Gene-Disease Predictor: Swarm Intelligence Based Biclustering Analysis for Moyamoya Disease in Taiwan

PROJECT SYNOPSIS

Moyamoya syndrome is a rare disease caused by blocked arteries in the brain and was first described in 1957. It primarily affects children, but it can also occur in adults. To the best of our knowledge, this disease tends to run in families and has inherited genetic abnormalities. We believe the studies that look for the abnormal gene (gene sets) may help reveal the bio-mechanisms of gene regulation for genetic disease. Previous studies have proven regional and temporal distinctions in epidemiological features of Moyamoya disease, but population-based studies in regions other than Japan are limited. Moreover there are a variety of available Moyamoya disease databases in Taiwan, but there is no tool that can help researchers analyze the gene-disease correlations and predict the most effective genes accurately, especially for facilities with limited computing sources or lack of bioinformatics processing experience. For this project, we will design and develop a standalone and efficient gene-disease predicting tool that allow general scientists to process large-scale biological data efficiently, and provide high level biclustering methods for discovering groups of correlated genes potentially associated to the disease or conditions under investigation. During this project I will come to better understand the relationships among affected genes with Moyamoya and associated diseases as well as the process of designing large-scale information technology and to create a framework to help Taiwanese researchers facilitate this kind of work. This work is not just limited to Moyamoya disease but also covers other types of genetic diseases. This project will also benefit American researchers to study inherited genetic conditions and help them to determine whether a disorder has a genetic component if we find an individual who carries Moyamoya disease in the United States.

TIME SCHEDULE

Week	Activity
Pre-EAPSI	Literature review on recent findings associated with Moyamoya disease. Prepare training datasets from National Taiwan University Hospital and other datasets provided by Host Advisor. Start algorithm design and implementation in English version of Gene-Disease Predicting (GDP) tool at home institute
1	Opening ceremony and orientation
2	Move to National Taiwan University Self-introduction to the Lab and provide a brief presentation of the project to lab mates. Apply basic statistical analysis on local patient records and identify potential marker genes
3	Conduct swarm intelligence based biclustering analysis on training data Improve the overall performance of the program in terms of time and space complexity using Lab computer facility
4	Prepare testing data and develop baseline method for comparison Develop Chinese version of the GDP tool with Dr. Chuang and lab mates

5	Evaluate SI based biclustering model and generate performance analysis Send results back to the United States for further assessment
6	Create web interface and test the overall performance Generate user manual in both English and Chinese
7	Finalize all source codes and prepare project report Project presentation to lab group and discuss future directions for collaboration
8	Closing ceremony and reprot presentation
Post-EAPSI	Build high-efficient and scalable version of the GDP tool to send back to host laboratory for evaluation and results analysis. Collaborate remotely with host supervisor on academic conference and publications

BACKGROUND

Moyamoya Disease: Moyamoya disease was first reported in Japanese literature in 1957 by Takeuchi and Shimizu [1]. It has since been found in individuals in Africa, Australia, Europe and North America [2]. The name "Moyamoya" in Japanese means "puff of smoke" and describes the look of the tangle of tiny vessels formed to compensate for the blockage. The cause of Moyamoya disease is unknown and there is no effective drug treatment. In the past decades, an annual incidence of 0.15 per 100,000 people appears to be carriers in Taiwan. Although Moyamoya disease is most prevalent in Asia, it has been diagnosed in people throughout the world. Stanford Moyamoya Center has observed that 57% of their patients treated at the center are Caucasian, 31% are Asian, and less than 12% are from other ethnic groups [3]. In USA, the risk of developing Moyamoya is significant increase in female and white, and it is most commonly diagnosed in children and may occur at any time [4].

Gene-Disease Associations: Understanding the role of genetics in human health and disease is one of the most important intentions of the biological sciences. However, the coverage of curated databases that provide information manually extracted from the literature is limited and to determine diseases related genes requires laborious experiments [5]. Therefore, predicting good candidate genes before experimental analysis will save time and effort.

Biclustering and Swarm Intelligence Techniques: Biclustering is a popular data mining method for analyzing gene expression data in Microarray technology. Any type of data can be represented as a matrix is amenable to biclustering [6]. In fact, most formulations of biclustering problems are non-deterministic polynomial-time hard. Hence, it is necessary to adopt an optimization technique for determining the set of potential solutions and defining the criteria which measures the quality of an individual solution. Swarm Intelligence is a robust and flexible optimization technique that mimics the behavior of swarms of social insects, flocks of birds and schools of fish [7]. The swarms follow very simple rules, and there is no centralized control structure. These properties make swarm intelligence a successful design paradigm to deal with increasingly complex problem, such as biclustering problems. Moyamoya disease is one of the 130 rare, intractable diseases at Japan Intractable Diseases information Center [8], to design and implement a scalable gene-disease predicting tool (not just limited to Moyamoya disease but also covers other types of genetic diseases) requires highly efficient methodology and powerful computing sources, such as supercomputer.

RESEARCH GOALS

The primary goals for this project are:

1) To understand the most central genes in an interaction network for Moyamoya disease;
2) To design a novel methodology to accurately predict gene-disease associations;
3) To develop a standalone gene-disease predicting tool for use with Moyamoya patients in Taiwan; and
4) To create a framework to help global researchers facilitate this kind of work in the future.

RESEARCH PLAN AND METHODS

While data mining and machine learning are flexible technologies that can be applied to numerous different tasks, interpreting significant genetic association requires domain knowledge in Biology. Dr. Eric Y. Chuang has been working in the joint field of bioinformatics and biostatistics. His expertise in both can guide me to better understand the high level gene interactions and mutation. In this research we will study the facts and analysis the data collected from National Taiwan University Hospital for Moyamoya disease. Our research plan consists of four major steps:

Data Pre-processing: There are many good databases in the United States for human diseases such as Gene Expression Omnibus (GEO). However, most of the data repositories are related to general human cancer; and Moyamoya patient samples are rarely available to the public. Dr. Chuang will help us to gather the most recent Moyamoya patients' records in Taiwan.

Bio-Statistical Analysis: We will conduct a statistical test for Moyamoya patient gene expression data with standard statistical methods such as t-test, Wilcoxon test and permutation test. Moreover, data size must be large enough to provide sufficient reliable data and help us to carry out significance testing. Hence, the guidance from Dr. Chuang to consolidate Moyamoya disease data is very important.

Swarm Intelligence based Biclustering Model: The traditional clustering is often grouping genes and conditions based on overall similarity. It sometimes experiences information lost during oversimplified similarity and grouping computation [9]. In our approach, we will design a novel biclustering method with swarm intelligence optimization algorithm. Biclustering technique generates a bicluster, subset of genes which exhibit similar behavior across a subset of features. This technique was first generalized and applied to biological gene expression data in 2000 [10]. The complexity of the biclustering problem is NP-complete. In our design, we will utilize the optimization power from Swarm Intelligence technique to achieve optimal gene-disease biclusters. We will also implement a standalone Gene-Disease Predicting (GDP) tool for global researchers to test on other types of genetic disease data. This application will be useful for Dr. Chuang and his lab for future research on genetic disease in Taiwan and benefit American scientists to study gene-disease relationships.

Evaluation and Assessment: To evaluate the approach proposed, we will test our model on different Moyamoya datasets collected in Taiwan and the United States. To illustrate the performance of our methodology, we will compare it to other studies of gene-disease associations and basic biclustering methods [11-13]. Results will be discussed with Dr. Robert Doerksen and Dr. Kuldeep Roy, who are our local collaborators in the Department of BioMolecular Science at the University of Mississippi.

SIGNIFICANCE OF PROPOSED WORK AND CAREER GOAL

My graduate study is focused on gene expression classification, gene mutation affects, and prediction of gene-disease relationships. This research experience will complement my dissertation and future research career by allowing me to gain a better perspective of the significance between genes and conditions and the factors in local genetic diseases. Through this summer research program, I will be more confident to utilize my strong technical skills and educational background. Beyond the benefits of engaging in international research, this project will also be a good way to initiate an academic research relationship between the Bioinformatics and Biostatistics Core Laboratory at the National Taiwan University and the Bioinformatics Group at the University of Mississippi. I will set up a workshop or seminar at the host and local universities to introduce the NSF research opportunity and the knowledge I have learned through this project.

RESEARCH EXPERIENCE

Graduate Research: As a fifth year Ph.D. student, I have already completed my coursework and have gained significant experience related to my proposed project. I have worked in close collaboration with scientists and researchers from the University of Mississippi designing and developing large-scale computational applications related to gene expression classification and gene-gene mutation. As the first and co-author, I have submitted my gene expression classification model and cancer driver genes identification model to the Intelligent Systems for Molecular Biology (ISMB) and the IEEE International Conference on Bioinformatics and Biomedicine (BIBM) conferences, respectively [14-15]. While preparing my publications I also collaborated with scientist from U.S. Army Engineer Research and Development Center in Mississippi to evaluate their regression model using my knowledge in Machine Learning. I am also a participant at the Mississippi EPSCOR/IDeA program funded by NSF for 5 years to develop technology to increase Mississippi's research competitiveness and support education growth.

Undergraduate and International Research: As an undergraduate research assistant I conducted research for two years in Machine Learning, Artificial Intelligence and Robotics. I also interacted and collaborated with researchers in China during the winter of my senior year in college and my second year of graduate school. I engaged in data preprocessing and analysis; this work has been published in the Journal of China Mechanical Engineering in 2013.

HOST RESEARCH TEAM

Selection of Host Institute: National Taiwan University is a national top-tier co-educational research university in Taipei, Taiwan. In particular, the Bioinformatics and Biostatistics Core Laboratory employs 5 faculty members, 5 research associates and many graduate level researchers who work in a variety of specializations including: statistical analysis, microarray data analysis, bioinformatics, computational biology and biochip technology. Their primary aim is to provide research services to all the investigators within NTU medical campus with excellent computer facility environment and clinical consultant services. Collaborating with researchers in this lab will help me understand the high-level explanation of the field and facilitate my future research. Moreover, the accessibility to their computer facility will allow me to build an efficient computational Chinese version of my Gene-Disease Predictor which will ultimately benefit Taiwanese students and scholars.

Biography of Host Advisor: Dr. Erci, Y. Chuang is a Professor in the Medical Engineering Group at the Department of Electrical Engineering and Principal Investigator of Bioinformatics and Biostatistics Core Laboratory at the National Taiwan University. Dr. Chuang received his Doctoral degree from Harvard University in Cancer Biology and a MS degree from Illinois Institute of Technology in 1997 and 1991, respectively. He played an important role in establishing microarray research projects at the U.S. National Cancer Institute (NCI) and National Institutes of Health (NIH). Moreover, Dr. Chuang has published over 100 articles in the leading peer review journals and top conferences. Dr Chuang possess an expertise knowledge and extensive experience in cancer biology, bioinformatics, molecular biology and biochip technologies.

BROADER IMPACT

Integration of Research and Education: This experience will be utilized to establish a relationship and opportunity between Taiwanese and American genetic disease researchers to promote professional development. Such an educational exchange will benefit both Taiwanese and Mississippi students to engage with multiple cultures. After the completion of my doctoral study, I plan to continue my research on genetic disease discovery and pursue an academic career teaching college-level courses involving fundamental knowledge of computational biology and genetic disease, which will encourage young researchers to gain interest in cross-disciplinary science in USA.

Integration Diversity into NSF Programs: Since 2012, I have been involved in the NSF Mississippi EPSCOR program to promote research in Modeling and Simulation of Complex Systems. This program encourages faculty member to collaborate with researchers at other universities and students to pursue graduate education in computer science, biology, chemistry and related field. This program also provide support to underrepresented groups to be successful in STEM. Upon returning to the US, I will prepare an outreach presentation and course session for a Summer Youth Academy program. The University of Mississippi offers students entering the 8th, 9th, and 10th grades an opportunity to experience early college life. The presentation can also be used for outreach efforts within other schools in Mississippi to showcase Taiwanese culture.

Dissemination and Publication Plan: This project will ultimately benefit patients with Moyamoya disease and prevent early childhood death. As I collaborate with researcher from National Taiwan University, it will strengthen the international tie between Taiwan and the United States. This project will allow me to experience Taiwanese culture and education system. I will create a personal blog to share my personal experience and research in Taiwan. The research results of this proposed project will be disseminated through international conferences, such as IEEE BIBM, ISMB, and ACM SigKDD or academic journal publications, such as PLOS ONE or BMC Bioinformatics.

LETTER OF REFERENCE

Dr. Dawn E. Wilkins, my current graduate research advisor. She is a Professor and Chair in the Department of Computer and Information Science at the University of Mississippi.