A COMPUTATIONAL ANALYSIS OF HUMAN GENETIC VARIATION

By

Asif Javed

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Prof. Petros Drineas, Thesis Adviser
Prof. Kristin P. Bennett, Member
Prof. Christopher Bystroff, Member
Dr. Peristera Paschou, Member
Prof. Mohammed J. Zaki, Member

Rensselaer Polytechnic Institute
Troy, New York

December 2008
(For Graduation December 2008)

ABSTRACT

The cost of assaying individuals for SNPs has decreased rapidly over the past few years. This has paved the way for a more statistical and computational analysis of human genetic variations. Fortunately the Linkage Disequilibrium (LD) structure of the genome, whereby neighboring SNPs exhibit varying degrees of correlation, facilitates the analysis. The thesis comprises of three such studies.

In the first study we exploit the LD structure to identify a smaller representative subset of SNPs (known as tagging SNPs) which can then be used to predict the remaining tagged SNPs. We propose greedy derandomized variants of recently developed matrix algorithms to address these issues. We evaluate them on genotypic data from 38 populations and four genomic regions (248 SNPs typed for approximately 2000 individuals). We also evaluate these algorithms on a second dataset consisting of genotypes available from the HapMap database (1336 SNPs for four populations) over the same genomic regions. Furthermore, we test these methods in the setting of a real association study, using a publicly available family dataset. Using a small set of carefully selected tSNPs we achieve very good reconstruction accuracy of "untyped" genotypes for most of the populations studied. Additionally, we demonstrate in a quantitative manner that the chosen tSNPs exhibit substantial transferability, both within and across different geographic regions. Finally, we show that reconstruction can be applied to retrieve significant SNP associations with disease, with important genotyping savings.

The skewed nature of modern genetic datasets, with hundreds of individuals genotyped for millions of SNPs, demands the development of novel algorithms for genome-wide data. In the second study we describe a novel window definition, which divides long genomic datasets into contiguous non-overlapping windows of high linear structure which allows efficient extension of our tSNP selection method to genome-wide datasets. We used the algorithms in conjunction for the analysis of 2.5 million SNPs and four populations from the HapMap database. We show that 10-25% of these SNPs suffice to predict genotypes in the remaining SNPs with

more than 95% accuracy. Replicating two real genome wide disease association studies (GWAS) made publicly available by Coriell institute, we demonstrate that carefully selecting less than half of these SNPs, would lead to the same association results. We also study the portability of our selection and prediction across different geographic regions using 1 million SNPs assayed for 1115 individuals from 11 diverse populations in HapMap phase 3 dataset. We compare the efficiency and accuracy of our approach to results obtained using the popular method implemented in Tagger.

Recombination rate plays a key role in determining the linkage structure within a region. In the third study, we use change in SNP pattern, among extant haplotypes, as evidence of recombinations. Using biological insight this evidence can be used to infer the recombinational history of DNA segments. This history can be represented as phylogenetic networks consisting of both mutational and recombinational events. We study a mathematical model to merge such networks into a single consensus network. Since the problem is NP-complete, we introduce a polynomial time approximate algorithm which reduces the number of new recombinations introduced in the merger within a factor of of the optimal. Furthermore experimental results, computed using the X-chromosome in HapMap, detect both continental and population specific recombinations. A statistical comparison with mutation based analysis reveals further support of the generally accepted 'Out of Africa' hypothesis, and can be viewed as an indirect validation of our approach.