Statement of Purpose

Tong He

As problem-solving and decision-making becomes ever more data-driven and computing-guided, I am thrilled at the prospect of entering your master program in Computer Science, preferably into the field of Big Data Genomics, an area full of opportunities for researchers to make incredible breakthroughs. After earning my Bachelor's degree from Sun Yat-Sen University in June 2013, I am now more than ready for this challenging and exciting endeavor at Simon Fraser *University*, along with considerable skills coupled with a keen curiosity to learn much more.

Before I entered the statistics program in the School of Mathematics in Sun Yat-Sen University, I participated in the Olympiad in Informatics in high school. I mastered classical algorithms and data structures, as well as coding ability. As I began my freshman and sophomore year in college, I gladly dove into the realm of upper level math and statistical courses such as mathematical analysis, advanced algebra, regression model and Bayesian inference. With strict training on mathematics and abundant programming experiences, I naturally became fond of a course involved Data Mining in my junior year. Although there exists plenty of related packages in R, I implemented all learned algorithms to reinforce my understanding. To meet my curiosity, I also read *Elements of Statistical Learning* and *Pattern Recognition And Machine Learning*. Using these learned algorithms, I gained lots of real world experiences in Data Mining competitions such as those on Kaggle.com.

Driven by my passion, I began to do research projects mainly concerning about Bayes model and microarray data in my junior year, including my undergraduate thesis. Researchers of biological data often encounter with the problem of small samples in the microarray studies, which makes many estimation inaccurate. The goal of my thesis was estimating FDR, so the problem of small samples haunted me as well. To enrich information using data, I borrowed the concept of Empirical Bayes methods from Efron Bradley. After careful observations and attempts, I decided to use hierarchical Bayes model to describe the empirical distribution of data and sample from it, then estimate FDR from the new sample. Optimization of parameters was a common difficulty in Bayes modeling. After works on literature survey, it was solved by Monte Carlo Expectation-Maximization algorithm. Finally, the estimated FDR was accurate in the experiment. When defending my thesis, I clearly and carefully introduced my design and results. Questions raised by the committee was satisfactorily answered as well. Therefore my thesis was rated as the top level.

With the help of my efforts in studying Data Mining, I found the opportunity to cooperate with the *Second Affiliated Hospital, Guangzhou University of Chinese Medicine* in a research program, and I was responsible for the modeling. Our goal was identifying the correlation of genetic polymorphism and adverse events. Medical experiments were expensive, our data only contained

nearly 40 test subjects. Firstly I turned it into a task of variable selection. A problem arose that there were plenty of well-developed models for large data, but only a few for small ones. Besides, medical research also requires the result to be robust. After days of comparison, I used lasso regression and decision tree to select variables from the small data, because their performances are both good with small samples. To make result reliable, I proposed the method which combined the output of two models. In the end my method showed excellent performance and the result is reliable. This paper was accepted by BIBM 2013 in Shanghai.

Along with special interests in research, I was also curious about the experiences of a career in information technology companies. Besides, I believe companies have many advanced demands and strong research ability as well. Therefore after my undergraduate study, I didn't immediately enter into a graduate program. Instead, I got a position in douban.com as an Algorithm Engineer intern. It is a famous SNS website in China with the largest database of Music, Books, Movies, etc. It was my first time analyzing massive datasets. During this half-year period, I had the opportunity to refine the inner parallel system for R, which is named Rpark, based on the famous parallel computing structure Spark. Besides, I also built up the user-gender database, which was generated by the ensemble of three logistic regression models. The biggest challenge was to analyze the folksonomy data generated by our users. After suffering from dirty raw text data, I creatively used association rules to detect synonyms which boosted manual operation. After manually cleansing, I constructed a tree representing levels of abstraction and relation of words in folksonomy. It was added as an important feature in our recommendation system. For instance, we could calculate similarity among books, movies and even musics with structured folksonomy information. This work also earned me a chance to give a talk on the Sixth useR China conference in Shanghai, Nov. 2013.

However, I cannot hold back the desire to learn deeper when realizing I am limited by my education level. And I strongly believe that Biological data is valuable not only because of its medical and environmental value but also due to its complex and noisy nature. The latter is pushing researchers to develop new technics for further improvements. It is precisely experiences and believes like these that have made me more determined and excited to pursue a master program in Big Data Genomics . I have selected Simon Fraser University as my primary choice based on the well-known strength of your department in this area.

In summary, confident of my skills, eager to enhance those skills even further, and committed to improving the common good regardless of any setbacks, I look forward to joining the Computer Science department of Simon Fraser University in Fall 2014.