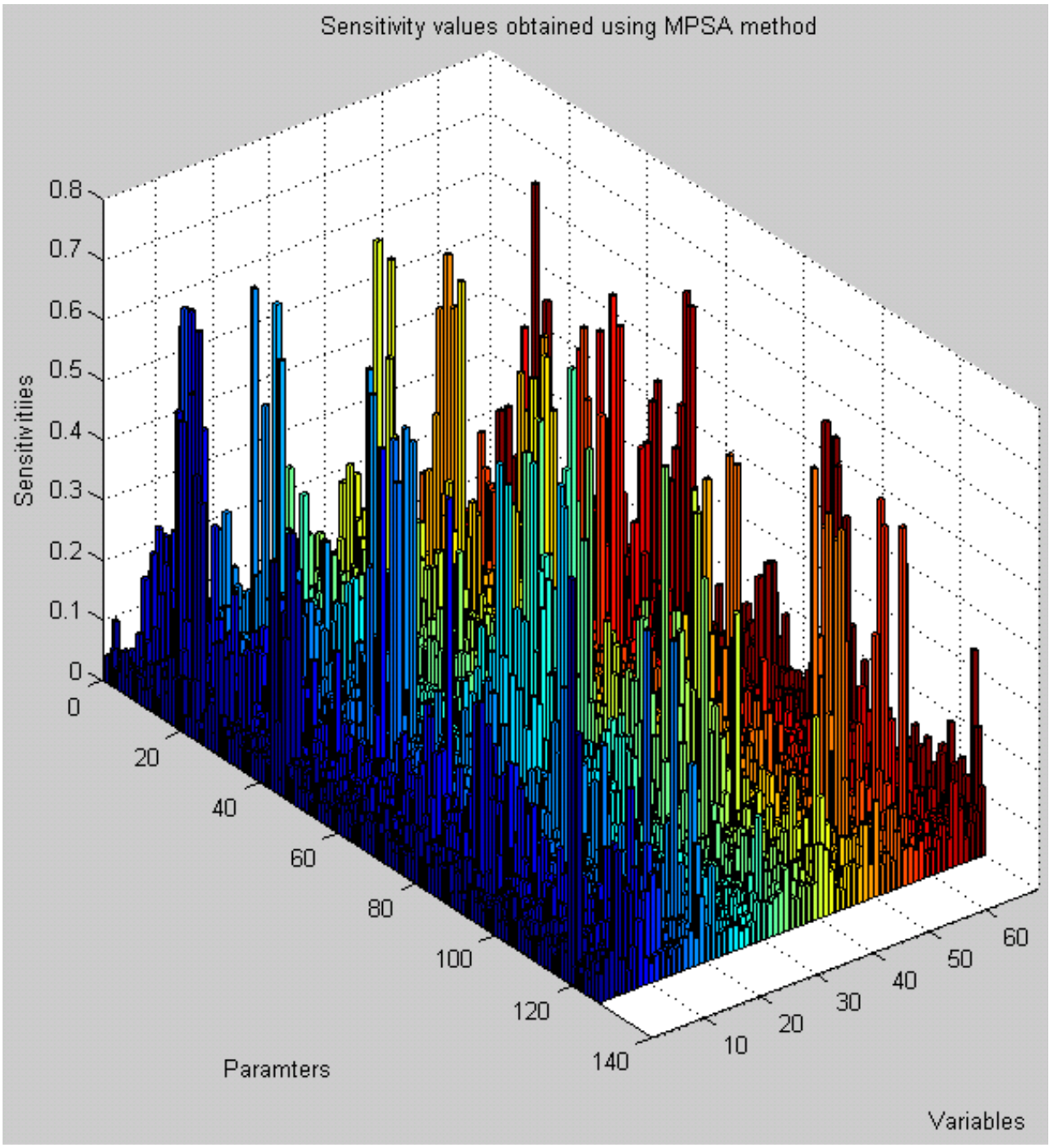

Figure 4.4: 3D bar plot of simulation results for multi-parametric sensitivity analysis (MPSA)

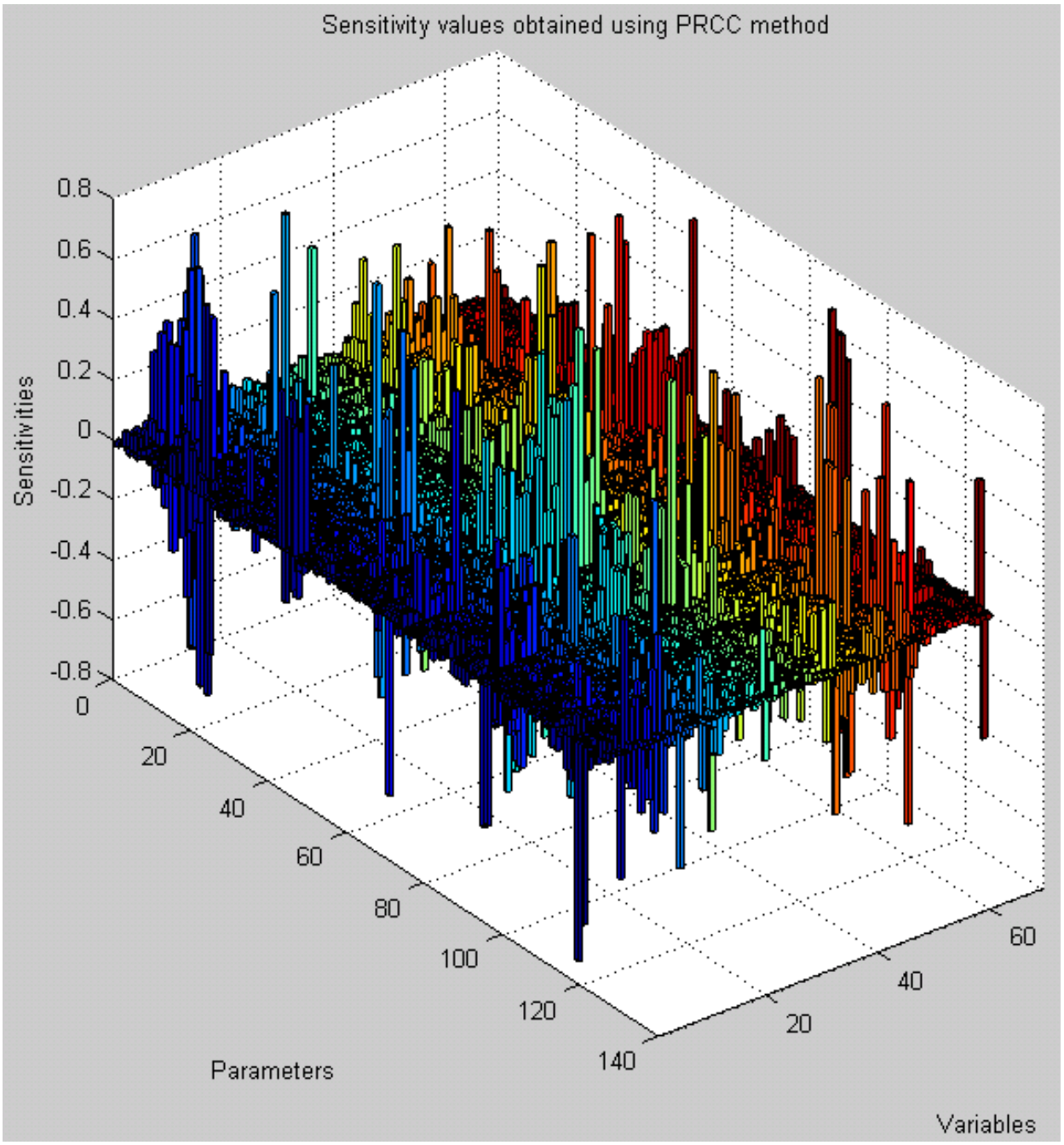

Figure 4.5: 3D bar plot of simulation results for partial rank correlation coefficient analysis (PRCC)

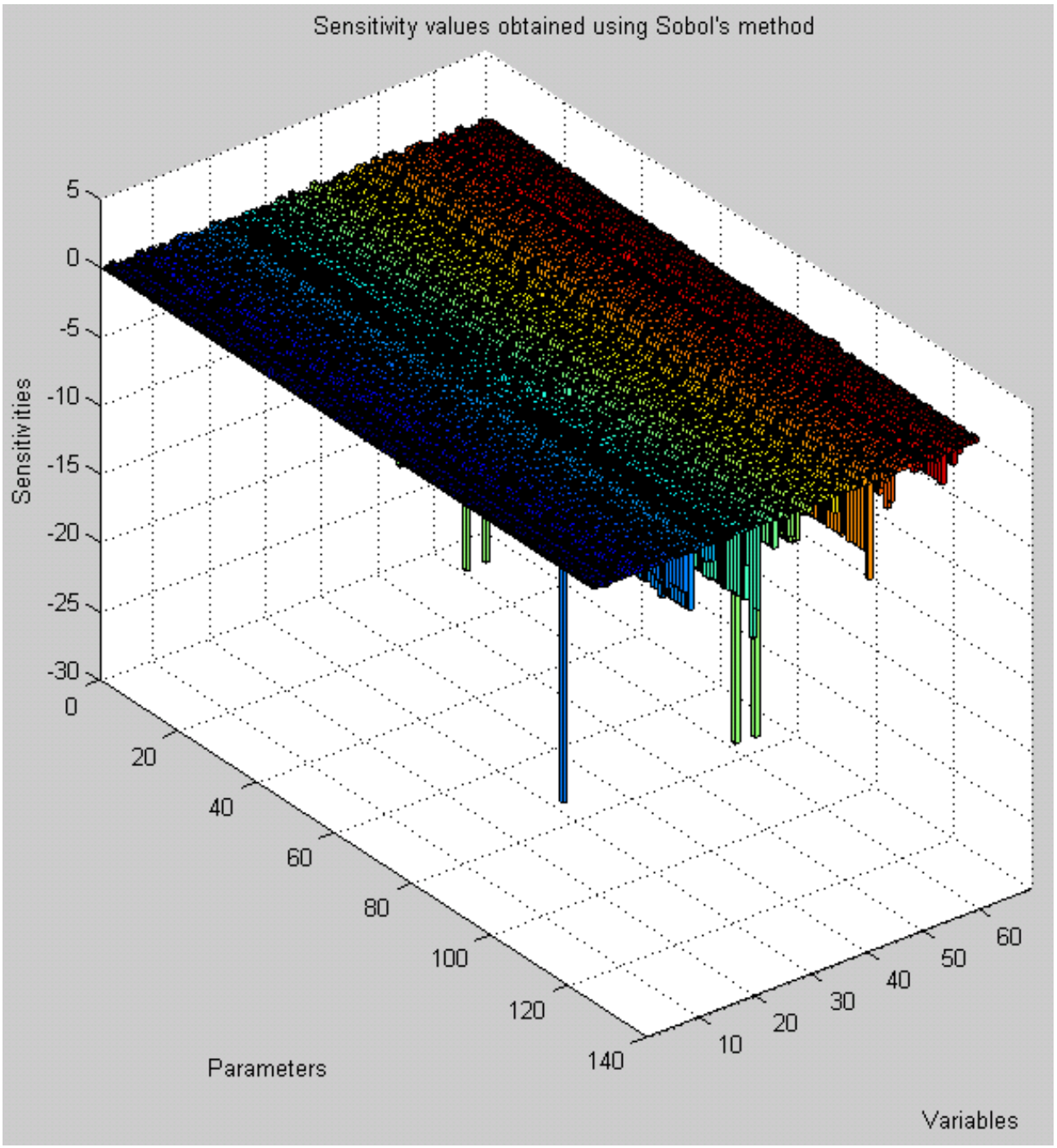

Figure 4.6: 3D bar plot of simulation results for Sobol's method analysis

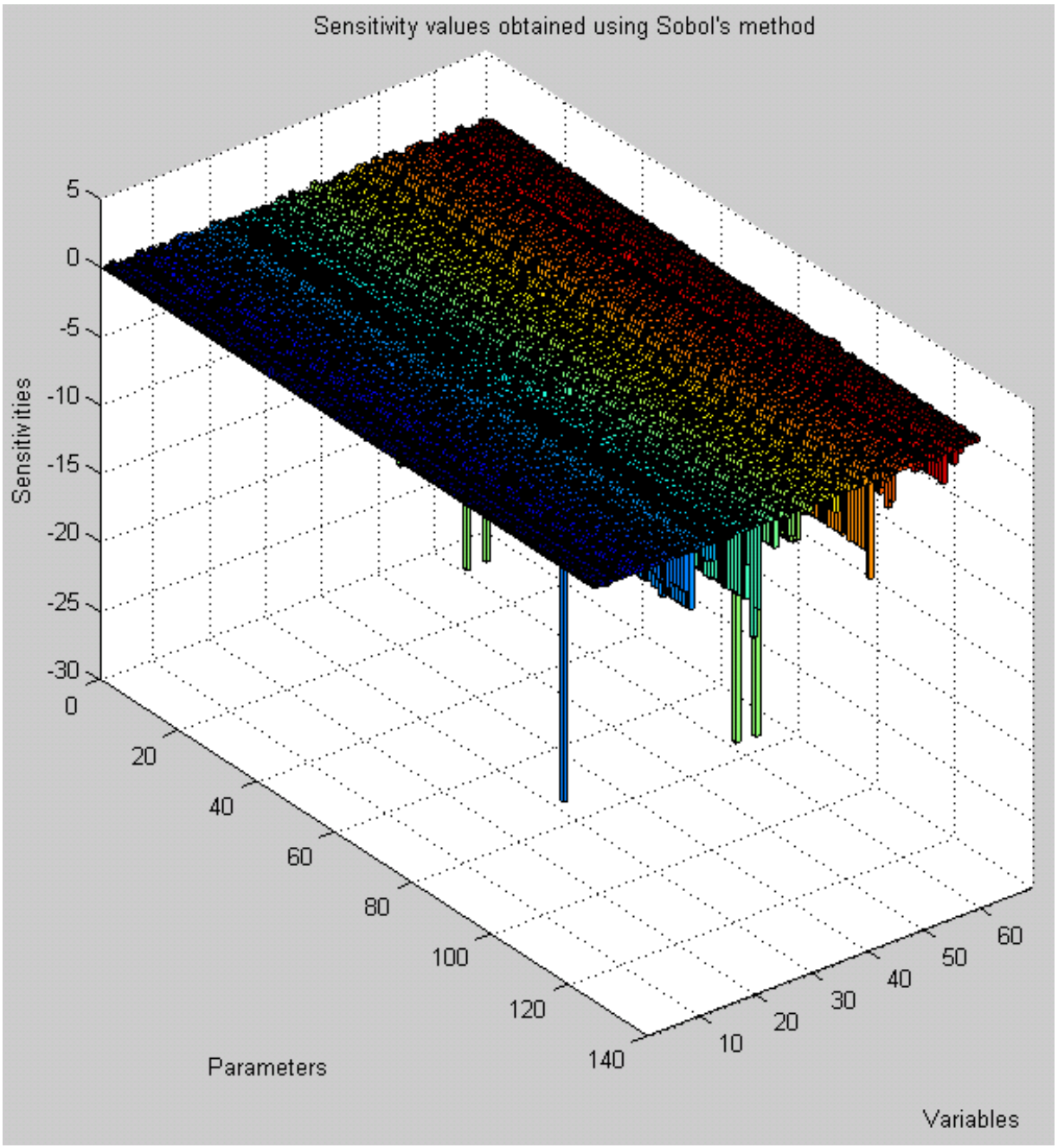

Figure 4.7: 3D bar plot of simulation Results for weighted average of local sensitivities analysis (WALS)

We also performed LSA using both SBML_SAT (Zi *et al.*, 2008) and COPASI software (Institute *et al.*). (Koh e Lee, 2011) used COPASI software to perform LSA and we wanted to see whether SBML_SAT software would perform better and output similar results. Resulting 3D plots are included in figures 8-10. Figure 11 indicates comparison of plots from Copasi and SBML-SAT tools. Figure 12 shows again 3D plot of Sobol method results, but from a different perspective, in order to better illustrate the variability of the sensitivity values. Corresponding data files containing all sensitivity values are available upon request. SBML_SAT software was difficult to set up, but easy to use. On the other hand, Copasi software installed quickly but contains the interface that's counter-intuitive and therefore was more difficult to use.

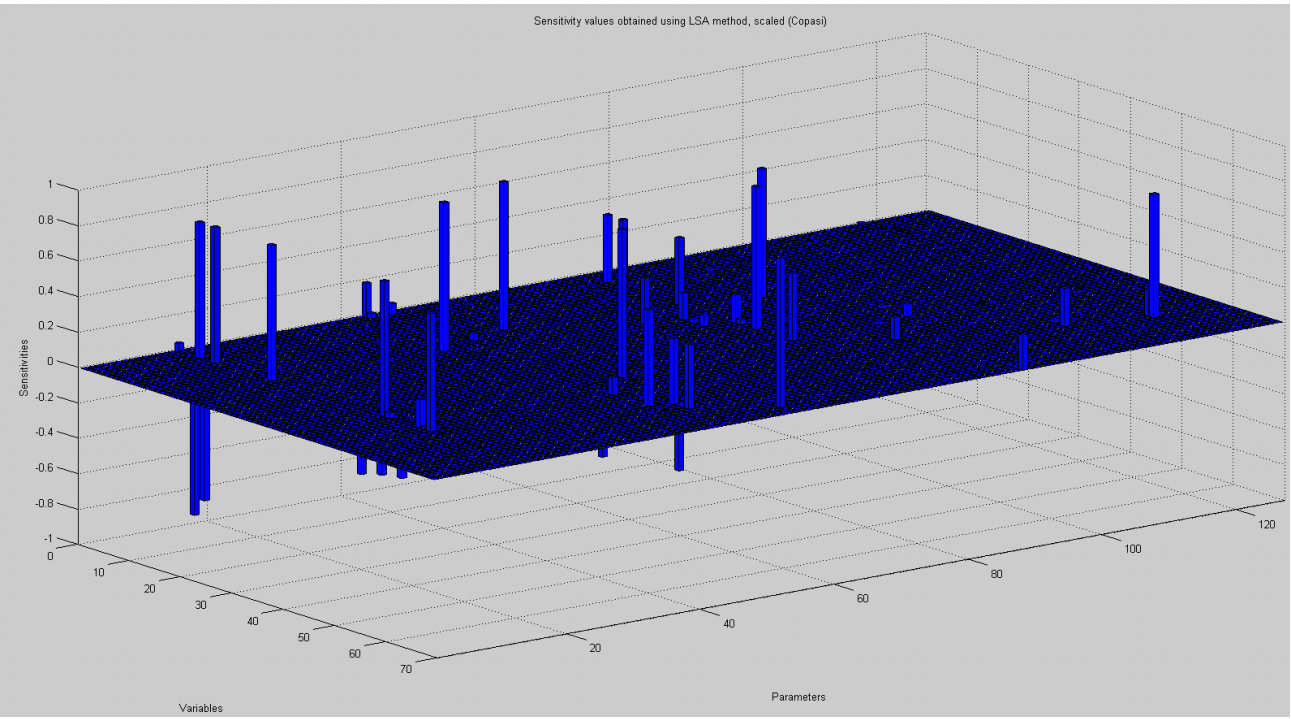

Figure 4.8: 3D bar plot of local sensitivity analysis results obtained using Copasi

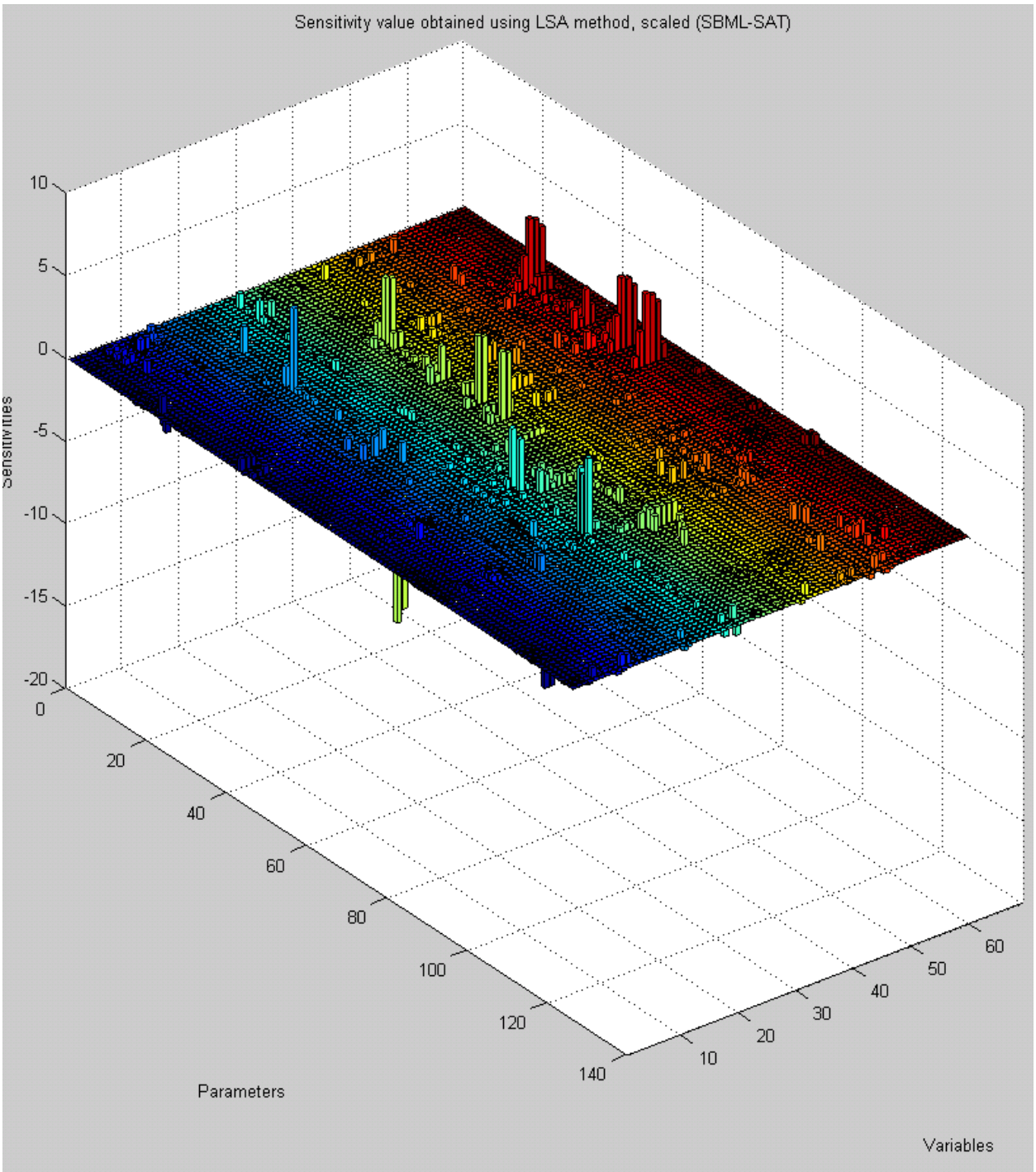

Figure 4.9: 3D bar plot obtained using local sensitivity analysis (LSA) in SBML-SAT

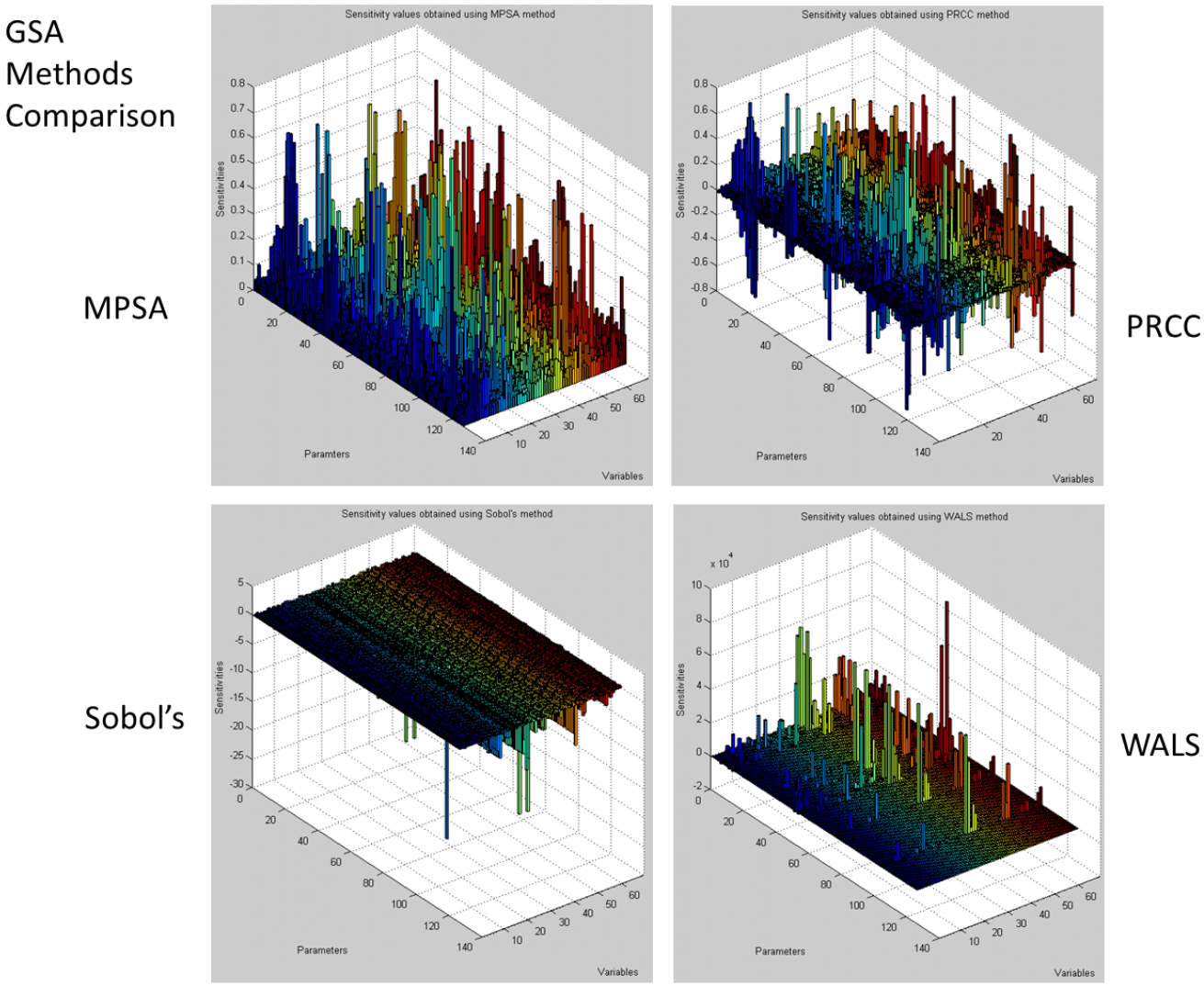

Figure 4.10: Comparison of plots for GSA and LSA methods

Figure 4.11: Comparison of plots from Copasi and SBML-SAT tools

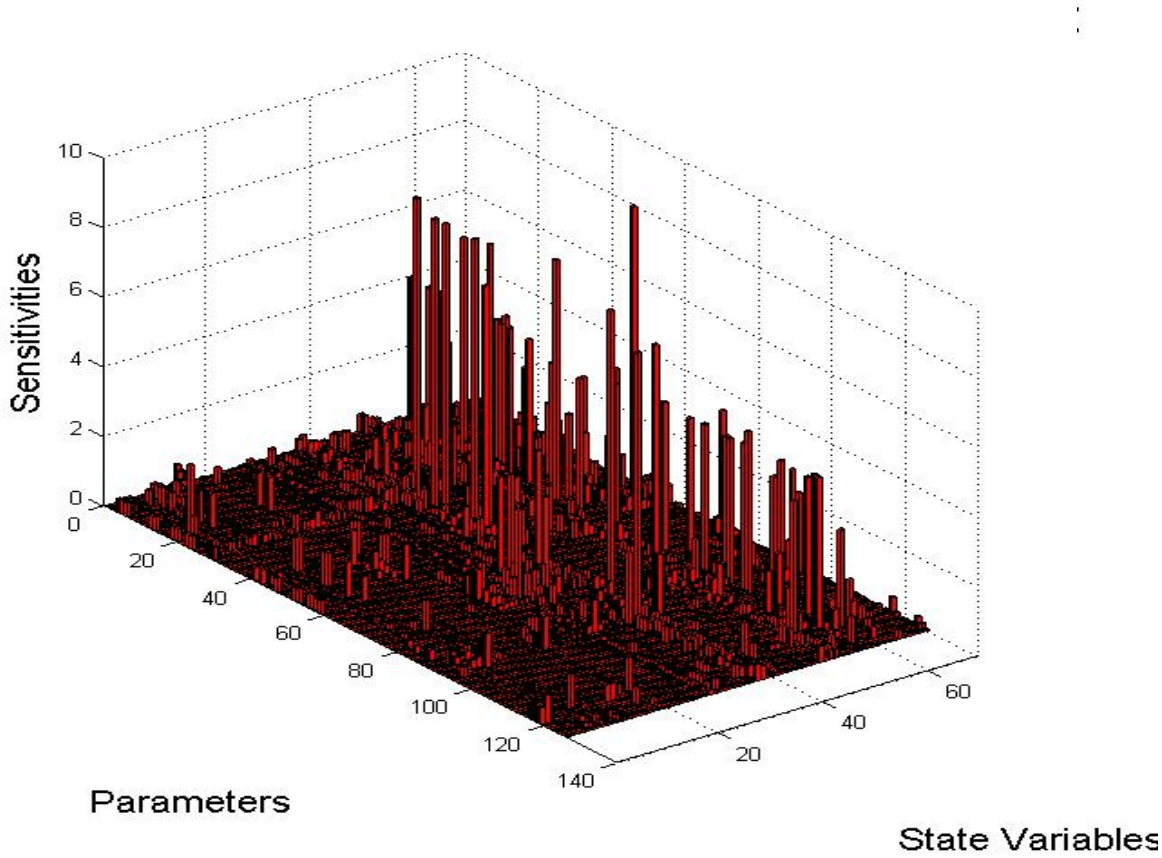

Figure 4.12: Sobol method results

4.5. 2 Augmentation of Human Thyroid Hormone Regulation Dynamics Model for Thyrotoxicosis

We modified the equations in computer model of human thyroid hormone (TH) regulation dynamics and applied computationally intensive global search method in Amigo software (Balsa-Canto e Banga, 2010) called GlobalM to fit the new model parameters. As a result, we obtained an augmented model that fits the available data for the state of thyrotoxicosis, which previous version of the model failed to do, as indicated in figure 13. The pharmacokinetic data we used for modeling thyrotoxicosis consisted of serum T4 and T3 concentrations collected over seven days in three subjects who were given single oral dose of 3mg of Levothroid® (T4) on one occasion and 3 mg of Syntrhoid®(T4) on another (Leboff *et al.*, 1982).

The main change in the model resulted to be the rate equation from compartment 2 to 5. The equation was changed from Michaelis-Menten function to Hill function.

Michaelis-Menten function:

$$v = \frac{v_{max}\, S}{S + K_m}$$

Hill function:

$$v = \frac{v_{max} S^n}{S^n + K_m^n}$$

As can be seen in the formula, Hill function is identical to Michaelis-Menten function when n=1, but as n increases, the graph becomes steeper (figures 14 and 15). Therefore the

search for the most appropriate value of the parameter n is of high importance for model improvement. The resulting parameters values after using global search method that fit the clinical data best are: n1=2000, VD1fast = 0.02, and KD1fast=2.8500 and the computational cost to accomplish this task was: 464.430577 s, as indicated in table 4.

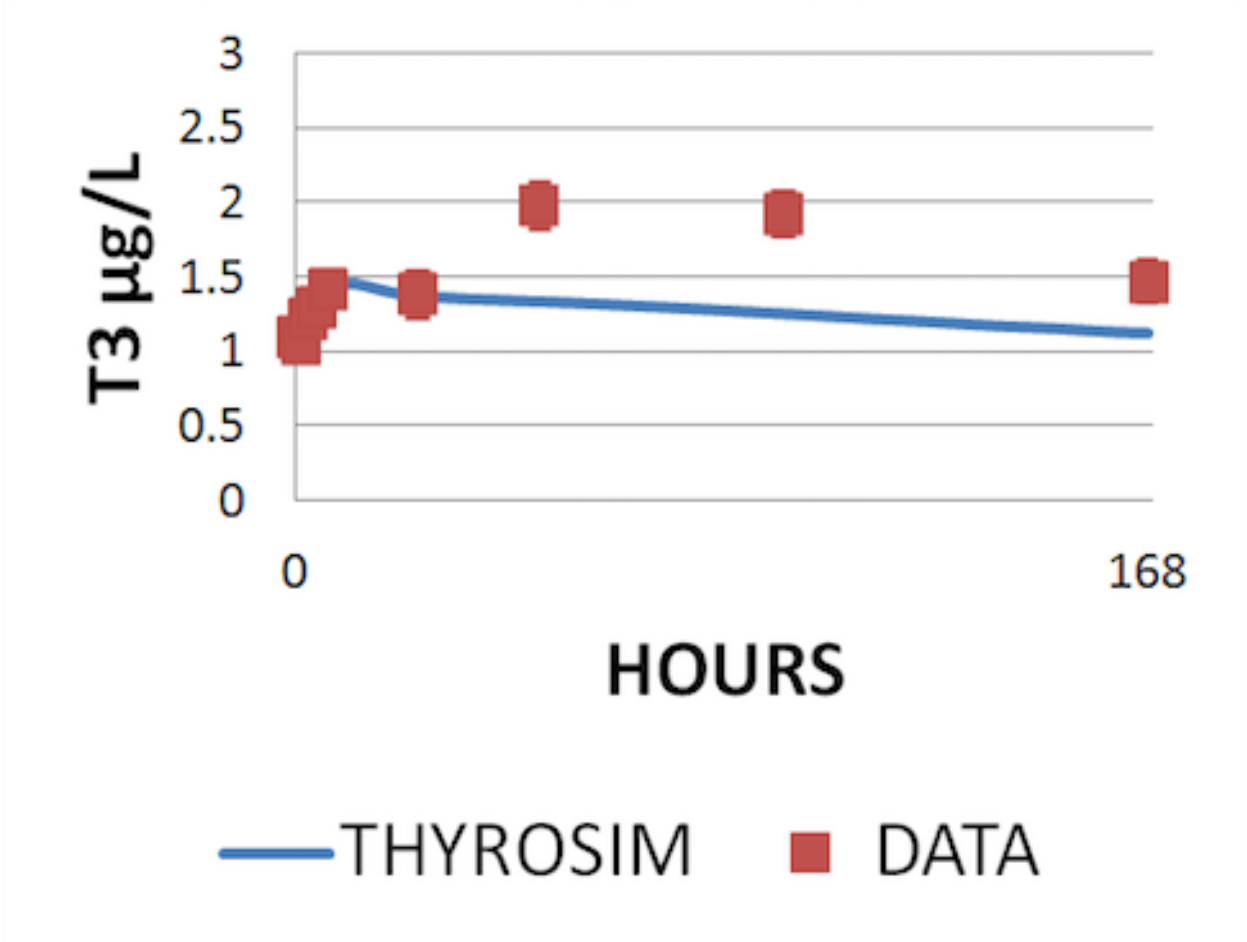

Figure 4.13: Model output in response to 3000ug T4 dose (thyrotoxic) thyroid hormone (TH) treatment. Simulated graph of the original TH model fits T4 but not T3 responses after 24 hours.

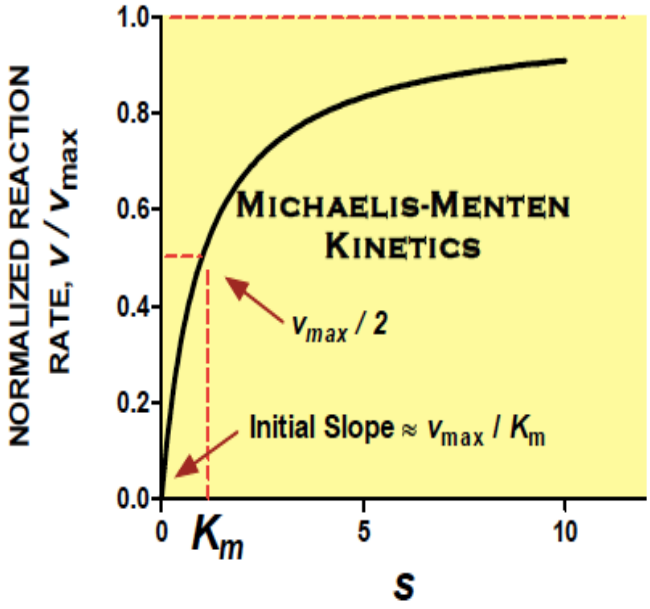

Figure 4.14: Michaelis-Menten function graph. Michaelis-Menten functions are in our simulation model to represent enzyme substrate interactions in cells. As substrate level goes up, reaction rate increases rapidly at first, and the increase rate slow down until it approaches its maximum rate.

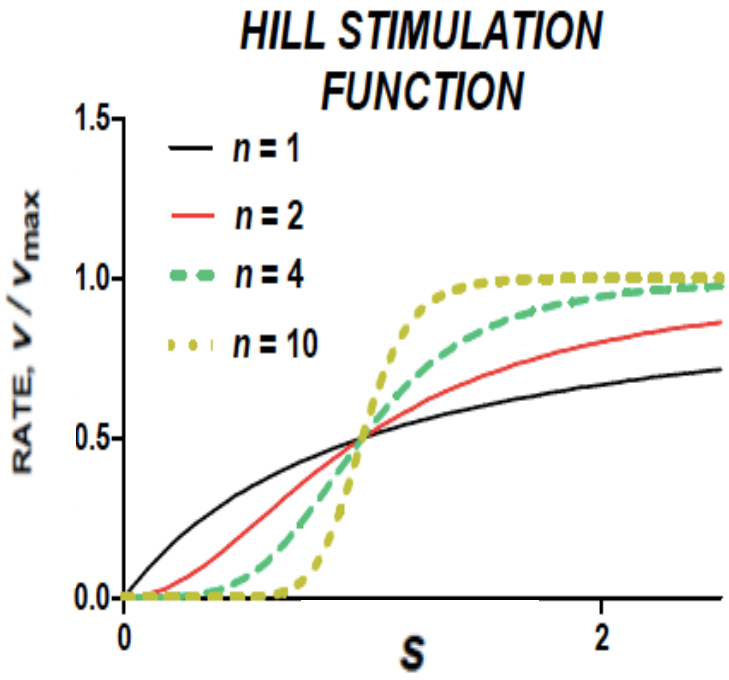

Figure 4.15: Hill function graph for different values of parameter n (n=1, n=2, n=4, n=10)

parameter	value
n	1.2000
VD1fast	0.0200
KD1fast	2.8500
Cost	464.430577 s

Table 4.4: Parameter Search Results using GlobalM Method in AMIGO SW for Thyroid Hormone (TH) model to handle state of thyrotoxicosis.

The resulting augmented model (figure 16) is able to successfully simulate TH body response to thyrotoxicosis, without losing any of its previously available capabilities. Resulting model output is shown in figure 17. The results can also be simulated using Thyrosim application (Han *et al.*, 2016). The snapshot of Thyrosim simulation output is shown in figure 18.

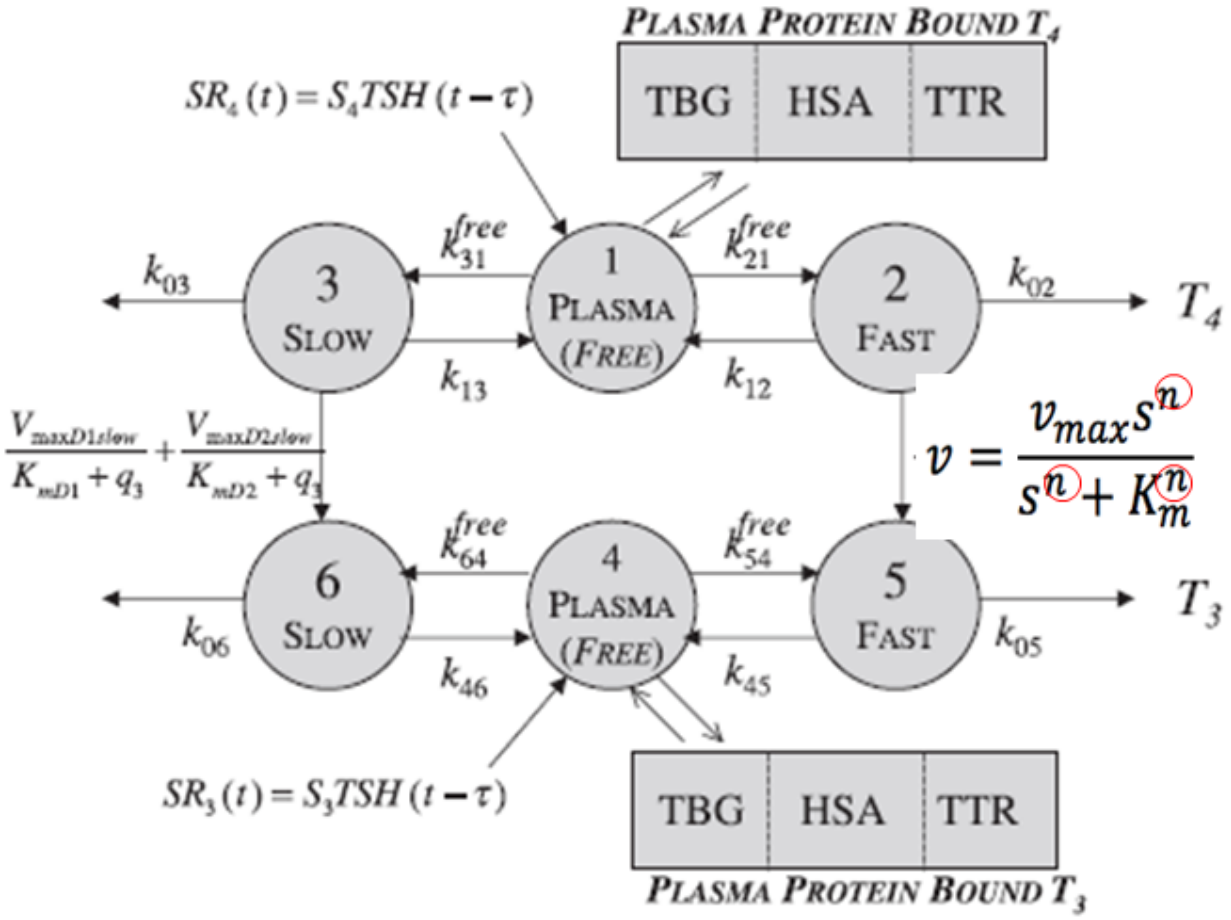

Figure 4.16: Augmented adult T3 and T4 D&E model to handle thyrotoxicosis. Non-linear, extravascular enzymatic T4 to T3 conversion from compartment 2 to 5 rate is based on higher-order Hill function instead of Michaelis-Menten (M-M) function.

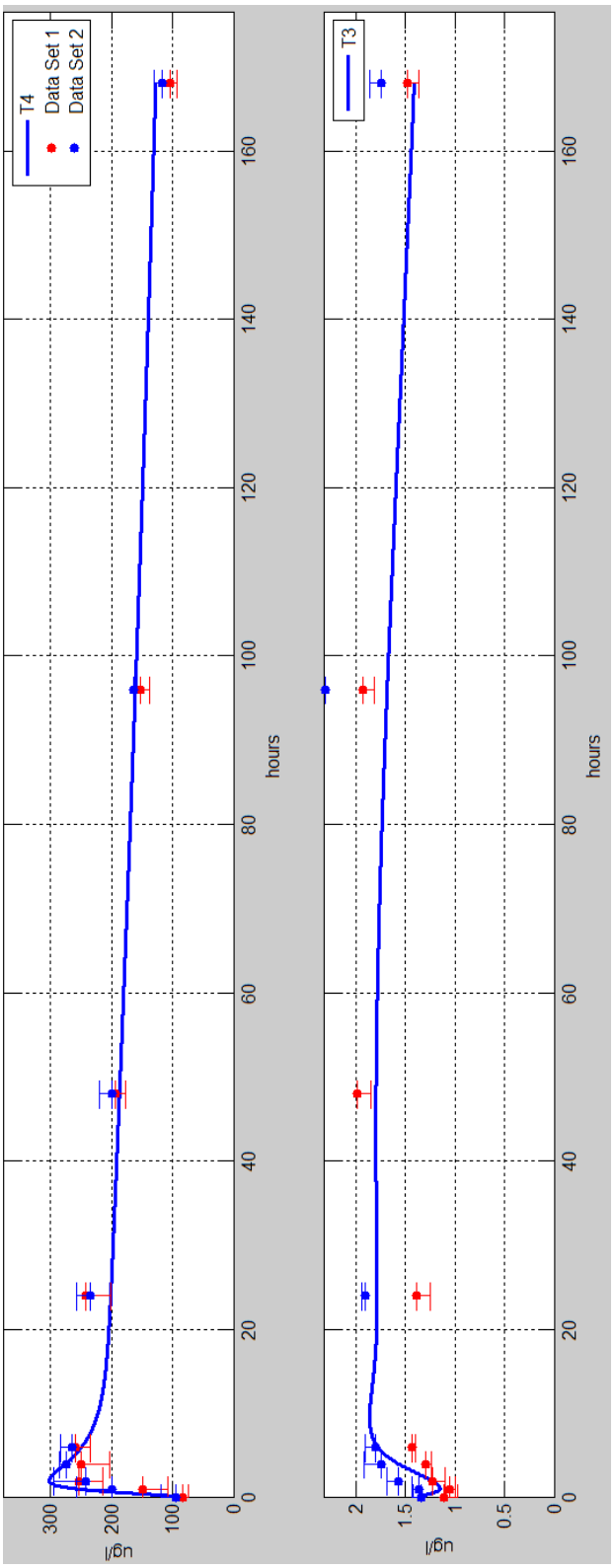

Figure 4.17: Augmented Model Simulation. T3 and T4 model response for 3000ug dose (thyrotoxicosis). Model output fits clinical data published in (Leboff *et al.*, 1982).

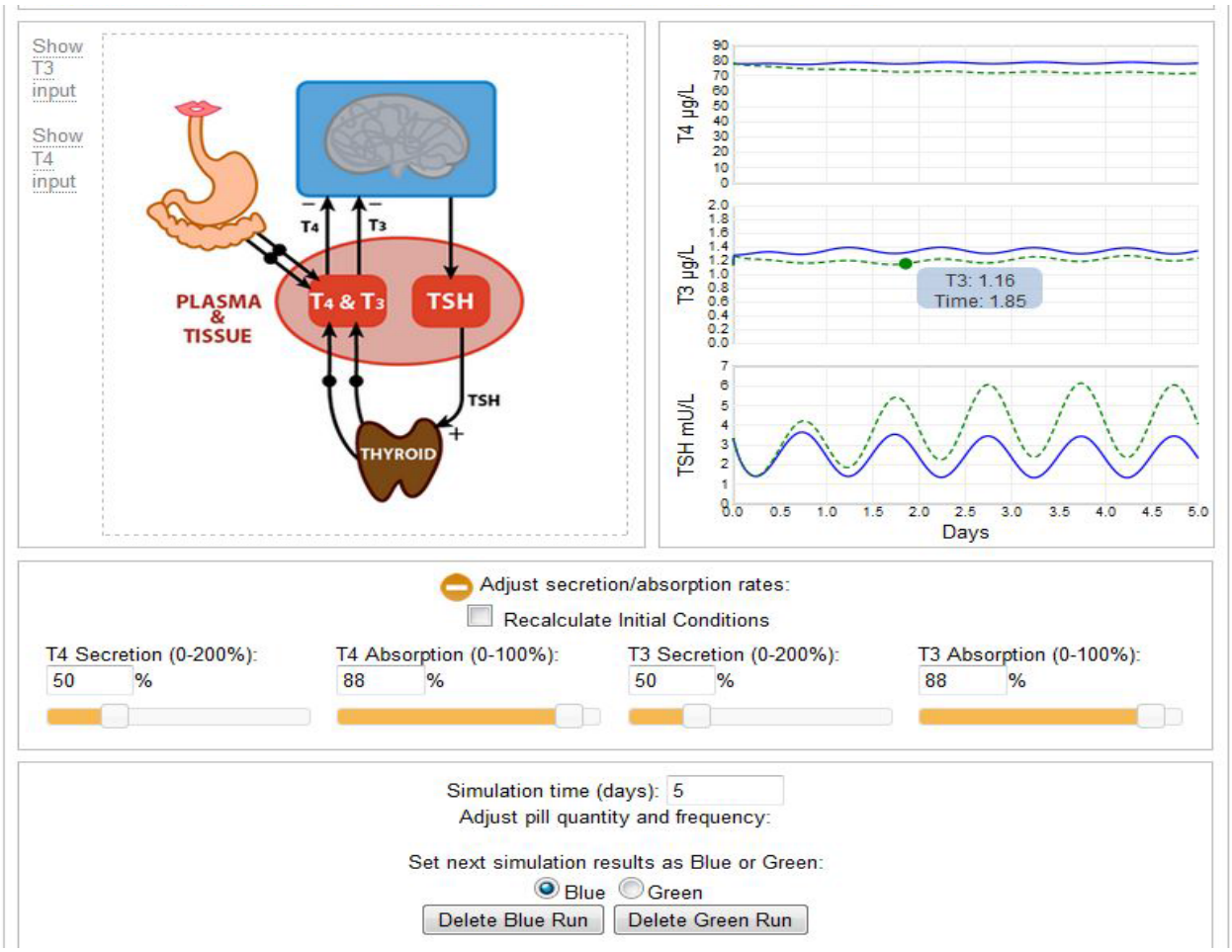

Figure 4.18: Thyrosim web-application used for TH model simulation (Distefano, 2014; Han *et al.*, 2016). With our improvement of TH model, it is possible to accurately simulate the state of thyrotoxicosis in Thyrosim

4.6. Discussion

4.6. 1 Discussion of Global Sensitivity Analysis of NF-kB Model

Our original hypothesis was that different methods for global sensitivity analysis would lead to similar results and accurate identification of the most sensitive parameters in the model.

Table 1 displays 10 most sensitive parameters obtained using PRCC, SOBOL's, and WALS methods in reference to a single state variable (s57). Different parameters are ranked in a different order by each method. There seems to be little correlation between results and no single parameter or parameter set stands out as most sensitive. Tables 2 and 3 also summarize top parameter rankings obtained using aforementioned GSA methods, but in reference to a different state variables (s194 and s222, respectively).

Second column of each table indicates top-ten parameter rankings based on sensitivity values obtained using PRCC method. Table 1 ranks parameters in decreasing order: ki18, ki22, ki1, ki2, ke9, ki4, ki5, kr4, ke3, kc5; Table 2: kr15, kn18, kn5, kn15, kr5, kr6, kn1, kn12, kn10, ki23; Table 3: kr2, kr1, kr4, kn18, ke9, kn5, kn10, kn15, kr17, kr19. Since objective of analysis is to determine most sensitive parameters in the system, one would expect same set would be ranked most-sensitive in majority of cases, but they are not. Similar comparison can be done row-wise by comparing parameter rankings obtained using different methods with respect to same state variable.

We concluded that all results are different and there is no evident pattern among parameter sensitivities obtained using different methods. There are also no apparent relationships among parameter rankings obtained by the same method in reference to a different state variable.

The lack of clear correlation is even more evident from the available plots. Figure 1 displays 3D-bar plot obtained by MPSA, reproduced from (Koh e Lee, 2011). Figures 2, 3, and 4 display 3D-bar plots obtained by PRCC, SOBOL's, and WALS methods, respectively. Sensitivities of all parameters appear fairly uniform (instead of high and low), especially for Sobol and WALS methods (a single color dominates each map). However, a different color dominates each plot (e.g. red for Sobol's, green for PRCC, and blue for WALS), even though the scale of each heat-map is the same (ranging from blue to red, blue indicating the least sensitive, and red indicating the most sensitive areas). In addition, ranges of sensitivity values seems to be far apart for each method, especially for SOBOL and WALS methods that are not scaled to a known interval, which warrants further investigation (there appears to be limitation in SBML-SAT package).

Even though LSA does not capture interactions/couplings of model parameters, which are likely to occur in nature, we were interested to verify if two software packages would provide similar results when performing same type of analysis on the exact same model. However, the results are different, as can be seen on 3D plots in figure 7, which contain different sensitivity values for the same set of parameter-state variable combinations.

Since the goal of sensitivity analysis is to identify key components of the same biological pathway, we would expect that these methods would produce similar results (allowing some, but little room for error). We have repeated each experiment several times in order to obtain more robust results, but results indicate something amiss. All methods should give similar results in order to provide useful information and aid in design of biological experiments and clinical studies. They do not. It is difficult to determine which parameters need to be estimated and which can be fixed to nominals or eliminated from the model. This remains an open research

area, and as with many computational methods, a user should choose one that is the most appropriate for the task at hand.

4.6.2. Augmentation of Human Thyroid Hormone Regulation Dynamics Model for Thyrotoxic Dose

Human thyroid hormone (TH) plays a role in metabolism and development of all tissues of the body. It is secreted by the thyroid gland, which is located below the neck and shaped in the form of a butterfly. TH refers to two hormones: thyroxine (T4) and triiodothyronine (T3). The pituary gland stimulates production of thyroid stimulating hormone (TSH), which then stimulates the thyroid gland to produce TH. T4 could be described as pro-hormone (precursor), while T3 is the product (derivative) of T4. T3 is major thyromimetric (active) hormone, and could be either bound or free, depending on whether it's bound or not to plasma proteins. Both T4 and T3 are unbound when secreted. Proteins thyroxine-binding globulin (TBG), human serum albumin (HAS), and transthyretin (TTR) act like capacitors and carry the hormone around the body in blood. In general, only free hormone is able to penetrate cellular membrane and flow into the cell.

Conversion from T4 to T3 happens in all organs, but each tissue makes different amount of it. For example, brain cells have a high conversion rate – they make the most T3 out of T4 (almost all they need). Overall, we can differentiate between slowly exchanging organs such as brain, muscle, and skin, and fast exchanging organs such as thyroid, kidney and liver. The reason for this differentiation is that it can be assumed that only one enzyme, called type 1 diodiniase (D1), catalyzes conversion of T4 to T3 (chemical reaction) in tissues with fast exchange, while two enzymes D1 and type 2 diodiniase (D2) are responsible for the same conversion in tissues

with slow exchange of TH. In fact, it is assumed that 80% of conversions occur via D2 in this case. Both D1 and D2 are considered activating enzymes, but a third enzyme type 3 diodiniase (D3), is deactivating, acting as a scavenger (Maia *et al.*, 2011). The mechanism of T3 interaction with these three enzymes is not fully understood, but it is known that D1 induces transcription of mRNA via dio1 gene and therefore increases production of T3 (Maia *et al.*, 2011; Darras e Van Herck, 2012).

The balance of TH in the body is maintained in large part via a closed negative feedback loop: when TH blood levels decrease, more TH gets produced, while the opposite occurs when TH blood levels are high. Plasma levels of all three hormones (TSH, T3, and T4) can be measured in blood. All hormone levels exhibit daily oscillations. Thyroid diseases are conditions affected by the amount of TH. Clinically, TSH values are used to discern among these conditions as the most sensitive to change. Normal TSH range (in euthyroid individuals) is 0.5 – 5 mU/l. Individuals who have TSH level greater than 4.5 mU/l are hypothyroid (thyroid gland doesn't produce enough TH), while those who have TSH level less than 0.5 mU/l are called hyperthyroid (thyroid gland produces too much TH). Normal ranges for T4 and T3 are: 5-12 ug/dl and 0.8-1.9 ng/ml, respectively. Graves disease is the most common type of hyperthyroidism, where thyroid gland produces excessive amounts of hormone as a response to antibodies activated by the autoimmune system (Thyroid Diseases). Hyperthyroidism can be severe hyperthyroidism and lead to thyrotoxicosis. Thyrotoxicosis occurs when toxic amount of T3 is present in the body causing the suppression of TSH and T4 to subnormal levels. It can also occur without hyperthyroidism, due to thyroiditis (inflammation of the thyroid gland) causing release of TH stored in the gland, or after ingestion of excessive amounts of TH hormone supplements

(Devereaux e Tewelde, 2014). Even mild thyrotoxicosis may have adverse effects on cardiac and bone health (Williams, 1997).

The default simulation model is challenged by published thyrotoxic data in adults, following a super-physiological dose of 3mg of oral T4 (either Levothroid® or Synthroid®) (Leboff *et al.*, 1982). It underestimates the observed T3 response in plasma, failing to reproduce ~30% rise in T3 concentrations at times greater than 24 hours, even though it correctly predicts T3 responses for the first 24 hours and T4 responses over the same 7 day period (similar peak values and areas under the curve (AUC's)). Our current model addresses this issue by incorporating update based on the hypothesis of the underlying biological mechanism. Data shows the increase in free T4 and free T3 fractions by 50% in the subjects who were given this super-physiological dose. Therefore, we hypothesize that conversion rates of T4 toT3 in both fast and slow tissues amplify between 24 and 48h under thyrotoxic conditions. Physiologically, due to abrupt increase of the amount of T4 (it increased to > 25 μg/dl) (Eisenberg *et al.*, 2008), TBG gets saturated (reaches its maximum binding capacity, which is only about 20 ug/dl of T4), and therefore the binding of both T4 and T3 is shifted to the lower affinity serum proteins (HAS and TTR). After those proteins get saturated as well, leftover (free) hormone escapes into the cell during the first 24 hours. After 24 hour delay, the dio1 gene is activated, makes mRNA, which subsequently makes D1 and increases T3 production from increased T4. This amplified increase in T3 level causes positive feedback (D1 interacting with T3 and making more of T3) until T4 level begins to fall (gut absorbs oral dose approximately after 48 hours). The result is a significantly greater and transient increase in total T3 (~30%) than would be expected for oral dosing with T4 within physiological range.

The original model assumed approximate Michaelis-Menten kinetics for T4 to T3 interconversion, via both D1 and D2 deiodinase activities, with enzyme concentrations assumed approximately constant. Even though, this was a good approximation for hypo- and euthyroid TH ranges, a more appropriate model for thyrotoxic TH levels was needed. The augmented model needed to simulate T3 production more substantially increased for extended periods, via positive product feedback of T3 on D1 production and thus the D1-T4 reaction. Simple Hill function with Hill constant $n > 1$ was able to accurately reflect this, with the challenge being to find the appropriate value of n. The overall mechanistic model underlying this biosystem involves coupled effects among variables and parameters of these subsystems and others, meaning significantly more effort was needed to fully refine model parameters to fit the extreme thyrotoxic response data, which we accomplished by using global search algorithm.

Chapter 5

Conclusion

5.1 Summary of Contribution of this Dissertation

At the time of writing this dissertation, it is estimated that only 4% of the human genome (including the aforementioned 1.5%) is exonic, while 96% of the human genome is non-exonic. Genomic and epigenomic datasets are large-scale data, which size is currently measured in gigabytes, terabytes, and possibly even petabytes. These datasets require statistical and computational methodology, in addition to incredible amount of computational resources (such as time, memory, and processing power) in order to be analyzed and produce scientific findings. This dissertation makes a contribution to developement and/or application of existing computational, statistical, numerical, and machine learning methods to analyze large-scale biomolecular and genomic data.

5.2 Summary of Methods

We presented computational and quantitative methods developed specifically for processing large-scale data and searching vast parameter space.

In particular, Predictor and Identifier of Conserved Regions (PICEL) method incorporates thousands of genomic datasets and provides a single resource to the scientific community. It uses supervised machine learning technique logistic regression for which the

details are given in chapter two. PICEL serves as a single genome-wide score for predicting whether the particular position on the human genome belongs to evolutionary conserved region.

Percent of variation statistics was used to identify how gene expression values vary between brain and blood and among various brain regions, and detect genes that had similar value of expression in both blood and brain tissues of the vervet monkeys. The dataset was large and required to be prepared and processed on a cluster computer using various scheduling techniques. Hierarchical clustering was used to characterize tissue specific gene expression differences among eight brain regions. Prior to obtaining results of the gene expression study, we successfully applied BLAST method to perform probe-sequence comparison and derived computational methodology to analyze large-scale output data, in order to verify whether gene expression studies for vervets could be performed using human probes.

Finally, we carried out various numerical methods in order to perform sensitivity analysis on NF-kB dynamic system model in an attempt to simplify the model and discern key molecular components of it. By using global search method, we were able to identify parameter values for thyroid hormone control flow system model in order to properly handle thyrotoxic case, as measured by clinical data.

5.3 Applications

With PICEL method, we have only scratched the surface of using its output to perform heritability analysis, researching false negatives results in gene ontology, and identifying false positive results as candidates for future biomolecular or genetic studies. The main purpose of PICEL is to summarize data and help discern the more interesting locations in the human

genome and the hope that a tool of this caliber may lead to discoveries in biological chemistry, human genetics, and medicine.

5.4 Future Work Methodology

PICEL method could be improved by incorporating even more (thousands of) data tracks, especially if new technologies get developed that measure areas of the open chromatin and also potentially if more accurate methods and data detecting conservation appear. In addition, model training could be performed using different form of supervised learning technique. Furthermore, model accuracy could be even more improved if more samples are used for training, which would be possible using stochastic logistic regression.

5.5 Future Work Grand Challenge

This work demonstrates that it is possible to use computers and algorithms in search for discoveries that would potentially lead to novel scientific findings. About a decade ago, when the author was introduced to this research area, the grand challenges were somewhat different. Computationally, processing power was reaching Moore's law and multi-core technologies, and the powerful cluster computers and process parallelization had been in its infancy successfully solving this problem. The idea of applying statistical and machine learning techniques to large-scale genomic datasets was only a few years old at a time.

In the field of human genetics, very little was known about human genome or genomes of many other species used as model organisms for studying complex traits and human diseases. The field of epigenetics was barely emerging. Less efficient wet lab techniques like western blot

and PCR were regularly used for isolating only small sequences of the genome, and massive parallel sequencing such as gene expression and next generation sequencing had just emerged. Furthermore, most types of cancer diseases at the time were in majority of the cases still incurable, and far less was known about mechanisms involving cancer or genes that could be contributing to its formation.

At the moment of concluding this dissertation computational methodology and quantitative techniques are closely tied to any work produced in the wet lab. More importantly, there are many forms of cancer that are now detectable and curable. Even though it is out of scope of this dissertation, it is fascinating to note that robotic surgery is a thing of the present (and not science-fiction), with many such surgeries successfully performed on daily basis.

The author hopes and firmly believes that this type of quantitative and computational methodology will prevail in the future as well, in order to aid in more efficient, less harmful and less invasive search for new therapies and medication, especially for finding cures for various neuro-psychological disorders (such as depression) that are starting to plague our society and are presenting the next grand challenge to diagnose, treat and cure.

References

BALSA-CANTO, E.; BANGA, J. R. **AMIGO: A model identification toolbox based on global optimization**. Computer Applications in Biotechnology, Leuven 2010.

BASAK, S.; BEHAR, M.; HOFFMANN, A. Lessons from mathematically modeling the NF-kappaB pathway. **Immunol Rev,** v. 246, n. 1, p. 221-38, Mar 2012. ISSN 0105-2896. Disponível em: < http://dx.doi.org/10.1111/j.1600-065X.2011.01092.x >.

BEN-SHACHAR, R. et al. Simulation of Post-thyroidectomy Treatment Alternatives for T3 or T4 Replacement in Pediatric Thyroid Cancer Patients. **Thyroid**, Dec 22 2011. ISSN 1050-7256. Disponível em: < http://dx.doi.org/10.1089/thy.2011-0355 >.

BERGER, S. L. The complex language of chromatin regulation during transcription. **Nature,** v. 447, n. 7143, p. 407-12, May 24 2007. ISSN 0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature05915 >.

BERNSTEIN, B. E.; MEISSNER, A.; LANDER, E. S. The mammalian epigenome. **Cell,** v. 128, n. 4, p. 669-81, Feb 23 2007. ISSN 0092-8674 (Print)0092-8674. Disponível em: < http://dx.doi.org/10.1016/j.cell.2007.01.033 >.

BIRD, A. Perceptions of epigenetics. **Nature,** v. 447, n. 7143, p. 396-8, May 24 2007. ISSN 0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature05913 >.

BUSH, W. S.; MOORE, J. H. Chapter 11: Genome-Wide Association Studies. In: (Ed.). **PLoS Comput Biol**, v.8, 2012. ISBN 1553-734X (Print)1553-7358 (Electronic).

CHANG, R. L. et al. Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. In: (Ed.). **Mol Syst Biol**, v.7, 2011. p.518. ISBN 1744-4292 (Electronic).

CHU, Y.; HAHN, J. **Parameter set selection for estimation of nonlinear dynamic systems.** AIChE J. 53**:** 2858-2870 p. 2007.

_____. Parameter Set Selection via Clustering of Parameters into Pairwise Indistinguishable Groups of Parameters. September 18, 2008 2008. Disponível em: < http://pubs.acs.org/doi/abs/10.1021/ie800432s >.

CINTRON-ARIAS, A., BANKS, H.T., CAPALDI, A. & LLOYD, A.L. **A sensitivity matrix based methodology for inverse problem formulation.** J. *Inv*. Il*l-p*osed *Problems*. 17**:** 545-564 p. 2009.

CONSORTIUM, T. E. P. An integrated encyclopedia of DNA elements in the human genome. **Nature,** v. 489, n. 7414, p. 57-74, 09/06/print 2012. ISSN 0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature11247 >.

COOPER, G. M. et al. Distribution and intensity of constraint in mammalian genomic sequence. **Genome Res,** v. 15, n. 7, p. 901-13, Jul 2005. ISSN 1088-9051 (Print)1088-9051. Disponível em: < http://dx.doi.org/10.1101/gr.3577405 >.

DARRAS, V. M.; VAN HERCK, S. L. Iodothyronine deiodinase structure and function: from ascidians to humans. **J Endocrinol,** v. 215, n. 2, p. 189-206, Nov 2012. ISSN 0022-0795. Disponível em: < http://dx.doi.org/10.1530/joe-12-0204 >.

DAUN, S. et al. An ensemble of models of the acute inflammatory response to bacterial lipopolysaccharide in rats: results from parameter space reduction. **J Theor Biol,** v. 253, n. 4, p. 843-53, Aug 21 2008. ISSN 0022-5193. Disponível em: < http://dx.doi.org/10.1016/j.jtbi.2008.04.033 >.

DAVYDOV, E. V. et al. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. **PLOS Comput Biol,** v. 6, n. 12, p. e1001025, 2010.

DE PAUW, D.; STEPPE, K.; DE BAETS, B. Identifiability analysis and improvement of a tree water flow and storage model. **ISSN: 0025-5564**, 2008 2008. Disponível em: < https://biblio.ugent.be/publication/423657 >.

DEVEREAUX, D.; TEWELDE, S. Z. Hyperthyroidism and thyrotoxicosis. **Emerg Med Clin North Am,** v. 32, n. 2, p. 277-92, May 2014. ISSN 0733-8627. Disponível em: < http://dx.doi.org/10.1016/j.emc.2013.12.001 >.

DISTEFANO, J. **Dynamic Systems Biology Modeling and Simulation**. 2014. Disponível em: < https://books.google.com/books/about/Dynamic_Systems_Biology_Modeling_and_Sim.html?id=nWoYAgAAQBAJ >.

EISENBERG, M.; DISTEFANO, J. J. TSH-based protocol, tablet instability, and absorption effects on L-T4 bioequivalence. **Thyroid,** v. 19, n. 2, p. 103-10, Feb 2009. ISSN 1050-7256. Disponível em: < http://dx.doi.org/10.1089/thy.2008.0148 >.

EISENBERG, M.; SAMUELS, M.; DISTEFANO, J. J., 3RD. L-T4 bioequivalence and hormone replacement studies via feedback control simulations. **Thyroid,** v. 16, n. 12, p. 1279-92, Dec 2006. ISSN 1050-7256 (Print)1050-7256. Disponível em: < http://dx.doi.org/10.1089/thy.2006.16.1279 >.

_____. Extensions, validation, and clinical applications of a feedback control system simulator of the hypothalamo-pituitary-thyroid axis. **Thyroid,** v. 18, n. 10, p. 1071-85, Oct 2008. ISSN 1050-7256. Disponível em: < http://dx.doi.org/10.1089/thy.2007.0388 >.

EISENBERG, M. C. et al. TSH regulation dynamics in central and extreme primary hypothyroidism. **Thyroid,** v. 20, n. 11, p. 1215-28, Nov 2010. ISSN 1050-7256. Disponível em: < http://dx.doi.org/10.1089/thy.2009.0349 >.

ENCODE, P. C. An integrated encyclopedia of DNA elements in the human genome. **Nature,** v. 489, n. 7414, p. 57-74, 09/06/print 2012. ISSN 0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature11247 >.

ERNST, J.; KELLIS, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. **Nat Biotech,** v. 28, n. 8, p. 817-825, 08//print 2010. ISSN 1087-0156. Disponível em: < http://dx.doi.org/10.1038/nbt.1662 >.

ERNST, J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types. **Nature,** v. 473, n. 7345, p. 43-49, 05/05/print 2011. ISSN 0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature09906 >.

ESTELLER, M. Epigenetics in cancer. **N Engl J Med,** v. 358, n. 11, p. 1148-59, Mar 13 2008. ISSN 0028-4793. Disponível em: < http://dx.doi.org/10.1056/NEJMra072067 >.

FAN, R.-E. et al. LIBLINEAR: A Library for Large Linear Classification. **Journal of Machine Learning Research,** v. 9, n. Aug, p. 1871-1874, 2008. ISSN ISSN 1533-7928. Disponível em: < http://www.jmlr.org/papers/volume9/fan08a/fan08a.pdf >.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, n. 27, p. 861-874, 2006.

FAY, J. C.; WYCKOFF, G. J.; WU, C.-I. Positive and Negative Selection on the Human Genome. 2001-07-01 2001. Disponível em: < http://www.genetics.org/content/158/3/1227 >.

FEINBERG, A. P.; TYCKO, B. The history of cancer epigenetics. **Nat Rev Cancer,** v. 4, n. 2, p. 143-53, Feb 2004. ISSN 1474-175X (Print)1474-175x. Disponível em: < http://dx.doi.org/10.1038/nrc1279 >.

FINUCANE, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. **Nat Genet,** v. 47, n. 11, p. 1228-35, Nov 2015. ISSN 1061-4036. Disponível em: < http://dx.doi.org/10.1038/ng.3404 >.

GAL-YAM, E. N. et al. Cancer epigenetics: modifications, screening, and therapy. **Annu Rev Med,** v. 59, p. 267-80, 2008. ISSN 0066-4219 (Print)0066-4219. Disponível em: < http://dx.doi.org/10.1146/annurev.med.59.061606.095816 >.

GARBER, M. et al. Identifying novel constrained elements by exploiting biased substitution patterns. 2009-06-15 2009. Disponível em: < http://bioinformatics.oxfordjournals.org/content/25/12/i54.full >.

GIBBS, R. A. et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. **Nature,** v. 428, n. 6982, p. 493-521, Apr 1 2004. ISSN 0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature02426 >.

GOLDBERG, A. D.; ALLIS, C. D.; BERNSTEIN, E. Epigenetics: a landscape takes shape. **Cell,** v. 128, n. 4, p. 635-8, Feb 23 2007. ISSN 0092-8674 (Print)0092-8674. Disponível em: < http://dx.doi.org/10.1016/j.cell.2007.02.006 >.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization, and Machine Learning**. Addison-Wesley Professional, 1989. 432 ISBN 0201157675. Disponível em: < http://www.amazon.com/Genetic-Algorithms-Optimization-Machine-Learning/dp/0201157675 >.

GRIFFON, A. et al. Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. 2015-02-27 2015. Disponível em: < http://nar.oxfordjournals.org/content/43/4/e27 >.

GULKO, B. et al. A method for calculating probabilities of fitness consequences for point mutations across the human genome. **Nat Genet,** v. 47, n. 3, p. 276-283, 03//print 2015. ISSN 1061-4036. Disponível em: < http://dx.doi.org/10.1038/ng.3196 >.

HAN, S. X. et al. THYROSIM App for Education and Research Predicts Potential Health Risks of Over-the-Counter Thyroid Supplements. **Thyroid,** v. 26, n. 4, p. 489-98, Apr 2016. ISSN 1050-7256. Disponível em: < http://dx.doi.org/10.1089/thy.2015.0373 >.

HINDORFF, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. **Proc Natl Acad Sci U S A,** v. 106, n. 23, p. 9362-7, Jun 9 2009. ISSN 0027-8424. Disponível em: < http://dx.doi.org/10.1073/pnas.0903103106 >.

HO, K. C. **Simulator of Newborn Thyroid Hormone (TH) Dynamics for Optimizing Treatment of Congenital Hypothyroidism (CH)**. Thesis for the degree Master of Science in Biomedical Engineering, UCLA Library 2013.

HOFFMANN, A. et al. The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. **Science,** v. 298, n. 5596, p. 1241-5, Nov 8 2002. ISSN 0036-8075. Disponível em: < http://dx.doi.org/10.1126/science.1071914 >.

INSTITUTE, V. B.; HEIDELBERG, U. O.; MANCHESTER, U. O. **COPASI Application for simulation and analysis of biochemical networks and their dynamics**.

IONITA-LAZA, I. et al. A spectral approach integrating functional genomic annotations for coding and noncoding variants. **Nat Genet,** v. 48, n. 2, p. 214-220, 02//print 2016. ISSN 1061-4036. Disponível em: < http://dx.doi.org/10.1038/ng.3477 >.

JAENISCH, R.; YOUNG, R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. **Cell,** v. 132, n. 4, p. 567-82, Feb 22 2008. ISSN 0092-8674. Disponível em: < http://dx.doi.org/10.1016/j.cell.2008.01.015 >.

JAEWOOK JOO, S. P., SHAWN MARTIN,LAURA SWILER,JEAN-LOUP FAULON. Sensitivity Analysis of a Computational Model of the IKK-NF-kappaB-IkappaBalpha-A20 Signal Transduction Network. **Annals of the New York Academy of Sciences,** v. 1115, n. 1, 2007. ISSN 1749-6632.

JASINSKA, A. J. et al. A non-human primate system for large-scale genetic studies of complex traits. In: (Ed.). **Hum Mol Genet**, v.21, 2012. p.3307-16. ISBN 0964-6906 (Print)1460-2083 (Electronic).

_____. Identification of brain transcriptional variation reproduced in peripheral blood: an approach for mapping brain expression traits. **Hum Mol Genet,** v. 18, n. 22, p. 4415-27, Nov 15 2009. ISSN 0964-6906. Disponível em: < http://dx.doi.org/10.1093/hmg/ddp397 >.

JAVIER, R. M. A. D. I., J.J. **Dynamical biocontrol systems: insights through mechanistic modeling**: Mathematical

and Computer Modeling of Dynamical Systems. 15**:** 1-16 p. 2009.

JENUWEIN, T.; ALLIS, C. D. Translating the histone code. **Science,** v. 293, n. 5532, p. 1074-80, Aug 10 2001. ISSN 0036-8075 (Print)0036-8075. Disponível em: < http://dx.doi.org/10.1126/science.1063127 >.

JONES, P. A.; BAYLIN, S. B. The epigenomics of cancer. **Cell,** v. 128, n. 4, p. 683-92, Feb 23 2007. ISSN 0092-8674 (Print)0092-8674. Disponível em: < http://dx.doi.org/10.1016/j.cell.2007.01.029 >.

KELLIS, M. et al. Sequencing and comparison of yeast species to identify genes and regulatory elements. **Nature,** v. 423, n. 6937, p. 241-54, May 15 2003. ISSN 0028-0836 (Print)0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature01644 >.

KENT, E.; HOOPS, S.; MENDES, P. Condor-COPASI: high-throughput computing for biochemical networks. **BMC Systems Biology,** v. 6, n. 1, p. 1, 2012-07-26 2012. ISSN 1752-0509. Disponível em: < http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-6-91 >.

KIRCHER, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. **Nat Genet,** v. 46, n. 3, p. 310-315, 03//print 2014. ISSN 1061-4036. Disponível em: < http://dx.doi.org/10.1038/ng.2892 >.

KOH, G.; LEE, D.-Y. Mathematical modeling and sensitivity analysis of the integrated TNFα-mediated apoptotic pathway for identifying key regulators. **Computers in Biology and Medicine,** v. 41, n. 7, p. 512-528, 07/01/2011 2011. ISSN 0010-4825. Disponível em: < http://dl.acm.org/citation.cfm?id=1999501.1999575 >.

KOUZARIDES, T. Chromatin modifications and their function. **Cell,** v. 128, n. 4, p. 693-705, Feb 23 2007. ISSN 0092-8674 (Print)0092-8674. Disponível em: < http://dx.doi.org/10.1016/j.cell.2007.02.005 >.

LAIRD, P. W. The power and the promise of DNA methylation markers. **Nat Rev Cancer,** v. 3, n. 4, p. 253-66, Apr 2003. ISSN 1474-175X (Print)1474-175x. Disponível em: < http://dx.doi.org/10.1038/nrc1045 >.

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature,** v. 409, n. 6822, p. 860-921, Feb 15 2001. ISSN 0028-0836 (Print)0028-0836. Disponível em: < http://dx.doi.org/10.1038/35057062 >.

LEBOFF, M. S. et al. Bioavailability of thyroid hormones from oral replacement preparations. **Metabolism,** v. 31, n. 9, p. 900-5, Sep 1982. ISSN 0026-0495 (Print)0026-0495. Disponível em: < http://dx.doi.org/ >.

LI, B.; CAREY, M.; WORKMAN, J. L. The role of chromatin during transcription. **Cell,** v. 128, n. 4, p. 707-19, Feb 23 2007. ISSN 0092-8674 (Print)0092-8674. Disponível em: < http://dx.doi.org/10.1016/j.cell.2007.01.015 >.

LINDBLAD-TOH, K. et al. A high-resolution map of human evolutionary constraint using 29 mammals. **Nature,** v. 478, n. 7370, p. 476-482, 10/27/print 2011. ISSN 0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature10530 >.

_____. Genome sequence, comparative analysis and haplotype structure of the domestic dog. **Nature,** v. 438, n. 7069, p. 803-19, Dec 8 2005. ISSN 0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature04338 >.

LIPNIACKI, T. et al. Mathematical model of NF-kappaB regulatory module. **J Theor Biol,** v. 228, n. 2, p. 195-215, May 21 2004. ISSN 0022-5193 (Print)0022-5193. Disponível em: < http://dx.doi.org/10.1016/j.jtbi.2004.01.001 >.

LU, Q. et al. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. **Sci Rep,** v. 5, p. 10576, 2015. ISSN 2045-2322. Disponível em: < http://dx.doi.org/10.1038/srep10576 >.

MAIA, A. L. et al. Deiodinases: the balance of thyroid hormone: type 1 iodothyronine deiodinase in human physiology and disease. **J Endocrinol,** v. 209, n. 3, p. 283-97, Jun 2011. ISSN 0022-0795. Disponível em: < http://dx.doi.org/10.1530/joe-10-0481 >.

MARBACH, D. et al. Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. **Nat Meth,** v. 13, n. 4, p. 366-370, 04//print 2016. ISSN 1548-7091. Disponível em: < http://dx.doi.org/10.1038/nmeth.3799 >.

MATHWORKS, T. SimBiology software by The MathWorks. A programmatic tool for computational systems biology and pharmacokinetics., Disponível em: < http://www.mathworks.com/products/simbiology/ >.

MATTICK, J. S.; MAKUNIN, I. V. Non-coding RNA. **Hum Mol Genet,** v. 15 Spec No 1, p. R17-29, Apr 15 2006. ISSN 0964-6906 (Print)0964-6906. Disponível em: < http://dx.doi.org/10.1093/hmg/ddl046 >.

MCLEAN, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. **Nat Biotechnol,** v. 28, n. 5, p. 495-501, May 2010. ISSN 1087-0156. Disponível em: < http://dx.doi.org/10.1038/nbt.1630 >.

MIAO, H. et al. ON IDENTIFIABILITY OF NONLINEAR ODE MODELS AND APPLICATIONS IN VIRAL DYNAMICS. **SIAM Rev Soc Ind Appl Math,** v. 53, n. 1, p. 3-39, Jan 1 2011. ISSN 0036-1445 (Print). Disponível em: < http://dx.doi.org/10.1137/090757009 >.

NEPH, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. **Nature,** v. 489, n. 7414, p. 83-90, 09/06/print 2012. ISSN 0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature11212 >.

Pathways | SABiosciences. 2016. Disponível em: < http://www.sabiosciences.com/pathwaymagazine/pathways8/ppage2.php >.

POLLARD, K. S. et al. Detection of nonneutral substitution rates on mammalian phylogenies. **Genome Res,** v. 20, n. 1, p. 110-21, Jan 2010. ISSN 1088-9051. Disponível em: < http://dx.doi.org/10.1101/gr.097857.109 >.

ROADMAP EPIGENOMICS, C. et al. Integrative analysis of 111 reference human epigenomes. **Nature,** v. 518, n. 7539, p. 317-330, 02/19/print 2015. ISSN 0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature14248 >.

ROSENBLOOM, K. R. et al. The UCSC Genome Browser database: 2015 update. **Nucleic Acids Res,** v. 43, n. Database issue, p. D670-81, Jan 2015. ISSN 0305-1048. Disponível em: < http://dx.doi.org/10.1093/nar/gku1177 >.

ROSOLEM, R. et al. **A**

fully-multiple criteria implementation of the Sobol' Method for parameter

sensitivity analysis. Journal

of Geophysical Research. 117**:** D07103 p. 2012.

SALTELLI, A. **Global Sensitivity Analysis: An Introduction**. Proc. 4th International Conference on Sensitivity Analysis of Model Output (SAMO '04)**:** 27-43 p. 2004.

SALTELLI, A., RATTO, M., ANDRES, T., CAMPOLONGO, F., CARIBONI, J., GATELLI, D., SAISANA, M. AND TARANTOLA, S. **Introduction to Sensitivity Analysis, in Global Sensitivity Analysis. The Primer**. Chichester, UK: Wiley & Sons, Ltd, 2007.

SCHMIDT, H.; JIRSTRAND, M. Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology. **Bioinformatics,** v. 22, n. 4, p. 514-5, Feb 15 2006. ISSN 1367-4803 (Print)1367-4803. Disponível em: < http://dx.doi.org/10.1093/bioinformatics/bti799 >.

SCHRIDER, D. R.; KERN, A. D. Inferring selective constraint from population genomic data suggests recent regulatory turnover in the human brain. 2015-11-19 2015. Disponível em: < http://gbe.oxfordjournals.org/content/early/2015/11/19/gbe.evv228 >.

SHALIZI, C. R. **Advanced Data Analysis from an Elementary Point of View**. Cambridge University Press, 2015.

SIEPEL, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. 2005-08-01 2005. Disponível em: < http://genome.cshlp.org/content/15/8/1034.long >.

SIN, C. **Algorithmic Parameter Space Reduction of a Systems Biology Model: A Case Study**. Thesis for the degree Master of Science in Biomedical Engineering, UCLA Library 2012.

SOBOL, I. M. **Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates**. Mathematics and computers in simulation. 55**:** 271-280 p. 2001.

STARK, A. et al. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. **Nature,** v. 450, n. 7167, p. 219-32, Nov 8 2007. ISSN 0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature06340 >.

Thyroid Diseases. Disponível em: < http://labtestsonline.org/understanding/conditions/thyroid?star=1 >.

WANG, Y.-M. et al. Correlation Between DNase I Hypersensitive Site Distribution and Gene Expression in HeLa S3 Cells. **PLoS ONE**, 2012. Disponível em: < http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0042414 >.

WANG, Y. V. et al. Quantitative analyses reveal the importance of regulated Hdmx degradation for p53 activation. **Proc Natl Acad Sci U S A,** v. 104, n. 30, p. 12365-70, Jul 24 2007. ISSN 0027-8424 (Print)0027-8424. Disponível em: < http://dx.doi.org/10.1073/pnas.0701497104 >.

WATERSTON, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. **Nature,** v. 420, n. 6915, p. 520-62, Dec 5 2002. ISSN 0028-0836 (Print)0028-0836. Disponível em: < http://dx.doi.org/10.1038/nature01262 >.

WEINHOLD, B. Epigenetics: The Science of Change. In: (Ed.). **Environ Health Perspect**, v.114, 2006. p.A160-7. ISBN 0091-6765 (Print)1552-9924 (Electronic).

WERNER, S. L. et al. Encoding NF-κB temporal control in response to TNF: distinct roles for the negative regulators IκBα and A20. **Genes Dev,** v. 22, n. 15, p. 2093-101, Aug 1 2008. ISSN 0890-9369 (Print)1549-5477 (Electronic). Disponível em: < http://dx.doi.org/10.1101/gad.1680708 >.

WILLIAMS, J. B. Adverse effects of thyroid hormones. **Drugs Aging,** v. 11, n. 6, p. 460-9, Dec 1997. ISSN 1170-229X (Print)1170-229x. Disponível em: < http://dx.doi.org/ >.

YUE, H. et al. Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: a case study of an NF-κB signalling pathway. 2006/10/23 2006. Disponível em: < http://pubs.rsc.org/en/content/articlelanding/2006/mb/b609442b#!divAbstract >.

ZI, Z. et al. In silico identification of the key components and steps in IFN-Y induced JAK-STAT

signaling pathway. FEBS

Letters. 579**:** 1101-1108 p. 2005.

_____. SBML-SAT: a systems biology markup language (SBML) based sensitivity analysis tool. **BMC Bioinformatics,** v. 9, n. 1, p. 1, 2008-08-15 2008. ISSN 1471-2105. Disponível em: < http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-342 >.