LOCALIZING FAULTS IN NUMERICAL SOFTWARE USING A VALUE-BASED CAUSAL MODEL

by

ZHUOFU BAI

Submitted in partial fulfillment of the requirements

For the degree of Doctor of Philosophy

Dissertation Advisor: Andy Podgurski

Department of Electrical Engineering and Computer Science

CASE WESTERN RESERVE UNIVERSITY

TBD

Contents

List of Tables

List of Figures

Acknowledgements

This dissertation would be impossible without the support of my advisors. I would like to thank my research advisor Dr. Rong Xu for her constant support and guidance. Dr. Xu led me into the field of translational biomedical research and guided me in every project. She shared innovative ideas with me, and contributed a lot of time to make my Ph.D. experience productive and exciting. Everything would be different without her. I would like to thank my advisor and dissertation committee chair Dr. Guo-qiang Zhang, who offered me the opportunity to join CCI, where I met great people and started my research. I appreciate all his insightful discussions and constructive suggestions to improve my dissertation and the overall research.

My sincere thanks also go to the members of my committee, Dr. Jing Li, Dr. Xiang Zhang and Dr. M. Cenk Cavusoglu, for their invaluable feedback, scientific suggestions and insightful discussions, which helped me improve this dissertation. I would like to give special thanks to Dr. Xiang Zhang, Dr. Xiaofeng Ren and Dr. Li Li, who generously offered advices and shared research experiences with me during my Ph.D. study.

I would like to thank all present and past members of CCI and all my friends in the EECS department for their love and friendship through all these years during my Ph.D. I would like to thank my family: my husband Zhuofu Bai, and my parents, for their unconditional love and support.

List of Abbreviations

- OMIM: Online mendelian inheritance in man
- GWAS: Genome-wide association study
- UMLS: Unified medical language system
- CBIR: Content-based image retrieval
- SIFT: Scale invariant feature transformation
- HOG: Histograms of oriented gradients
- SVM: support vector machine
- HPRD: Human protein reference database
- PPI: Protein-protein interaction
- HDN: Human disease network
- CRC: Colorectal cancer
- PD: Parkinson's disease
- DMN: Disease manifestation network
- IMPC: International mouse phenotyping consortium
- FDA: Food and drug administration

Development of computational approaches for medical image retrieval, disease gene prediction, and drug discovery

Abstract

by

Yang Chen

With the deluge of biomedical data, developing computational approaches for data analysis and interrogation has become a key step in translational biomedical research. It is critical to leverage existing data to ask the right question and design algorithms for specific biomedical applications. In this dissertation, I propose using domain knowledge to guide the data gathering, data fusion and algorithm design in solving specific biomedical problems. I demonstrate the strategy with applications in three distinct contexts.

The first application is retrieving disease manifestation images from the web for supporting patients' self-education and decision making. The challenge is threefold: heterogeneous irrelevant web images need to be filtered; the positive examples of disease images contain diverse objects and complex backgrounds; and large amounts of manual efforts in generating training data are unaffordable. We observe that detecting disease-affected abnormal organs may greatly reduce the manual labeling efforts. In our approach, we extract the disease-organ semantic relationships from ontologies to guide the organ detection with pre-trained detectors. Comparing with a standard supervised method, we improve the average precision by 4% while reduce the manual efforts by 85%.

In the second application, we develop three disease-specific models to detect genetic basis for human diseases. For parasitic infectious diseases, we construct

a cross-species genetic network to model host-pathogen interactions and analyze the network to predict disease associated genes. We apply the approach on malaria and demonstrate the potential of the top-ranked genes in guiding anti-malaria drug discovery. For multifactorial diseases, we assume that phenotypic similarity reflects common genetic basis between diseases. We explore a new disease phenotype data source in medical ontologies and construct the Disease Manifestation Network (DMN). Then we integrate multiple phenotype networks with genetic networks to predict genes. We apply the approach on Crohn's disease, and demonstrate the translational potential of the predicted genes in drug discovery. Last, we identify the mutual comorbidity for colorectal cancer and obesity in the comorbidity network to detect genetic basis for the link between the two diseases.

Finally, I present a drug repositioning approach combining disease genetics and phenotypic descriptions for mouse genetic mutations. Disease associated genes have the potential to guide drug discovery. On the other hand, the mouse phenotypes provide knowledge on gene functions, which is impossible to be obtained in human. In our approach, we identify disease-specific mouse phenotypes using well-studied disease genes, and search all FDA-approved drugs for the candidates that share similar mouse phenotype profiles with the disease. We used the approach to predict drugs for Parkinson's disease, and demonstrate significantly improvements comparing with a state-of-art approach based on mouse phenotype data. Overall, I demonstrate the effectiveness of the domain knowledge guided computational approaches in concrete biomedical applications.

In summary, I demonstrate the effectiveness of computational algorithms in translational biomedical research. I demonstrate that my computation-based work have great potential in elucidating disease genetic basis, finding innovative drugs, and improving patient health education.

Chapter 1

Introduction

1.1 Domain knowledge guided strategy for developing computational approaches for biomedical applications

Biomedicine and healthcare have become data intensive fields [**?**]. Currently, researchers have generated and shared access to vast amounts of genetic, genomic, and phenomic data. With the increase in amount and heterogeneity of biomedical data, developing approaches for data integration, analysis, and interrogation has become a key step to fulfill the translational needs of understanding human diseases, discovering new treatment options and facilitating medically-relevant decision making [**?**, **?**].

One of the major challenges in designing computational approaches for biomedical applications is to ask the right question, gather relevant data and develop algorithms based on a deep understanding of the problem. In this dissertation, I present a domain knowledge guided strategy towards addressing this challenge. I use problem-specific motivations based on domain knowledge to guide the pro-

cess of (1) gathering relevant data from massive amounts of existing biomedical data, (2) connecting heterogenous data, and (3) designing algorithms to discover knowledge from the data (Fig. 1.1).

I demonstrate the effectiveness of the strategy using applications in three distinct contexts: (1) retrieving medically-related images from massive web images based on their contents, (2) detecting genetic basis for human diseases towards genetics-based drug discovery, and (3) repositioning drug treatments based on disease genetics. Among them, the goal of web medical image retrieval is to support patients' self-education and decision-making. Disease-associated gene prediction and drug repositioning are fundamental components of translational biomedical research. The rest of this chapter will describe the background, challenges and the application of the knowledge guided strategy in each context.

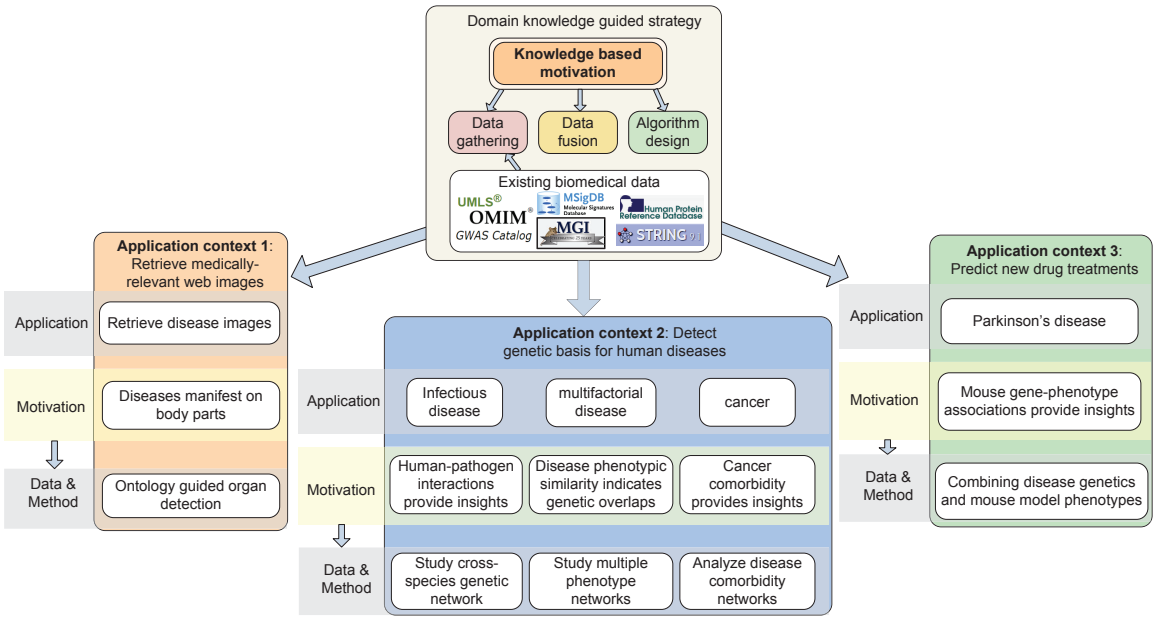

Figure 1.1: Domain knowledge guided strategy in designing approaches for biomedical application. The strategy is applied in three contexts: (1) retrieving medically-related web images, (2) detecting genetic basis for human diseases, and (3) repositioning drug treatments.

1.2 Retrieving medically-relevant web images

Medical knowledge in both textual and visual format is important for health information retrieval and clinical applications. A number of comprehensive textual knowledge bases have been constructed and made available in the medical domain, such as the Unified Medical Language System (UMLS) [**?**]. In comparison, fewer studies have attempted to systematically organize medical knowledge in a visual format. Many medical image bases concentrate on specific domains, such as lung CT images [**?**], cardiovascular MRI images [**?**], and human anatomy images [**?**]. The scale of these databases is limited, largely because the image collection processes are manual and laborious. Also, they annotate images by natural language sentences, which introduce ambiguities in image retrieval. Last but not least, most existing image bases are not freely available.

Our eventual goal is to build a freely accessible, large scale and patient oriented health image base, which contains images of human disease manifestations, organs, drugs and other medical entities. Unlike existing databases, we plan to build up our image base in line with the UMLS structure and annotate images by terms from standard medical ontologies, such as the FMA (Foundational Model of Anatomy) [**?**], ICD9 (International Classification of Diseases, 9th revision) [**?**] and RxNorm [**?**]. For each medical term, we seek to provide a set of high quality images with relevant contents, creating a rich and reusable information resource for patient education, patient selfcare and web-content illustration. Since the image base is designed for consumers, we collect photographic images, which are a significant subset of all biomedical images.

The most challenging problem in building the image base is how to collect a large number of credible images for tens of thousands of medical terms. The web is a readily available source: it is free, it contains billions of images and is fast growing, and search engines such as Google can already do reasonable image re-

trieval based on text queries. However, the web is heterogeneous, and most of the images are non-medical and need to be filtered. Generic image retrieval engines such as Google are not specialized for medical applications. For example, the top Google results for UMLS concepts "heart," "ear deformities, acquired," and "ibuprofen" do not only contain images of the heart organs, ear deformities and ibuprofen tablets, but also include other items such as cartoon symbols, paper snapshots, and molecular formulae. In particular, image retrieval for disease terms is highly challenging, since disease manifestation images contain diverse objects and complex backgrounds. To collect medically-relevant images from the web, we clearly need a content-based image retrieval (CBIR) method, which requires minimal manual effort, as the number of disease terms is large.

This application focuses on developing an automatic approach to retrieving web images on human diseases. Traditional supervised methods need a training image set for each disease, thus will not scale when the number of diseases is large. Our key observation is that although the number of diseases is in the tens of thousands, most disease manifestations are shown on body parts, and the number of body parts is much smaller. I develop an ontology-guided approach to retrieve disease images from the web. In this approach, I extract the knowledge of the affected body parts for a given disease term from UMLS. Then I use this knowledge to guide the selection of pre-trained organ detectors, and combined the organ detection outputs to retrieve disease images.

1.3 Detecting novel genetic basis for human diseases

Identifying genetic basis for human diseases plays an important role in elucidating disease mechanisms and discovering targets of drug treatments [**?**, **?**]. For computational strategies to predict disease genes, mining relevant data for specific disease

types can lead to new discoveries [**?**, **?**, **?**]. Traditional approaches exploited human genomic data and prioritized genes for a disease if the genes are functionally similar to the known disease genes [**?**, **?**, **?**, **?**]. A few recent studies incorporated clinical phenotype data to increase the ability of identifying new disease genes [**?**, **?**, **?**, **?**, **?**, **?**] and assumed that similar disease phenotypes reflect overlapping genetic causes [**?**, **?**, **?**].

In the first application, I develop an approach to predicting genes for parasitic infectious diseases. Traditional disease gene discovery methods that exploit human protein interactome are insufficient for infectious diseases, which naturally involve human-pathogen protein interactions. I hypothesize that the study on human-parasite protein interactions can provide insights into the molecular signatures for disease-specific host immune responses [**?**, **?**, **?**]. I construct a cross-species network to integrate human-human, parasite-parasite and human-parasite protein interactions. Then I use known disease genes as the seeds to find novel candidate disease associated genes. I apply the approach on *Plasmodium falciparum* malaria, which is the most deadly parasitic infectious disease and killed six millions people worldwide in 2012 [**?**]. I demonstrate that the top-ranked candidate genes are not only associated with malaria, but also have the potential to guide genetics-based anti-malaria drug discovery.

In the second application, I develop a phenotype-driven approach to predicting disease-associated genes. For syndromes and many multifactorial diseases, systematically analyzing disease phenotype networks in combination with protein functional interaction networks have great potential in illuminating disease pathophysiological mechanisms [**?**, **?**, **?**]. However, disease phenotype networks remain largely incomplete, and most current disease gene discovery studies used only one data source of human disease phenotypes [**?**, **?**, **?**, **?**, **?**]. Incorporating more comprehensive phenotype data can enhance the performance of disease gene

prediction. Therefore, I explore a new disease phenotype data source–the disease-manifestation semantic relationships in the UMLS, and constructed a Disease Manifestation Network (DMN). I demonstrate through comparative analysis that the phenotype clustering in DMN reflects common disease genetics and contains different knowledge from mimMiner, which is a widely-used phenotype database. Then I develop an innovative and generic strategy to combine DMN, mimMiner, and a genetic network, and predict disease-gene associations from the integrated network. I apply the approach on Crohn's disease and demonstrate that the predicted genes have the translational potential in drug discovery by integrating with drug-target associations.

In the third application, I develop a comorbidity network analysis approach to infer novel genetic basis for the link between two diseases, and apply the approach on colorectal cancer (CRC) and obesity. Phenotype-driven approaches to predicting novel disease genes may not be suitable for cancers, which usually have non-specific disease manifestations, such as pain, fever and ascites. Disease comorbidity often leads to unexpected disease links [**?**] and offers novel insights into disease genetic mechanisms [**?**, **?**]. Specially, studying cancer comorbidity has impacted the understanding of cancer mechanisms [**?**]. The common comorbidity between CRC and obesity in the context of comorbidity network provides insights into the novel molecular evidence underlying both diseases. Traditional comorbidity studies usually focus on pairwise disease links [**?**, **?**, **?**, **?**], and the results are often biased due to noises and intrinsic bias in the patient data. I explore new patient data, which are not biased towards patients of certain ages and genders, and develop a comorbidity mining approach to reduce the bias towards rare diseases. Instead of studying pairwise disease comorbdity, I construct a disease comorbidity network and design a network analysis approach to identify common comorbidity between two diseases. Gene expression analysis guided by the detected common

comorbidity identifies a few genes that have the potential to explain the link between CRC and obesity.

1.4 Predicting novel drug treatments based on disease genetics

Computational drug repositioning approaches lead to rapid drug discovery. Previous studies have predicted new indications for existing drugs by analyzing multiple types of data, such as drug side effects [**?**], drug response gene expressions [**?**], and disease similarities [**?**]. Recent studies demonstrate that disease genetics in genome-wide association studies (GWAS) [**?**] and Online Mendelian Inheritance in Man (OMIM) [**?**] has great potential to guide drug discovery. On the other hand, International Mouse Phenotyping Consortium (IMPC) [**?**] has made available large amounts of phenotypic descriptions for mouse genetic mutations based on systematic gene knockouts, which are impossible on human. The mouse phenotype data enrich the knowledge on disease genetic basis, and has facilitated the detection of new disease genes [**?**] and drug targets [**?**]. Combining human disease genetics and mouse phenotype data will provide novel insights into the genetics of many complex diseases, which can guide the discovery of novel drug options.

In this application, I develop a novel drug repositioning approach leveraging both disease genetics and mouse model phenotypes. I apply the approach on drug repositioning for Parkinson's disease (PD). I first identify PD-specific mouse phenotypes using well-studied human disease genes. Then I search all FDA-approved drugs for candidates that share similar mouse phenotype profiles with PD. I also compare the approach with pure genetics-based approaches and a state-of-art drug repositioning approach based on mouse phenotypes [**?**] to demonstrate the importance of combining these two kinds of data.

1.5 Contribution and organization of the dissertation

In this dissertation, I use five applications to demonstrate that the knowledge guided strategies of combining unique data and novel computational approaches effectively contribute in solving specific medical problems. This dissertation makes the following contributions: the development of a novel ontology-guided approach to retrieving disease manifestation images from the web; the development of three computational approaches to predicting genetic basis for parasitic infectious diseases, multifactorial diseases, and cancers, respectively; the demonstration of the translational potential of predicted disease genes in drug discovery; and development of a novel drug repositioning approach based on disease genetics and mouse model phenotypes.

The remainder of the dissertation is organized as follows:

Chapter **??** presents an ontology-guided image retrieval method for identifying disease web images.

Chapter **??** presents a disease gene prediction approach for malaria based on studying the cross-species genetic networks.

Chapter **??** describes the construction of a novel disease phenotype network and development of a generalizable disease gene prediction approach based on multiple disease phenotype data sources.

Chapter **??** introduces the construction of a disease comorbidity network and development of a novel network analysis approach to detecting genetic basis for the link between colorectal cancer and obesity.

Chapter **??** presents a computational drug repositioning approach combining disease genetics and mouse model phenotypes and its application on Parkinson's disease.

Chapter 2.6 concludes this dissertation and discusses the possible improvements for future work.

Chapter 2

Causal Inference Based Fault Localization for Numerical Software with NUMFL

2.1 Introduction

The benefits of using BART is:

The main contributions of this paper are:

The rest of the chapter is organized as follows:

2.2 BART Model Algorithm

The BART model algorithm consists of three pars: a sum of regression trees model, a regularization prior and a fitting algorithm Marcov Chain Mote Carlo (MCMC)

2.2.1 A Single Regression Tree model

Regression tree is one type of decision tree that predicts the value of a target variable based on several input variables. Regression tree handles the situation when the target variable is continuous. Figure shows a single regression tree model. All the interior nodes of a regression tree have decision rules which send the input data set to either left or right side. After the input data set go through the interior nodes and reach the bottom of the tree, the data set is divided into several disjoint subgroup. Each group of data is represent by a leaf node.

A single regression tree model is denoted as: $y = g(T, R, M) + /epsilon$

Here R denotes a binary regression tree consisting a set of interior node decision rules and a set of terminal nodes, and let $M = \left\{ \mu_1, \mu_2, ..., \mu_b \right\}$ denotes a set of parameter value associated with each of the b terminal nodes of R. Each terminal node represent a regression model of outcome Y on Treatment T

2.2.2 A sum of regression trees model

2.2.3 A regularization prior

2.2.4 Bayesian Backfitting MCMC Algorithm

2.3 TWO VERSIONS OF NUMFL

2.3.1 NUMFL-GPS

Control of Confounding.

Failure-causing Effect Estimation.

2.4 EMPIRICAL EVALUATION

2.4.1 Limitations

2.5 RELATED WORK

2.6 CONCLUSION