6-29-2007

Discrete Algorithms for Analysis of Genotype Data

Dumitru Brinza

DISCRETE ALGORITHMS FOR ANALYSIS OF GENOTYPE DATA

by

DUMITRU BRINZA

Under the Direction of Alexander Zelikovsky

ABSTRACT

Accessibility of high-throughput genotyping technology makes possible genome-wide association studies for common complex diseases. When dealing with common diseases, it is necessary to search and analyze multiple independent causes resulted from interactions of multiple genes scattered over the entire genome. The optimization formulations for searching disease-associated risk/resistant factors and predicting disease susceptibility for given case-control study have been introduced. Several discrete methods for disease association search exploiting greedy strategy and topological properties of case-control studies have been developed. New disease susceptibility prediction methods based on the developed search methods have been validated on datasets from case-control studies for several common diseases. Our experiments compare favorably the proposed algorithms with the existing association search and susceptibility prediction methods.

INDEX WORDS: algorithm, SNP, genotype, phasing, case-control study, disease association, risk factor, optimization, haplotype, combinatorial methods

DISCRETE ALGORITHMS FOR ANALYSIS OF GENOTYPE DATA

by

DUMITRU BRINZA

A Dissertation Submitted in Partial Fulfillment of the Requirements of the Degree of

Doctor of Philosophy

in the College of Arts and Science

Georgia State University

2007

DISCRETE ALGORITHMS FOR ANALYSIS OF GENOTYPE DATA

by

DUMITRU BRINZA

Major Professor:	Alexander Zelikovsky
Committee:	Yi Pan
	Robert Harrison
	Ion Mandoiu

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Science
Georgia State University
August 2007

ACKNOWLEDGMENTS

Many thanks to my advisor Dr. Alexander Zelikovsky for being my oracle and best friend during my Ph.D life. I am grateful to the Chair Prof. Yi Pan for his constant encouragement and great advises through out my stay at Georgia State University. I would like to give special thanks to Dr. Raj Sunderraman for his invaluable guidance throughout my Ph.D program. I thank Dr. Robert Harrison for sharing his valuable knowledge and reviewing my dissertation work.

Special thanks to Dr. Ion Mandoiu for sharing his knowledge and giving helpful advices. Thanks a lot to Dr. Andrei Perelygin and Dr. Tommaso Dragani for sharing with me their deep biological knowledge and their biological datasets.

Special thanks to my friends Gulsah, Kelly, Stefan, Hae-Jin, and Vanessa for their help and big support in my personal life during my Ph.D program. I also thank Jim, Weidong, Irina, Nisar, Sunsook, Stephen, Bernard, Qiong, Navin, Akshaye, Liang, Diego, Faheem, Mourad for spending their time discussing with me various research problems.

Thanks to Mrs. Tammie, Mrs. Adrienne, Mrs. Venette, and Mrs. Celena for helping me with all my paper work.

Special thanks with all my heart to my girlfriend Diana for being my driving force during the last year of my Ph.D program.

Special thanks to my Father Constantin and my Mother Svetlana who encourage me to go into the Ph.D program and have patiently been my moral support thorough out my education and career.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Recent improvement in accessibility of high-throughput DNA sequencing brought a great deal of attention to disease association and susceptibility studies. Successful genome-wide searches for disease-associated gene variations have been recently reported [140, 144]. However, complex diseases can be caused by combinations of several unlinked gene variations. This work addresses computational challenges of genotype data analysis studies including haplotype inference, informative SNPs se-lection, searching for diseases associated SNPs, and predicting of genotype susceptibility.

Disease association studies analyze genetic variation across cases (diseased individuals) and controls (healthy individuals). The difference between individual DNA sequences occurs at single-base sites, in which more than one allele is observed across population. Such variations are called single nucleotide polymorphisms (SNPs). The number of simultaneously typed SNPs for association and linkage studies is reaching 106 for SNP Mapping Arrays [131]. High density maps of SNPs as well as massive DNA data with large number of individuals and number of SNPs become publicly available [156].

Diploid organisms, like human, have two near-identical copies of each chromosome. Most genotyping techniques (e.g., SNP Mapping Arrays [131]) do not provide separate SNP sequences (*haplotypes*) for each of the two chromosomes. Instead, they provide SNP sequences (*genotypes*) representing mixtures of two haplotypes -- each site is defined by an unordered pair of allele readings, one from each haplotype -- while haplotypes are computationally inferred from genotypes [101].

Given the likely complexity of trait determination, it is widely assumed that the genetic basis (if any) of important traits (e.g., diseases) can be best understood by assessing the association between the occurrence of particular haplotypes and particular traits. Hence, one of the challenges addressed in this work is computational inferring of the haplotypes (Phasing).

1.1 Haplotype Inference

The input to the phasing problem consists of *n* genotype vectors each with *m* coordinates corresponding to SNPs. The phasing problem asks for explaining each genotype with two haplotypes corresponding to chromosomes. In general as well as in common biological setting, there is exponential number of possible haplotype pairs for the same input genotype. Indeed, an individual genotype with *k* heterozygous sites can have 2^{k-1} haplotype pairs that can resolve this genotype. Without additional biological insight, one cannot deduce which of the exponential number of solutions is the most biologically meaningful and, therefore, would serve as a good guess for the real underlying haplotypes.

Several methods have been explored and some are intensively used for this task [6, 7, 12, 14, 20, 18, 19]. None of these methods are presently fully satisfactory, although many give impressively accurate results.

Although the main concern of phasing algorithms is the accuracy of the inferred haplotypes, the emerging volume of collected genotype data brought more attention to the running time of such methods. Indeed, the leading phasing software tool PHASE [95] is getting hard time in competing with generally less accurate but much faster tools such as HAPLOTYPER [41] and GERBIL [87].

In this work, we first explore phasing of genotypes with two SNPs which have ambiguity when the both sites are heterozygous. Then there are two possible phasings (so called *cis*- and *trans*-) and we assume that the true phasing tends to choose the most frequent pair of haplotypes observed in the population sample. Complete haplotypes for a given genotype can be inferred based on the maximum spanning tree of a complete graph with vertices corresponding to heterozygous sites and edge weights given by inferred 2-SNP frequencies. On datasets across 23 chromosomal regions from HapMap [86], proposed (2SNP) method is several orders of magnitude faster than GERBIL and PHASE while matching them in quality measured by the number of correctly phased genotypes, single-site and switching errors.

Frequently, genotype data represent family trios consisting of the two parents and their child since that allows to recover haplotypes with higher confidence. A simple logical analysis allows to substantially decrease uncertainty of phasing. For example, for two SNPs in a trio with parent genotypes f = 22 and m = 02, and the child genotype k = 01, there is a unique feasible phasing of the parents: f_1 = 10, f_2 = 01, m_1 = 01, m_2 = 00 such that the haplotypes f_2 and m_1 are inherited by the child. In fact, it is not difficult to check that logical ambiguity exists only if all three genotypes have 2's in the same SNP site.

Although, there exist many phasing methods for unrelated adults or pedigrees, phasing and missing data recovery for trios is lagging behind. We have tried several well-known computational methods for phasing Daly et al. [138] family trio data, but, surprisingly, all of them give infeasible solutions with high inconsistency rate.

In this work we propose an integer linear programming and greedy methods for solving phasing problem for family trios by finding most Parsimonious solution. We have also enhanced our 2SNP algorithm to phase family trio data. 2SNP trio phasing have been compared with four other well-known phasing methods on simulated data from [90]. 2SNP is much faster than all of them while loosing in quality only to PHASE.

1.2 Disease Association Search

Several challenges in genome-wide association studies of complex diseases have not yet been adequately addressed [137]: interaction between non-linked genes, multiple independent causes, multiple testing adjustment, etc. Since complex common diseases can be caused by multi-loci interactions two-locus analysis can be more powerful than traditional one-by-one SNP association analysis [143]. Multi-loci analysis is expected to find even deeper disease-associated interactions. The computational challenge (as pointed in [137]) is caused by the dimension catastrophe. Indeed, two-SNP interaction analysis (which can be more powerful than traditional one-by-one SNP association analysis [143]) for a genome-wide scan with 1 million SNPs (3 kb coverage) will afford 10^{12} possible pair wise tests. Multi-SNP interaction analysis reveals even deeper disease-associated interactions but is usually computationally infeasible and its statistical significance drastically decreases after multiple testing adjustment [121, 125].

Disease association analysis searches for Risk Factors (RF) modeled as multi-SNP combinations (MSC) (,i.e., subsets of SNP's with specified alleles) with frequency among diseased individuals (cases) considerably higher than among non-diseased individuals (controls). Only statistically significant MSCs (whose frequency distribution has p-value less than 0.05) are reported. Successful as well as unsuccessful searches for SNPs with statistically significant association have been recently reported for different diseases and different suspected human genome regions (see e.g. [151]). Unfortunately, reported findings are frequently not reproducible on different populations. It is believed that this happens because the p-values are unadjusted to multiple testing -- indeed, if the reported SNP is found among 100 SNPs then the probability that the SNP is associated with a disease by mere chance becomes roughly 100 times larger. This work focuses on optimization approach to resolve these issues instead of traditionally used statistical and computational intelligence methods.

In order to handle data with huge number of SNPs, one can extract informative (indexing) SNPs that can be used for (almost) lossless reconstructing of all other SNPs[130]. To avoid information loss, index SNPs are chosen based on how well the other non-index SNPs can be reconstructed. The corresponding **informative SNP selection problem (ISSP)** can be formulated as follows.

Given a sample S of a population P of *individuals* (either haplotypes or genotypes) on *m* SNPs, select positions of *k* ($k < m$) SNPs such that for any individual, one can predict non-selected SNPs from these *k* selected SNPs. The *Multiple Linear Regression based* MLR-tagging algorithm [112] solves the optimization version of ISSP which asks for *k* informative SNPs *minimizing the prediction error* measured by the number of incorrectly predicted SNPs. The number of tags (informative SNPs) *k* depends on the desirable data size. More tags will keep more genotype information while less tags allows deeper analysis and search. In the reduced set of SNPs one can search for deeper disease association.

We next discuss the optimization **problem of finding the most disease-associated risk factor (RF)** for given case-control data. Since it is plausible that common diseases can have also genetic resistance factors, one can also search for *the most disease-resistant risk factor*. Association of risk or resistance factors with the disease can be measured in terms of p-value of the skew in case and control frequencies, risk rates or odds rates. Here we concentrate on three association measurements: p-value of the skew in case and control frequencies, *odds ratio* (OR) of corresponding risk factor, and *positive predictive value* (PPV) which is the frequency of case individuals among all individuals with a given multi-SNP combination.

This optimization problem is NP-hard and can be viewed as a generalization of the maximum independent set problem. Therefore, we have applied our exhaustive, combinatorial and complimentary greedy search heuristics [132, 133], alternating combinatorial, and randomized complimentary greedy searches to solve this problem. Although complimentary greedy search and randomized complimentary greedy search cannot guarantee finding of close to optimum RFs, in the experiments with real data, they find RFs with non-trivially high OR and PPV. For example, for Crohn's disease data [138], complimentary greedy search finds in less than second a case-free (disease-resistant) RF containing 24 controls, while exhaustive and combinatorial searches need more than 1 day to find case-free RFs with at most 17 controls.

We next model *atomic* risk factors (ARF's) (i.e., SNP risk factors that cannot be split into simpler factors) for common diseases as a *multi-SNP combination* (MSC), i.e., subsets of SNP's with specified alleles. Our first optimization formulations ask for a MSC the most tightly associated with the disease (i.e., minimizing p-value) [132] and having the highest positive predictive value which is the case frequency among exposed to ARF individuals [133]. In contrast, epidemiologists measure the quality of risk factors in case-control studies by *odds ratio* defined as the ratio of the odds of disease occurring in the exposed (to the risk factor) group to the odds of it occurring in unexposed group. Thus, we ask for an ARF with the maximum odds ratio. We show connection of this problem with the known Red-Blue Set Cover problem [166] and apply the complementary greedy search (CGS) algorithm [133].
We next propose to consider more complex but also more relevant SNP risk factors, so called *k*-relaxed atomic risk factors, for which exposed individuals can deviate in at most *k* sites from a given MSC. We generalize the complimentary greedy algorithm (*k*-CGS) to find *k*-relaxed

atomic risk factors with fixed *k*. Our experiments show advantage of the new method over CGS in finding risk factors with significantly higher odds ratio and association with the disease.

We then introduce even more general risk factors, so called weighted relaxed atomic risk factors. The individuals exposed to such factors should be within *weighted* Hamming distance *k* from a given MSC. We also proposed a novel heuristic (WCGS) for finding weighted relaxed ARF with odds ratio.

We have applied and cross-validated CGS [133], *k*-CGS, and WCGS methods for finding atomic risk factors (ARF), *k*-relaxed ARF's and weighted relaxed ARF's, respectively, with large odds ratios on real case-control studies for several diseases (Crohn's disease [138], autoimmune disorder [163], tick-born encephalitis [134], lung cancer [144], and rheumatoid arthritis [135]). New proposed methods found SNP risk factors that are statistically significant even after multiple-testing adjustment on all data while CGS could not find significant ARF on two of these data. The found relaxed atomic risk factors explain much higher number of cases, 1.5-4 times larger than atomic ARFs found by CGS.

The next challenge commonly facing disease-association studies [136] is finding *reproducible* associations. This challenge have been traditionally addressed in statistics [143] while here we apply computational approaches -- permutation tests and cross-validation. To measure and compare the significance of found risk factors we use traditional permutation tests. To measure and compare ability of search methods to find reproducible risk factors, we propose to apply cross-validation scheme usually used for prediction validation.

We have applied our search methods to real case-control studies for several diseases (Crohn's disease [138], autoimmune disorder [163], tick-born encephalitis [134], lung cancer [144], and rheumatoid arthritis [135]). Proposed methods are compared favorably to the exhaustive search -- they are faster, find more frequently statistically significant risk factors and have significantly higher leave-half-out cross-validation rate.

1.3 Disease Susceptibility Prediction

Some complex diseases, such as psychiatric disorders, are characterized by a non mendelian, multifactorial genetic contribution with a number of susceptible genes interacting with each other [161, 147]. In general, a single SNP or gene may be impossible to associate because a disease may be caused by completely different modifications of alternative pathways. Furthermore, there are no reliable tools applicable to large genome ranges that could rule out or confirm association with a disease. It is even difficult to decide if a particular disease is genetic, e.g., the nature of Crohn's disease has been disputed [146]. Although answers to above questions may not explicitly help to find specific disease-associated SNPs, they may be critical for disease prevention. Indeed, knowing that an individual is (or is not) susceptible to (or belong to a risk group for) a certain disease will allow greatly reduce the cost of screening and preventive measures or even help to completely avoid disease development, e.g., by changing a diet.

This study is devoted to the problem of assessing accumulated information targeting to predict genotype susceptibility to complex diseases with significantly high accuracy and statistical power. We first formulate **disease susceptibility prediction problem** (see [142, 158, 159, 162, 164]) and describe several universal classifiers and discrete optimization based algorithms for prediction disease susceptibility. We then compare leave-one- and leave-many-out tests demonstrating that prediction accuracy of suggested methods is sufficiently resilient to discarding case-control data implying that leave-one-out test is a trustworthy accuracy measure. The permutation tests have been used for computing the statistical significance level of proposed methods and resulted prediction weights.

The proposed methods have been favorably compared with well known universal classifiers ,e.g., SVM, Random Forest, etc., on two publicly available datasets of: Crohn's disease [153] and autoimmune disorder [163]. In the leave-one-out cross-validation tests the proposed linear programming (LP) based method achieves prediction rate of 69.5%(p-value below 2%) and 61.3%(p-value below 62%) and the risk rates of 2.23 and 0.98, respectively.

The last problem addressed in this work is the **disease susceptibility prediction problem** exploiting the developed methods for searching associated risk and resistance factors. A novel optimum clustering problem formulation has been proposed [133]. Also a model-fitting method has been suggested which transforms a clustering algorithm into the corresponding model-fitting susceptibility prediction algorithm. Since common diseases can be caused by multiple independent and coexisting factors, an association-based clustering of case/control population has been proposed [133]. The resulted association-based combinatorial prediction algorithm significantly outperforms existing prediction methods. For all three real data sets that were available to us (Crohn's disease [138], autoimmune disorder [163], and tick-borne encephalitis [100]) the accuracy of the prediction based on combinatorial search is respectively, 76%, 74%, and 80%, which is higher by 7% compared to the accuracy of all previously proposed methods implemented in [142, 118]. The accuracy of the prediction based on complimentary greedy search almost matches the best accuracy but is much more scalable.

1.4 Overview

Chapter 2 introduces SNPs, haplotypes, genotypes, and formally describes the phasing of complete unrelated genotypes and family trios. Several previously know methods are presented and motivation for designing new more accurate and faster methods is given. Here we introduce optimization formulation for family trios phasing problem and propose integer linear programming and greedy methods for solving it. We also describe our accurate and scalable combinatorial method for genotype phasing (2SNP). Each of the proposed methods is favorably compared with existing methods.

Chapter 3 introduces case-control study and gives motivation behind disease-association search (DAS). The DAS problem is formally defined and several previously known approaches are presented. The motivation for designing new robust methods for DAS which can be applied to diseases caused by multiple genes is given. The DAS problem is formulated as an optimization problem for finding the maximum control-free cluster corresponded or maximum odds ratio atomic risk factor and the fast greedy heuristic to solve this problem is proposed. A complex risk factor (RF) is proposed to model as close (non-weighted or weighted) match to a MSC (e.g., no more than *k* mismatches) *k*-relaxed atomic risk factor or weighted relaxed atomic risk factor and the optimization formulation asks for RF with the maximum odds ratio. For measuring and comparing the significance of found risk factors we describe traditional permutation tests. Then we describe proposed cross-validation scheme for measuring and comparing ability of search methods to find reproducible risk factors. Next we present our extensive experimental studies which consist of searching for disease-associated risk factor in several real case-control study data. We also present comparison of searching methods by ability to find disease-associated risk factors and reproducibility of these factors.

Chapter 4 introduces the problem of disease-susceptibility prediction (DSP) and motivation behind it. Then two general approaches for DSP are described -- universal classifier approaches and disease association-based combinatorial approaches. Next we present several universal classifier methods and our new LP-based prediction method for DSP. Also in this chapter we formulate optimum disease clustering problem and describe our disease-association based model-fitting prediction method for DSP. We finalize this chapter with experimental study of existing and proposed methods on several real case-control study datasets.

Chapter 5 describes our software packages: TrioPhasing -- the code which implements the integer linear programming method for trio phasing; 2SNP -- the code which implements our fast, scalable and accurate 2SNP phasing algorithm; DACS -- software which implements several algorithms for solving DAS and DSP problems; GeneSuscept -- software which implements the LP-based algorithm for DSP problem.

Chapter 6 describes our future work followed by related publications and bibliography.

CHAPTER 2

POPULATION-BASED PHASING AND MISSING DATA RECOVERY

A "haplotype" is a DNA sequence that has been inherited from one parent. Each person possesses two haplotypes for most regions of the genome. The most common type of variation among haplotypes possessed by individuals in a population is the single nucleotide polymorphism (SNP), in which different nucleotides (alleles) are present at a given site (locus). Almost always, there are only two alleles at a SNP site among the individuals in a population. Given the likely complexity of trait determination, it is widely assumed that the genetic basis (if any) of important traits (e.g., diseases) can be best understood by assessing the association between the occurrence of particular haplotypes and particular traits. Hence, one of the current priorities in human genomics is the development of a full *Haplotype Map* of the human genome [1, 46, 47, 17], to be used in large-scale screens of populations [16, 53]. In this endeavor, a key problem is to infer haplotype pairs and/or haplotype frequencies from genotype data, since collecting haplotype data is generally more difficult than collecting genotype data. Here, we review the haplotype inference problem (inferring pairs and inferring frequencies), the major combinatorial and statistical methods proposed to solve these two problems, and the genetic models that underlie these methods.

2.1 Introduction to Variation, SNPs, Genotypes, and Haplotypes

Now that high-throughput genomic technologies are available, the dream of assessing DNA sequence variation at the population level is becoming a reality. The processes of natural selection, mutation, recombination, gene-conversion, genome rearrangements, lateral gene transfer, admixture of populations, and random drift have mixed and remixed alleles at many loci so as to create the large variety of genotypes found in many populations. The challenge is to find those genotypes that have significant and biologically meaningful associations with important traits of interest. A key technological and computational part of this challenge is to infer "haplotype information" from "genotype information". In this section, we explain the basic biological and computational background for this "genotype to haplotype" problem.

In many diploid organisms (such as humans) there are two (not completely identical) "copies" of almost all chromosomes. Sequence data from a single copy is called a haplotype, while a description of the conflated (mixed) data on the two copies is called a genotype. When assessing the genetic contribution to a trait, it may often be much more informative to have haplotype data than to have only genotype data. The underlying data that form a haplotype are either the full DNA sequence in the region, the number of repeats at microsatellite markers, or more commonly the *single nucleotide polymorphisms* (SNPs) in that region. A SNP is a single nucleotide site where more than one (usually two) nucleotides occur with a population frequency above some threshold (often around 5-10%). The SNP-based approach is the dominant one, and high-density SNP maps have been constructed across the human genome with a density of about one SNP per thousand nucleotides [47, 17].

2.1.1 The Biological Problem

In general, it is not easy to examine the two copies of a chromosome separately, and genotype data rather than haplotype data are obtained, although it is the haplotype data that may be of greater use. The data set typically consists of *n* genotype vectors, each of length *m*, where each value in the vector is either 0, 1, or 2. The variable *n* denotes the number of individuals in the sample, and *m* denotes the number of SNP sites for which one has data. Each site in the genotype vector has a value of 0 (respectively 1) if the associated site on the chromosome has state 0 (respectively 1) on both copies (it is a *homozygous* site); it has a value of 2 otherwise (the chromosome site is *heterozygous*). The goal is to extract haplotype information from the given genotype information.

A variety of methods have been developed and used to do this (e.g., [14, 15, 28, 36, 42, 58, 61, 63, 66, 69]). Some of these methods give very accurate results in some circumstances, particularly when identifying common haplotypes in a population. However, research on haplotype inference continues because no single method is considered fully adequate in all applications, the task of identifying rare haplotypes remains difficult, and the overall accuracy of present methods has not been resolved.

2.1.2 The Computational Problems

The *haplotype inference* (HI) problem can be abstractly posed as follows. Given a set of *n* genotype vectors, a *solution* to the HI problem is a set of *n* pairs of binary vectors, one pair for each genotype vector. For any genotype vector *g*, the associated binary vectors $v_1; v_2$ must both have value 0 (or 1) at any position where *g* has value 0 (or 1); but for any position where *g* has value 2, exactly one of $v_1; v_2$ must have value 0, while the other has value 1.

A site in *g* is considered "resolved" if it contains 0 or 1, and "ambiguous" if it contains a 2. If a vector *g* has zero ambiguous positions, it is called "resolved" or "unambiguous"; otherwise it is called "ambiguous". One can also say that the *conflation* of *v1* and *v2* produces the genotype vector *g*, which will be ambiguous unless *v1* and *v2* are identical.

For an individual with h heterozygous sites there are 2^{h-1} possible haplotype pairs that could underlie its genotype. For example, if the observed genotype g is 0212, then one possible pair of vectors is 0110, 0011, while the other is 0111, 0010. Of course, we want to infer the pair that gave rise to the genotype of each of the n individuals.

A related problem is to estimate the frequency of the haplotypes in the sample. We call this the HF problem. It is important to note that a solution to the HI problem necessarily solves the HF problem, but the converse is not true.

2.1.3 The Need for a Genetic Model

Non-experimental haplotype inference (the HI and HF problems) would likely be inaccurate without the use of some genetic model of haplotype evolution to guide an algorithm in constructing a solution. The choice of the underlying genetic model can influence the type of algorithm used to solve the associated inference problem.

2.1.4 Family Trio Phasing

Frequently, genotype data represent family trios consisting of the two parents and their child since that allows to recover haplotypes with higher confidence. A simple logical analysis allows to substantially decrease uncertainty of phasing. For example, for two SNPs in a trio with parent genotypes f = 22 and m = 02, and the child genotype k = 01, there is a unique feasible phasing of the parents: f_1 = 10, f_2 = 01, m_1 = 01, m_2 = 00 such that the haplotypes f_2 and m_1 are inherited by the child. In fact, it is not difficult to check that logical ambiguity exists only if all three genotypes have 2's in the same SNP site.

Although there exist many phasing methods for unrelated adults or pedigrees, phasing and missing data recovery for trios is lagging behind. We have tried several well known computational methods for phasing Daly et al. [138] family trio data, but, surprisingly, all of them give infeasible solutions with high inconsistency rate.

In this section we deal with problems of phasing and missing data recovery on family trios data. Formally, given a set of genotypes partitioned into family trios, the Trio Phasing Problem (TPP) requires to find for each trio a quartet of parent haplotypes which agree with all three genotypes. Trio Missing Data Recovery Problem (TMDRP) asks for the SNP values missed in given genotype data.

2.2 Previous Work

There are two major approaches to solving the inference problem: combinatorial methods and statistical methods. Combinatorial methods often state an explicit objective function that one tries to optimize in order to obtain a solution to the inference problem. Statistical methods are usually based on an explicit model of haplotype evolution; the inference problem is then cast as a maximum-likelihood or a Bayesian inference problem. In the section 2.2.1 we discuss combinatorial approaches, and statistical approaches are discussed in section 2.2.2.

2.2.1 Combinatorial approaches

2.2.1.1 Maximum Parsimony

The earliest algorithm for haplotype reconstruction (from genotype data) was described by Clark [1990], based on the principle of maximum parsimony. This algorithm resolves the haplotypes following three steps:

1) identifying all unambiguous haplotypes (all homozygotes and single-site

heterozygotes) and considering them as "resolved;"

2) determining whether each of the resolved haplotypes could be one of the alleles in the remaining yet-to-be-phased genotypes;

3) each time a new haplotype is identified as one of the resolved ones, this new haplotype is assumed to be known, and the remaining haplotype is added to the resolved haplotype set.

The rationale for this algorithm is that homozygous haplotypes are probably common, and that a phase-ambiguous genotype is likely to contain known common haplotypes. Clark [1990] stated that when all haplotypes are resolved based on maximum parsimony, the solution is unique and correct, and the results of Clarks algorithm based on certain empirical data are shown to be reliable by comparison with haplotypes obtained by direct molecular methods [Clark et al., 1998; Rieder et al., 1999].

Gusfield [2001] reformulated Clarks statement into a "maximum resolution" **(MR) problem**: given a set of vectors (some ambiguous and some resolved), what is the maximum number of ambiguous vectors that can be resolved by successive application of Clarks inference rule? Gusfield [2001] proved that the MR problem is NP-hard and Max-SNP-complete, which can be reduced to an integer linear programming problem. The advantages of Clarks algorithm are that it is a relatively straightforward procedure, and it can handle a large number of loci when haplotype diversity is rather limited in the population. The disadvantages of Clarks algorithm are that:

1) the algorithm does not start when there are no homozygotes or single-site heterozygotes in the population;

2) the algorithm does not give unique solutions, because the phasing results are

dependent on the order of genotypes that need to be phased (therefore, when there is a large number of distinct haplotypes compared to the sample size due to the presence of recombination hotspots, Clarks algorithm sometimes cannot resolve a relatively large fraction of heterozygous individuals);

3) although Clarks algorithm does not explicitly assume Hardy-Weinberg equilibrium (HWE), its performance is still relatively sensitive to the extent of deviation from HWE [Niu et al., 2002].

2.2.1.2 Perfect and Near-Perfect Phylogeny

A phylogeny-based algorithm intends to deterministically deduce haplotype phases based on phylogenetic reconstruction [Gusfield, 2002]. The coalescent model of haplotype evolution tells us that without recombination, the evolutionary history for *w* distinct haplotypes can be displayed as a "Perfect PHylogeny" (PPH) with *w* leaves, and each of the SNP sites labels exactly one edge of the tree. Rooted in the coalescent model, there are two key assumptions of PPH: 1) no recombination; and 2) the infinite-sites model. Under these two assumptions, a PPH problem is stated as:

PPH: *Given a set S of n genotype vectors, we would like to find a PPH T (S), and a pairing of the 2n leaves of T (S) that explains S .*

A PPH problem can be reduced to a classical problem of recognizing graphic matroids [Tutte, 1960]. Although the general PPH problem is NP-hard [Steel, 1992] and multiple solutions can be possible [Gusfield, 2002], for binary phylogenies, PPH is shown to be linear-time solvable [Bixby and Wagner, 1988]. A program named "Perfect Phylogeny Haplotyper" [Chung and Gusfield, 2003] was developed. More recently, Halperin and Eskin [2004] developed an "near-perfect" phylogeny method that extends the framework of

PPH by allowing for both recurrent mutations and recombinations. In their approach, the multiple linked SNPs are partitioned into blocks, and for each block, they predict each individual's haplotype. Then the block-based haplotypes are assembled together to form the long-range haplotype utilizing the PL idea of Niu et al. [2002]. The imperfect phylogeny method appears to be more robust than PPH, and allows the handling of missing data and resolution of a large number of SNPs. Overall, PPH appears to be an interesting application of the graphical model in haplotype inference, although their accuracies remain to be benchmarked by using extensive simulations and real datasets.

2.2.2 Statistical approaches

2.2.2.1 Expectation-Maximization (EM)

The expectation-maximization (EM) algorithm [Dempster et al., 1977] estimates population haplotype probabilities based on maximum likelihood, finding the values of the haplotype probabilities which optimize the probability of the observed data, based on the assumption of HWE. Excoffier and Slatkin [1995] were the first to discuss the use of the EM algorithm in this context. The likelihood function in the EM algorithm can be written as

$$L(\Theta) = P(G \mid \Theta) = \prod_{i=1}^{n} \sum_{(a,b):a\otimes b=g_i} \theta_a \theta_b \qquad (2.1)$$

where G denotes the observed unphased genotype data for n individuals, g_i denotes the observed unphased genotype data for the *i*-th individual, Θ denotes the overall haplotype frequency, θ_a and θ_b denote the respective haplotype frequencies for haplotypes *a* and *b*, such that $a \otimes b = g_i$ denotes that the haplotype pair (*a, b*) is compatible with the *i*-th observed genotype data-g_i. The EM algorithm, under the assumption of HWE, is an iterative procedure: $\theta_a^{(k+1)} = E_{\Theta(k)}(n_a \mid G)/2n$, where $\Theta^{(k)}$ is the current estimate of haplotype frequencies, and n_a is the count of haplotype a that exists in *G*. It can be important that the initial values of haplotype frequencies are reasonably close to the true population frequencies. The advantages of the EM algorithm are that: 1) it is based on solid statistical theory; and 2) although the EM algorithm makes an explicit assumption of HWE, simulation studies demonstrate that its performance is not strongly affected by the departures from HWE, particularly when the direction of departure is towards an excess of homozygosity [Niu et al., 2002]. The disadvantages are that: 1) the performance is sensitive to the initial value of Θ; 2) if there exist local maxima, the iteration may lead to locally optimal maximum likelihood estimates (MLEs), which becomes most serious when there are many distinct haplotypes (one sensible way to employ the EM algorithm is to use a good initial guess on Θ [e.g., the product of the allele frequencies, as suggested by Excoffier and Slatkin, 1995]); and 3) the standard EM algorithm cannot handle a large number of loci. Recently, a variant on the EM algorithm, the stochastic-EM algorithm, was applied to the problem of estimation haplotypes from unphased genotypes, by Tregouet et al. [2004]. This algorithm can be useful for avoiding convergence to local maxima.

2.2.2.2 Coalescent-Based Algorithm

Stephens et al. [2001] proposed a coalescence based Markov-chain Monte Carlo (MCMC) approach: a pseudo-Gibbs sampler (PGS) for reconstructing haplotypes from genotype data. PGS uses Gibbs sampling to obtain an approximate sample from the posterior distribution, $P\ (Z|G)$, where $Z = (z_1, z_2, \ldots, z_n)$ and $G = (g_1, g_2, \ldots, g_n)$ denote the phased and unphased (i.e., observed) genotype data for n individuals. The major piece of this iterative sampling algorithm is that at the (k + 1)-th iteration, we aim to sample $z_i^{(k+1)}$ from $P(z_i \mid G, Z_{-i}^k)$, where Z_{-i}^k is a set of phased genotypes for all the remaining (n - 1) subjects excluding the haplotype pair z_i for individual i, at the k-th iteration. Then, we have $P(Z_i \mid G, Z_{-i}^k) \propto P(z_i = \alpha \otimes \beta \mid Z_{-i}^k) \quad \propto \pi(\alpha \mid Z_{-i})\pi(\beta \mid Z_{-i}, \alpha)$, where α and β denote the two respective haplotypes that form z_i : Here, $\pi(\alpha \mid Z_{-i})$ is essentially a prior for a future sampled haplotype, which is not known to the investigators a priori. Instead of using the Dirichlet prior [based on the "parent-independent mutation" model in Stephens et al., 2001], Stephens et al. [2001] suggested using an approximate coalescent prior: equation 17 of Stephens and Donnelly [2000], which is a stationary Markov chain with transition matrix

$$T_{ah} = \frac{\theta}{r+\theta} P_{ah} + \frac{r_\alpha}{r+\theta}$$

, where r and r_α are the total number of haplotypes and the total number of type α haplotypes in Z, respectively, θ is the scale mutation rate, and P is the transition matrix. The approximate coalescent prior is based on the assumption that "the genetic sequence of a mutant offspring will differ only slightly from the progenitor sequence (often by a single-base change)" [Stephens and Donnelly, 2003]. Recently, Stephens and Donnelly [2003] modified the implementation of the PGS algorithm by incorporating a variant of the partitionligation (PL) idea [Niu et al., 2002] and by allowing for recombination and decay of linkage disequilibrium (LD) with distance. The key advantage of the PGS is that it incorporates the coalescence theory into its prior, and although the induced Markov chain has a stationary distribution that may depend on the order of g_is, it was shown to perform well in simulations based on a coalescent model: a constant-size population evolving for a long period of time without recombination or recurrent mutations [Stephens et al., 2001; Stephens and Donnelly, 2003]. The disadvantages are that:

1. PGS is not a fully Bayesian model and it lacks a measure of the overall "good ness" of the constructed haplotypes;

2. because this algorithm makes only local moves in each iteration (i.e., a "piece-by-piece" strategy) to update a new haplotype that closely resembles an existing haplotype, PGS is quite slow and it takes millions of iterations for the algorithm (2 million iterations are suggested as the default value for PHASE version 1.0 in Stephens and Donnelly [2003]) to start to converge to the right answer [e.g., the ACE data from Rieder et al., 1999];

3. it remains unclear whether the algorithm performs favorably compared to other algorithms (e.g., standard EM algorithm) for admixed or rapidly expanding populations when the coalescent model does not hold, which is often the case for cosmopolitan US cohorts (e.g., a random sample from downtown Los Angeles).

For example, although Stephens et al. [2001] demonstrated that PHASE outperformed EM by a significant margin under certain conditions, Zhang et al. [2001] and Xu et al. [2002] revealed that this was not the case: PHASE and EM-based methods exhibited similar performances in their simulated datasets.

2.2.3 Family Trio Phasing Approaches

In this section we overview previous research and on phasing based on statistical methods (Phamily and PHASE, HAPLOTYPER, HAP and greedy algorithms). The ILP based approaches will be discussed in the next section.

Stephens et al. [95] introduced a Bayesian statistical method PHASE for phasing genotype data. It exploits ideas from population genetics and coalescent theory that make phased haplotypes to be expected in natural populations. It also estimates the uncertainty associated with each phasing. The software can deal with SNP in any combination, any size of population and missing data are allowed. The drawback of this method is that it takes long time for large population

Acherman et al. [27] described the tool Phamily for phasing the trio families based on well-known phasing tool PHASE [95]. It first uses the logical method described above to infer the SNPs in the parental haplotypes. Then children genotypes are discarded while the parental genotypes and known haplotypes are passed to PHASE. Because the children genotypes are discarded, PHASE no longer can maintain parent-child trio constraints resulting in 8.02% error rate for phasing Daly et al [138] data.

Niu et al [41] proposed a new Monte Carlo approach HAPLOTYPER for phasing genotype data. It first partition the whole haplotype into smaller segments then use the Gibbs sampler both to construct the partial haplotypes of each segment and to assemble all the segments together. This method can accurately and rapidly infer haplotypes for a large number of linked SNPs. The drawback of HAPLOTYPER is that it can not handle lengthy genotype with large population. It limits 100 SNPs and 500 population.

Halperin et al. [35] used the greedy method for phasing and missing data recovery. For each trio the author introduce four partially resolved haplotypes with the coordinates 0, 1 and ?. The values of 0 and 1 correspond to fully resolved SNPs which can be found via logical resolution from the section 2.4.1, while the ?'s corresponds to ambiguous and missing positions. The greedy algorithm iteratively finds the complete haplotype which covers the maximum possible number of partial haplotypes, removes this set of resolved partial haplotypes and continues in that manner. The authors replace each genotype in Daly et al [138] data with a pair of logically partial resolved haplotypes referring to each ambiguous SNP value as a ?. The ?'s constitute 16% of all data. Then extra 10% of data are erased (i.e., replaced with ?'s) and the resulted 26% of ambiguous SNP values are inferred by the greedy algorithm minimizing haplotype variability within blocks. When measured on the additionally erased 10% of data, the error rate for the greedy algorithm is 2.8% [35] which has been independently confirmed in our computational experiments. Unfortunately, the error rate o for the original 16% of ?'s is at least 25% which has been measured by the number of inconsistently phased SNPs. This may lead to a conclusion that the complexity of missing genotype data is considerably higher than the complexity of the successfully genotyped data.

2.3 2SNP: Scalable Phasing Method for Unrelated Individuals

Emerging microarray technologies allow affordable typing of very long genome sequences. A key challenge in analyzing of such huge amount of data is scalable and accurate computational inferring of haplotypes (i.e., splitting of each genotype into a pair of corresponding haplotypes). In this work, we propose a phasing method which first phase genotypes consisting only of two SNPs using genotypes frequencies adjusted to the random mating model and then extend phasing of two-SNP genotypes to phasing of complete genotypes using maximum spanning trees. Runtime of the proposed 2SNP algorithm is $O(nm(n + log\ m)$, where n and m are the numbers of genotypes and SNPs, respectively, and it can handle genotypes spanning entire chromosomes in a matter of hours.

On datasets across 23 chromosomal regions from HapMap [86], 2SNP is several orders of magnitude faster than GERBIL(combinatorial method) and PHASE(statistical method) while matching them in quality measured by the number of correctly phased genotypes, single-site and switching errors. For example, 2SNP requires 4 s on Pentium 4 3Ghz processor to phase 30 genotypes with 1500 SNPs (ENm010.7p15:2 data from HapMap) versus GERBIL and PHASE requiring more than a week of runtime and admitting no less errors than 2SNP. For genotypes with more than 1500 SNPs

PHASE and GERBIL fall into infinite loop. However, 2SNP software phases entire chromosome (10^5 SNPs from HapMap) for 30 individuals in 2 hours with average switching error 7.7%.

2.3.1 Introduction

The difference between individual DNA sequences mostly occurs at single-base sites, in which more than one nucleic acid or gap is observed across the population. Such variations are called single nucleotide polymorphisms (SNPs). The number of sufficiently frequent SNPs in the human population is estimated to be around 10 million [88]. For complex diseases caused by more than a single gene, it is important to identify a set of alleles inherited together. Identification of haplotypes, the sequences of alleles in contiguous SNP sites along a chromosomal region, is a central challenge of the International HapMap project [86]. The number of simultaneously typed SNPs for association and linkage studies is reaching 500,000 for SNP Mapping Arrays from Affymetrix [131].

Diploid organisms, like human, have two near-identical copies of each chromosome. Most experimental techniques for determining SNPs do not provide the haplotype information separately for each of the two chromosomes. Instead, they generate for each site an unordered pair of allele readings, one from each copy of the chromosome, which is called a genotype.

The input to the phasing problem consists of n genotype vectors each with m coordinates corresponding to SNPs. The phasing problem asks for explaining each genotype with two haplotypes corresponding to chromosomes. In general as well as in common biological setting, there are 2^{k-1} possible haplotype pairs for the same input genotype with k heterozygous sites.

Computational inferring of haplotypes from the genotypes (or *phasing*) has been initiated by Clark[79] who proposed a parsimony-based approach. It has been shown later that the likelihood based expectation-maximization (EM) is more accurate [92]. Markov chain Bayesian haplotype reconstruction methods have been used in PHASE [95], PLEM [91], and HAP2 [89]. A combinatorial model based on the perfect phylogeny tree assumption was suggested in [82]. HAP [83] exploits perfect phylogeny model and block structure showing good performance on real genotypes with low error rates. Recently, GERBIL [87] has combined block identification and phasing steps for reliable phasing of long genotypes.

In this work, we first explore phasing of genotypes with two SNPs which have ambiguity when the both sites are heterozygous. Then there are two possible phasings (so called *cis*- and *trans*-) and we assume that the true phasing tends to choose the most frequent pair of haplotypes observed in the population sample. Complete haplotypes for a given genotype can be inferred based on the maximum spanning tree of a complete graph with vertices corresponding to heterozygous sites and edge weights given by inferred 2-SNP frequencies. On datasets across 23 chromosomal regions from HapMap[86], 2SNP is several orders of magnitude faster than GERBIL and PHASE while matching them in quality measured by the number of correctly phased genotypes, single-site and switching errors. For example the 2SNP software phases entire chromosome (10^5 SNPs from HapMap) for 30 individuals in 2 hours with average switching error 7.7%. A brief description of 2SNP software can be found in the application note [77].

The rest of the section is organized as follows. The next subsection describes phasing of genotypes with only two SNPs. Subsection 2.3.3 describes the phasing of complete unrelated genotypes. Subsection 2.4.6 compares the proposed phasing 2SNP algorithm with PHASE, GERBIL, and PLEM for unrelated individuals.

2.3.2 Phasing of 2-SNP genotypes

In this section we first formally introduce the phasing problem and suggest a LD-based formula for the expected frequencies of cis- or trans-phasing of 2-SNP genotypes. We conclude with adjusting of expected haplotype frequencies to deviation from the random mating model.

The input to the phasing problem consists of *n* genotype vectors each with *m* coordinates corresponding to SNPs. SNP values belong to {0, 1, 2, ?}, where 0's and 1's denote homozygous sites with major allele and minor allele, respectively; 2's stand for heterozygous sites, and ?'s denote missed SNP values. Phasing replaces each genotype vector by two haplotype vectors with SNP values belonging to {0, 1}. Feasible phasing requires each 0 (resp. 1 or 2) in a genotype to be replaced with two 0's (resp. two 1's or 0 and 1) in the haplotypes. A 2-SNP genotype 22 can be cis-phased, i.e., represented as 00 and 11 haplotypes, or trans-phased, i.e. represented as 01 and 10 haplotypes.

Tendency of cis- or trans- phasing. The random mating model (i.e., assumption that any two haplotypes are equally likely to match in a genotype) implies that two homozygous SNPs *i* and *j* tend to phase cis- or trans- according to odds ratio

$$\lambda = \frac{F_{00} \times F_{11}}{F_{01} \times F_{10}} \tag{2.2}$$

where F_{00}, F_{01}, F_{10}, F_{11} are unknown true frequencies of haplotypes with the first and the second binary index denoting alleles of the *i*-th and *j*-th SNP, respectively. In other words, the larger odds ratio corresponds to more frequent cis-phasing. On the other hand, we have noticed that the *additive odds ratio*

$$\lambda' = \frac{F_{00} + F_{11}}{F_{01} + F_{10}}$$

describes true cis- or trans- phasing more accurately than the odds ratio (2.2) on real datasets.

It is known that many SNPs are in linkage disequilibrium (LD) and the higher LD corresponds to the higher tendency towards cis- or trans-phasing. We can estimate how far LD pushes additive odds ratio λ' away from $E(\lambda')$, which is expected in absence of LD,

$$E(\lambda') = \frac{F_{0\times} \times F_{\times 0} + F_{1\times} \times F_{\times 1}}{F_{0\times} \times F_{\times 1} + F_{1\times} \times F_{\times 0}}$$

where × in the indices denote "do not care", e.g., $F_{0\times} = F_{01} + F_{00}$. Thus we can measure the effect of LD between SNPs i and j by the ratio of the additive odds ratio λ' over the expected value of $E(\lambda')$,

$$LD_{ij} = \lambda' / E(\lambda')$$

Finally, it has been observed [91, 92, 87] the higher LD between pairs of closer SNPs. In order to discard falsely encountered LD between non-linked SNPs which are far apart, we divide logarithm of LD by the square of the distance between the SNPs. The complete formula for the tendency t_{ij} of cis- or trans-phasing of two homozygous SNPs i and j expressed in observed haplotype frequencies is as follows

$$t_{ij} = \frac{\log LD_{ij}}{(i-j)^2} \qquad (2.3)$$

$$= \frac{\log\left(\dfrac{n + (F_{00}F_{11} - F_{01}F_{10})/(F_{01} + F_{10})}{n - (F_{00}F_{11} - F_{01}F_{10})/(F_{00} + F_{11})} \right)}{(i-j)^2}$$

where n is number of input genotypes, and F_{00}, $F_{01,}$ F_{10}, F_{11} are frequencies of haplotypes with the first and the second binary index denoting alleles of the *i*-th and *j*-th SNP, respectively. Haplotype frequencies are computed based on all genotype frequencies except 22. For 22 genotypes, the haplotype frequencies are chosen to fit best Hardy-Weinberg equilibrium adjusted to observed deviation in single-site genotype distribution.

Adjusting observed frequencies to the random mating model. The formula for tendency (2.3) is based on unknown *true* haplotype frequencies F_{ij}'s. One can calculate only *observed* haplotype frequencies F^*_{ij} 's which can be extracted from all 2-SNP genotypes except heterozygous in the both SNPs.

The distribution of cis- and trans-phasing of 22-genotypes (i.e., 2-SNP genotypes heterozygous in the both sites) can be adjusted to the random mating model as follows. Let C_{22} and P_{22} denote unknown numbers of trans- and cis-phasings, then $C_{22} + P_{22} = G_{22}$, where G_{22} is the number of 22-genotypes. Then the adjusted odds ratio λ' can be expressed as

$$\lambda' = \frac{F^*_{00} + F^*_{11} + P_{22}}{F^*_{01} + F^*_{10} + C_{22}}$$

where F^*_{ij} denote observed haplotype frequencies. The best-fit values of C_{22} and P_{22} should minimize the sum of differences between expected and observed genotype frequencies. It is easy to compute the expected genotype frequencies for the random mating model, which unfortunately can significantly deviate from the real mating.

According to the random mating model, the best values of C_{22} and P_{22} minimize the sum of differences between expected and observed genotype frequencies. Unfortunately, the true mating frequently has statistically significant deviation from random mating. Based on observed real data set we have found that genotypes heterozygous in one or both positions should be assumed to be under-represented -- empirically they are assumed 3 times less frequent than it follows from the random mating model.

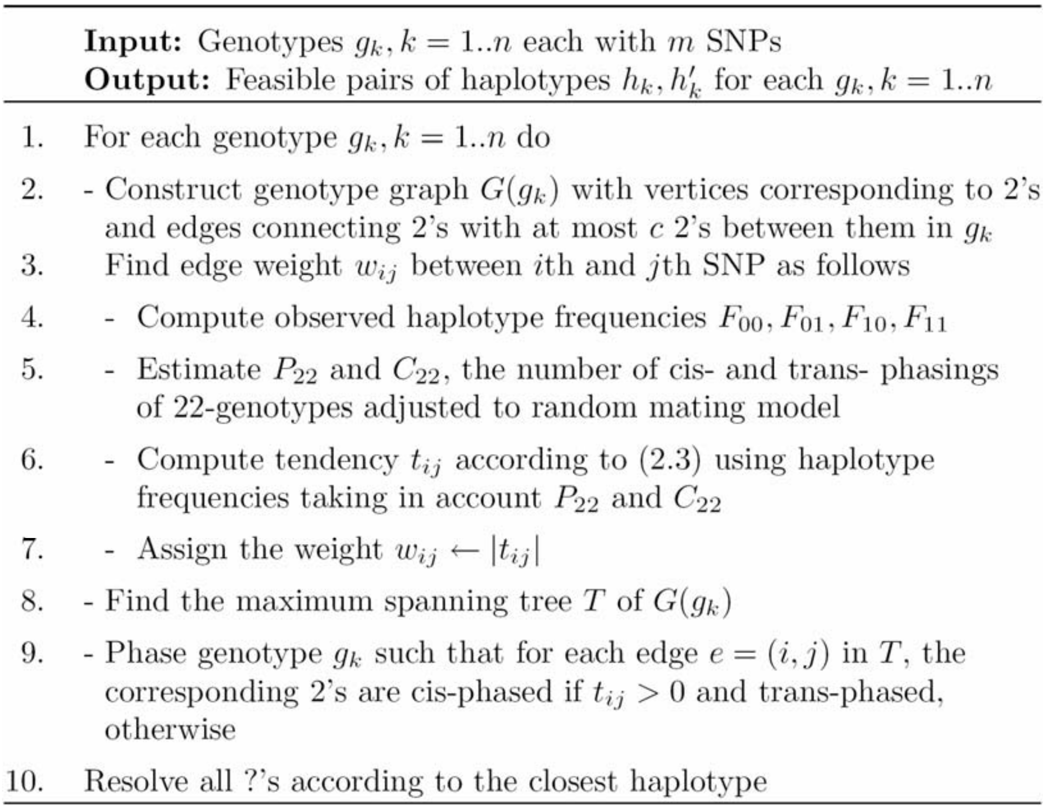
2.3.3 Phasing of Complete Genotypes

In this section we first describe phasing of complete genotypes using tendency of cis- or trans- phasing of pairs of 2's followed by resolving of missing data. We then show how to modify 2SNP algorithm to phase family trios and conclude with analysis of the runtime.

Genotype Graph. For each genotype *g*, 2SNP constructs a *genotype graph*, which is a weighted complete graph with vertices corresponding to 2's (i.e., heterozygous sites) of *g*. The edge weight represents the tendency of the corresponding 2's being cis- or trans-phased according to formula (2.3).

The maximum spanning tree of the genotype graph uniquely determines the phasing of the corresponding genotype since it gives cis-/trans- phasing for any two 2's. Obviously, if for any pair of 2's we know if they are cis- or trans-phased, then the entire phasing is known. Note that [87] have applied the same construction for preliminary estimation of haplotype frequencies rather than phasing *per se*. Therefore, for the edge weight, they have chosen LD-based formula over probabilities of *full* (*i* - *j*)-haplotypes given by maximum-likelihood solution. Instead, edge weights in 2SNP do not account for SNPs between *i* and *j*.

Our computational experience with multiple data sets shows that the edges connecting a pair of 2's that have at least $c = 20$ 2's between them in the genotype *g* never have large enough weight to be included in the maximum spanning tree. This may be result of fast decreasing of LD and tendency with the distance between 2's. Therefore, the edges between 2's that are *c* 2's apart are safely disregarded while drastically improving runtime for computing of *G*.



```
Input: Genotypes $g_k, k = 1..n$ each with $m$ SNPs
Output: Feasible pairs of haplotypes $h_k, h'_k$ for each $g_k, k = 1..n$

1.  For each genotype $g_k, k = 1..n$ do
2.    - Construct genotype graph $G(g_k)$ with vertices corresponding to 2's
        and edges connecting 2's with at most $c$ 2's between them in $g_k$
3.      Find edge weight $w_{ij}$ between $i$th and $j$th SNP as follows
4.        - Compute observed haplotype frequencies $F_{00}, F_{01}, F_{10}, F_{11}$
5.        - Estimate $P_{22}$ and $C_{22}$, the number of cis- and trans- phasings
            of 22-genotypes adjusted to random mating model
6.        - Compute tendency $t_{ij}$ according to (2.3) using haplotype
            frequencies taking in account $P_{22}$ and $C_{22}$
7.        - Assign the weight $w_{ij} \leftarrow |t_{ij}|$
8.    - Find the maximum spanning tree $T$ of $G(g_k)$
9.    - Phase genotype $g_k$ such that for each edge $e = (i, j)$ in $T$, the
        corresponding 2's are cis-phased if $t_{ij} > 0$ and trans-phased,
        otherwise
10. Resolve all ?'s according to the closest haplotype
```

Figure 2.1 2SNP Phasing Algorithm

Missing data recovery. Missing data (?'s) are inferred after phasing of 2's. Each haplotype is partitioned into segments of length 40. For each segment h we find the closest (with respect to Hamming distance) segment h' and infer ?'s in h using the corresponding values from h'.

2.3.4 Runtime

The runtime of Step 3 (see Figure 2.1) is $O(n)$; therefore, the total runtime of Step 2 is $O(cnm)$, since Step 3 is repeated at most cm times. The runtime of Step 8 is $O(cm \log m)$ since the graph $G(g_k)$ has maximum degree c. Therefore, the total runtime of Step 1 is $O(cn^2m +cnm \log m)$. Missing data recovery (Step 10) computes all pairwise Hamming distances between $2n$ haplotypes each with m SNPs requiring $O(n^2m)$ time. As a result, the total runtime of the algorithm is $O(cnm(n + \log m))$, where n and m are the number of genotypes and SNPs, respectively. Figure 2.2 shows average runtime of 2SNP on datasets from HapMap.

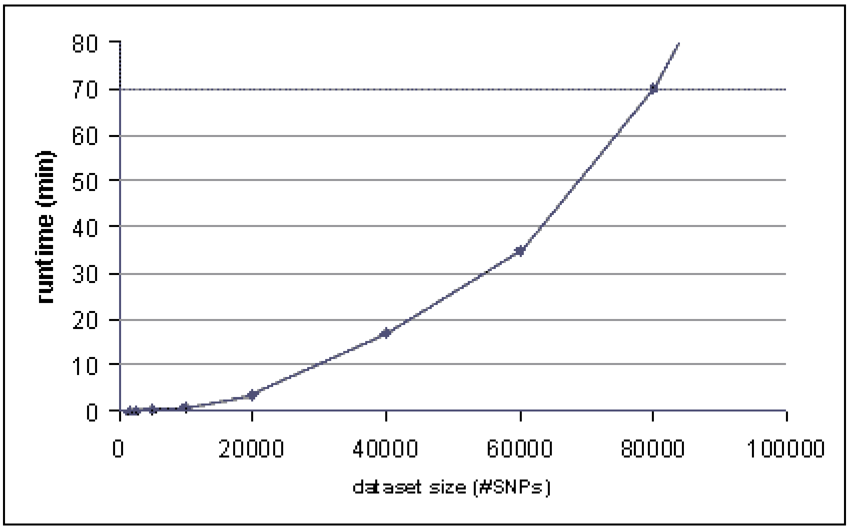


Figure 2.2 Average runtime of 2SNP algorithm on datasets from HapMap build 35, July 2006 over all chromosomes and 3 races (European, African, and Japanese).

2.3.5 Experimental Results

In this section we first describe the datasets and quality measures. Then, we compare our 2SNP method with PHASE-2.1.1[96], PLEM[91] and GERBIL[87] on unrelated data.

Real data sets. The comparison of phasing methods was performed on 46 real datasets from 79 different genomic regions and on 4 simulated datasets. All real datasets represent family trios -- the computationally inferred offspring haplotypes for offspring have been compared with haplotypes inferred from parental genotypes.

- *Chromosome* 5*q*31: 129 genotypes with 103 SNPs derived from the 616 KB region of human Chromosome 5*q*31 [138].

- *Yoruba population (D)*: 30 genotypes with SNPs from 51 various genomic regions, with number of SNPs per region ranging from 13 to 114 [81].

- *HapMap datasets I*: 30 genotypes of Utah residents and Yoruba residents available on HapMap by December 2005. The number of SNPs varies from 52 to 1381 across 40 regions including ENm010, ENm013, ENr112, ENr113 and ENr123 spanning 500 KB regions of chromosome bands 7p15:2, 7q21:13, 2p16:3, 4q26 and 12q12 respectively, and two regions spanning the gene STEAP and TRPM8 plus 10 KB upstream and downstream.

- *HapMap datasets II*: entire 23 chromosomes of 30 individuals of Utah residents, Yoruba residents, and Japanese residents available on HapMap by July 2006, build 35.

Simulated data sets.

- *Random matching* 5*q*31: 128 genotypes each with 89 SNPs from 5q31 cytokine gene cluster generated by random matching from 64 haplotypes of 32 West African reported by [85].

- *MS-simulated data*: 258 populations have been generated by MS[84] haplotype generator (using recombination rates 0,4 and 16). From each population of 100 haplotypes with 103 SNPs we have randomly chosen one haplotype and generated 129 genotypes by random matching.

Cis-/trans- odds ratio on real and simulated datasets. Here we point out to an important difference between real data and data simulated using coalescent model. The odds ratio (2.2) is one of the measures of LD between two SNPs. The odds ratio much larger than 1 or much smaller than 1 indicates high LD. When computed for *all* pairs of SNPs *i* and *j*, the *averaged odds ratio*

$$\overline{\lambda} = \frac{\sum_{i \neq j} F_{00} \times F_{11}}{\sum_{i \neq j} F_{01} \times F_{10}} \tag{2.4}$$

averaged over real data is 1.9 while for all simulated data is very close to 1. This can be partially explained by the fact that each SNP can be flipped over, i.e., 0 and 1 notations for major and minor alleles. Each such flipping swaps the products $F_{00}F_{11}$ and $F_{01}F_{10}$ and may increase $\overline{\lambda}$ (2.4). Therefore, we greedily flip SNPs increasing $\overline{\lambda}$. For real data sets the averaged greedily increased $\overline{\lambda}$ equals 2.5 while for MS-simulated data $\overline{\lambda}$ heavily depends on the recombination rate. The haplotype sets generated with recombination rate r between 40 and 100 are the closest to the real datasets with respect to $\overline{\lambda}$. On the other hand, the average $\overline{\lambda}$ for ST 1, ST 2, and ST 3 are 1.95, 1.37, and 1.65 showing a significant deviation from the real datasets.

Error measures. A *single-site error* [96] is the percent of erroneous SNPs among all SNPs in phased haplotypes. An *individual error* [91] is the percent of genotypes phased with at least one error among all genotypes. A *switching error* [87] is the percent of switches (among all possible switches) between inferred haplotypes necessary to obtain a true haplotype. For each dataset we bootstraped phasing result 100 times, and for each bootstrap sample we computed an error. The 95% confidence interval for the error mean was computed based on the 100 error values.

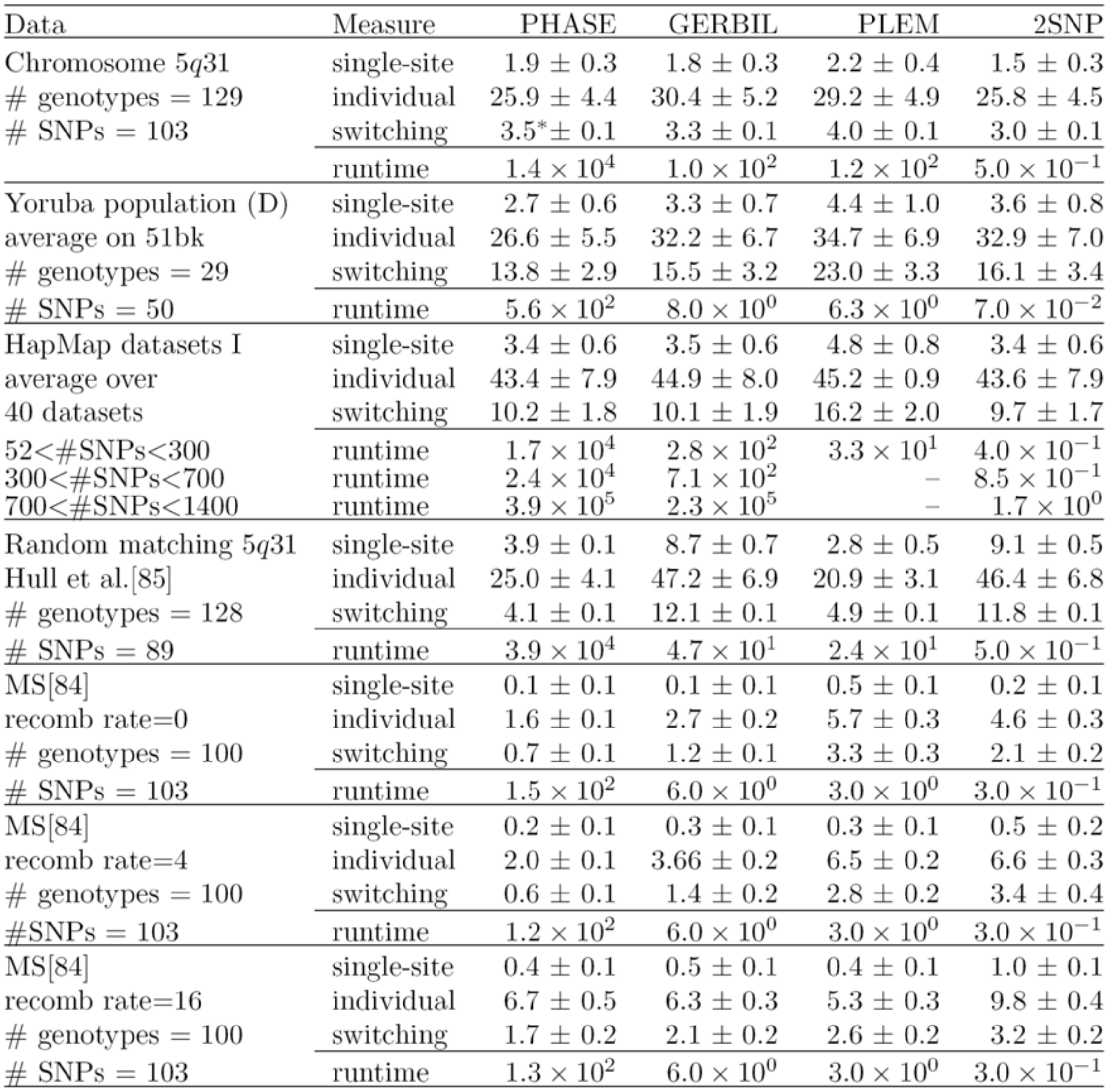
Comparison of phasing methods. Table 2.1 shows performance of four phasing methods on real and simulated datasets. Average runtimes for *HapMap datasets I* are separately reported for three ranges of genotype length in SNPs. All runs were performed on computer with Intel Pentium 4, 3.0Ghz processor and 2 Gigabytes of Random Access Memory. PLEM was run with 20 rounds and the 2.1.1 version of PHASE was run with the default parameters. The *-marked switching error is 2.9% for the earlier version PHASE-2.0.2. The dashes in PLEM columns correspond to the cases when it does not output valid phasing. The 95% confidence intervals for the error means are computed using bootstrapping.

From Table 2.1, one can see that 2SNP is several orders of magnitude faster than two other phasing methods handling large datasets in a matter of seconds. The reported mean errors with the respective 95% confidence intervals show that GERBIL, PHASE, and 2SNP have the same accuracy for real data (Chromosome 5q31, Yoruba(D), HapMap datasets). On the other hand, 2SNP and GERBIL are considerably outperformed by PHASE and PLEM on some simulated data. Poor performance of 2SNP can be caused by wrong recombination rate distribution on simulated data. For the mixed populations one can see deterioration of all phasing methods.

2.4 Phasing and Missing data recovery in Family Trios

Although there exist many phasing methods for unrelated adults or pedigrees, phasing and missing data recovery for data representing family trios is lagging behind. This work is an attempt to fill this gap by considering the following problem. Given a set of genotypes partitioned into family trios, find for each trio a quartet of parent haplotypes which agree with all three genotypes and recover the SNP values missed in given genotype data. Our contributions include (i) formulating the pure-parsimony trio phasing and the trio missing data recovery problems, (ii) proposing two new greedy and integer linear programming based solution methods, and (iii) extensive experimental validation of proposed methods showing advantage over the previously known methods.

Table 2.1 Mean single-site, individual, and switching errors with 95% confidence intervals and runtimes of PHASE, GERBIL, PLEM, and 2SNP on real and simulated datasets.



Data	Measure	PHASE	GERBIL	PLEM	2SNP
Chromosome $5q31$	single-site	1.9 ± 0.3	1.8 ± 0.3	2.2 ± 0.4	1.5 ± 0.3
# genotypes = 129	individual	25.9 ± 4.4	30.4 ± 5.2	29.2 ± 4.9	25.8 ± 4.5
# SNPs = 103	switching	$3.5^* \pm 0.1$	3.3 ± 0.1	4.0 ± 0.1	3.0 ± 0.1
	runtime	1.4×10^4	1.0×10^2	1.2×10^2	5.0×10^{-1}
Yoruba population (D)	single-site	2.7 ± 0.6	3.3 ± 0.7	4.4 ± 1.0	3.6 ± 0.8
average on 51bk	individual	26.6 ± 5.5	32.2 ± 6.7	34.7 ± 6.9	32.9 ± 7.0
# genotypes = 29	switching	13.8 ± 2.9	15.5 ± 3.2	23.0 ± 3.3	16.1 ± 3.4
# SNPs = 50	runtime	5.6×10^2	8.0×10^0	6.3×10^0	7.0×10^{-2}
HapMap datasets I	single-site	3.4 ± 0.6	3.5 ± 0.6	4.8 ± 0.8	3.4 ± 0.6
average over	individual	43.4 ± 7.9	44.9 ± 8.0	45.2 ± 0.9	43.6 ± 7.9
40 datasets	switching	10.2 ± 1.8	10.1 ± 1.9	16.2 ± 2.0	9.7 ± 1.7
52<#SNPs<300	runtime	1.7×10^4	2.8×10^2	3.3×10^1	4.0×10^{-1}
300<#SNPs<700	runtime	2.4×10^4	7.1×10^2	–	8.5×10^{-1}
700<#SNPs<1400	runtime	3.9×10^5	2.3×10^5	–	1.7×10^0
Random matching $5q31$	single-site	3.9 ± 0.1	8.7 ± 0.7	2.8 ± 0.5	9.1 ± 0.5
Hull et al.[85]	individual	25.0 ± 4.1	47.2 ± 6.9	20.9 ± 3.1	46.4 ± 6.8
# genotypes = 128	switching	4.1 ± 0.1	12.1 ± 0.1	4.9 ± 0.1	11.8 ± 0.1
# SNPs = 89	runtime	3.9×10^4	4.7×10^1	2.4×10^1	5.0×10^{-1}
MS[84]	single-site	0.1 ± 0.1	0.1 ± 0.1	0.5 ± 0.1	0.2 ± 0.1
recomb rate=0	individual	1.6 ± 0.1	2.7 ± 0.2	5.7 ± 0.3	4.6 ± 0.3
# genotypes = 100	switching	0.7 ± 0.1	1.2 ± 0.1	3.3 ± 0.3	2.1 ± 0.2
# SNPs = 103	runtime	1.5×10^2	6.0×10^0	3.0×10^0	3.0×10^{-1}
MS[84]	single-site	0.2 ± 0.1	0.3 ± 0.1	0.3 ± 0.1	0.5 ± 0.2
recomb rate=4	individual	2.0 ± 0.1	3.66 ± 0.2	6.5 ± 0.2	6.6 ± 0.3
# genotypes = 100	switching	0.6 ± 0.1	1.4 ± 0.2	2.8 ± 0.2	3.4 ± 0.4
#SNPs = 103	runtime	1.2×10^2	6.0×10^0	3.0×10^0	3.0×10^{-1}
MS[84]	single-site	0.4 ± 0.1	0.5 ± 0.1	0.4 ± 0.1	1.0 ± 0.1
recomb rate=16	individual	6.7 ± 0.5	6.3 ± 0.3	5.3 ± 0.3	9.8 ± 0.4
# genotypes = 100	switching	1.7 ± 0.2	2.1 ± 0.2	2.6 ± 0.2	3.2 ± 0.2
# SNPs = 103	runtime	1.3×10^2	6.0×10^0	3.0×10^0	3.0×10^{-1}

2.4.1 Pure-Parsimony Trio Phasing

Note that it is easy to find a feasible solution to TPP but the number of feasible solutions is exponential and it is necessary to choose criteria for comparing such solutions. In [39] for haplotyping pedigree data, the objective is to minimize recombinations. That objective is not suitable for TPP since the trios are not full-fledged pedigree data and contain no clues to evidence recombination reconstruction. Thus, following [29, 82], we have decided to pursue parsimonious objective, i.e., minimization of the total number of haplotypes.

The drawback of pure parsimony is that when the number of SNPs becomes large (as well as the number of recombinations), then the quality of pure parsimony phasing is diminishing [82]. Therefore, following the approach in [33], we suggest to partition the genotypes into blocks, i.e., substrings of bounded length, and find solution for the pure parsimony problem for each block separately. Note that in case of family trios we have great advantage over the method of [33] since we do not need to solve the problem of joining blocks. Indeed, for each family trio we can make four haplotype templates (partially resolved by logic means of haplotypes) that imply unique way of gluing together blocks to arrange complete haplotypes for the entire sequence of SNPs.

Formally, let *genotype* be a vector with *m* coordinates each corresponding to an SNP and having one of the following values: 0 (homozygote with major allele), 1 (homozygote with minor allele), 2 (heterozygote), or ? (missing SNP value). Let *haplotype* be a vector with m coordinates where each coordinate is either 0 or 1. We say that two haplotypes *explain* a genotype if

- for any 0 (resp. 1) in the genotype vector, the corresponding coordinates in the both haplotype vectors are 0's (resp. 1's),
- for any 2 in the genotype vector, the corresponding coordinates in the two haplotype vectors are 0 and 1,
- for any ? in the genotype vector, the corresponding coordinates in the haplotypes are unconstrained (can be arbitrary).

We say that four haplotypes h_1, h_2, h_3, h_4 explain a family trio of genotypes (f, m, k), if h_1 and h_2 explain the genotype f , h_3 and h_4 explain the genotype m, and h_1 and h_3 explain the genotype k.

Pure-Parsimony Trio Phasing (PPTP). Given $3n$ genotypes corresponding to n family trios find minimum number of distinct haplotypes explaining all trios.

2.4.2 Greedy Method for Trio Phasing

We apply the greedy algorithm from Halperin [35] for trio phasing. For each trio we introduce four partial haplotypes with the coordinates 0, 1 and ?. The values of 0 and 1 correspond to fully resolved SNPs which can be found via logical resolution from the section 4.1, while the ?'s corresponds to ambiguous and missing positions. The greedy algorithm iteratively finds the complete haplotype which covers the maximum possible number of partial haplotypes, removes this set of resolved partial haplotypes and continues in that manner. The drawback of this greed method is to introduce error to trio constraint, even for phasing with error $O(m)$ to the maximum concentration. In the future, we will try to modify the greedy algorithm to overcome its shortcoming.

2.4.3 Integer Linear Program for Trio Phasing

We have implemented the following integer linear program (ILP) formulation (2.5)-(2.8) for pure-parsimony trio phasing. It uses 0-1 variable x_i for each possible haplotype with the minimization objective:

$$Minimize \sum x_i \tag{2.5}$$

For each trio we introduce four template haplotypes, i.e., haplotypes with the coordinates 0,1,2 and ?: 0's and 1's correspond to fully resolved SNPs, 2's come in pairs corresponding to the genotype 2's and ?'s correspond to unconstrained SNPs. For each 2 in each template we introduce a 0-1-variable y and a constraint binding it with the variable y' corresponding to the complimentary 2:

$$y + y' = 1$$

Instead of completely resolving templates, we can resolve only 2's. Then several haplotypes can fit partially resolved templates and at least one of the corresponding *x*-variables should be set to 1. This results in the following constraint: For any *y*-assignment of 2's in each template *T*,

$$\sum_{x\_fits\_all\_y'\_in\_template\_T} x \geq 1 + \sum_{i \in I_1} y_i + |I_1| + \sum_{i \in I_2} (1 - y_i) - |I_2| \tag{2.7}$$

We should guarantee that each template is resolved. This is done by the following

constraint: For each template T ,

$$\sum_{x\_fits\_template\_T} x \geq 1 \tag{2.8}$$

2.4.4 Experimental Study of ILP and Greedy Methods

In this section we compare our greedy and ILP based methods suggested in Section 2.4.1 with previously known phasing methods such as Phamily [27], PHASE[95] and HAPLOTYPER[41] applied to phasing and missing recovery on family trio data. We first describe the test data sets then give experimental results of five methods for phasing and then for missing data recovery of family trio data.

Our algorithms are evaluated on real and simulated data. The data set collected by Daly et al. [138] is derived from the 616 kilobase region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohn disease by genotyping 103 SNPs for 129 trios. The missing data in this genotype data is about 16%. The data set was partitioned [138] into eleven blocks with only a few common haplotypes inside.

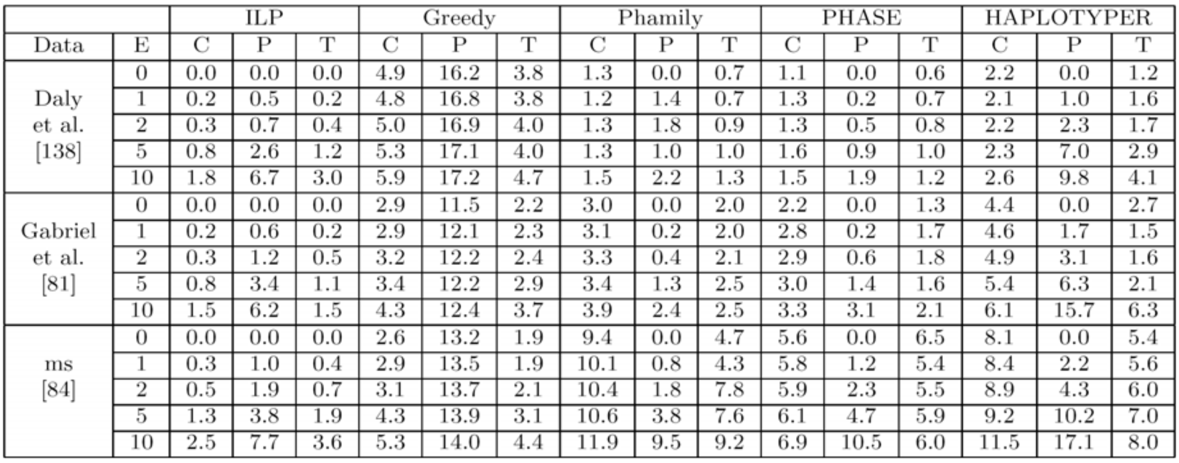
The another real data set is collect by Gabriel et al. [81]. This data consists of genotypes of SNPs from 62 region. We use population D which contains of 30 trois from Yoruba. This data set contains about 10% missing data.

The simulated data is generated using ms [84], a well-known haplotype generator based on the coalescent model of SNP sequence evolution. The ms generator emits a haplotype population for the given number of haplotypes, number of SNPs, and the recombination rate. We have simulated Daly et al. [138] data by generating 258 populations, each population with 100 individuals and each haplotype with 103 SNPs, then randomly choosing one haplotype from each population. We only simulate parents' haplotypes, then we obtain family trio haplotypes and genotypes by random matching the parental haplotypes.

It is clear how to validate a phasing method on simulated data since the underlying haplotypes are known. The validation on real data is usually performed on the trio data. E.g., a phasing method is applied to parents (respectively, to children) genotypes and the resulted haplotypes are validated on children's (respectively, on parents') genotypes. Unfortunately, in our case, one can not apply such validation since a trio phasing method may rely on both children and parents' genotypes. Therefore, we suggest to validate trio phasing by erasing randomly chosen SNP values and recording the errors in the erased SNP sites. In Tables 2.2, 2.3, 2.4, each row corresponds to an instance of real data (Daly et al. or Gabriel et al.) or simulated data (ms) and the column (E) shows the percent of erased data (0% - no data erased, 1%-10% - percent of SNP values erased) .

The value of phasing errors is measured by the Hamming distance from the method's solution to the closest feasible phasing. In Tables 2.2 and 2.3, for parents (P) we report the percent of SNP values that should be inverted out of the total number of SNP values that should be inferred (i.e., number of 2 plus number of unknown values). For children (C), we report the percent of SNP which should be inverted with respect to the total number of SNPs. The total number of errors (T) is the percent of SNP's that should be inverted in order to obtain a feasible phasing solution.

Table 2.2 The results for five phasing methods on the real data sets of Daly et al.[138] and Gabrile et al. [81] and on simulated data. The second column corresponds to the ratio of erased data. The C corresponds to the error of child. The P corresponds to the error of parents. The T corresponds to the total error.

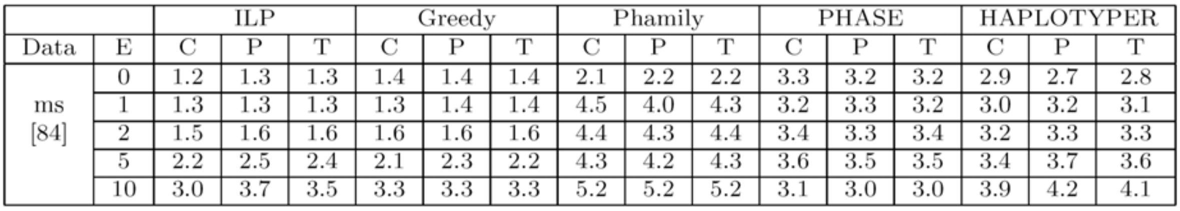


		ILP			Greedy			Phamily			PHASE			HAPLOTYPER		
Data	E	C	P	T	C	P	T	C	P	T	C	P	T	C	P	T
Daly et al. [138]	0	0.0	0.0	0.0	4.9	16.2	3.8	1.3	0.0	0.7	1.1	0.0	0.6	2.2	0.0	1.2
	1	0.2	0.5	0.2	4.8	16.8	3.8	1.2	1.4	0.7	1.3	0.2	0.7	2.1	1.0	1.6
	2	0.3	0.7	0.4	5.0	16.9	4.0	1.3	1.8	0.9	1.3	0.5	0.8	2.2	2.3	1.7
	5	0.8	2.6	1.2	5.3	17.1	4.0	1.3	1.0	1.0	1.6	0.9	1.0	2.3	7.0	2.9
	10	1.8	6.7	3.0	5.9	17.2	4.7	1.5	2.2	1.3	1.5	1.9	1.2	2.6	9.8	4.1
Gabriel et al. [81]	0	0.0	0.0	0.0	2.9	11.5	2.2	3.0	0.0	2.0	2.2	0.0	1.3	4.4	0.0	2.7
	1	0.2	0.6	0.2	2.9	12.1	2.3	3.1	0.2	2.0	2.8	0.2	1.7	4.6	1.7	1.5
	2	0.3	1.2	0.5	3.2	12.2	2.4	3.3	0.4	2.1	2.9	0.6	1.8	4.9	3.1	1.6
	5	0.8	3.4	1.1	3.4	12.2	2.9	3.4	1.3	2.5	3.0	1.4	1.6	5.4	6.3	2.1
	10	1.5	6.2	1.5	4.3	12.4	3.7	3.9	2.4	2.5	3.3	3.1	2.1	6.1	15.7	6.3
ms [84]	0	0.0	0.0	0.0	2.6	13.2	1.9	9.4	0.0	4.7	5.6	0.0	6.5	8.1	0.0	5.4
	1	0.3	1.0	0.4	2.9	13.5	1.9	10.1	0.8	4.3	5.8	1.2	5.4	8.4	2.2	5.6
	2	0.5	1.9	0.7	3.1	13.7	2.1	10.4	1.8	7.8	5.9	2.3	5.5	8.9	4.3	6.0
	5	1.3	3.8	1.9	4.3	13.9	3.1	10.6	3.8	7.6	6.1	4.7	5.9	9.2	10.2	7.0
	10	2.5	7.7	3.6	5.3	14.0	4.4	11.9	9.5	9.2	6.9	10.5	6.0	11.5	17.1	8.0

In Table 2.3, we also report true error for phasing simulated genotype data which is the Hamming distance between inferred and actual simulated underlying haplotypes for children (C), for parents (P) and the total error (T).

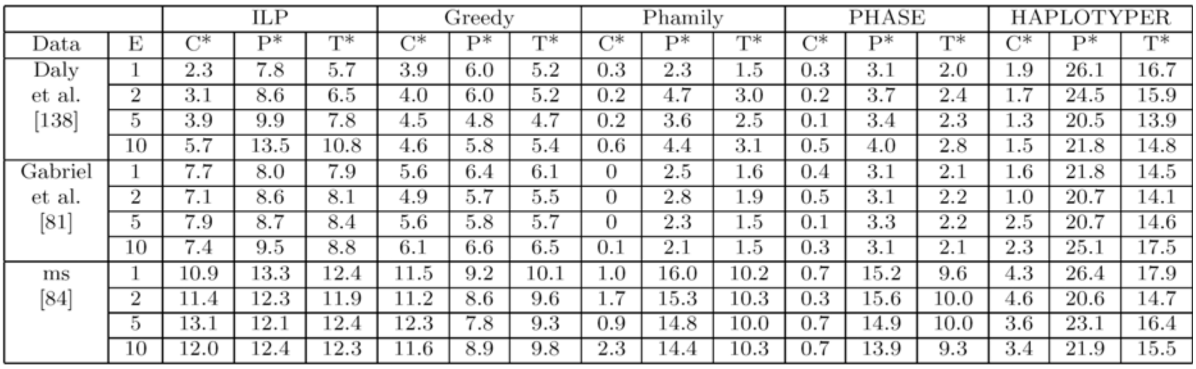
Table 2.4 compares five methods (ILP, Greedy, Phamily, PHASE and HAPLOTYPER) on trio missing data recovery on the real data sets (Daly [138] and Gabriel [81]) and simulated data. We erase random data in trio genotypes with certain amount (1%; 2%; 5% and 10%) of the entire data. Instead, We report the error as the number of incorrectly recovered erased positions of the genotypes on child (C*), parents (P*) and trios (T*) divided the total number of erased positions in parent genotypes in percentage. We count only half error if the compared paired SNP is 2 and 0 (or 1).

Table 2.3 The results for five phasing methods on the simulated data sets. The column E represents the percent of erased data. The C corresponds to the true error of child. The P corresponds to the true error of parents. The T corresponds to the true total error.



		ILP			Greedy			Phamily			PHASE			HAPLOTYPER		
Data	E	C	P	T	C	P	T	C	P	T	C	P	T	C	P	T
ms [84]	0	1.2	1.3	1.3	1.4	1.4	1.4	2.1	2.2	2.2	3.3	3.2	3.2	2.9	2.7	2.8
	1	1.3	1.3	1.3	1.3	1.4	1.4	4.5	4.0	4.3	3.2	3.3	3.2	3.0	3.2	3.1
	2	1.5	1.6	1.6	1.6	1.6	1.6	4.4	4.3	4.4	3.4	3.3	3.4	3.2	3.3	3.3
	5	2.2	2.5	2.4	2.1	2.3	2.2	4.3	4.2	4.3	3.6	3.5	3.5	3.4	3.7	3.6
	10	3.0	3.7	3.5	3.3	3.3	3.3	5.2	5.2	5.2	3.1	3.0	3.0	3.9	4.2	4.1

Table 2.4 The results for missing data recovery on the real and simulated data sets with five methods. The second column corresponds to the ratio of erased data. The C* corresponds to the error of child. The P* corresponds to the error of parents. The T* corresponds to the total error.



		ILP			Greedy			Phamily			PHASE			HAPLOTYPER		
Data	E	C*	P*	T*	C*	P*	T*	C*	P*	T*	C*	P*	T*	C*	P*	T*
Daly et al. [138]	1	2.3	7.8	5.7	3.9	6.0	5.2	0.3	2.3	1.5	0.3	3.1	2.0	1.9	26.1	16.7
	2	3.1	8.6	6.5	4.0	6.0	5.2	0.2	4.7	3.0	0.2	3.7	2.4	1.7	24.5	15.9
	5	3.9	9.9	7.8	4.5	4.8	4.7	0.2	3.6	2.5	0.1	3.4	2.3	1.3	20.5	13.9
	10	5.7	13.5	10.8	4.6	5.8	5.4	0.6	4.4	3.1	0.5	4.0	2.8	1.5	21.8	14.8
Gabriel et al. [81]	1	7.7	8.0	7.9	5.6	6.4	6.1	0	2.5	1.6	0.4	3.1	2.1	1.6	21.8	14.5
	2	7.1	8.6	8.1	4.9	5.7	5.5	0	2.8	1.9	0.5	3.1	2.2	1.0	20.7	14.1
	5	7.9	8.7	8.4	5.6	5.8	5.7	0	2.3	1.5	0.1	3.3	2.2	2.5	20.7	14.6
	10	7.4	9.5	8.8	6.1	6.6	6.5	0.1	2.1	1.5	0.3	3.1	2.1	2.3	25.1	17.5
ms [84]	1	10.9	13.3	12.4	11.5	9.2	10.1	1.0	16.0	10.2	0.7	15.2	9.6	4.3	26.4	17.9
	2	11.4	12.3	11.9	11.2	8.6	9.6	1.7	15.3	10.3	0.3	15.6	10.0	4.6	20.6	14.7
	5	13.1	12.1	12.4	12.3	7.8	9.3	0.9	14.8	10.0	0.7	14.9	10.0	3.6	23.1	16.4
	10	12.0	12.4	12.3	11.6	8.9	9.8	2.3	14.4	10.3	0.7	13.9	9.3	3.4	21.9	15.5

2.4.5 2SNP Phasing Method for Family Trios

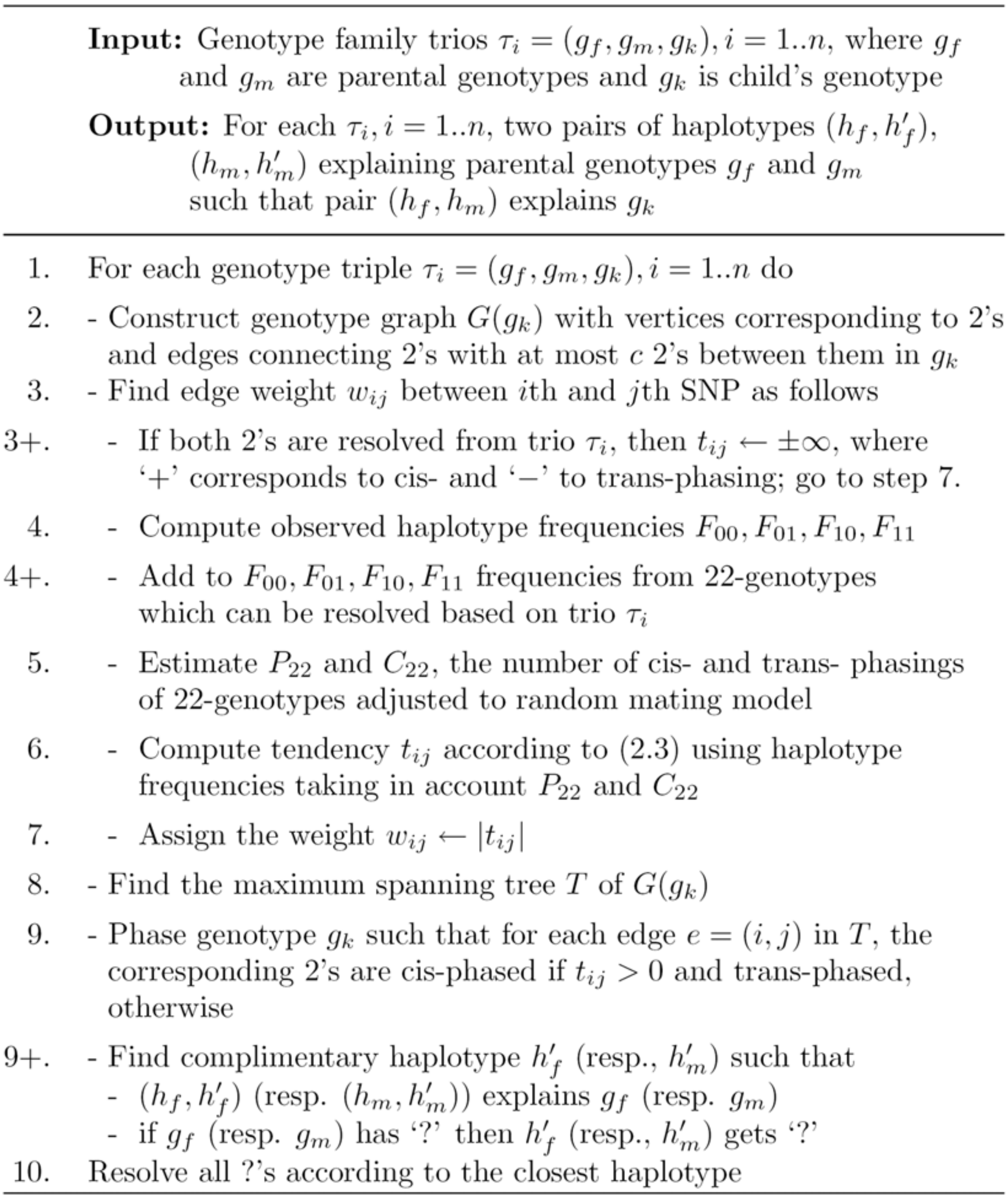
We have enhanced 2SNP algorithm to phase family trio data. 2SNP trio phasing have been compared with four other well-known phasing methods on simulated data from [90]. 2SNP is much faster than all of them while loosing in quality only to PHASE.

We have enhanced 2SNP algorithm as follows (see Figure 2.3). First we modify observed haplotype frequencies F_{00}, F_{01}, F_{10}, F_{11} with haplotype frequencies in 22's which can be imputed from trio constraints. For example if we have three genotypes 22,01,22 (first parent, second parent, and child) then we can impute first parent and child genotypes (10+01 and 01+10). Only positions with 2's in all three genotypes can be left unresolved. For these genotypes we perform our standard estimation of P22 and C22.

The genotype graph g is constructed only for the children's genotypes. The edges connecting resolved 2's are forced into MST by setting their weight to a large number. Obtained maximum spanning tree uniquely determines the phasing of the corresponding children's genotype. The parental haplotypes are inferred assuming no recombinations, i.e., one of the parental haplotypes is the haplotype transmitted to the child and another (non-transmitted) haplotype is complimentary.

If the parental genotype has a ? which can not be resolved from child's haplotypes, then we keep ?'s in the complimentary haplotype. Finally, the left missing data are recovered the same way as in Fig. 2.1.

For the family trios, the 2SNP is considerably faster than the 2SNP for the same trios if they are assumed to be unrelated. Indeed, the size of each genotype graph is reduced and, essentially, only children genotypes are phased.



```
Input: Genotype family trios $\tau_i = (g_f, g_m, g_k), i = 1..n$, where $g_f$
       and $g_m$ are parental genotypes and $g_k$ is child's genotype
Output: For each $\tau_i, i = 1..n$, two pairs of haplotypes $(h_f, h'_f)$,
        $(h_m, h'_m)$ explaining parental genotypes $g_f$ and $g_m$
        such that pair $(h_f, h_m)$ explains $g_k$

1.   For each genotype triple $\tau_i = (g_f, g_m, g_k), i = 1..n$ do
2.   - Construct genotype graph $G(g_k)$ with vertices corresponding to 2's
       and edges connecting 2's with at most $c$ 2's between them in $g_k$
3.   - Find edge weight $w_{ij}$ between $i$th and $j$th SNP as follows
3+.    - If both 2's are resolved from trio $\tau_i$, then $t_{ij} \leftarrow \pm\infty$, where
         '+' corresponds to cis- and '$-$' to trans-phasing; go to step 7.
4.     - Compute observed haplotype frequencies $F_{00}, F_{01}, F_{10}, F_{11}$
4+.    - Add to $F_{00}, F_{01}, F_{10}, F_{11}$ frequencies from 22-genotypes
         which can be resolved based on trio $\tau_i$
5.     - Estimate $P_{22}$ and $C_{22}$, the number of cis- and trans- phasings
         of 22-genotypes adjusted to random mating model
6.     - Compute tendency $t_{ij}$ according to (2.3) using haplotype
         frequencies taking in account $P_{22}$ and $C_{22}$
7.     - Assign the weight $w_{ij} \leftarrow |t_{ij}|$
8.   - Find the maximum spanning tree $T$ of $G(g_k)$
9.   - Phase genotype $g_k$ such that for each edge $e = (i, j)$ in $T$, the
       corresponding 2's are cis-phased if $t_{ij} > 0$ and trans-phased,
       otherwise
9+.  - Find complimentary haplotype $h'_f$ (resp., $h'_m$) such that
       - $(h_f, h'_f)$ (resp. $(h_m, h'_m)$) explains $g_f$ (resp. $g_m$)
       - if $g_f$ (resp. $g_m$) has '?' then $h'_f$ (resp., $h'_m$) gets '?'
10.  Resolve all ?'s according to the closest haplotype
```

Figure 2.3 2SNP Trio Phasing Algorithm. The steps added to 2SNP phasing algorithm are marked with +'s -- 3+, 4+, and 9+.

2.4.6 Experimental Study of 2SNP for Family Trios

In this section we first describe the datasets and quality measures. Then, we compare our 2SNP method with PHASE[96], HAP[83], HAP2[89], and PLEM[91] on simulated trio data.

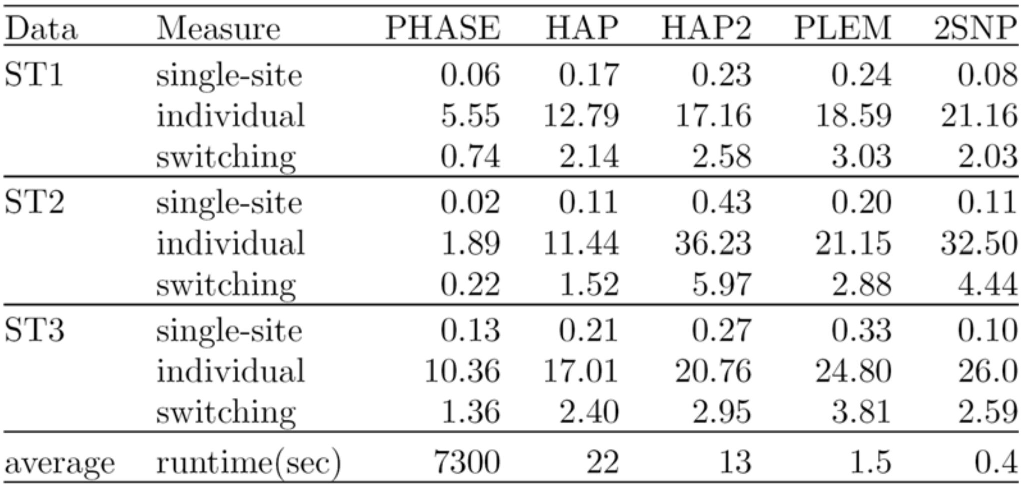
Error measures. A *single-site error* [96] is the percent of erroneous SNPs among all SNPs in phased haplotypes. An *individual error* [91] is the percent of genotypes phased with at least one error among all genotypes. A *switching error* [87] is the percent of switches (among all possible switches) between inferred haplotypes necessary to obtain a true haplotype. For each dataset we bootstraped phasing result 100 times, and for each bootstrap sample we computed an error. The 95% confidence interval for the error mean was computed based on the 100 error values.

Simulated data sets.

- *ST1, ST2, ST3 - family trio data* [93]: Trio datasets generated using a coalescent model that incorporates variation in recombination rates and demographic events. Specific simulation parameters used can be found in [94]. ST1 - 20 data sets of 30 trios simulated with constant recombination rate across the region, constant population size, and random mating. Each of the 20 data sets consisted of 1 Mb of sequence. ST2 - same as ST1, but with the addition of a variable recombination rate across the region. ST3 - same as ST2, except a model of demography consistent with white Americans was used.

Table 2.5 compares the 2SNP algorithm for family trios with PHASE[96], HAP[83], HAP2[89], and PLEM[91](The results for PHASE, HAP, HAP2, and PLEM are taken from [90]) on simulated family trio datasets ST1, ST2, ST3. As for unrelated individuals, 2SNP is much faster than other phasing methods. For ST1 is losing in switching error only to PHASE, for ST2 it loses to 3 methods, and for ST3 it loses to PHASE and HAP. This can be explained by the fact that ST3 and especially ST2 strongly deviate from real datasets in the averaged odds ratio $\overline{\lambda}$.

Table 2.5 Comparison of 5 phasing methods on simulated trio data.



Data	Measure	PHASE	HAP	HAP2	PLEM	2SNP
ST1	single-site	0.06	0.17	0.23	0.24	0.08
	individual	5.55	12.79	17.16	18.59	21.16
	switching	0.74	2.14	2.58	3.03	2.03
ST2	single-site	0.02	0.11	0.43	0.20	0.11
	individual	1.89	11.44	36.23	21.15	32.50
	switching	0.22	1.52	5.97	2.88	4.44
ST3	single-site	0.13	0.21	0.27	0.33	0.10
	individual	10.36	17.01	20.76	24.80	26.0
	switching	1.36	2.40	2.95	3.81	2.59
average	runtime(sec)	7300	22	13	1.5	0.4

2.5 Conclusions

In conclusion, we have developed a new extremely fast and simultaneously highly accurate phasing algorithm 2SNP based on 2-SNP haplotypes. We hope that it will be very useful for high-throughput genotype data processing, e.g., SNP Mapping Arrays (Affymetrix).

CHAPTER 3

DISEASE ASSOCIATION SEARCH

Accessibility of high-throughput genotyping technology allows genome/chromosome-wide association studies for common complex diseases. This chapter addresses two challenges commonly facing such studies [136]: (i) searching an *enormous amount* of possible gene interactions and (ii) finding *reproducible* associations. These challenges have been traditionally addressed in statistics [143] while here we apply computational approaches -- optimization and permutation tests. A complex risk factor is modeled as a subset of SNP's with specified alleles (MSC) and the optimization formulation asks for the one with the minimum p-value or the maximum odds ratio. We have applied our search methods to real case-control studies for several diseases (Crohn's disease [138], autoimmune disorder [163], tick-born encephalitis [134], lung cancer [144], and rheumatoid arthritis [135]). Proposed methods are compared favorably to the exhaustive search -- they are faster, find more frequently statistically significant risk factors.

3.1 Introduction

Recent improvement in accessibility of high-throughput DNA sequencing brought a great deal of attention to disease association and susceptibility studies. Successful genome-wide searches for disease-associated gene variations have been recently reported [140, 144]. However, complex diseases can be caused by combinations of several unlinked gene variations.

This chapter addresses computational challenges of genotype data analysis in epidemiological studies including selecting of informative SNPs, searching for diseases associated SNPs, and predicting of genotype susceptibility.

Disease association studies analyze genetic variation across exposed to a disease (diseased) and healthy (non-diseased) individuals. The difference between individual DNA sequences occurs at a single-base sites, in which more than one allele is observed across population. Such variations are called single nucleotide polymorphisms (SNPs). The number of simultaneously typed SNPs for association and linkage studies is reaching 10^6 for SNP Mapping Arrays [131]. High density maps of SNPs as well as massive DNA data with large number of individuals and number of SNPs become publicly available [156]. Diploid organisms, like human, have two near-identical copies of each chromosome. Most genotyping techniques (e.g., SNP Mapping Arrays [131]) do not provide separate SNP sequences (*haplotypes*) for each of the two chromosomes. Instead, they provide SNP sequences (*genotypes*) representing mixtures of two haplotypes -- each site is defined by an unordered pair of allele readings, one from each haplotype -- while haplotypes are computationally inferred from genotypes [101]. To genotype data we refer as unphased data and to haplotype data we refer as phased data. The disease association study analyze data given as genotypes or haplotypes with disease status.

Several challenges in genome-wide association studies of complex diseases have not yet been adequately addressed [137]: interaction between non-linked genes, multiple independent causes, multiple testing adjustment, etc. Since complex common diseases can be caused by multi-loci interactions two-loci analysis can be more powerful than traditional one-by-one SNP association analysis [143]. Multi-loci analysis is expected to find even deeper disease-associated interactions. The computational challenge (as pointed in [137]) is caused by the dimension catastrophe. Indeed, two-SNP interaction analysis (which can be more powerful than traditional one-by-one SNP association analysis [143]) for a genome-wide scan with 1 million SNPs (3 kb cover age) will afford 10^{12} possible pair wise tests. Multi-SNP interaction analysis reveals even deeper disease-associated interactions but is usually computationally infeasible and its statistical significance drastically decreases after multiple testing adjustment [121, 125]

Disease association analysis searches for Risk Factors (RF) modeled as Multi-SNP combinations with frequency among diseased individuals (cases) considerably higher than among non-diseased individuals (controls). Only statistically significant SNPs (whose frequency distribution has p-value less than 0.05) are reported. Successful as well as unsuccessful searches for SNPs with statistically significant association have been recently reported for different diseases and different suspected human genome regions (see e.g. [151]). Unfortunately, reported findings are frequently not reproducible on different populations. It is believed that this happens because the p-values are unadjusted to multiple testing -- indeed, if the reported SNP is found among 100 SNPs then the probability that the SNP is associated with a disease by mere chance becomes roughly 100 times larger.

3.1.1 Optimization Approach

This chapter discusses optimization approach to resolve these issues instead of traditionally used statistical and computational intelligence methods. In order to handle data with huge number of SNPs, one can extract informative (indexing) SNPs that can be used for (almost) lossless reconstructing of all other SNPs[130]. To avoid information loss, index SNPs are chosen based on how well the other non-index SNPs can be reconstructed. The corresponding **informative SNP selection problem (ISSP)** can be formulated as follows

(See Figure 3.1). Given a sample *S* of a population P of *individuals* (either haplotypes or genotypes) on *m* SNPs, select positions of *k* ($k < m$) SNPs such that for any individual, one can predict non-selected SNPs from these *k* selected SNPs. The *Multiple Linear Regression based* MLR-tagging algorithm

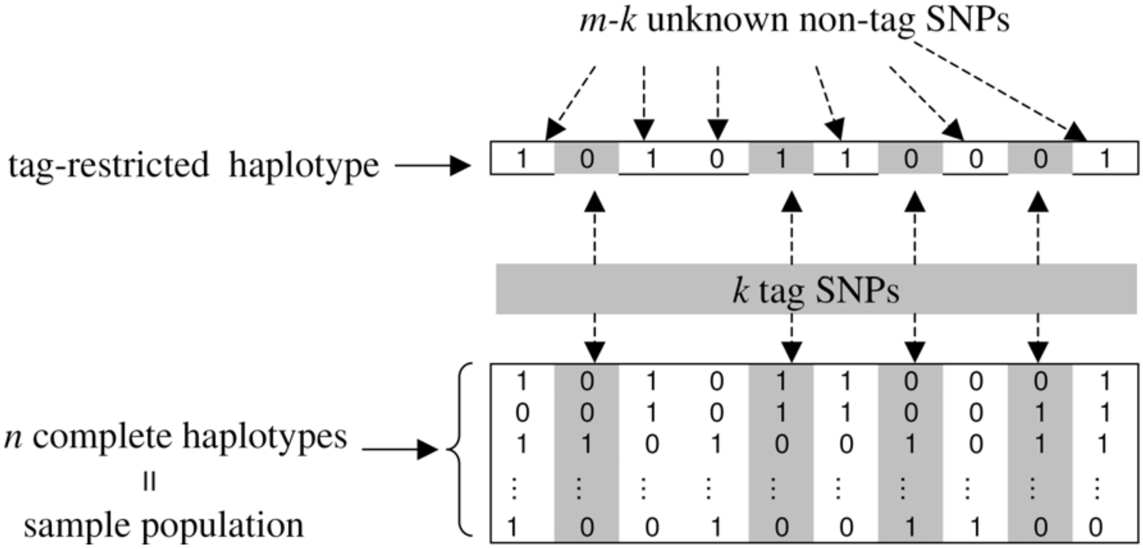


Figure 3.1 Informative SNP Selection Problem. The shaded columns correspond to k tag SNPs and the clear columns correspond to non-tag SNPs. The unknown (m – k) non-tag SNP values in tag-restricted individual (top) are predicted based on the known k tag values and complete sample population.

[112] solves the optimization version of ISSP which asks for k informative SNPs *minimizing the prediction error* measured by the number of incorrectly predicted SNPs. The number of tags (informative SNPs) k depends on the desirable data size. More tags will keep more genotype information while less tags allows deeper analysis and search.

In the reduced set of SNPs one can search for deeper disease association. In this chapter we discuss the **optimization problem of finding the most disease-associated multi-SNP combination** for given case-control data. Since it is plausible that common diseases can have also genetic resistance factors, one can also search for the *most disease-resistant multi-SNP* *combination*. Association of risk or resistance factors with the disease can be measured in terms of p-value of the skew in case and control frequencies, risk rates or odds rates. Here we first concentrate on two association measurements: p-value of the skew in case and control frequencies and *positive predictive value* (PPV) which is the frequency of case individuals among all individuals with a given multi-SNP combination. This optimization problem is NP-hard and can be viewed as a generalization of the maximum independent set problem. A fast *complimentary greedy search* proposed in [133] is compared with the *exhaustive search* and *combinatorial search* that has been discussed in [132]. Although complimentary greedy search cannot guarantee finding of close to optimum MSCs, in the experiments with real data, it finds MSCs with non-trivially high PPV. For example, for Crohn's disease data [138], complimentary greedy search finds in less than second a case-free MSC containing 24 controls, while exhaustive and combinatorial searches need more than 1 day to find case-free MSCs with at most 17 controls.

We have applied our search methods (exhaustive, combinatorial and complimentary greedy searches [132, 133], alternating combinatorial search and a randomized of complimentary greedy search) for finding MSC with largest odds ratio on real case-control studies for several diseases (Crohn's disease [138], autoimmune disorder [163], tick-born encephalitis [134], lung cancer [144], and rheumatoid arthritis [135]). Proposed methods are compared favorably to the exhaustive search -- they are faster, find more frequently statistically significant risk factors.

We next model *atomic* risk factors (ARF's) (i.e., SNP risk factors that cannot be split into simpler factors) for common diseases as diplotypes, i.e., subsets of SNP's with specified alleles. Our first optimization formulations ask for a diplotype the most tightly associated with the disease (i.e., minimizing p-value) [132] and having the highest positive predictive [133]. In contrast, epidemiologists measure the quality of risk factors in case-control studies by *odds* *ratio* defined as the ratio of the odds of disease occurring in the exposed (to the risk factor) group to the odds of it occurring in unexposed group. Thus, we ask for an ARF with the maximum odds ratio. We show connection of this problem with the known Red-Blue Set Cover problem [166] whose special case is equivalent to the weighted set cover problem. The best known heuristics for the weighted set cover problem is a well-known greedy algorithm. Thus, in order to find ARF with the maximum odds ratio, we propose to apply the complementary greedy search (CGS) algorithm [133] originally designed for finding ARF's with highest positive predictive value.

We next propose to consider more complex but also more relevant SNP risk factors, so called *k*-relaxed atomic risk factors, for which exposed individuals can deviate in at most k sites from a given diplotype. We generalize the complimentary greedy algorithm (*k*-CGS) to find *k*-relaxed atomic risk factors with fixed *k*. Our experiments show advantage of the new method over CGS in finding risk factors with significantly higher odds ratio and association with the disease.

We then introduce even more general risk factors, so called weighted relaxed atomic risk factors. The individuals exposed to such factors should be within *weighted* Hamming distance *k* from a given diplotype. We also proposed a novel heuristic (WCGS) for finding weighted relaxed ARF with odds ratio.

We have applied and cross-validated CGS [133], k-CGS, and WCGS methods for finding atomic risk factors (ARF), k-relaxed ARF's and weighted relaxed ARF's, respectively, with large odds ratios on real case-control studies for several diseases (Crohn's disease [138], autoimmune disorder [163], tick-born encephalitis [134], lung cancer [144], and rheumatoid arthritis [135]). The CGS algorithm finds non-trivially ARF's with significantly high odds ratios. For the smallest dataset (tick-born encephalitis) CGS found an optimal solution which has been confirmed using an exact integer linear program. The k-CGS and WCGS methods found SNP risk factors that are statistically significant even after multiple-testing adjustment on all data while CGS could not find significant ARF on two of these data. The found relaxed atomic risk factors explain much higher number of cases, 1.5-4 times larger than atomic ARFs found by CGS. The new methods also have significantly higher leave-half-out cross-validation rate.

3.1.2 Overview

In the next section the disease association search problem is formulated, the searching algorithms and their quality measures are described, the optimization version of disease association search for most disease associated MSC is reformulated as an independent set problem and the fast complimentary greedy search and randomized complimentary greedy search algorithms are given. Then we formally introduce the problem of finding a risk factor with maximum odds ratio and apply our searching methods to solve this problem. We further model genetic RFs as atomic RFs (ARFs), then we generalize it to *k*-relaxed ARFs (*k*-RARF) and weighted RARF (WRARF). For the introduced models we formulate optimization problems and give heuristics for finding maximum odds ratio atomic, k-relaxed atomic and weighted relaxed atomic risk factors, respectively. We next describe available real datasets and conclude with comparison and cross-validation of described heuristics.

3.2 Methods for Disease Association Search

In this section the search of statistically significant disease-associated multi-SNP combinations is formally described. Then the corresponding optimization problem is formulated and its complexity is discussed. The combinatorial search introduced in [132] and the fast complementary greedy search introduced in [133] are described.

The typical case-control or cohort study results in a sample population *S* consisting of *n* individuals represented by values of *m* SNPs and the disease status. Since it is expensive to obtain individual chromosomes, each SNP value attains one of three values 0, 1 or 2, where 0's and 1's denote homozygous sites with major allele and minor allele, respectively, and 2's stand for heterozygous sites. SNPs with more than 2 alleles are rare and can be conventionally represented as biallelic. Thus the sample S is an (0, 1, 2)-valued $n \times (m + 1)$-matrix, where each row corresponds to an individual, each column corresponds to a SNP except last column corresponding to the disease status (0 stands for case and 1 stands for control). Let *H* and *D* be the subsets of rows with controls (healthy) and cases (disease), respectively.

The standard measure of disease association in case-control studies is the *odds ratio* (OR) – the odds of a risk factor occurring in cases to the odds of it occurring in controls. Formally, let *C* be the subset of genotypes of *S* exposed to a risk factor (further referred as a *cluster*) and let $h(C) = C \cap H$ be the set of controls and $d(C) = C \cap D$ be the set of cases in cluster *C*. Then the odds ratio of a risk factor which forms cluster *C* is computed as

$$OR = \frac{|\,d(C)\,|\,/\,|\,d(D-C)\,|}{|\,h(C)\,|\,/\,|\,h(H-C)\,|}$$

The odds ratio is ill-defined when $|h(C)| = 0$, i.e., when the corresponding cluster is control free. In order to define and compare OR for such cases, we follow standard practice of adding 0.5 to all 4 values $|d(C)|$, $|d(D - C)|$, $|h(C)|$, $|h(H - C)|$.

The association of a risk factor with the disease status can be also measured with the following parameters:

- relative risk $RR = (|\,d(C)\,|*|\,S - C\,|)/(|\,D - d(C)\,|*|\,C\,|)$ (for cohort studies) positive predictive value $PPV =|\,d(C)\,|\,/\,|\,C\,|$ (for susceptibility prediction)
- p-value of the partition of the cluster into cases and controls:

$$p = \sum_{k=0}^{|d(C)|} \binom{|C|}{k} \left(\frac{|D|}{|S|}\right)^{k} \left(\frac{|H|}{|S|}\right)^{|C|-k}$$

A complex risk factor in SNP case-control study is modeled as a subset of SNP's with specified alleles (diplotype, multi-SNP combination (MSC)), and general disease association searches for all risk factors with OR above (or below) a certain threshold. The common formulation is to find all risk factors with p-value below 0.05. An optimization version can be formulated as follows.

Most Associated Risk Factor Problem. *Given a genotype case/control study on a sample population S, find an risk factor most associated with the disease.*

Since associated risk factors are searched among exponentially many risk factors (SNP combinations), the computed p-value (significance) requires adjustment for multiple testing (MT) which can be done with simple but overly pessimistic Bonferroni correction or computationally extensive but more accurate randomization method. In our study we use randomization method which computes the p-value of a risk factor by comparing its OR with OR's of 10^4 risk factors found by searching method on 10^4 randomized datasets. A randomized dataset is obtained from original dataset by random swapping of the disease statuses. If OR of a risk factor is larger than 9500 (95%) of OR's from randomized datasets then the MT-adjusted p-value of this risk factor is <5%.

3.2.1 Full Searches for Associated MSCs

Risk and resistance factors representing gene variation interaction can be defined in terms of SNPs as a multi-SNP combination (MSC) which is a subset of SNP with fixed alleles values. General disease association searches for all MSCs with one of the parameters above (or below) a certain threshold. The common formulation is to find all MSCs with adjusted p-value below 0.05.

3.2.1.1 Exhaustive Search (ES)

The search for disease-associated MSCs among all possible combinations can be done by the following *exhaustive search*. In order to find a MSC with the p-value of the frequency distribution below 0.05, one should check all one-SNP, two-SNP, ..., m-SNP combinations. The checking procedure takes runtime $\mathrm{O}\left(n\sum_{k=1}^{m}\binom{m}{k}3^{k}\right)$ for unphased combinations since there are 3 possible SNP values {0, 1, 2}. Similarly, for phased combination, the runtime is $\mathrm{O}\left(n\sum_{k=1}^{m}\binom{m}{k}3^{k}\right)$ since there only 2 possible SNP values. The exhaustive search is infeasible even for small number of SNPs, therefore the search is limited to the small number of SNPs, i.e., instead of searching all MSCs, one can search only containing at most k = 1, 2, 3 SNPs. We refer to k as *search level* of exhaustive search. Also one can reduce the depth (number of simultaneously interacting SNPs) or reduce m by extracting informative SNPs from which one can reconstruct all other SNPs. The MLR-tagging is used to choose maximum number of index SNPs that can be handled by ES in a reasonable computational time.

3.2.1.2 Combinatorial Search (CS)

Here we discuss suggested in [132] search method for disease-associated MSCs which avoids insignificant MSCs or clusters without loosing significant ones. CS searches only for closed MSCs, where closure is defined as follows. The closure $\overline{C}$ of MSC with cluster C is an MSC whose cluster has minimum control elements $h(\overline{C})$ and the same case elements $d(\overline{C}) = d(C)$. $\overline{C}$ can be easily found by incorporating into its set of SNPs all SNPs with common values among all case individuals in C .

The combinatorial search proposed [132] finds the best p-value of frequency distribution of the closure of each single-SNP, after that it searches for the best frequency distribution p-value among closure of all 2-SNP combinations and so on. The procedure stops after all closure of all k-SNP combinations ($k < m$) are checked. The corresponding *search level* is the number of SNPs selected for closuring, e.g., on the level 2 of searching combinatorial search will test closure of all 2-SNP combinations for association with a disease. Because of the closure, for the same level of searching combinatorial search finds better association than exhaustive search. However, the above combinatorial search is as slow as exhaustive search.

A faster implementation of this method avoids checking MSCs which are not (and cannot lead to) statistically significant ones. Formally, a MSC I is called an intersection of MSCs with clusters C_1 and C_2 if $d(I) = d(C_1) \cap d(C_2)$ and $|h(I)|$ is minimized. An MSC is called trivial if its unadjusted p-value is larger than 0.05 even if the set $h(C)$ would be empty. Note that intersection of a trivial MSC with another is trivial.

A faster implementation of the combinatorial search is as follows:

1. Compute a set G_1 of all 1-SNP closed MSCs, exclude trivial combinations.

2. Compute sets G_k of all pair wise intersections of the MSCs from G_{k-1}, exclude trivial combinations and already existing in $G_1 \cup G_2 \cup ... \cup G_{k-1}, k = 2..N$.

3. For each G_k output MSCs whose unadjusted $p < 0.05$.

Still, in order to find all MSCs associated with a disease one has to check all possible SNP combinations with all possible SNP values. This searching approach is also computationally intensive and step 2 from the algorithm can generate an exponential number of MSCs. However, closure avoids generation and checking of non-significant MSCs. Additionally, removing of trivial MSCs at each iteration of step 2 considerably reduces the number of newly generated MSCs. CS has been shown much faster than ES and capable of finding more significant MSCs than ES for equivalent search level.

3.2.1.3 Alternating Combinatorial Search (ACS)

This is a modification of CS which tries to further decrease the number of controls in a cluster C obtained by CS repeatedly using the following alternating closing and complimenting:

$$C \leftarrow \overline{(C - \overline{(C \cap D)})}$$

These iterations stop as soon as C stops changing.

3.2.2 Maximum Control-Free Cluster

Following [133], we next consider another optimization formulation corresponding to the general association search problem, e.g., find MSC with the minimum adjusted p-value. In particular, we focus on maximization of PPV. Obviously, the MSC with maximum PPV should not contain control individuals in its cluster and the problem can be formulated as follows:

Maximum Control-Free Cluster Problem (MCFCP). *Find a cluster C which does not contain control individuals and has the maximum number of case individuals.*

It is not difficult to see that this problem includes the maximum independent set problem. Indeed, given a graph $G = (V , E)$, for each vertex v we put into correspondence a case individual v' and for each edge $e = (u, v)$ we put into correspondence a control individual e' such that any cluster containing u' and v' should also contain e' (e.g., u' , v' , and e' are identical except one SNP where they have 3 different values 0,1, and 2). Obviously, the maximum independent set of G corresponds to the maximum control-free cluster and vice versa. Thus, one cannot reasonably approximate MCFCP in polynomial time for an arbitrary sample S. On the other hand, the sample S is not “arbitrary” -- it comes from a certain disease association study. Therefore, one may have hope that simple heuristics (particularly greedy algorithms) can perform much better than in the worst arbitrary case.

3.2.3 Maximum Odds Ratio Atomic Risk Factor

Since there are exponentially many genetic risk factors, in this section we concentrate only on basic atomic risk factors. We then give the corresponding optimization formulations and describe their relationship with the red-blue set cover problem. Finally, the complementary greedy algorithm inspired by the greedy algorithm is described.

In general, an arbitrary SNP risk factor can be expressed as a boolean formula over SNPs as follows. A genotype SNP value is a function over x and y which are SNP values in haplotypes (0 and 1 in haplotypes stand for major and minor alleles, respectively). These function are boolean -- a genotype equals 0 iff $g_0 = \overline{x \vee y}$ is true, equals 1 iff $g_1 = x \wedge y$ is true, and equals 2 iff $g_2 = x \oplus y$ is true. Then a risk factor is a disjunctive normal form (DNF) over g_0's, g_1's and g_2's for all SNPs. Substituting each negation with $\overline{g}_i = g_{i-1} \vee g_{i+1}, i = 0,1,2$, we can make sure that DNF does not contain negations. An atomic risk factor (ARF) corresponds to a single clause of DNF and cannot be further split into simpler risk factors. An atomic risk factor may be also viewed as a diplotype which is a subset of SNP-columns of S with fixed values.

An ARF can model multiple gene interactions resulting in a single cause of a disease (see Figure 3.2 (a)) -- several pathways may result in generating a certain substance (e.g., protein) P necessary for a normal cell functions. Several mutations should cut all such pathways to create a lack of P -- any pathway that is not destroyed by mutations can produce sufficient amount of P preventing disease development.

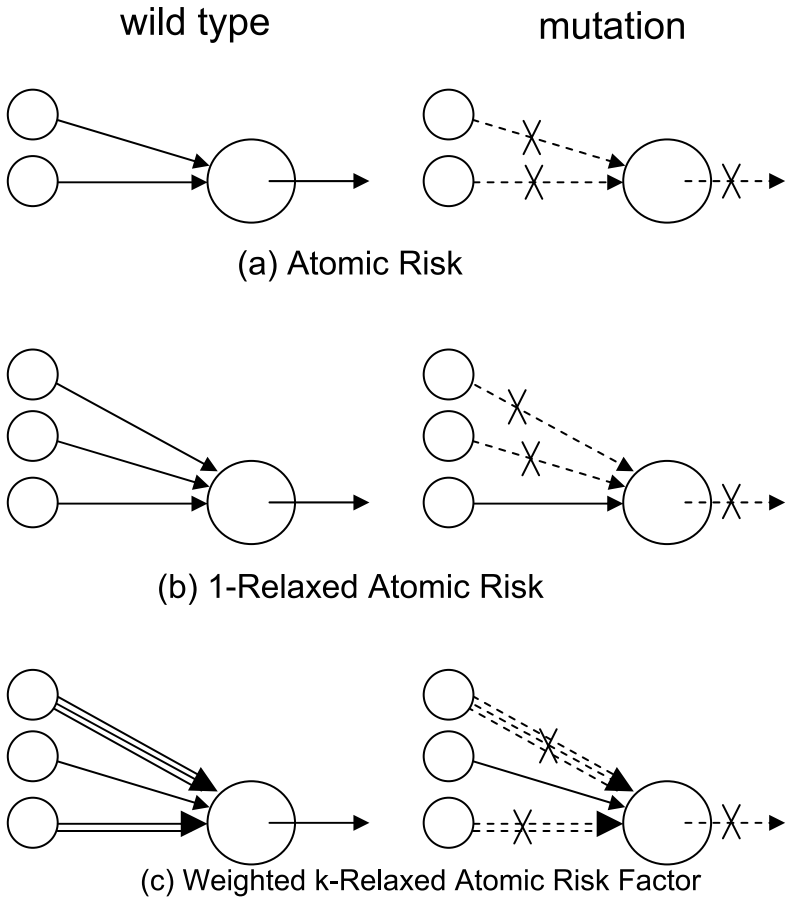

wild type
mutation
(a) Atomic Risk
(b) 1-Relaxed Atomic Risk
(c) Weighted k-Relaxed Atomic Risk Factor

Figure 3.2 Pathway motivation of Atomic Risk Factors, *k*-Relaxed ARFs, and Weighted Relaxed ARFs. (a) ARF models the case when all n pathways producing the substance P should be broken by mutations to cause the disease. (b) *k*-RARF models the case when it is necessary to have arbitrary $k + 1$ unaffected pathways out of total n pathways to produce sufficient amount of P and, therefore, breaking of any (n - k) pathways causes the disease. (c) WRARF models the same case as k-RARF but when different pathways can produce different integer number w of units of P.

Here we suggest to consider optimization formulation :

Maximum Odds Ratio Atomic Risk Factor (MORARF): *Given a genotype case-control study on a sample population S, find atomic risk factor with the maximum odds ratio.*

Our experiments with real and simulated genotype case-control studies showed that the solution for the MORARF problem is usually control-free, i.e., for the corresponding cluster C , $h(C) = 0$. The corresponding problem has been formulated in [133].

Maximum Control-Free Cluster (MCFC) Problem. *Given a genotype case-control study on a sample population S, find a maximum size control-free cluster* C .

It has been shown in [133] that this problem contains the independent set problem and therefore is NP-hard and is hard to approximate. Alternatively, instead of maximizing the number of cases in the cluster we can pursue the complimentary objective -- minimize number of cases outside of the cluster. The corresponding problem has been previously studied under the name red-blue set cover problem [166].

Red-Blue Set Cover (RBSC) Problem. *Given a set of "red" elements R, a set of "blue" elements B and a family of subsets* $\Phi \subseteq 2^{R \cup B}$ *, find a subfamily* $X \subseteq \Phi$ *covering all blue elements and minimizes the minimum number of covered red elements.*

When searching MCFC, the red elements are cases and blue elements are controls. The RBSC problem minimizes the number of red elements covered by a family of X while the MCFC problem maximizes non-covered red elements. The difference between RBSC and MCFC is the same as between the minimum vertex cover and maximum independent set problems.

The complexity of the RBSC problem has been studied in [166]. They show that even the case where every set contains only one blue and two red elements cannot be approximated to within $O(2^{\log^{1-\delta} n})$, where $\delta = 1/\log\log^{c} n$, for any constant c < 1/2 (where n = Φ). Unfortunately, the approximation algorithms proposed in [166] for RBSC do not have a non-trivial bounds in the MCFC context.

If restrictions on R of all sets from Φ coincide or do not intersect then the RBSC problem is equivalent to the weighted set cover problem where the cost of each set is equal to the number of red elements in it. The best known approximation algorithm for the weighted set cover problem is a well-known greedy algorithm.

3.2.3.1 Complimentary Greedy Search (CGS)

In graphs, instead of the maximum independent set one can search for its complement, the minimum vertex cover -- repeat picking and removing vertices of maximum degree until no edges left. In this case one can minimize the relative cost of covering (or removal) of control individuals, which is the number of removed case individuals. The corresponding heuristic for MCFCP and MORARF is the following

It can be inductively applied to the general RBSC problem -- iteratively choose the next set minimizing the ratio of newly covered red elements over newly covered blue element. This algorithm applied to the MCFC problem has been proposed in [133] as the Complimentary Greedy Search (CGS) (see Figure 3.3). CGS iteratively adds SNP value that minimizes the relative cost of removal of controls, where the cost is the number of removed cases by the same SNP value.

CGS can be directly applied to the MORARF problem rather than to MCFCP -- at each iteration update and output the diplotype with currently best OR. Our experiments show that each iteration of CGS always improved OR for all real dataset although sometimes first iterations can slightly violate monotonicity for randomized (bootstrapped) data (see Figure 3.4).

```
**Input:** Sample population $S$ partitioned into subsets $H$ and $D$
**Output:** Control-free cluster $C$ of ARF with diplotype $x$

$C \leftarrow S, x \leftarrow \emptyset$
Repeat until $h(C) > 0$
    Find a SNP $s$ and its value $v \in \{0, 1, 2\}$ with the cluster $C_{s=v}$
    minimizing $\frac{\triangle D}{\triangle H} = \frac{d(C)-d(C \cap C_{s=v})}{h(C)-h(C \cap C_{s=v})}$
    $C \leftarrow C \cap C_{s=v}, x \leftarrow x \cup \{s = v\}$
Output $C$ and $x$
```

Figure 3.3 Complimentary Greedy Search

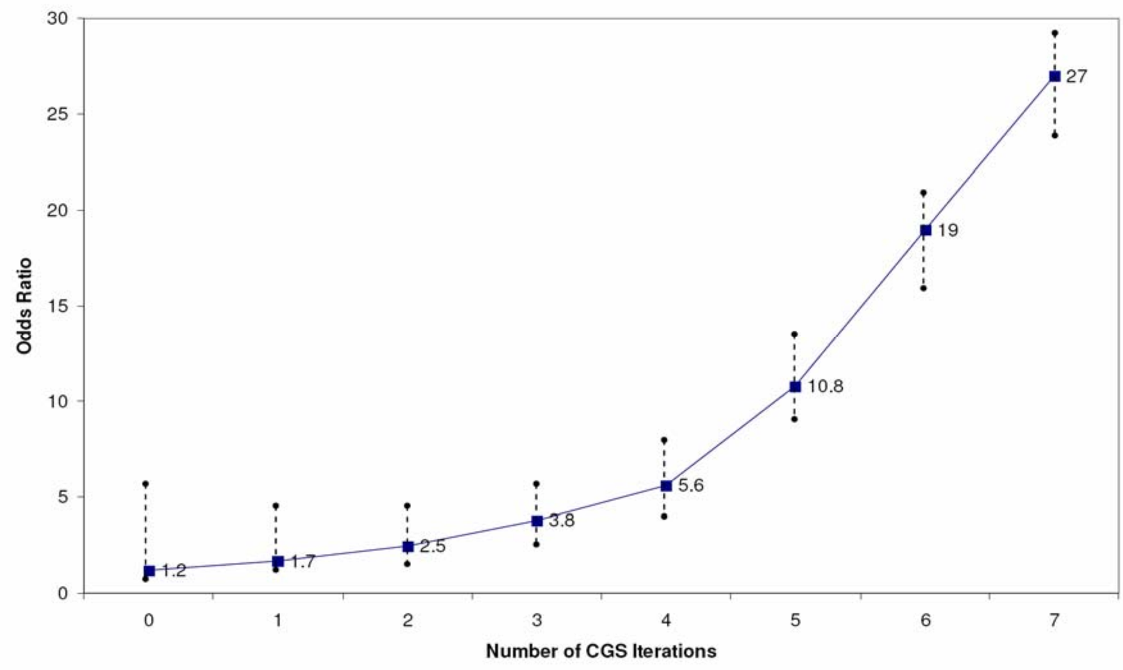


Figure 3.4 The value of the odds ratio of atomic risk factors (with vertical bars representing 95% confidence interval (CI)) on i-th iteration of complimentary greedy search on lung-cancer dataset (see Section 2.4.4).

Our experiments with several available real data sets (see Section 3.3) show that the complimentary greedy search can find atomic risk factors with no-trivially high OR. For the tick-borne encephalitis dataset (see Section 3.3.1) CGS found an optimal solution which has been confirmed using an exact integer linear program. Similarly to the maximum control-free cluster corresponding to the most expressed risk factor, one can also search for the maximum diseased-free cluster corresponding to the most expressed resistance factor. The experiments with three real data sets (see Section 3.3) show that the complimentary greedy search can find nontrivially large control-free and case-free clusters.

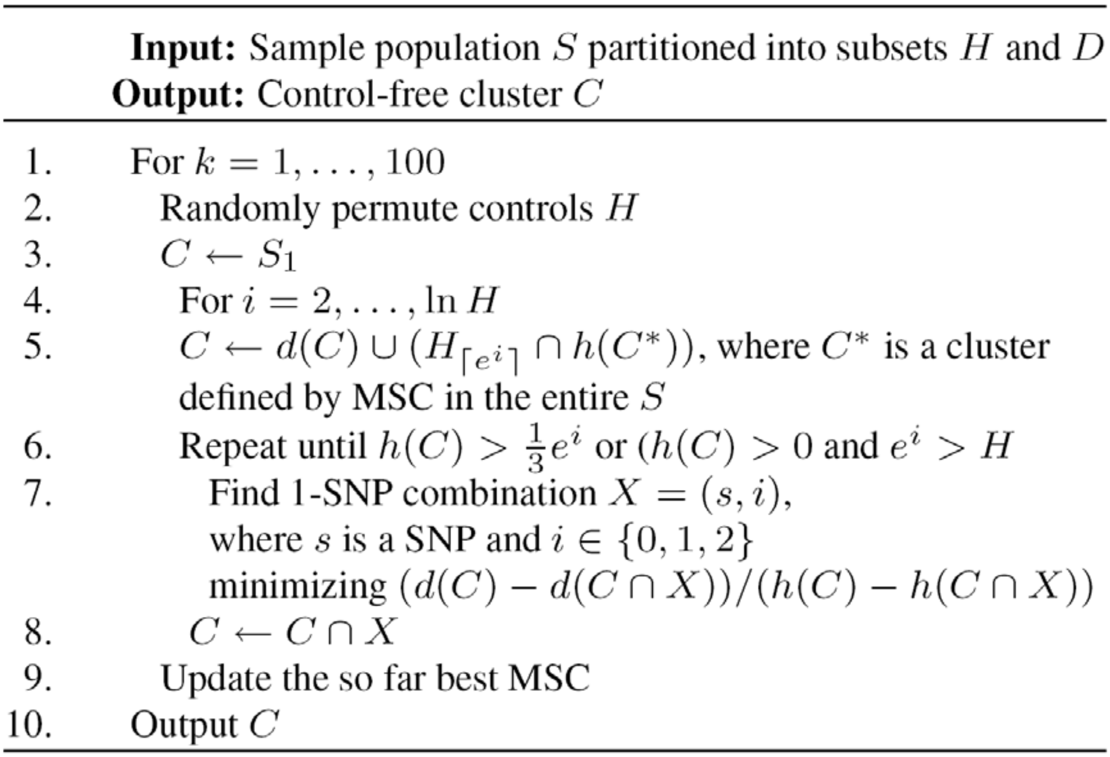
It is known that randomization improves quality of greedy algorithms. In this paper, we propose the following randomization enhancement of CGS

3.2.3.2 Randomized CGS (RCGS).

We expect randomization to change strictly greedy behavior of the CGS allowing to avoid the local maximum.

Let H_i be a subset of controls S_0 with indices $1,\dots i$. RCGS (see Figure 3.5) repeatedly permutes controls and gradually covers them with SNPs -- first it covers two thirds of $H_{\lceil e^2 \rceil}$ and then it covers two thirds of $H_{\lceil e^3 \rceil}$ and so on, finally on the iteration covers the entire H. RCGS outputs the maximum OR atomic risk factor out of 100 runs. It also outputs several atomic risk factors in case if adjusted p-value of their ORs are below 0.05. RCGS always finds risk factor with OR larger than OR of the risk factor found by CGS. However, it is not always more significant after multiple testing adjustment.

Experimental results (see Sections 3.3, 3.4) show that greedy heuristics are faster, find more frequently statistically significant risk factors and have significantly higher cross-validation rate.



```
Input: Sample population $S$ partitioned into subsets $H$ and $D$
Output: Control-free cluster $C$

For $k = 1, \ldots, 100$
  Randomly permute controls $H$
  $C \leftarrow S_1$
   For $i = 2, \ldots, \ln H$
   $C \leftarrow d(C) \cup (H_{\lceil e^i \rceil} \cap h(C^*))$, where $C^*$ is a cluster
   defined by MSC in the entire $S$
   Repeat until $h(C) > \frac{1}{3}e^i$ or $(h(C) > 0$ and $e^i > H$
     Find 1-SNP combination $X = (s, i)$,
     where $s$ is a SNP and $i \in \{0, 1, 2\}$
     minimizing $(d(C) - d(C \cap X))/(h(C) - h(C \cap X))$
    $C \leftarrow C \cap X$
  Update the so far best MSC
Output $C$
```

Figure 3.5 Randomized Complimentary Greedy Search

Let H_i be a subset of controls S_0 with indices 1,…*i*. RCGS (see Figure 3.5) repeatedly permutes controls and gradually covers them with SNPs -- first it covers two thirds of $H_{\lceil e^2 \rceil}$ and then it covers two thirds of $H_{\lceil e^3 \rceil}$ and so on, finally on the iteration covers the entire *H*. RCGS outputs the maximum OR atomic risk factor out of 100 runs. It also outputs several atomic risk factors in case if adjusted p-value of their ORs are below 0.05. RCGS always finds risk factor with OR larger than OR of the risk factor found by CGS. However, it is not always more significant after multiple testing adjustment.

Experimental results (see Sections 3.3, 3.4) show that greedy heuristics are faster, find more frequently statistically significant risk factors and have significantly higher cross-validation rate.

3.2.4 Maximum Odds Ratio k-Relaxed and Weighted Relaxed Atomic Risk Factors

In this section we give intuition and formally introduce a generalization of atomic risk factors with potentially larger OR, so called k-relaxed atomic risk factors. We then describe generalization of CGS to the problem of finding the k-relaxed atomic risk factor with the largest odds ratio.

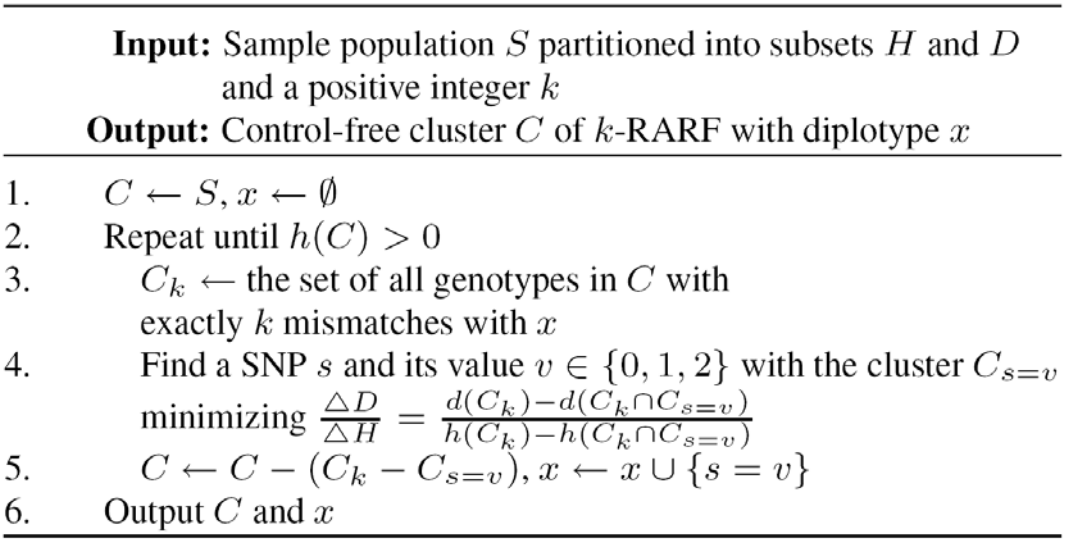
Below we give two different intuitions leading to the notion of *k*-relaxed atomic risk factor. Recall that ARF models a single complex cause of a disease -- several simultaneous mutations should cut all pathways producing a protein *P* (see Figure 3.2 (a)). Assume now that several individual causes are closely related. This may happen if the disease is caused by insufficient amount of a protein *P* rather than complete absence of *P*. Then breaking of sufficiently many pathways (but not necessarily all of them) can cause a disease. For example, Figure 3.2 (b) illustrates the case when out of 3 different pathways it is sufficient to break any 2 to cause a disease.

On the other hand, in the Hamming metric space of all diplotypes, one can expect that there exists a “representative" diplotype x associated with the disease that is completely surrounded by other diplotypes having significantly high OR or being control-free. The maximum radius ball around x consisting of diplotypes that are underrepresented in controls corresponds to such a risk factor. We say that a genotype g is exposed to a *k*-relaxed atomic risk factor (*k*-RARF) defined by an atomic risk factor with the diplotype x, if there are at most *k* SNPs in *g* mismatching corresponding SNPs in *x*. A cluster *C* of individuals exposed to a *k*-RARF with the diplotype *x* is a set of all genotypes in *S* whose restriction on all but *k* SNPs of *x* coincide with SNP-values in *x*.

Maximum Odds Ratio k-Relaxed Atomic Risk Factor (MORRARF) Problem: *Given a genotype case-control study on a sample population S and a positive integer k, find a k-relaxed atomic risk factor with the maximum odds ratio.*

3.2.4.1 *k*-Complimentary Greedy Search (*k*-CGS)

For the MORRAF problem, we have adjusted the complimentary greedy search applied to the MORARF problem. The adjusted algorithm is called *k*-CGS (see Figure 3.6).



```
 Input: Sample population $S$ partitioned into subsets $H$ and $D$
        and a positive integer $k$
Output: Control-free cluster $C$ of $k$-RARF with diplotype $x$

$C \leftarrow S, x \leftarrow \emptyset$
Repeat until $h(C) > 0$
    $C_k \leftarrow$ the set of all genotypes in $C$ with
    exactly $k$ mismatches with $x$
    Find a SNP $s$ and its value $v \in \{0, 1, 2\}$ with the cluster $C_{s=v}$
    minimizing $\frac{\triangle D}{\triangle H} = \frac{d(C_k)-d(C_k \cap C_{s=v})}{h(C_k)-h(C_k \cap C_{s=v})}$
    $C \leftarrow C - (C_k - C_{s=v}), x \leftarrow x \cup \{s = v\}$
Output $C$ and $x$
```

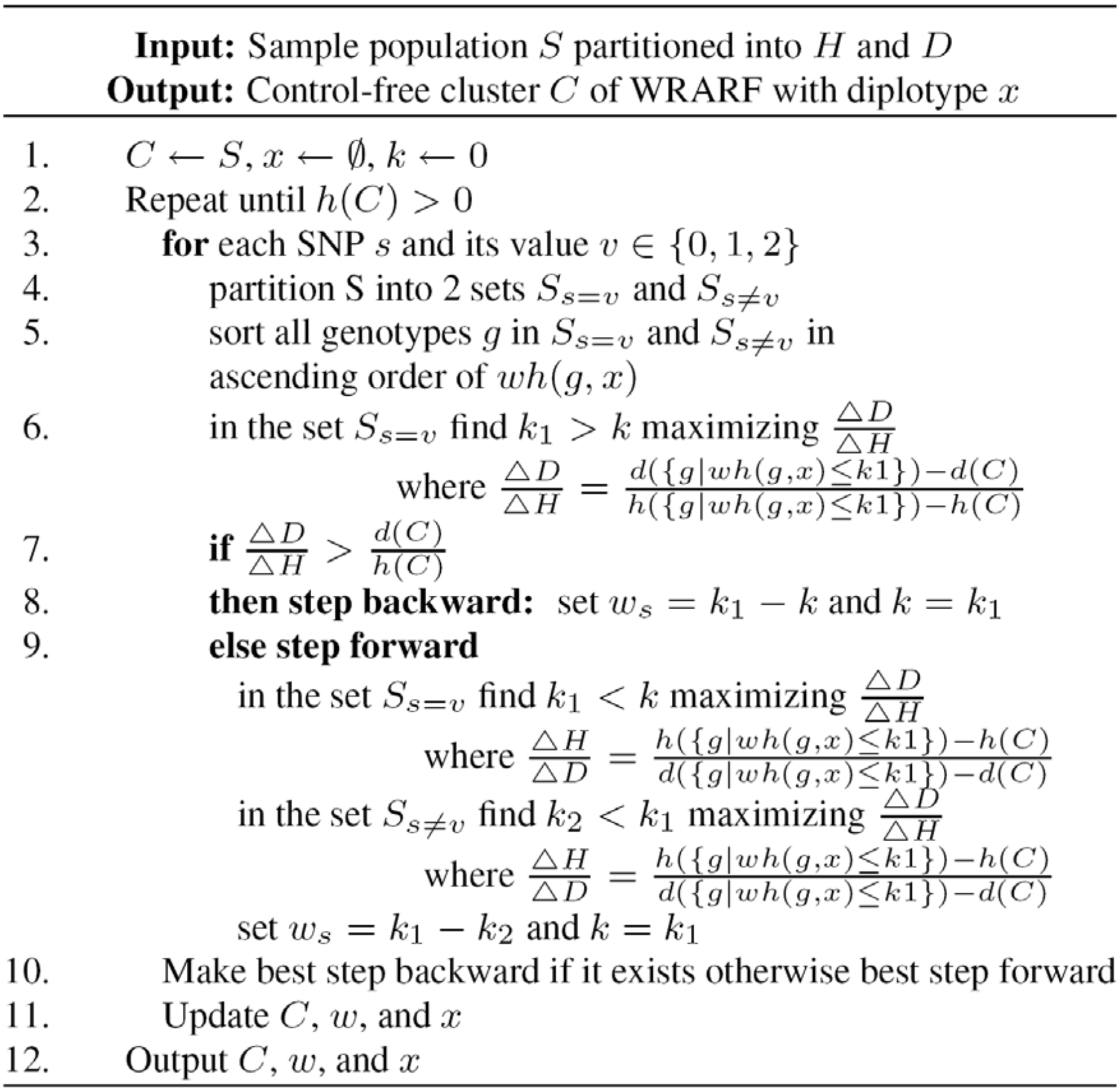
Figure 3.6 *k*-Complimentary Greedy Search (*k*-CGS)

3.2.4.2 Weighted Complimentary Greedy Search (WCGS)

In this section we further relax k-relaxed ARF's to weighted relaxed ARF's, formulate the corresponding optimization problem and give the corresponding complimentary greedy algorithm.

Similarly to k-RARF, we give two intuitions leading to the notion of weighted relaxed atomic risk factor. Recall that k-RARF models the case when breaking of sufficiently many pathways can cause a disease (see Figure 3.2 (b)). The weighted relaxed ARF models the case when different pathways produce different amount of the resulted protein P , and, therefore, different SNPs cancel different amount of P. For example, Figure 3.2 (c) illustrates the case when the first pathway produces 3 units of P while the second and the third pathways produce 1 and 2 units of P, respectively. Therefore, a SNP breaking the first pathway already cancel 3 units of P and causes a disease while breaking only the second or only the third pathway is not sufficient to disease development.

On the other hand, in the Hamming metric space of all diplotypes, a ball around a “representative” diplotype x consisting of diplotypes underrepresented may be stretched or shrunk along different axes (SNPs). Formally, given a diplotype x with positive integer weight w_s for each SNP s in x, the *w-weighted Hamming distance* from a genotype g to x equals the weighted number of mismatches between them, $wh(g,x)=\sum_{s\in x} w_s\delta_s$, where $\delta_s = 1$ if there is a mismatch in s and 0, otherwise. A cluster C of individuals exposed to a weighted k-relaxed risk factor (WRARF) with the diplotype x and weights w on SNPs of s is a set of all genotypes within w-weighted Hamming distance k from x.



```
Input: Sample population $S$ partitioned into $H$ and $D$
Output: Control-free cluster $C$ of WRARF with diplotype $x$

$C \leftarrow S, x \leftarrow \emptyset, k \leftarrow 0$
Repeat until $h(C) > 0$
    for each SNP $s$ and its value $v \in \{0, 1, 2\}$
        partition S into 2 sets $S_{s=v}$ and $S_{s\neq v}$
        sort all genotypes $g$ in $S_{s=v}$ and $S_{s\neq v}$ in
        ascending order of $wh(g, x)$
        in the set $S_{s=v}$ find $k_1 > k$ maximizing $\frac{\triangle D}{\triangle H}$
                where $\frac{\triangle D}{\triangle H} = \frac{d(\{g|wh(g,x) \leq k1\}) - d(C)}{h(\{g|wh(g,x) \leq k1\}) - h(C)}$
        if $\frac{\triangle D}{\triangle H} > \frac{d(C)}{h(C)}$
        then step backward: set $w_s = k_1 - k$ and $k = k_1$
        else step forward
          in the set $S_{s=v}$ find $k_1 < k$ maximizing $\frac{\triangle D}{\triangle H}$
                where $\frac{\triangle H}{\triangle D} = \frac{h(\{g|wh(g,x) \leq k1\}) - h(C)}{d(\{g|wh(g,x) \leq k1\}) - d(C)}$
          in the set $S_{s\neq v}$ find $k_2 < k_1$ maximizing $\frac{\triangle D}{\triangle H}$
                where $\frac{\triangle H}{\triangle D} = \frac{h(\{g|wh(g,x) \leq k1\}) - h(C)}{d(\{g|wh(g,x) \leq k1\}) - d(C)}$
          set $w_s = k_1 - k_2$ and $k = k_1$
    Make best step backward if it exists otherwise best step forward
    Update $C$, $w$, and $x$
Output $C$, $w$, and $x$
```

Figure 3.7 Weighted Complimentary Greedy Search

Weighted Complimentary Greedy Search (WCGS) on Figure 3.7 is a fast greedy heuristic for finding maximum odds ratio WRARF. It consists of forward and backward steps (see Figure 3.8). The forward step is similar to the iterations in CGS and *k*-CGS -- it finds the best thresholds $k_{s=v}$ in $S_{s=v}$ and $k_{s\neq v}$ in $S_{s\neq v}$, i.e., thresholds that reduce controls paying the least amount of removed cases per control and equalizing these thresholds by setting the weight to difference $k_{s=v}$ - $k_{s\neq v}$ (see Figure 3.8 (a)). The backward step does not have analogy -- it increases the number of cases so that the number of added controls per case is smaller than in the last forward step (see Figure 3.8 (b)). Whenever possible, WCGS makes step backward and if it is impossible it makes the step forward.

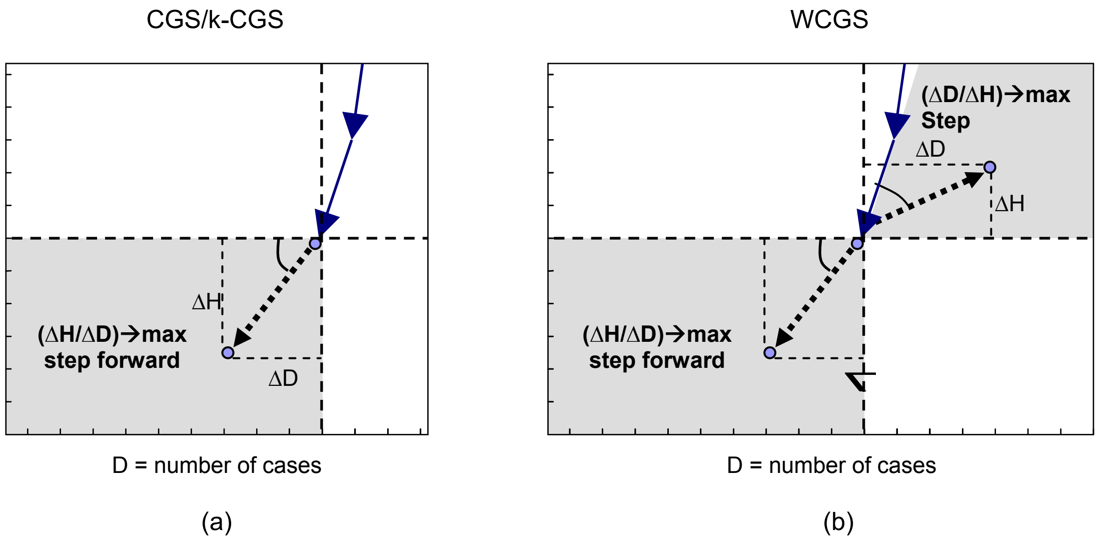


Figure 3.8 One greedy iteration of the CGS, k-CGS, and WCGS algorithms.

3.3 Experimental Results

In this section we discuss the results of methods for searching disease-associated multi-SNP combinations and report the number of found statistically significant atomic risk factors on real datasets. We first describe several datasets, then overview search methods and conclude with description and discussion of their performance. All experiments were ran on Processor Pentium 4 3.2Ghz, RAM 2Gb, OS Linux.

3.3.1 Data Sets

- *Crohn's disease (5q31)*: The data set Daly et al [138] is derived from the 616 kilobase region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohn's disease by genotyping 103 SNPs for 129 trios. All offspring belong to the case population, while almost all parents belong to the control population. In entire data, there are 144 case and 243 control individuals.

- *Autoimmune disorder*: The data set of Ueda et al [163] are sequenced from 330kb of human DNA containing gene CD28, CTLA4 and ICONS which are proved related to autoimmune disorder. A total of 108 SNPs were genotyped in 384 cases of autoimmune disorder and 652 controls.

- *Tick-borne encephalitis*: The tick-borne encephalitis virus-induced dataset of Barkash et al [100] consists of 41 SNPs genotyped from DNA of 21 patients with severe tickborne encephalitis virus-induced disease and 54 patients with mild disease.

- *Rheumatoid arthritis:* dataset [135] represents a dense panel of 2300 SNPs genotyped by Illumina for an approximately 10 kb region of chromosome 18q that showed evidence for linkage in the U.S. and French linkage scans. These markers were individually genotyped on 460 cases and 460 controls. Controls were recruited from a New York City population.

- *Lung cancer:* dataset [144] was obtained using genome-wide DNA pooling strategy in a group of German smokers clinically diagnosed with lung cancer and agematched healthy smokers. 83.715 SNPs had beed screened. 141 SNPs showed putative allelic imbalance between case and control DNA pools and were eventually genotyped in individual samples. Finally, dataset includes 322 male smokers with lung cancer and 273 healthy smokers genotyped in 141 SNPs.

- *HapMap datasets*: Regions ENr123 and ENm010 from 2 population: 45 Han Chinese from Beijing (HCB) and 44 Japanese from Tokyo (JPT) for three regions (ENm013, ENr112, ENr113) from 30 CEPH family trios obtained from HapMap ENCODE Project [156]. Two gene regions STEAP and TRPM8 from 30 CEPH family trios were obtained from HapMap.

The datasets have been phased using inhouse 2SNP software [101]. The missing data (16% in [138] and 10% in [163]) have been imputed in genotypes from the resulted haplotypes. We have also created corresponding haplotype datasets in which each individual is represented by a haplotype with the disease status inherited from the corresponding individual genotype.

3.3.2 Search Methods

Here we discuss the results of four methods for searching disease-associated MSCs on real phased and unphased datasets. The p-values of the frequency distribution of the found MSCs are used as a quality measurement. Then we report the number of found statistically significant atomic risk factors by six proposed searching methods.

Search Methods. We have compared the following 5 methods for search disease-associated MSCs.

- Exhaustive Search (**ES**)
- Indexed Exhaustive Search (**IES$_{30}$**): exhaustive search on the indexed datasets obtained by extracting 30 indexed SNPs with MLR based tagging method [113]
- Combinatorial Search (**CS**)
- Indexed Combinatorial Search (**ICS$_{30}$**): combinatorial search on the indexed datasets obtained by extracting 30 indexed SNPs with MLR based tagging method [113]
- Complimentary Greedy Search (**CGS**): approximate search for the maximum

control-free or maximum odds ration cluster and corresponding Atomic Risk Factor.

- *k*-Relaxed Complimentary Greedy Search (***k*-CGS**): approximate search for the maximum odds ration k-Relaxed Atomic Risk Factor.
- Weighted Complimentary Greedy Search (**WCGS**): approximate search for the maximum odds ration Weighted *k*-Relaxed Atomic Risk Factor.

Significant MSCs have been found only on levels 1 and 2 because adjusted p-value grows with the level. The size of the datasets is enough large to make exhaustive search impossible even for a combination of 6 SNPs.

3.3.3 Comparison of Full Searches

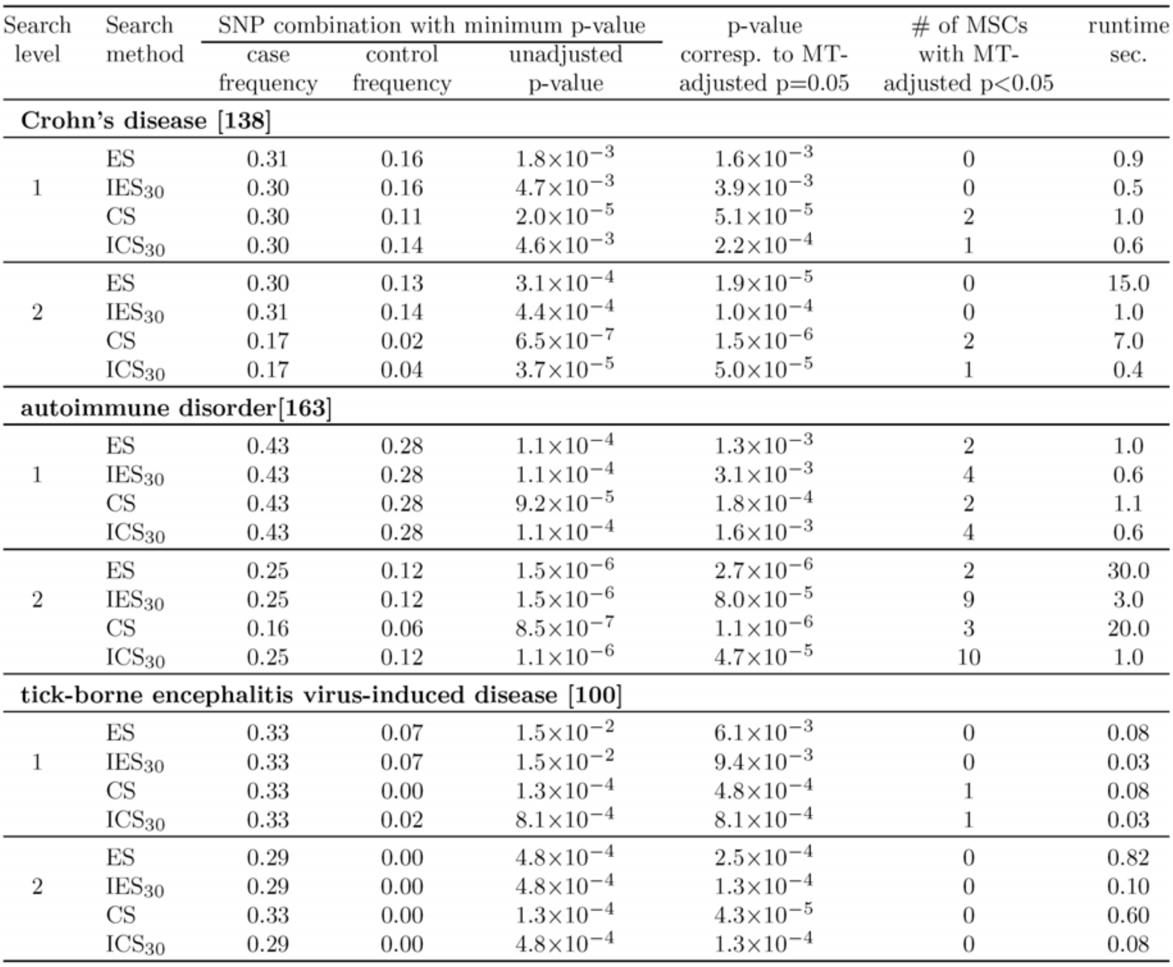
The quality of searching methods is compared by the number of found statistically significant MSCs (see the 7th column) in genotypes (see Table 3.1) and haplotypes (see Table 3.2). Since statistical significance should be adjusted to multiple testing, we report for each method and data set the 0.05 threshold adjusted for multiple testing (this threshold is computed by randomization and given in the third column of Tables 3.1 and 3.2). In the 3d, 4th and 5th columns, we give the frequencies of the best MSC among case and control population and the unadjusted p-value, respectively

3.3.4 Comparison of Approximate and Full Searches

We have compared IES_{30} and ICS_{30} with CGS (see Section 3.2) for search disease associated multi-SNP combinations with the largest PPV.

The quality of searching methods is compared by the PPV of found clusters as well as their statistical significance Table 3.3.

Table 3.1 Comparison of 4 methods for searching disease-associated multi-SNPs combinations for unphased genotype. Index value denote number of indexed SNPs. datasets.

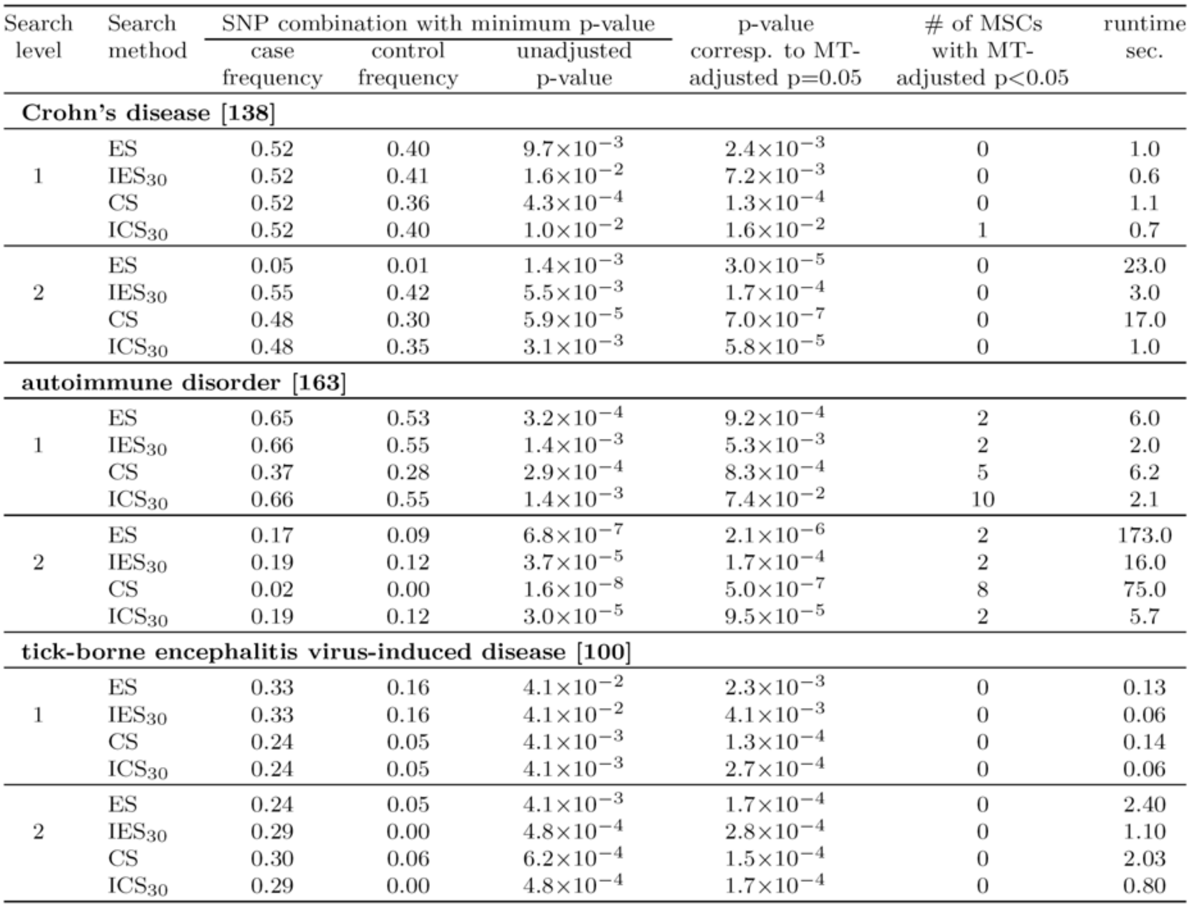


Search	Search	SNP combination with minimum p-value			p-value	# of MSCs	runtime
level	method	case	control	unadjusted	corresp. to MT-	with MT-	sec.
		frequency	frequency	p-value	adjusted p=0.05	adjusted p<0.05	
Crohn's disease [138]							
1	ES	0.31	0.16	1.8×10^{-3}	1.6×10^{-3}	0	0.9
	IES_{30}	0.30	0.16	4.7×10^{-3}	3.9×10^{-3}	0	0.5
	CS	0.30	0.11	2.0×10^{-5}	5.1×10^{-5}	2	1.0
	ICS_{30}	0.30	0.14	4.6×10^{-3}	2.2×10^{-4}	1	0.6
2	ES	0.30	0.13	3.1×10^{-4}	1.9×10^{-5}	0	15.0
	IES_{30}	0.31	0.14	4.4×10^{-4}	1.0×10^{-4}	0	1.0
	CS	0.17	0.02	6.5×10^{-7}	1.5×10^{-6}	2	7.0
	ICS_{30}	0.17	0.04	3.7×10^{-5}	5.0×10^{-5}	1	0.4
autoimmune disorder[163]							
1	ES	0.43	0.28	1.1×10^{-4}	1.3×10^{-3}	2	1.0
	IES_{30}	0.43	0.28	1.1×10^{-4}	3.1×10^{-3}	4	0.6
	CS	0.43	0.28	9.2×10^{-5}	1.8×10^{-4}	2	1.1
	ICS_{30}	0.43	0.28	1.1×10^{-4}	1.6×10^{-3}	4	0.6
2	ES	0.25	0.12	1.5×10^{-6}	2.7×10^{-6}	2	30.0
	IES_{30}	0.25	0.12	1.5×10^{-6}	8.0×10^{-5}	9	3.0
	CS	0.16	0.06	8.5×10^{-7}	1.1×10^{-6}	3	20.0
	ICS_{30}	0.25	0.12	1.1×10^{-6}	4.7×10^{-5}	10	1.0
tick-borne encephalitis virus-induced disease [100]							
1	ES	0.33	0.07	1.5×10^{-2}	6.1×10^{-3}	0	0.08
	IES_{30}	0.33	0.07	1.5×10^{-2}	9.4×10^{-3}	0	0.03
	CS	0.33	0.00	1.3×10^{-4}	4.8×10^{-4}	1	0.08
	ICS_{30}	0.33	0.02	8.1×10^{-4}	8.1×10^{-4}	1	0.03
2	ES	0.29	0.00	4.8×10^{-4}	2.5×10^{-4}	0	0.82
	IES_{30}	0.29	0.00	4.8×10^{-4}	1.3×10^{-4}	0	0.10
	CS	0.33	0.00	1.3×10^{-4}	4.3×10^{-5}	0	0.60
	ICS_{30}	0.29	0.00	4.8×10^{-4}	1.3×10^{-4}	0	0.08

3.3.5 Significance of Found Atomic Risk Factors

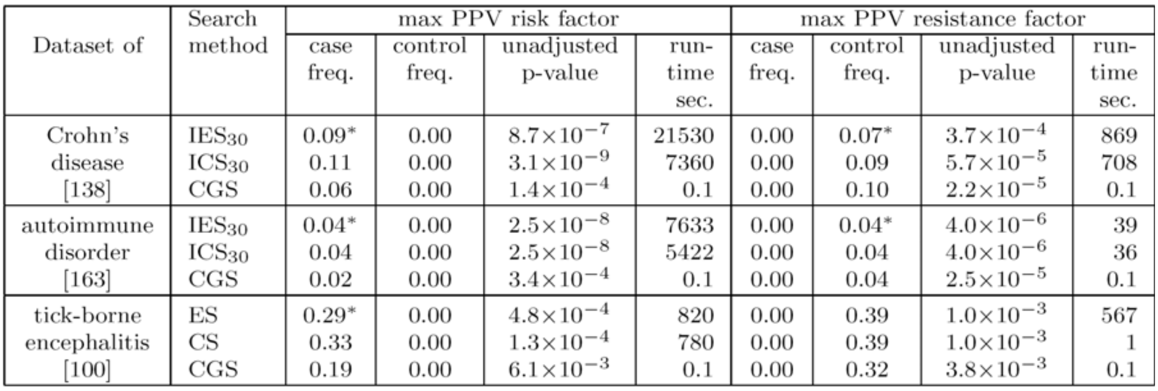
The Table 3.4 compares search heuristics ES(1), CS(1), ACS(1) (1-level exhaustive, combinatorial and alternating combinatorial searches), their 2-level counterparts, and CGS with RCGS applied to all 5 real datasets. For (almost) all datasets the OR of methods are ordered as follows: ES(1), CS(1), ES(2), ACS(1), CS(2), ACS(2), RCGS. The place of CGS is mostly between CS(2) and RCGS. The same order for heuristics holds for the lower limit of 95% CI of OR and unadjusted p-value. All methods except CGS also can report all found atomic risk factors with multiple-testing adjusted p-value less than 5%. On the most of data our new methods can find more statistically significant atomic risk factors. Note that ES, CS, and ACS are much slow than greedy heuristics.

Table 3.2 Comparison of 4 methods for searching disease-associated multi-SNPs combinations for phased genotype. Index value denote number of indexed SNPs. datasets.



Search level	Search method	SNP combination with minimum p-value			p-value corresp. to MT-adjusted p=0.05	# of MSCs with MT-adjusted p<0.05	runtime sec.
		case frequency	control frequency	unadjusted p-value			
Crohn's disease [138]							
1	ES	0.52	0.40	9.7×10^{-3}	2.4×10^{-3}	0	1.0
	IES_{30}	0.52	0.41	1.6×10^{-2}	7.2×10^{-3}	0	0.6
	CS	0.52	0.36	4.3×10^{-4}	1.3×10^{-4}	0	1.1
	ICS_{30}	0.52	0.40	1.0×10^{-2}	1.6×10^{-2}	1	0.7
2	ES	0.05	0.01	1.4×10^{-3}	3.0×10^{-5}	0	23.0
	IES_{30}	0.55	0.42	5.5×10^{-3}	1.7×10^{-4}	0	3.0
	CS	0.48	0.30	5.9×10^{-5}	7.0×10^{-7}	0	17.0
	ICS_{30}	0.48	0.35	3.1×10^{-3}	5.8×10^{-5}	0	1.0
autoimmune disorder [163]							
1	ES	0.65	0.53	3.2×10^{-4}	9.2×10^{-4}	2	6.0
	IES_{30}	0.66	0.55	1.4×10^{-3}	5.3×10^{-3}	2	2.0
	CS	0.37	0.28	2.9×10^{-4}	8.3×10^{-4}	5	6.2
	ICS_{30}	0.66	0.55	1.4×10^{-3}	7.4×10^{-2}	10	2.1
2	ES	0.17	0.09	6.8×10^{-7}	2.1×10^{-6}	2	173.0
	IES_{30}	0.19	0.12	3.7×10^{-5}	1.7×10^{-4}	2	16.0
	CS	0.02	0.00	1.6×10^{-8}	5.0×10^{-7}	8	75.0
	ICS_{30}	0.19	0.12	3.0×10^{-5}	9.5×10^{-5}	2	5.7
tick-borne encephalitis virus-induced disease [100]							
1	ES	0.33	0.16	4.1×10^{-2}	2.3×10^{-3}	0	0.13
	IES_{30}	0.33	0.16	4.1×10^{-2}	4.1×10^{-3}	0	0.06
	CS	0.24	0.05	4.1×10^{-3}	1.3×10^{-4}	0	0.14
	ICS_{30}	0.24	0.05	4.1×10^{-3}	2.7×10^{-4}	0	0.06
2	ES	0.24	0.05	4.1×10^{-3}	1.7×10^{-4}	0	2.40
	IES_{30}	0.29	0.00	4.8×10^{-4}	2.8×10^{-4}	0	1.10
	CS	0.30	0.06	6.2×10^{-4}	1.5×10^{-4}	0	2.03
	ICS_{30}	0.29	0.00	4.8×10^{-4}	1.7×10^{-4}	0	0.80

Table 3.3 Comparison of three methods for searching the disease-associated and disease-resistant multi-SNPs combinations with the largest PPV. The starred values refer to results of the runtime-constrained exhaustive search

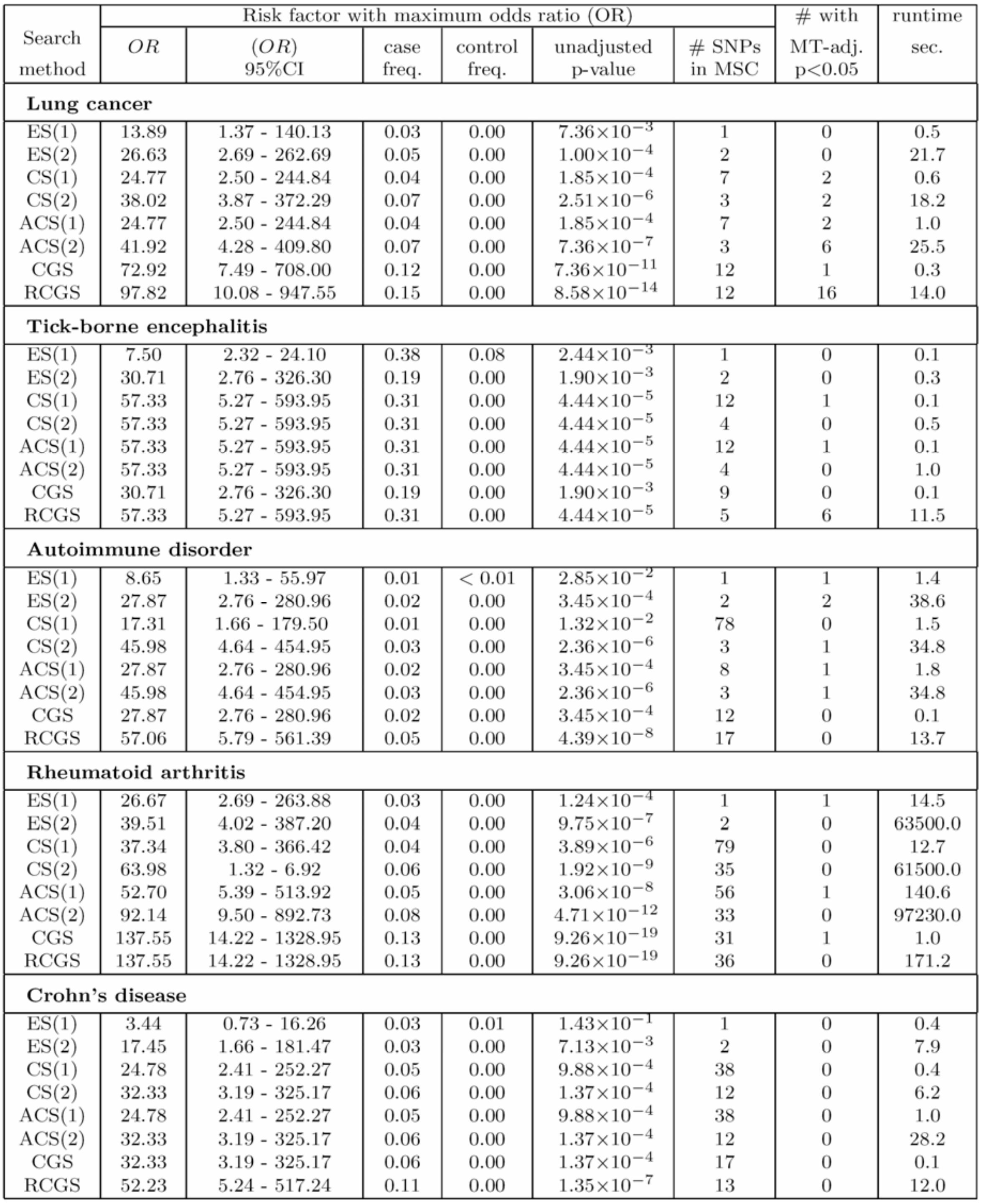


Dataset of	Search method	max PPV risk factor				max PPV resistance factor			
		case freq.	control freq.	unadjusted p-value	run-time sec.	case freq.	control freq.	unadjusted p-value	run-time sec.
Crohn's	IES_{30}	0.09*	0.00	8.7×10^{-7}	21530	0.00	0.07*	3.7×10^{-4}	869
disease	ICS_{30}	0.11	0.00	3.1×10^{-9}	7360	0.00	0.09	5.7×10^{-5}	708
[138]	CGS	0.06	0.00	1.4×10^{-4}	0.1	0.00	0.10	2.2×10^{-5}	0.1
autoimmune	IES_{30}	0.04*	0.00	2.5×10^{-8}	7633	0.00	0.04*	4.0×10^{-6}	39
disorder	ICS_{30}	0.04	0.00	2.5×10^{-8}	5422	0.00	0.04	4.0×10^{-6}	36
[163]	CGS	0.02	0.00	3.4×10^{-4}	0.1	0.00	0.04	2.5×10^{-5}	0.1
tick-borne	ES	0.29*	0.00	4.8×10^{-4}	820	0.00	0.39	1.0×10^{-3}	567
encephalitis	CS	0.33	0.00	1.3×10^{-4}	780	0.00	0.39	1.0×10^{-3}	1
[100]	CGS	0.19	0.00	6.1×10^{-3}	0.1	0.00	0.32	3.8×10^{-3}	0.1

Figure 3.9 shows the results of CGS, *k*-CGS, and WCGS algorithms after each iteration. The results are shown for all available datasets except autoimmune disorder datset. For this dataset results are similar to the results for rheumatoid arthritis (see Figure 3.9 (b)). The plots in the Figure 3.9 should be read from right to left. The top-right corner corresponds to the starting point of each algorithm. The bottom axis corresponds to the stopping point. The algorithm for best method stops with the largest number of cases in the cluster. For all compared datasets the OR of RFs found by the methods are ordered as follows: CGS, *k*-CGS, WCGS. The WCGS finds clusters by 1.5 to 4 times larger than CGS.

Table 3.5 compares 3 search heuristics applied to all 5 real datasets. For *k*-CGS method we consider *k* = 5, that is the best found value of k which fits all 5 datasets. The third column in the table shows that for all datasets CGS is outperformed by *k*-CGS which is inferior to WCGS when compared to the number of cases in the cluster formed by found RF (equivalent to maximum OR). The forth column shows the multiple testing adjusted p-value of the found RFs. The k-CGS finds statistically more significant RFs than CGS and WCGS finds even more significant RFs.

Table 3.4 Comparison of 5 methods searching atomic risk factors represented by multi-SNP combinations on five real datasets. For accurate search methods the number in parenthesizes denote searching level.



Search method	Risk factor with maximum odds ratio (OR)						# with MT-adj. $p<0.05$	runtime sec.
	OR	(OR) 95%CI	case freq.	control freq.	unadjusted p-value	# SNPs in MSC		
Lung cancer								
ES(1)	13.89	1.37 - 140.13	0.03	0.00	7.36×10^{-3}	1	0	0.5
ES(2)	26.63	2.69 - 262.69	0.05	0.00	1.00×10^{-4}	2	0	21.7
CS(1)	24.77	2.50 - 244.84	0.04	0.00	1.85×10^{-4}	7	2	0.6
CS(2)	38.02	3.87 - 372.29	0.07	0.00	2.51×10^{-6}	3	2	18.2
ACS(1)	24.77	2.50 - 244.84	0.04	0.00	1.85×10^{-4}	7	2	1.0
ACS(2)	41.92	4.28 - 409.80	0.07	0.00	7.36×10^{-7}	3	6	25.5
CGS	72.92	7.49 - 708.00	0.12	0.00	7.36×10^{-11}	12	1	0.3
RCGS	97.82	10.08 - 947.55	0.15	0.00	8.58×10^{-14}	12	16	14.0
Tick-borne encephalitis								
ES(1)	7.50	2.32 - 24.10	0.38	0.08	2.44×10^{-3}	1	0	0.1
ES(2)	30.71	2.76 - 326.30	0.19	0.00	1.90×10^{-3}	2	0	0.3
CS(1)	57.33	5.27 - 593.95	0.31	0.00	4.44×10^{-5}	12	1	0.1
CS(2)	57.33	5.27 - 593.95	0.31	0.00	4.44×10^{-5}	4	0	0.5
ACS(1)	57.33	5.27 - 593.95	0.31	0.00	4.44×10^{-5}	12	1	0.1
ACS(2)	57.33	5.27 - 593.95	0.31	0.00	4.44×10^{-5}	4	0	1.0
CGS	30.71	2.76 - 326.30	0.19	0.00	1.90×10^{-3}	9	0	0.1
RCGS	57.33	5.27 - 593.95	0.31	0.00	4.44×10^{-5}	5	6	11.5
Autoimmune disorder								
ES(1)	8.65	1.33 - 55.97	0.01	< 0.01	2.85×10^{-2}	1	1	1.4
ES(2)	27.87	2.76 - 280.96	0.02	0.00	3.45×10^{-4}	2	2	38.6
CS(1)	17.31	1.66 - 179.50	0.01	0.00	1.32×10^{-2}	78	0	1.5
CS(2)	45.98	4.64 - 454.95	0.03	0.00	2.36×10^{-6}	3	1	34.8
ACS(1)	27.87	2.76 - 280.96	0.02	0.00	3.45×10^{-4}	8	1	1.8
ACS(2)	45.98	4.64 - 454.95	0.03	0.00	2.36×10^{-6}	3	1	34.8
CGS	27.87	2.76 - 280.96	0.02	0.00	3.45×10^{-4}	12	0	0.1
RCGS	57.06	5.79 - 561.39	0.05	0.00	4.39×10^{-8}	17	0	13.7
Rheumatoid arthritis								
ES(1)	26.67	2.69 - 263.88	0.03	0.00	1.24×10^{-4}	1	1	14.5
ES(2)	39.51	4.02 - 387.20	0.04	0.00	9.75×10^{-7}	2	0	63500.0
CS(1)	37.34	3.80 - 366.42	0.04	0.00	3.89×10^{-6}	79	0	12.7
CS(2)	63.98	1.32 - 6.92	0.06	0.00	1.92×10^{-9}	35	0	61500.0
ACS(1)	52.70	5.39 - 513.92	0.05	0.00	3.06×10^{-8}	56	1	140.6
ACS(2)	92.14	9.50 - 892.73	0.08	0.00	4.71×10^{-12}	33	0	97230.0
CGS	137.55	14.22 - 1328.95	0.13	0.00	9.26×10^{-19}	31	1	1.0
RCGS	137.55	14.22 - 1328.95	0.13	0.00	9.26×10^{-19}	36	0	171.2
Crohn's disease								
ES(1)	3.44	0.73 - 16.26	0.03	0.01	1.43×10^{-1}	1	0	0.4
ES(2)	17.45	1.66 - 181.47	0.03	0.00	7.13×10^{-3}	2	0	7.9
CS(1)	24.78	2.41 - 252.27	0.05	0.00	9.88×10^{-4}	38	0	0.4
CS(2)	32.33	3.19 - 325.17	0.06	0.00	1.37×10^{-4}	12	0	6.2
ACS(1)	24.78	2.41 - 252.27	0.05	0.00	9.88×10^{-4}	38	0	1.0
ACS(2)	32.33	3.19 - 325.17	0.06	0.00	1.37×10^{-4}	12	0	28.2
CGS	32.33	3.19 - 325.17	0.06	0.00	1.37×10^{-4}	17	0	0.1
RCGS	52.23	5.24 - 517.24	0.11	0.00	1.35×10^{-7}	13	0	12.0

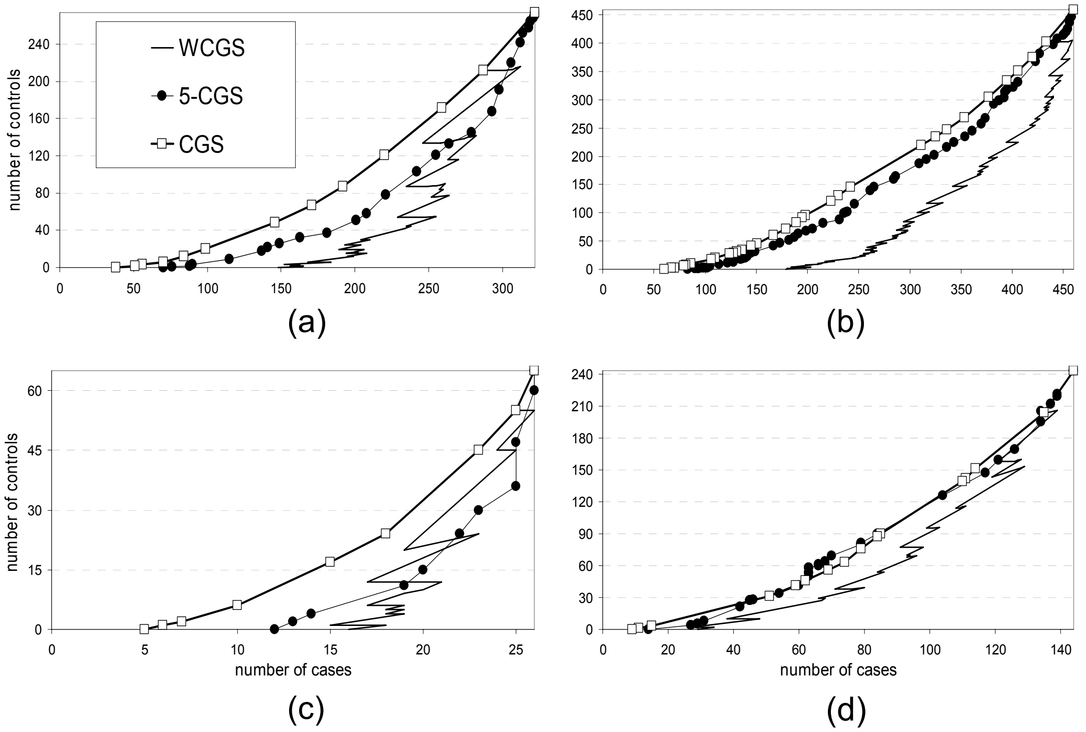


Figure 3.9 Number of cases and controls in the cluster after each iteration of CGS, k-CGS, and WCGS algorithms run on 4 datasets: (a) - lung cancer, (b) rheumatoid arthritis, (c) - tick-borne encephalitis, (d) - Crohn's disease. Starting point corresponds to the top-right corner where clusters contains entire S . Algorithms stop when number of controls in the cluster is zero (bottom axis).

Table 3.5 Comparison of CGS, k-CGS, and WCGS methods for searching ARF, k-ARF, and WRARF with maximum OR cluster on five real datasets.

Dataset	Search Method	Risk factor with maximum OR			
		# of cases with RF (%)	MT-adjusted p-value (%)	# of SNPs in RF	runtime (sec)
Lung cancer	CGS	11.8	4.0	12	1
	5-CGS	21.7	0.2	34	15
	WCGS	46.0	0.01	69	27
Tick-borne encephalitis	CGS	19.2	6.0	8	0.1
	5-CGS	46.2	4.0	17	0.1
	WCGS	61.5	2.0	20	0.1
Crohn's disease	CGS	6.3	34.0	17	0.5
	5-CGS	9.8	20.0	35	4
	WCGS	20.3	5.0	40	4
Autoimmune disorder	CGS	2.1	24.0	12	1
	5-CGS	4.7	18.0	30	3
	WCGS	9.4	4.1	47	20
Rheumatoid arthritis	CGS	13.0	4.3	31	40
	5-CGS	18.0	4.1	74	870
	WCGS	38.9	0.7	117	1365

For datasets of Crohn's disease [138] and autoimmune disorder [163] WCGS finds significant WRARFs while no other methods were able to find anything significant. The number of SNPs in the diplotypes corresponding to RFs are presented in the fifth column. The last column in the Table 3.5 shows the runtime of each method.

3.4 Cross-Validation of Searching Methods

This section addresses problems of validation of methods searching for disease associated risk factors in genotype data analysis.

This challenge is of statistical nature. Only statistically significant risk factors (whose frequency distribution has p-value less than 0.05) are believed to be reproducible. Successful as well as unsuccessful searches for SNPs with statistically significant association have been reported for different diseases and different suspected human genome regions (see, e.g., [136]). Unfortunately, reported findings are frequently not reproducible in independent case-control studies. It can be caused by confounding factors or the lack of significance -- the reported p-values are frequently unadjusted to multiple testing. Nevertheless, validation of findings on independent data is usually very expensive and additional evidence may help. Still it would be preferable to estimate how frequently a search method finds reproducible MSC. Here we propose to directly validate searching methods using cross-validation scheme, i.e., in the same way as predictions are validated.

We have cross-validated our search methods on real case-control studies for several diseases (Crohn's disease [138], autoimmune disorder [163], tick-born encephalitis [134], and lung cancer [144]). Proposed methods are compared favorably to the exhaustive search -- they have significantly higher leave-half-out cross-validation rate.

3.4.1 Permutation Test for Multiple-Testing Adjustment

Since statistically significant MSCs are searched among many such combinations, the computed p-value requires adjustment for multiple testing. The standard Bonferroni correction adjusts p-value by multiplying it by the number of the number of tests, i.e., number of tested MSCs. However, the Bonferroni correction is overly pessimistic, e.g., for finding one significant SNP among 100 we should multiply its p-value by 100; as a result, SNP should have $p < 0.0005$ in order to be statistically significant. Similarly, this factor grows to 10^4 for 2-SNP combinations. Instead, we compute multiple testing adjustment using more accurate but computationally extensive randomization method. 10^4 times, we repeat the following:

1. randomize the status of individuals (by random swapping);
2. find the 500th smallest p-value of MSCs.

This p-value corresponds to the multiple testing adjusted $p = 0.05$.

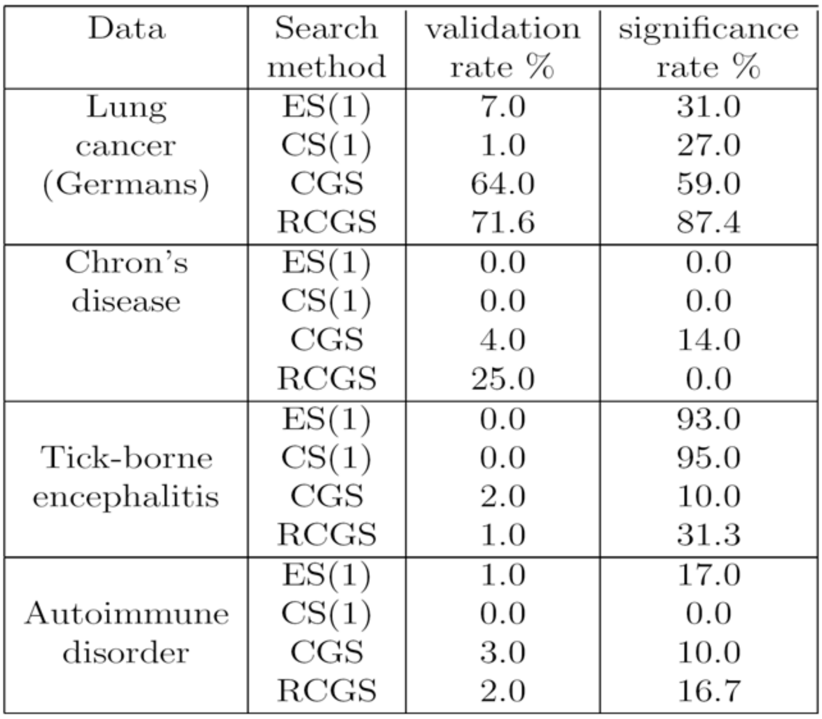
3.4.2 Experimental Results

3.4.2.1 Data Sets

- *Crohn's disease (5q31):* The data set Daly et al [138] is derived from the 616 kilobase region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohn's disease by genotyping 103 SNPs for 129 trios. All offspring belong to the case population, while almost all parents belong to the control population. In entire data, there are 144 case and 243 control individuals.

- *Autoimmune disorder:* The data set of Ueda et al [163] are sequenced from 330kb of human DNA containing gene CD28, CTLA4 and ICONS which are proved related to autoimmune disorder. A total of 108 SNPs were genotyped in 384 cases of autoimmune disorder and 652 controls.

Table 3.6 Leave-half-out cross-validation of 4 disease-association search methods on 4 real datasets. The validation rate for a method is the portion of MCS found on the training half that have been validated (i.e., have p-value < 5%) on the testing half. The significance rate is the portion of MSC on the training half that stay significant after multiple testing adjustment. For accurate search methods the number in parenthesizes denote searching level.



Data	Search method	validation rate %	significance rate %
Lung cancer (Germans)	ES(1)	7.0	31.0
	CS(1)	1.0	27.0
	CGS	64.0	59.0
	RCGS	71.6	87.4
Chron's disease	ES(1)	0.0	0.0
	CS(1)	0.0	0.0
	CGS	4.0	14.0
	RCGS	25.0	0.0
Tick-borne encephalitis	ES(1)	0.0	93.0
	CS(1)	0.0	95.0
	CGS	2.0	10.0
	RCGS	1.0	31.3
Autoimmune disorder	ES(1)	1.0	17.0
	CS(1)	0.0	0.0
	CGS	3.0	10.0
	RCGS	2.0	16.7

- *Tick-borne encephalitis:* The tick-borne encephalitis virus-induced dataset of Barkash et al [100] consists of 41 SNPs genotyped from DNA of 21 patients with severe tick-borne encephalitis virus-induced disease and 54 patients with mild disease.

- *Lung cancer* : dataset [144] was obtained using genome-wide DNA pooling strategy in a group of German smokers clinically diagnosed with lung cancer and age matched healthy smokers. 83.715 SNPs had beed screened. 141 SNPs showed putative allelic imbalance between case and control DNA pools and were eventually genotyped in individual samples. Finally, dataset includes 322 male smokers with lung cancer and 273 healthy smokers genotyped in 141 SNPs.

3.4.2.2 Leave-Half-Out Cross-Validation of Searching Methods

The Table 3.6 describes the results of leave-half-out cross-validation which has been organized as follows: 100 times we randomly choose one half of cases and one half of controls for the training set keeping the rest for the testing set. The table reports two rates:

- the validation rate for a method is the portion of MCS found on the training half that have been validated (i.e., have p-value < 5%) on the testing half;

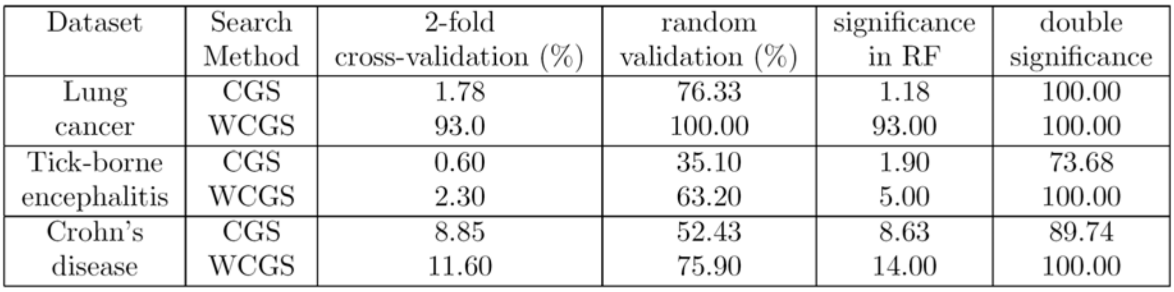
- the significance rate is the portion of MSC on the training half that stay significant after multiple testing adjustment.

All methods have low validation rate but greedy heuristics on all sets are more frequently validated. Also the greedy heuristics frequently find significant risk factors while ES also finds a lot of significant factors that are rarely validated on testing data.

Only statistically significant risk factors (whose frequency distribution has p-value less than 0.05) are believed to be reproducible. Nevertheless, validation of findings on independent data is usually very expensive and additional evidence may help. However, it would be preferable to estimate how frequently a search method finds reproducible RFs. He we directly validate searching methods using cross-validation scheme, i.e., in the same way as predictions are validated.

In Table 3.7 WCGS is compared favorably to the CGS -- it finds statistically significant risk factors more frequently and has significantly higher 2-fold cross-validation rate.

Table 3.7 Validation of CGS and WCGS on 3 real datasets. The 2-fold cross-validation column corresponds to % best RFs on the training half validated on testing half (p-value < 5%). The random-validation columns is the same but testing is allowed to overlap with training. The significance corresponds to % best RFs on the training half significant after MT-adjustment. The double significance column shows % of best RFs on the training half significant after MT-adjustment that are also significant on the testing half.



Dataset	Search Method	2-fold cross-validation (%)	random validation (%)	significance in RF	double significance
Lung cancer	CGS WCGS	1.78 93.0	76.33 100.00	1.18 93.00	100.00 100.00
Tick-borne encephalitis	CGS WCGS	0.60 2.30	35.10 63.20	1.90 5.00	73.68 100.00
Crohn's disease	CGS WCGS	8.85 11.60	52.43 75.90	8.63 14.00	89.74 100.00

3.5 Conclusions

Comparing indexed counterparts with exhaustive and combinatorial searches shows that indexing is quite successful. Indeed, indexed search finds the same MSCs as nonindexed search but it is much faster and its multiple-testing adjusted 0.05-threshold is higher and easier to meet.

Comparing combinatorial searches with the exhaustive counterparts is advantageous to the former. Indeed, for unphased data [138] the exhaustive search on the first and second search levels is unsuccessful while the combinatorial search finds several statistically significant MSCs for the same searching level. Similarly, for unphased and phased data of [163] the combinatorial search found much more statistically significant MSCs than the exhaustive search for the same searching level.

Results show (see Tables 3.6 and 3.2) that the indexing approach and the combinatorial search method are very promising techniques for searching statistically significant diseases-associated MSCs which can lead to discovery disease causes. The next step is biological validation of statistically significant MSCs discovered by proposed searching methods.

The comparison of three association searches (see Table 3.3) shows that combinatorial search always finds the same or larger cluster than exhaustive search and is significantly faster. The search method runtime is a critical in deciding whether it can be used in in clustering and susceptibility prediction. Note that the both exhaustive and combinatorial searches are prohibitively slow on the first two datasets and, therefore, we reduce these datasets to 30 index SNPs while complementary greedy search is fast enough to handle the complete datasets. This resulted in improvement of the complementary greedy over combinatorial search for the first dataset when search for the largest case-free cluster - after compression to 30 tags the best cluster simply disappears.

On the most of data our new methods can find more statistically significant atomic risk factors. Note that ES, CS, and ACS are much slow than greedy heuristics. All methods have low validation rate but greedy heuristics on all sets are more frequently validated. Also the greedy heuristics frequently find significant risk factors while ES also finds a lot of significant factors that are rarely validated on testing data.

The k-CGS and WCGS are significant improvement over CGS since they found risk factors that are statistically significant even after multiple-testing adjustment on all data while CGS could not find significant ARF on two of these data (see Table 3.5). The found relaxed atomic risk factors explain much higher number of cases, 1.5-4 times larger than atomic ARFs found by CGS. The *k*-CGS and WCGS methods also have significantly higher leave-half-out crossvalidation rate (see Table 3.7).

We conclude that the indexing approach, the combinatorial and complementary greedy search methods as well as *k*-Relaxed and Weighted *k*-Relaxed CGS are very promising techniques that can possibly help (i) to discover gene interactions causing common diseases and (ii) to create diagnostic tools for genetic epidemiology of common diseases.

CHAPTER 4

DISEASE SUSCEPTIBILITY PREDICTION

This chapter explores possibility of applying discrete optimization methods to predict the genotype susceptibility for complex disease. We propose two approaches for predicting genotype susceptibility: (i) universal classifier combinatorial approach and (ii) association based combinatorial approach. The proposed combinatorial method have been favorably compared with existing methods on several publicly available genotype data.

4.1 Introduction

Recent improvement in accessibility of high-throughput genotyping brought a great deal of attention to disease association and susceptibility studies [165]. High density maps of single nucleotide polymorphism (SNPs) [156] as well as massive genotype data with large number of individuals and number of SNPs become publicly available[153, 154, 157]. A catalogue of all human SNPs is hoped to allow genome-wide search of SNPs associated with genetic diseases.

Success stories when dealing with diseases caused by a single SNP or gene were reported. But some complex diseases, such as psychiatric disorders, are characterized by a non mendelian, multifactorial genetic contribution with a number of susceptible genes interacting with each other [161, 147]. In general, a single SNP or gene may be impossible to associate because a disease may be caused by completely different modifications of alternative pathways.

Furthermore, there are no reliable tools applicable to large genome ranges that could rule out or confirm association with a disease. It is even difficult to decide if a particular disease is genetic, e.g., the nature of Crohn's disease has been disputed [146]. Although answers to above questions may not explicitly help to find specific disease-associated SNPs, they may be critical for disease prevention. Indeed, knowing that an individual is (or is not) susceptible to (or belong to a risk group for) a certain disease will allow greatly reduce the cost of screening and preventive measures or even help to completely avoid disease development, e.g., by changing a diet.

4.1.1 Universal Classifier Approaches

This study is devoted to the problem of assessing accumulated information targeting to predict genotype susceptibility to complex diseases with significantly high accuracy and statistical power. We first give several discrete optimization based algorithms for prediction disease susceptibility. We then compare leave-one- and leave-many-out tests demonstrating that prediction accuracy of suggested methods is sufficiently resilient to discarding case-control data implying that leave-one-out test is a trustworthy accuracy measure. The randomization techniques have been used for computing the statistical significance level of proposed methods and resulted prediction weights. We show that prediction rate and statistical significance are well correlated.

The proposed methods are applied to two publicly available data: Crohn's disease [153] and autoimmune disorder [163]. In the leave-one-out cross-validation tests the proposed linear programming (LP) based method achieves prediction rate of 69.5%(p-value below 2%) and 61.3%(p-value below 62%) and the risk rates of 2.23 and 0.98, respectively. We also show that SVM methods used in [159, 164] are not much worse than our proposed LP-based method.

4.1.2 Association Based Combinatorial Approaches

The second part of the chapter shows how to apply association search methods to disease susceptibility prediction following [133]. First the problem and cross-validation schemes are discussed. Then the relevant formulation of the optimum clustering problem is given and the general method how any clustering algorithm can be transformed into a prediction algorithm is described. We conclude with description of two association search-based prediction algorithms.

Below is the formal description of the problem from [133].

Disease Susceptibility Prediction Problem (DSP). *Given a sample population S (a training set) and one more individual t $\notin$ S with the known SNPs but unknown disease status (testing individual), find (predict) the unknown disease status.*

The main drawback of such problem formulation that it cannot be considered as a standard optimization formulation. One cannot directly measure the quality of a prediction algorithm from the given input since it does not contain the predicted status.

A standard way to measure the quality of prediction algorithms is to apply a cross-validation scheme. In the leave-one-out cross-validation, the disease status of each genotype in the population sample is predicted while the rest of the data is regarded as the training set. There are many types of leave-many-out cross-validations where the testing set contains much larger subset of the original sample. Any cross-validation scheme produces a confusion table (see Table 4.1). The main objective is to maximize prediction accuracy while all other parameters also reflect the quality of the algorithm.

The next section formally defines the problem and describes several universal and ad hoc methods for predicting genotype susceptibility to complex diseases. Section 4.3 introduces our LP-based prediction algorithm. Then we introduce an optimum disease clustering problem and describe association based combinatorial methods for disease susceptibility prediction.

Section 4.6 describes real case-control study data sets, discusses prediction and risk rate measures and compares results for several susceptibility prediction methods. We draw the conclusion in the last section.

4.2 Previous Work

In this section we first describe the input and the output of prediction algorithms and how to predict genotype susceptibility. We the describe several universal and adhoc prediction methods.

Specifications of prediction algorithms. Data sets have n genotypes and each has m SNPs. The input for a prediction algorithm includes:

(G1) Training genotype set $g_i = (g_{i,j})$; $i = 0,.., n-1, j = 1,.., m$, $g_{i,j} \in \{0, 1, 2\}$

(G2) Disease status $s(g_i) \in \{-1, 1\}$, indicating if $g_{i,}$ $i = 0,.., n - 1$, is in case (1) or in control (-1) , and

(G3) Testing genotype g_n without any disease status.

The input data can also be phased, then each genotype is represented by a pair of haplotypes. We will refer to the parts (G1-G2) of the input as *training* set and to the part (G3) as the test case. The output of prediction algorithms is the disease status of the genotype

g_n, i.e., $s(g_n)$.

Below we describe several universal prediction methods. These methods are adaptations of general computer-intelligence classifying techniques.

Closest Genotype Neighbor (CN). For the test genotype g_t, find the closest (with respect to Hamming distance) genotype g_i in the training set, and set the status $s(g_t)$ equal to $s(g_i)$.

Support Vector Machine Algorithm (SVM). Support Vector Machine (SVM) is a generation learning system based on recent advances in statistical learning theory. It maps input vectors to a higher dimensional space where a maximal separating hyperplane is constructed. Then 2 parallel hyperplanes are constructed on each side of the hyperplane that separates the data. The separating hyperplane is the hyperplane that maximises the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the smaller the classification error is. SVMs deliver state-of-the-art performance in real-world applications and have been used in case/control studies [159, 164]. We use SVM-light [150] with the radial basis function with $\gamma= 0.5$.

Random Forest (RF). A random forest is a collection of CART-like trees following specific rules for tree growing, tree combination, self-testing, and post-processing. We use Leo Breiman and Adele Cutler's original implementation of RF version 5.1 [148]. This version of RF handles unbalanced data to predict accurately. RF tries to perform regression on the specified variables to produce the suitable model. RF uses bootstrapping to produce random trees and it has its own cross-validation technique to validate the model for prediction/classification.

CDPG: Tomita [162] introduced the Criterion of Detecting Personal Group (CDPG) for extracting risk factor candidates (RFCs). RFCs are extracted using binomial test and random permutation tests. CDPG performs exhaustive combination analysis using case/control data and assumes the appearance of case and control subjects belonging to a certain rule as a series of Bernoulli trials, where two possible outcomes are case and control subjects with some probabilities.

We now describe two ad hoc prediction methods (i.e., classifying techniques taking in account the nature of the classification problem). The first method is 2-SNP method [158] and the second method is a variation of the LP-based method [160].

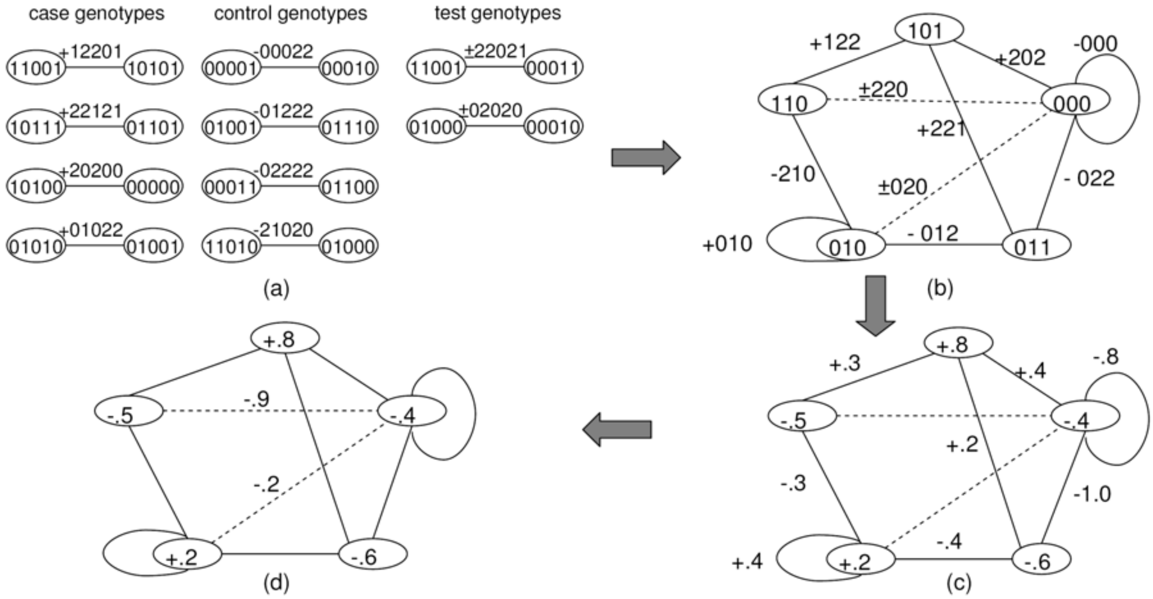
Most Reliable 2 SNP Prediction [158] (MR). This method chooses a pair of adjacent SNPs (site of s_i and s_{i+1}) to predict the disease status of the test genotype *gt* by voting among genotypes from training set which have the same SNP values as gt at the chosen sites s_i and s_{i+1}. They chose the 2 adjacent SNPs with the highest prediction rate in the training set.

4.3 LP-based Prediction Algorithm (LP).

This method are based on the following genotype graph $X = \{H, G\}$, where the vertices H are distinct haplotypes and the edges G are genotypes each connecting its two haplotypes (vertices) (see Figure 4.1(a)).

When applying graph heuristics to X, we found that it is necessary to increase the density of X. This can be achieved by dropping certain SNPs (or, equivalently, keeping only certain tag SNPs). Indeed, dropping a SNP may result in collapsing of certain vertices in X, i.e., different vertices become identical. Collapsing vertices may also result in collapsing certain edges (genotypes). Discarding a SNP is not allowed if it results in collapsing edges from case and control populations, but collapsing of edges from the same population is allowed (see Figure 4.1(b)).

Figure 4.1 LP-based Prediction Method. (a) The set of case, control and test genotypes are phased resulting in the sparse graph with vertices-haplotypes and edges genotypes. (b) The last two SNPs are dropped without collapsing case and control edges resulting in a denser graph. (c) The LP finds optimal weights for vertices haplotypes. (d) The status of test genotypes is predicted from the sign of the sum of weights of their endpoints.



A simple greedy strategy consists of traversing all the SNPs and dropping a SNP if it is allowed. The resulted set of SNPs is a minimal subset of SNPs which do not collapse genotypes from opposite disease status. Unfortunately, in the original graph X we may already have collapsed edges from opposite populations - in fact, Daly et al data contain such pair of genotypes. Only such original collapsing is allowed -- the status of such edges is assumed to be the one of majority of genotypes. Our experiments show that on average, we are left with 21 tag SNP's out of 103 for Daly et al [153] data and 29 tag SNP's out of 108 for Ueda et al [163] data (see description of the next section). The selected set tag SNPs are better candidates for being disease associated, in fact only such tag SNPs were used in the prediction methods with the highest accuracy.

After collapsing the graph *X* we add the edge corresponding to the test-case genotype g_n. If the edge g_n collapses with another edge g_i, then we set the predicted disease status $s(g_n) = s(g_i)$. Otherwise, we apply one of the following two methods for computing the disease status $s(g_n)$. The LP-based method assumes that certain haplotypes are susceptible to the disease while others are resistant to the disease. The genotype susceptibility is then assumed to be a sum of susceptibilities of its two haplotypes.

We want to assign a positive weight to susceptible haplotypes and a negative weight to resistant haplotypes such that for any control genotype the sum of weights of its haploptypes is negative and for any case genotype it is positive (see Figure 4.1 (c)). We would also like to maximize the confidence of our weight assignment which can be measured by the absolute values of the genotype weights. In other words, we would like to maximize the sum of absolute values of weights over all genotypes.

Formally, for each vertex $\mathrm{h_i}$ (corresponding to haplotype) of the graph *X* we wish to assign the weight p_i, $-1 < p_i <1$ such that for any genotype-edge $e_{ij} = (h_i, h_j)$, $s(e_{ij})(p_i + p_j)$ ¸ 0 where $s(e_{ij}) \in \{-1, 1\}$ is the disease status of genotype represented by edge e_{ij}.

The total sum of absolute values of genotype weights is maximized

$$\sum_{e_{ij}=(h_i,h_j)} s(e_{ij})(p_i + p_j) \tag{4.1}$$

The above formulation with the objective (4.1) is the linear program which can be efficiently solved by a standard linear program solver such as GNU Linear Programming Kit (GLPK) [155].

For the left-out testing genotype g_n, we compute the sum of weights of its haplotypes. If the sum is strictly positive, the genotype is attributed to the case, if the sum is strictly negative, it is attributed to the control (see Figure 4.1(d)), otherwise $s(g_n)$ is assigned according to 2-SNP prediction algorithm [158].

4.4 Optimum Disease Clustering Problem

Here we proposes to avoid cross-validation and instead suggests a different objective by restricting the ways how prediction can be made. It is reasonable to require that every prediction algorithm should be able to predict the status inside the sample. Therefore, such algorithms is supposed to be able to partition the sample into subsets based only on the values of SNPs, i.e., partition of S into clusters defined by MSCs. Of course, a trivial clustering where each individual forms its own cluster can always perfectly distinguish between case and control individuals. On the other hand such clustering carries minimum information. Ideally, there should be two clusters perfectly distinguishing diseased from control individuals. There is a trade-off between number of clusters and the information carried by clustering which results in trade-off between number of errors (i.e., incorrectly clustered individuals) and informativeness which was proposed to measure by information entropy instead of number of clusters [133].

Optimum Disease Clustering Problem. *Given a population sample S, find a partition P of S into clusters* $S = S_1 \cup .. \cup S_k$*, with disease status 0 or 1 assigned to each cluster* S_i*, minimizing*

$$entropy(P) = -\sum_{i=1}^{k} \frac{|S_i|}{|S|} \ln \frac{|S_i|}{|S|}$$

for a given bound on the number of individuals who are assigned incorrect status in clusters of the partition P, $error(P) < \alpha * |P|$.

The above optimization formulation is obviously NP-hard but has a huge advantage over the prediction formulation that it does not rely on cross-validation and can be studied with combinatorial optimization techniques. Still, in order to make the resulted clustering algorithm useful, one needs to find a way ho to apply it to the original prediction problem.

4.5 Model-Fitting Prediction

The following general approach has been proposed by us [133]. Assuming that the clustering algorithm indeed distinguishes real causes of the disease, one may expect that the major reason for erroneous status assignment is in biases and lack of sampling. Then a plausible assumption is that a larger sample would lead to a lesser proportion of clustering errors. This implies the following transformation of clustering algorithm into prediction algorithm:

Clustering-based Model-Fitting Prediction Algorithm

```
Set disease status 0 for the testing individual t and
        Find the optimum (or approximate) clustering $P_0$ of $S \cup \{t\}$
Set disease status 1 for the testing individual t and
        Find the optimum (or approximate) clustering $P_1$ of $S \cup \{t\}$
Find which of two clusterings $P_0$ or $P_1$ better fits model, and accordingly predict status of $t$,
```

$$status(t) = \arg(\min_{i=0,1} error(P_i))$$

Two clustering algorithms based the combinatorial and complementary greedy association searches has been proposed in [133]. This clustering finds for each individual an MSC or its cluster that contains it and is the most associated according to a certain characteristic (e.g., RR, PPV or lowest p-value)) with disease-susceptibility and disease-resistance. Then each individual is attributed the ratio between these two characteristic values -- maximum disease-susceptibility and disease-resistance. Although the resulted partition of the training set S is easy to find, it is still necessary to decide which threshold between case and control clusters should be used. The threshold can be chosen to minimize the clustering error.

The *combinatorial search-based prediction algorithm* (CSP) exploits combinatorial search to find the most-associated cluster for each individual. Empirically, the best association characteristic is found to be the relative risk rate RR. The *complimentary greedy search-based prediction algorithm* (CGSP) exploits complimentary greedy search to find the most-associated cluster for each individual. Empirically, the best association characteristic is found to be the positive predictive value PPV. The leave-one-out cross-validation (see Section 4.6.3) show significant advantage of CSP and GCSP over previously known prediction algorithms for all considered real datasets.

4.6 Experemental Study

In this section we discuss the results of methods for susceptibility prediction on real datasets. We first describe available real datasets, then overview search and prediction methods and conclude with description and discussion of their performance on leave-one-out and leave-many-out cross-validation tests. All experiments were ran on Processor Pentium 4 3.2Ghz, RAM 2Gb, OS Linux.

4.6.1 Data Sets

- *Crohn's disease (5q31):* The data set Daly et al [138] is derived from the 616 kilobase region of human Chromosome 5q31 that may contain a genetic variant responsible for Crohn's disease by genotyping 103 SNPs for 129 trios. All offspring belong to the case population, while almost all parents belong to the control population. In entire data, there are 144 case and 243 control individuals.

- *Autoimmune disorder*: The data set of Ueda et al [163] are sequenced from 330kb of human DNA containing gene CD28, CTLA4 and ICONS which are proved related to autoimmune disorder. A total of 108 SNPs were genotyped in 384 cases of autoimmune disorder and 652 controls.

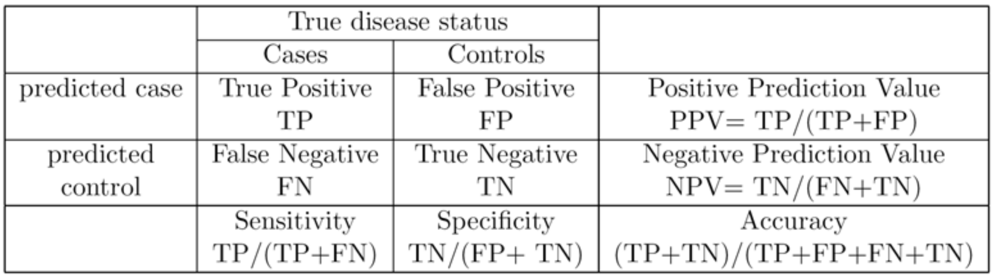
- *Tick-borne encephalitis*: The tick-borne encephalitis virus-induced dataset of Barkash et al [100] consists of 41 SNPs genotyped from DNA of 21 patients with severe tickborne encephalitis virus-induced disease and 54 patients with mild disease.

The datasets have been phased using 2SNP software [101]. The missing data (16% in [138] and 10% in [163]) have been imputed in genotypes from the resulted haplotypes. We have also created corresponding haplotype datasets in which each individual is represented by a haplotype with the disease status inherited from the corresponding individual genotype.

4.6.2 Results for Universal Classifiers Methods

Cross-validation Tests. In the leave-one-out cross-validation, the disease status of each genotype in the data set is predicted while the rest of the data is regarded as the training set. In the leave-many-out cross-validation, n individuals are uniformly at random picked up from the data set, marked and put back, where n is the size of the data set. This way, approximately 2/3 of the individuals are picked at least once and marked while the rest will not be marked. The training set consists of marked data and the testing set consists of unmarked data.

Table 4.1 Confusion table



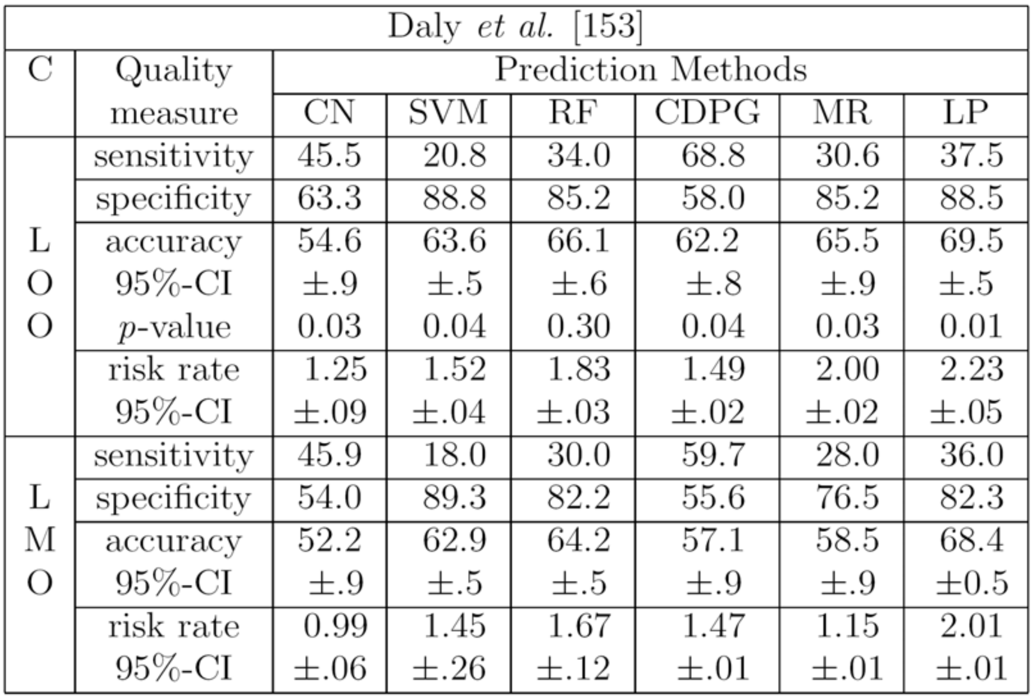
	True disease status		
	Cases	Controls	
predicted case	True Positive TP	False Positive FP	Positive Prediction Value PPV= TP/(TP+FP)
predicted control	False Negative FN	True Negative TN	Negative Prediction Value NPV= TN/(FN+TN)
	Sensitivity TP/(TP+FN)	Specificity TN/(FP+ TN)	Accuracy (TP+TN)/(TP+FP+FN+TN)

Quality Measures. In cross-validation tests, the predicted and the actual disease statuses are compared and the standard confusion matrix is filled (see Table 4.1). We report sensitivity, specificity, and accuracy of the prediction methods. We also report the the risk rate of the corresponding integrated risk factor associated with each prediction method. It is computed as the the ratio of the probability of developing disease among those predicted susceptible to the probability of developing disease among those predicted non-susceptible [152]:

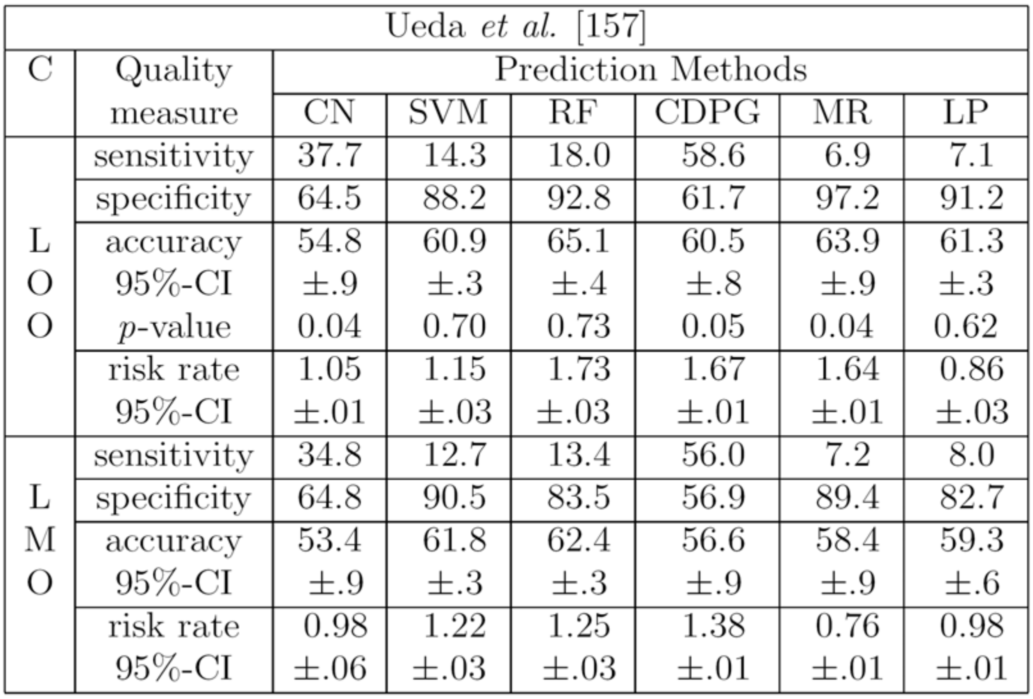
$$Risk\_Rate = \frac{TP}{TP+FP} / \frac{FN}{TN+FN}$$

We report the 95% confidence intervals (CI) for accuracy and risk rate, for leave-one-out test 95% CI is computed using bootstrapping. We also compute significance level, p-value, for the accuracy of prediction algorithms computed using 5000 randomized instances. On the randomized instances, the average prediction rate for SVM and RF has been 60% and for all other methods except has been 50%. Results. Table 4.2 compares 6 different prediction methods for both data sets. Column C denotes performed cross-validation tests, LOO stays for leave-one-out test and LMO stays for leave-many-out test. For leave-one-out test, the best accuracy is achieved by LP -- 69.5% on Daly et al. [153] data and by MR -- 63.9% on Ueda et al. [157] data. For leave-many-out test, the accuracy only slightly degrades showing resiliency to the size of the data. The risk rates for the integrated risk factor associated with prediction methods are comparable with risk rates for individual SNPs -- for the first data set, 2.23 (LP method) vs 2.7 and for the second data set, 1.73 (RF method) and 1.64 (MR method) vs 3.2. The good performance of SVM and certain other universal methods indicate that they can possibly be adjusted to improve specific ad hoc methods for prediction of susceptibility to complex diseases.

Table 4.2 The comparison of sensitivity, specificity, accuracy and risk rate with 95% confidence intervals (CI) and p-value for 6 prediction methods for two real data sets.



Daly *et al.* [153]							
C	Quality measure	Prediction Methods					
		CN	SVM	RF	CDPG	MR	LP
	sensitivity	45.5	20.8	34.0	68.8	30.6	37.5
	specificity	63.3	88.8	85.2	58.0	85.2	88.5
L	accuracy	54.6	63.6	66.1	62.2	65.5	69.5
O	95%-CI	±.9	±.5	±.6	±.8	±.9	±.5
O	p-value	0.03	0.04	0.30	0.04	0.03	0.01
	risk rate	1.25	1.52	1.83	1.49	2.00	2.23
	95%-CI	±.09	±.04	±.03	±.02	±.02	±.05
	sensitivity	45.9	18.0	30.0	59.7	28.0	36.0
L	specificity	54.0	89.3	82.2	55.6	76.5	82.3
M	accuracy	52.2	62.9	64.2	57.1	58.5	68.4
O	95%-CI	±.9	±.5	±.5	±.9	±.9	±0.5
	risk rate	0.99	1.45	1.67	1.47	1.15	2.01
	95%-CI	±.06	±.26	±.12	±.01	±.01	±.01



Ueda *et al.* [157]							
C	Quality measure	Prediction Methods					
		CN	SVM	RF	CDPG	MR	LP
	sensitivity	37.7	14.3	18.0	58.6	6.9	7.1
	specificity	64.5	88.2	92.8	61.7	97.2	91.2
L	accuracy	54.8	60.9	65.1	60.5	63.9	61.3
O	95%-CI	±.9	±.3	±.4	±.8	±.9	±.3
O	p-value	0.04	0.70	0.73	0.05	0.04	0.62
	risk rate	1.05	1.15	1.73	1.67	1.64	0.86
	95%-CI	±.01	±.03	±.03	±.01	±.01	±.03
	sensitivity	34.8	12.7	13.4	56.0	7.2	8.0
L	specificity	64.8	90.5	83.5	56.9	89.4	82.7
M	accuracy	53.4	61.8	62.4	56.6	58.4	59.3
O	95%-CI	±.9	±.3	±.3	±.9	±.9	±.6
	risk rate	0.98	1.22	1.25	1.38	0.76	0.98
	95%-CI	±.06	±.03	±.03	±.01	±.01	±.01

4.6.3 Results for Association Based Prediction Methods

We compare the prediction algorithms based on combinatorial and complimentary greedy searches (see Section 3.2.3.1) from previous chapter [133] with three universal classifiers prediction methods. We have chosen SVM, RF, and LP since they have best prediction results for two real data sets [138] and [163] (see Table 4.2).

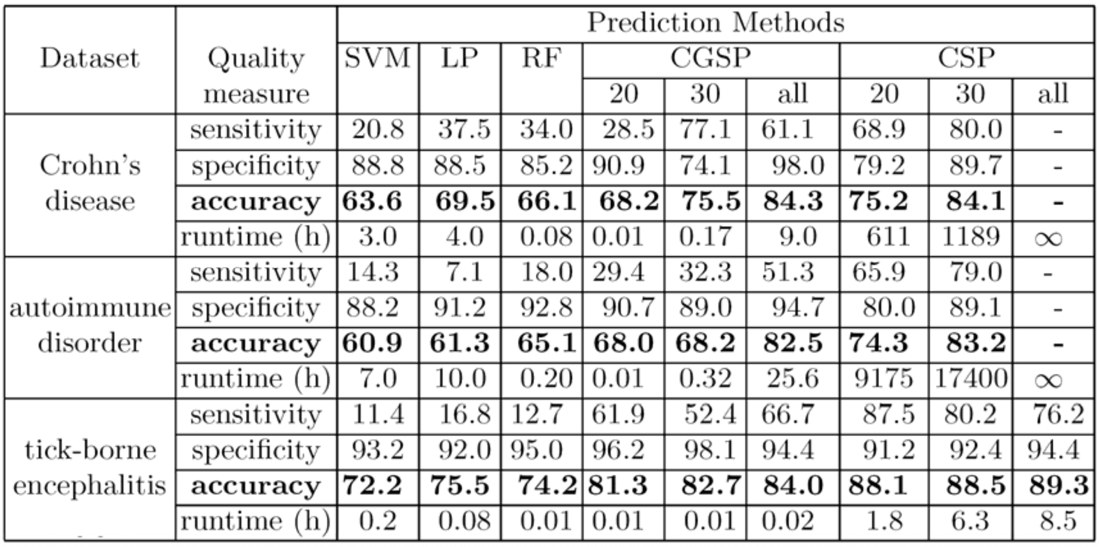
Table 4.3 reports comparison of all considered prediction methods. Their quality is measured by sensitivity, specificity, accuracy and runtime. Since prediction accuracy is the most important quality measure, it is given in bold.[1] Figure 4.2 shows the receiver operating characteristics (ROC) representing the trade-off between specificity and sensitivity. ROC is computed for all five prediction methods applied to the tick-borne encephalitis data [100].

4.7 Conclusions

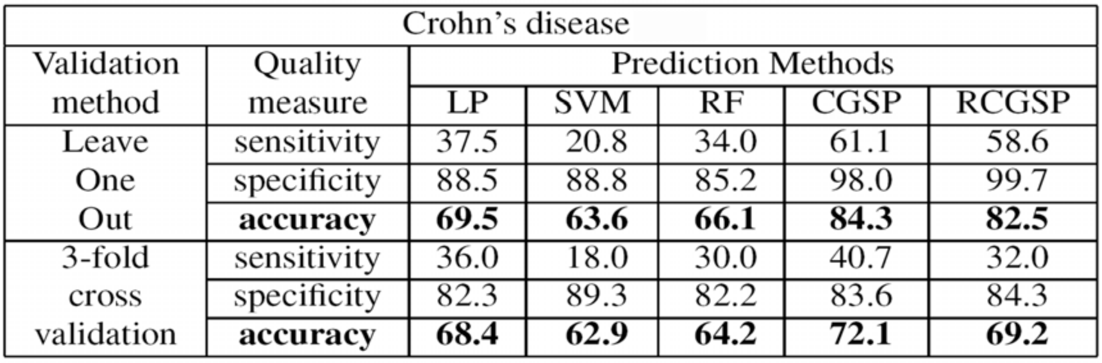
In this chapter, we discuss motivation behind the genotype susceptibility studies. The proposed susceptibility prediction method based on linear programming is shown to have high prediction rates and high relevant risk rate for associated integrated risk factors for two completely different case-control studies for Crohn's disease [153] and autoimmune disorders [157].

The comparison of the proposed association search-based and previously known susceptibility prediction algorithms (see Table 4.3) shows a considerable advantage of new methods. Indeed, for the first dataset the best proposed method (CGSP) beats the previously best method (LP) in prediction accuracy 76.3% to 69.5%. For the second dataset, the respective numbers are 74.2% (CSP(30)) to 65.1% (RF), and for the third dataset, they are 80.3% (CSP) to 75.5% (LP). It is important that this lead is the result of much higher sensitivity of new methods -- the specificity is almost always very high since all prediction methods tend to be biased toward non-diseased status.

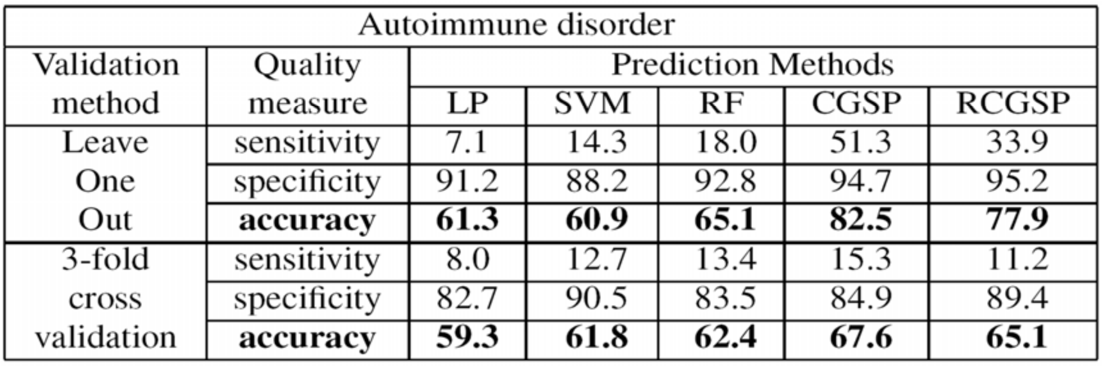
Table 4.3 Leave-one-out cross validation results of four prediction methods for three real data sets. Results of combinatorial search-based prediction (CSP) and complimentary greedy search-based prediction (CGSP) are given when 20, 30, or all SNPs are chosen as informative SNPs.



Dataset	Quality measure	Prediction Methods								
		SVM	LP	RF	CGSP			CSP		
					20	30	all	20	30	all
Crohn's disease	sensitivity	20.8	37.5	34.0	28.5	77.1	61.1	68.9	80.0	-
	specificity	88.8	88.5	85.2	90.9	74.1	98.0	79.2	89.7	-
	accuracy	**63.6**	**69.5**	**66.1**	**68.2**	**75.5**	**84.3**	**75.2**	**84.1**	**-**
	runtime (h)	3.0	4.0	0.08	0.01	0.17	9.0	611	1189	∞
autoimmune disorder	sensitivity	14.3	7.1	18.0	29.4	32.3	51.3	65.9	79.0	-
	specificity	88.2	91.2	92.8	90.7	89.0	94.7	80.0	89.1	-
	accuracy	**60.9**	**61.3**	**65.1**	**68.0**	**68.2**	**82.5**	**74.3**	**83.2**	**-**
	runtime (h)	7.0	10.0	0.20	0.01	0.32	25.6	9175	17400	∞
tick-borne encephalitis	sensitivity	11.4	16.8	12.7	61.9	52.4	66.7	87.5	80.2	76.2
	specificity	93.2	92.0	95.0	96.2	98.1	94.4	91.2	92.4	94.4
	accuracy	**72.2**	**75.5**	**74.2**	**81.3**	**82.7**	**84.0**	**88.1**	**88.5**	**89.3**
	runtime (h)	0.2	0.08	0.01	0.01	0.01	0.02	1.8	6.3	8.5



Crohn's disease						
Validation method	Quality measure	Prediction Methods				
		LP	SVM	RF	CGSP	RCGSP
Leave One Out	sensitivity	37.5	20.8	34.0	61.1	58.6
	specificity	88.5	88.8	85.2	98.0	99.7
	accuracy	**69.5**	**63.6**	**66.1**	**84.3**	**82.5**
3-fold cross validation	sensitivity	36.0	18.0	30.0	40.7	32.0
	specificity	82.3	89.3	82.2	83.6	84.3
	accuracy	**68.4**	**62.9**	**64.2**	**72.1**	**69.2**



Autoimmune disorder						
Validation method	Quality measure	Prediction Methods				
		LP	SVM	RF	CGSP	RCGSP
Leave One Out	sensitivity	7.1	14.3	18.0	51.3	33.9
	specificity	91.2	88.2	92.8	94.7	95.2
	accuracy	**61.3**	**60.9**	**65.1**	**82.5**	**77.9**
3-fold cross validation	sensitivity	8.0	12.7	13.4	15.3	11.2
	specificity	82.7	90.5	83.5	84.9	89.4
	accuracy	**59.3**	**61.8**	**62.4**	**67.6**	**65.1**

The ROC curve also illustrates advantage of CSP and GCSP over previous methods. Indeed the area under ROC curve for CSP is 0.81, for SVM is 0.52 compared with random guessing area of 0.5. Another important issue is how proposed prediction algorithms tolerate data compression. The prediction accuracy (especially sensitivity) is increases for CGSP when more SNPs are made available -- e.g., for the second dataset, the sensitivity grows from 29.4% (20 SNPs) to 32.3% (30 SNPs) to 46.3% (all 108 SNPs).

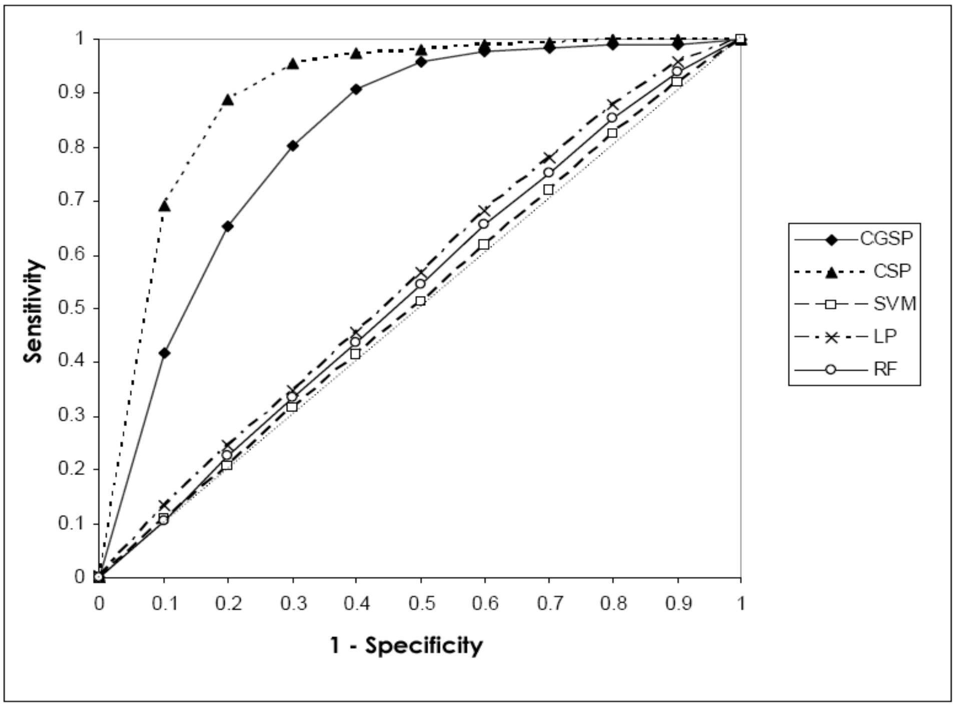


Figure 4.2 The receiver operating characteristics (ROC) for the five prediction methods applied to the tick-borne encephalitis data [100]. All SNPs are considered tags for CGSP and CSP.

The extensive computational results show great potential of the proposed association search based susceptibility prediction algorithms.

CHAPTER 5

SOFTWARE PACKAGES

1. **2SNP:** fast and scalable phasing software for trios and unrelated individuals, based on 2-SNP haplotypes, *http://alla.cs.gsu.edu/~software/2SNP*.

2. **DACS:** software for searching statistically significant multi-SNP combinations associated with a disease and predicting disease susceptibility of a given individual, *http://alla.cs.gsu.edu/~software/DACS*.

3. **TrioPhasing:** integer linear programming based software for phasing Trio data, requires LP solver, *http://alla.cs.gsu.edu/»software/trioPhasing.html*.

4. **GeneSuscept:** software for predicting disease susceptibility. It implements LP-based prediction algorithm.

CHAPTER 6

FUTURE WORK

In our future work we concentrate on six challenges arising in disease-association and disease-susceptibility prediction studies (i) design of more robust, fast and scalable algorithms for disease association search, (ii) development of faster algorithms for adjusting to the multiple testing the significance of the risk factors, (iii) developing of fast and acceptable methods for validation of disease association search methods, (iv) developing of the scalable indexing methods for genome-wide study, (v) design and testing of several disease models, (vi) development of faster, scalable and more accurate disease susceptibility prediction methods.

In the current work we had addressed the challenge of **developing fast and scalable methods for disease association search** (see chapter 3) and several good solutions had been proposed. The complimentary greedy search method (CGS) and its randomized implementation (RCGS) give very impressive results by finding the significant risk factors on the case-control study datasets where no associations were found before. Some of those risk factors had been confirmed by independent biological studies. The CGS method is very fast and scalable. However, it is too straightforward, and for several datasets, was unable to find risk factors significantly associated with a disease while slower RCGS found some of them. The problem with RCGS is that it is not scalable for genome-wide study. Therefore, one of our future goals is speeding-up of the RCGS method by limiting its randomness. We plan to implement randomization process using simulated annealing

We also plan to develop new **searching methods for wide-genome scan**. These methods will localize the DNA segments that possibly contain markers associated with a disease. Then on the extracted pieces of DNA we can run our exhaustive combinatorial search to find the multi-SNP combinations which mark disease.

When the disease-association methods will be accurate enough, we will move to our next goal which is **development of many disease models** and their evaluation on different diseases using our search methods. Successful models will allow us to understand how complex diseases are embedded within the DNA.

The models which we will first analyze consider **multiplicative and additive integration of SNPs** in forming complex risk factors for a complex diseases. In chapter 3 we search for a risk factor associated with a disease modeled as an MSC,

i.e., set of SNPs with fixed allele values (multiplicative integration). However, the same disease can be partially explained by several MSCs. Therefore, we should also consider additive integration of MSCs. The next question arisen in model selection is **haplotypes or genotypes are basis of the disease causes**.

One of the next goals is the design of the **indexing methods** which will allow us to reduce large datasets to smaller ones without losing markers which marks disease. In the reduced set we will be able to run more robust but slow searching methods. Currently we use MLR-Tagging method (see chapter 3) for indexing which helps to perform deeper disease association search. However, this method is not extendable for wide-genome scan study.

The next challenge that we address in our future work is design of the **fast methods for computing the significance of the risk factors**. The significance of the risk factor should be adjusted to the multiple testing by considering the total number of risk factors among which the search was performed and the complexity of the searching method. Currently we use permutation test which is very slow and completely useless for medium or large datasets. For large datasets we have to perform Bonferroni adjustment which is overly pessimistic and might disregard significant risk factors. Therefore, there is a great need for a method that adjusts the significance in more accurate than Bonferroni way and it is fast and scalable.

The next challenge which we address in our future work is **methods for validation of searching methods**. Because of the lack of available genotype case-control study data and their small size there no established methods for validation of the disease-association search methods. In chapter 3 we propose to use cross-validation schemes usually applied for prediction methods. However, this method is not very well tested and could be significantly modified in our future research.

Based on the new developed searching methods we address in our future work the challenge of developing of **faster, scalable and more accurate disease susceptibility prediction methods**. The methods described in chapter 4 are acceptably accurate but extremely slow for large datasets. Therefore, there is a great need for faster disease susceptibility prediction methods.

CHAPTER 7

RELATED PUBLICATIONS

Refereed Journal Articles and Book Chapters

1. Brinza, D., He, J., and Zelikovsky, A. (2007) `Optimization Methods for Genotype Data Analysis in Epidemiological Studies', book chapter in Bioinformatics Algorithms: Techniques and Applications, January 2007.

2. Brinza, D. and Zelikovsky, A. (2007) `2SNP: Scalable Phasing Method for Trios and Unrelated Individuals', Journal of IEEE/ACM Transactions on Computa tional Biology and Bioinformatics (TCBB), invited, January 2007.

3. Brinza, D. and Zelikovsky, A. (2006) `2SNP: Scalable Phasing Based on 2-SNP Haplotypes', BIOINFORMATICS , Pub 2006 Feb 1;22(3):371--373. Epub 2005 Nov 15.

4. Brinza, D., He, J., Mao, W. and Zelikovsky, A. (2005) `Family Trio Phasing and Missing Data Recovery', International Journal Bioinformatics Research and Applications, IJBRA'05 Vol. 1, No.2 pp. 221--229.

Refereed Conference Articles

5. Brinza, D. and Zelikovsky, A. (2007) `Design and Validation of Methods Searching for Risk Factors in Genotype Case-Control Studies', The Third RECOMB Satellite Conference on SNPs and Haplotype Analysis (SNPHAP 2007).

6. Brinza, D. and Zelikovsky, A. (2006) `Combinatorial Analysis of Disease Association and Susceptibility for Rheumatoid Arthritis SNP Data', Proc. of Genetic Analysis Workshop (GAW15), 13:6--11.

7. Andre R. de Vries, Ilja M. Nolte, Geert T. Spijker, Dumitru Brinza, Alexander Zelikovsky, and Gerard J. te Meerman (2006) `Cross Haplotype Sharing Statistic: Haplotype length based method for whole genome association testing', Proc. of Genetic Analysis Workshop (GAW15), 13:16--21.

8. Brinza, D., He, J. and Zelikovsky, A. (2006) `Combinatorial Search Methods for Multi-SNP Disease Association', Proc. International Conf. of the IEEE Engineering in Medicine and Biology (EMBC'06), pp. 5802--5805.

9. Brinza, D. and Zelikovsky, A. (2006) `Combinatorial Methods for Disease Association Search and Susceptibility Prediction', 6th Workshop on Algorithms in BioInformatics (WABI 2006), Lecture Notes in Bioinformatics 4175, pp. 286--297.

10. Brinza, D., Perelygin, A., Brinton, M. and Zelikovsky, A. (2006) `Search for multi-SNP Disease Association', The Fifth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'06), pp. 122-125.

11. Brinza, D. and Zelikovsky, A. (2006) `Phasing of 2-SNP Genotypes based on Non-Random Mating Model', International Workshop on Bioinformatics Research and Applications (IWBRA'06), Proc. of ICCS 2006, LNCS 3992, pp.767-774.

12. Mao, W., Brinza, D., Hundewale, N., Gremalschi, S. and Zelikovsky, A. (2006) `Genotype Susceptibility and Integrated Risk Factors for Complex Diseases', Proc. IEEE Intl Conf on Granular Computing (GRC 2006), pp. 754--757.

13. Altun, G., Hu, H.-J., Brinza, D., Harrison, R.W., Zelikovsky, A. and Pan, Y. (2006) `Hybrid SVM kernels for protein secondary structure prediction', Proc. IEEE Intl Conf on Granular Computing (GRC 2006), pp. 762--765.

14. Mao, W., He, J., Brinza, D. and Zelikovsky, A. (2005) `A Combinatorial Method for Predicting Genetic Susceptibility to Complex Diseases', International Conference of the IEEE Engineering In Medicine and Biology Society (EMBC'05), pp. 224--227.

15. Brinza, D., He, J., Mao, W. and Zelikovsky, A. (2005) `Phasing and Missing Data Recovery in Family Trios', International Workshop on Bioinformatics Research and Applications (IWBRA'05),Proc. of ICCS 2005 LNCS 3515, pp.1011--1019.

Poster and Oral Presentations

1. Brinza, D. and Zelikovsky, A. (November 2006) `Case(Control)-Free Multi-SNP Combinations in Case-Control Studies', Algorithmic Biology 2006, at University of California, San Diego.

2. Brinza, D. and Zelikovsky, A. (November 2006) `Combinatorial Analysis of Disease Association and Susceptibility for Rheumatoid Arthritis SNP Data', Genetic Analysis Workshop 15, St Pete Beach, FL.

3. Brinza, D. and Zelikovsky, A. (October 2006) `Combinatorial methods for disease association search and susceptibility prediction', Proc. of 9-th Conference of Computational Genomics (CG'06), Baltimore, MD.

4. Brinza, D., He, J. and Zelikovsky, A. (September 2006) `Combinatorial Search Methods for Multi-SNP Disease Association', International Conf. of the IEEE Engineering in Medicine and Biology (EMBC'06), New-York, NY.

5. Barkhash, A., Perelygin, A., Brinza, D., Pilipenko, P., Bogdanova, YU., Romaschenko, A., Voevoda, M. and Brinton, M. (July 2006) `GENETIC RESISTANCE TO FLAVIVIRUSES', Genomics Proteomics Bioinformatics and Nanotechnologies for Medicine (GPBM'06), Novosibirsk, RU.

6. Brinza, D. and Zelikovsky, A. (Jun 2006) `Combinatorial Search Methods for Genotypes Associated with Lung Cancer', Mollecular Basis of Disease Symphosium (MBD'06), Atlanta, GA.

7. Barkhash, A., Perelygin, A., Brinza, D., Pilipenko, P., Bogdanova, YU., Romaschenko, A., Voevoda, M. and Brinton, M. (May 2006) `VARIABILITY IN THE 2-5 OLIGOADENYLATE SYNTHETASE (OAS) GENE CLUSTER IS ASSOCIATED WITH SEVERITY OF TICK-BORNE ENCEPHALITIS VIRUS-INDUCED DISEASE IN RUSSIAN PATIENTS', 9-th Southeastern Regional Virology Conference 2006, Atlanta, GA.

8. Brinza, D. and Zelikovsky, A. (November 2005) `2SNP: New Scalable Phasing Method', The Fifth Georgia TECH International Conference on Bioinformatics (GTICB'05), Atlanta, GA.

9. Brinza, D. and Zelikovsky, A. (October 2005) `2SNP: New Scalable Phasing Method', The Second SECABC Fall Workshop ON Biocomputing (SECABC'05), Atlanta, GA.

10. Brinza, D. and Zelikovsky, A. (August 2005) `New 2-SNP Statistics method for Phasing and Missing data recovery', Georgia State Biotech Symposium, Atlanta, GA.

11. Brinza, D., He, J., Mao, W. and Zelikovsky, A. (May 2005) `Family Trio Phasing and Missing Data Recovery', Research in Computational Molecular Biology (RECOMB'05), Boston, MA.

BIBLIOGRAPHY

[1] V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as perfect phylogeny: A direct approach. Technical report, UC Davis, Department of Computer Science, 2002.

[2] V. Bafna, D.Gusfield, G. Lancia, and S. Yooseph. Haplotyping as perfect phylogeny: A direct approach. J. Computational Biology, 10:323--340, 2003.

[3] R. E. Bixby and D. K. Wagner. An almost linear-time algorithm for graph realization. Mathematics of Operations Research, 13:99--123, 1988.

[4] R.H. Chung and D. Gusfield. Empirical exploration of perfect phylogeny haplotyping and haplotypers. In Proceedings of COCOON 03 The 9'th International Conference on Computing and Combinatorics, volume 2697 of LNCS, pages 5--19, 2003.

[5] R.H. Chung and D. Gusfield. Perfect phylogeny haplotyper: Haplotype inferral using a tree model. Bioinformatics, 19(6):780--781, 2003.

[6] A. Clark. Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol, 7:111--122, 1990.

[7] A. Clark, K. Weiss, and D. Nickerson et. al. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am. J. Human Genetics, 63:595--612, 1998.

[8] M. Daly, J. Rioux, S. Schaffner, T. Hudson, and E. Lander. High resolution haplotype structure in the human genome. Nature Genetics, 29:229--232, 2001.

[9] P. Donnelly. Comments made in a lecture given at the DIMACS conference on Computational Methods for SNPs and Haplotype Inference, November 2002.

[10] E. Eskin, E. Halperin, and R. Karp. Large scale reconstruction of haplotypes from genotype data. Proceedings of RECOMB 2003, April 2003.

[11] E. Eskin, E. Halperin, and R. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. Technical report, UC Berkeley, Computer Science Division (EECS), 2002.

[12] M. Fullerton, A. Clark, Charles Sing, and et. al. Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. Am. J. of Human Genetics, pages 881--900, 2000.

[13] Gabriel, G., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E., Daly, M. and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. Science, 296:2225-2229.

[14] D. Gusfield. Inference of haplotypes from samples of diploid populations: complexity and algorithms. Journal of computational biology, 8(3), 2001.

[15] D. Gusfield. Haplotyping as Perfect Phylogeny: Conceptual Framework and Efficient Solutions (Extended Abstract). In Proceedings of RECOMB 2002: The Sixth Annual International Conference on Computational Biology, pages 166--175,2002.

[16] E. Halperin and E.Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. Bioinformatics. Advance Access published on February 26, 2004.

[17] R. Hudson. Gene genealogies and the coalescent process. Oxford Survey of Evolutionary Biology, 7:1--44, 1990.

[18] S. Lin, D. Cutler, M. Zwick, and A. Cahkravarti. Haplotype inference in random population samples. Am. J. of Hum. Genet., 71:1129--1137, 2003.

[19] T. Niu, Z. Qin, X. Xu, and J.S. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am. J. Hum. Genet, 70:157--169, 2002.

[20] S. Orzack, D. Gusfield, and V. Stanton. The absolute and relative accuracy of haplotype inferral methods and a consensus approach to haplotype inferral. Abstract Nr 115 in Am. Society of Human Genetics, Supplement 2001.

[21] Patil, N., Berno, A., Hinds, D., Barrett, W., Doshi, J., Hacker, C., Kautzer, C., Lee, D., Marjoribanks, C., McDonough, D., Nguyen, B., Norris, M., Sheehan, J., Shen, N., Stern, D., Stokowski, R., Thomas, D., Trulson, M., Vyas, K., Frazer, K., Fodor, S. and Cox, D. (2001). Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science, 294, 171923.

[22] M. Stephens, N. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. Am. J. Human Genetics, 68:978--989, 2001.

[23] D.Brinza, J.He, W.Mao, and A.Zelikovsky (2005) `Family trio phasing and missing data recovery', Proc. International Workshop on Bioinformatics Research and Application (IWBRA'05).

[24] Kimmel,G. and Shamir,R. (2005) `GERBIL: Genotype resolution and block identification using likelihood', Proceedings of the National Academy of Sciences , 102: 158-162.

[25] Zhang, K., Calabrese, P., Nordborg, M., Sun, F. (2002) `Haplotype block structure and its applications in association studies: power and study design', The American Journal of Human Genetics, 71: 1836--1894.

[26] Forton, J., Kwiatkowski, D., Rockett, K., Luoni, G., Kimber, M. and Hull J.(2005) `Accuracy of haplotype reconstruction from haplotype-tagging singlenucleotide polymorphisms', Am. J. Hum. Genet., 76(3):438-48.

[27] Ackerman, H. et al (2003) `Haplotypic analysis of the TNF locus by association efficiency and entropy', Genome Biology, 4:R24.

[28] Brown, D.G. and Harrower, I.M. (2004) `A new integer programming formulation for the pure parsimony problem in the haplotype association', Workshop on Algorithms in Bioinformatics, v.3240, 3-540-23018-1

[29] Clark, A. (1990) `Inference of haplotypes from PCR-amplified samples of diploid populations', Mol. Biol., Evol, 7:111--122.

[30] Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) `Highresolution haplotype structure in the human genome', Nature Genetics, 29:229--232.

[31] Gabriel, G., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E., Daly, M. and Altshuler, D. (2002) `The structure of haplotype blocks in the human genome', Science, 296:2225-2229.

[32] Gusfield, D. (2003) `Haplotype inference by pure parsimony', In R. Baeza-Yates, E.

Chavez, and M. Chrochemore,ed. 14'th Annual Symposium on Combinatorial Pattern Matching, v. 2676 of Springer LNCS, 144--155.

[33] Halperin, E. and Eskin, E. (2004) `Haplotype reconstruction from genotype data using imperfect phylogeny', Bioinformatics, 20(12):1842-9.

[34] Halperin, E. and Karp, R.M. (2003) `Large Scale Reconstruction of Haplotypes from Genotype Data', International Conference on Research in Computational Molecular Biology, 104--113.

[35] Halperin, E. and Karp, R.M. (2004) `Perfect phylogeny and haplotype assignment', International Conference on Research in Computational Molecular Biology, 1-58113-755-9.

[36] Hudson, R. (1990) `Gene genealogies and the coalescent process', Oxford Survey of Evolutionary Biology, 7:1--44.

[37] He, J. and Zelikovsky, A. (2004) `Linear Reduction for Haplotype Inference', Proc. Workshop on Algorithms in Bioinformatics, September 2004, Lecture Notes in Bioinformatics, 3240:242-253.

[38] He, J. and Zelikovsky, A. (2004) `Linear Reduction Methods for Tag SNP Selection', Proc. International Conf. of the IEEE Engineering in Medicine and Biology, 2840-2843.

[39] Li, J. and Jiang, J. (2003) `Efficient Rule-Based Haplotyping Algorithm for Pedigree Data. In Proc.', International Conference on Research in Computational Molecular Biology, 197-206

[40] Lin, S., Chakravarti, A. and Cutler, D.J. (2004) `Haplotype and Missing Data

Inference in Nuclear Families', Genome Res,14(8):1624-32.

[41] Niu, T., Qin, Z., Xu, X. and Liu, J.S. (2002) `Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms', Am. J. Hum. Genet, 70:157--169.

[42] Stephens, M., Smith, N. and Donnelly, P. (2001) `A new statistical method for haplotype reconstruction from population data', Am. J. Human Genetics, 68:978–989.

[43] Avi-Itzhak, H.I., Su, X. and de la Vega, F.M. (2003) `Selection of minimum subsets of single nucleotide polymorphism to capture haplotype block diversity', Proceedings of Pacific Symposium on Biocomputing, Vol. 8, pp. 466--477.

[44] Bafna, V., Halldorsson, B.V., Schwartz, R.S., Clark, A.G. and Istrail, S. (2003) `Haplotypes and informative SNP selection algorithms: don't block out information', Proceedings of the Seventh International Conference on Research in Computational Molecular Biology, pp. 19--27.

[45] Brinza, D., He, J., Mao, W. and Zelikovsky, A. (2005). `Phasing and Missing data recovery in Family', International Workshop on Bioinformatics Research and Applications (IWBRA 2005).

[46] Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004) `Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium', American Journal of Human Genetics, Vol. 74, No. 1, pp. 106--120.

[47] Clark, A., Weiss, K., Nickerson, D., Taylor, S., Buchanan, A., Stengard, J., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E., et al. (1998) `Haplotype

structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase', American Journal of Human Genetics, Vol. 63, pp. 595--612.

[48] Clark, A. (2003) `Finding genes underlying risk of complex disease by linkage disequilibrium mapping', Current Opinion in Genetics & Development, Vol. 13, No. 3, pp. 296--302.

[49] Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2003). `Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power', Human Heredity, Vol. 56, pp. 18--31.

[50] Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) `High resolution haplotype structure in the human genome', Nature Genetics, Vol. 29, pp. 229--232.

[51] Eskin, E., Halperin, E. and Karp, R. (2003) `Efficient reconstruction of haplotype structure via perfect phylogeny', Journal of Bioinformatics and Computational Biology, Vol. 1, No. 1, pp. 1--20.

[52] Forton, J., Kwiatkowshi, D., Rockett, K., Luoni, G., Kimber, M. and Hull, J. (2005) `Accuracy of Haplotype Reconstruction from Haplotype-Tagging Single-Nucleotide Polymorphisms', American Journal of Human Genetics Vol. 76, pp 438--448.

[53] Halldorsson, B.V., Bafna, V., Lippert, R., Schwartz, R., de la Vega, F.M., Clark, A.G. and Istrail, S. (2004) `Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies', Genome Research Vol. 14, pp. 1633--1640.

[54] He, J. and Zelikovsky, A. (2004) `Linear Reduction Methods for Tag SNP Selection', Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology (EMBC'04), pp. 2840--2843.

[55] He, J. and Zelikovsky, A. (2004) `Linear Reduction for Haplotype Inference', Proceedings of the Workshop on Algorithms in Bioinformatics (WABI'04), Vol. 3240, pp. 242--253.

[56] Hudson, R. (1990) `Gene genealogies and the coalescent process', Oxford Survey of Evolutionary Biology, Vol. 7, pp. 1--44.

[57] Judson, R., Salisbury, B., Schneider, J., Windemuth, A. and Stephens, J.C. (2002) `How many SNPs does a genome-wide haplotype map require?', Pharmacogenomics, Vol. 3, pp. 379--391.

[58] Patil, N., Berno, A., Hinds, D., Barrett, W., Doshi, J., Hacker, C., Kautzer, C., Lee, D., Marjoribanks, C., McDonough, D., Nguyen, B., Norris, M., Sheehan, J., Shen, N., Stern, D., Stokowski, R., Thomas, D., Trulson, M., Vyas, K., Frazer, K., Fodor, S. and Cox, D. (2001) `Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome', Science, Vol. 294, pp. 1719--1723.

[59] Qin, Z., Niu, T., and Liu, J. (2002) `Partitioning-Ligation-Expectation-Maximization algorithm for haplotype inference with single-nucleotide polymorphisms', American Journal of Human Genetics, Vol. 71, pp. 1242--1247.

[60] Sebastiani, P., Lazarus, R., Weiss, S., Kunkel, L., Kohane, I., and Ramoni, M. (2003) `Minimal haplotype tagging', Proceedings of the National Academy of Sciences, Vol. 100, pp. 9900--9905.

[61] Stram, D., Haiman, C., Hirschhorn, J., Altshuler, D., Kolonel, L., Henderson, B. and Pike, M. (2003). `Choosing haplotype-tagging SNPs based on unphased genotype data using as preliminary sample of unrelated subjects with an example from the multiethnic cohort study', Human Heredity, Vol. 55, pp. 27--36.

[62] Zhang, K., Qin, Z., Liu, J., Chen, T., Waterman, M., and Sun, F. (2004) `Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies', Genome Research, Vol. 14, pp. 908--916.

[63] Zhang, K., Calabrese, P., Nordborg, M., Sun, F. (2002) `Haplotype Block Structure and Its Applications in Association Studies: Power and Study Design', The American Journal of Human Genetics, 71:1836-1894.

[64] The International HapMap Project, http://www.hapmap.org

[65] Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) `High Resolution Haplotype Structure in the Human Genome', Nature Genetics, 29:229-232.

[66] Gabriel, G., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., et al. (2002) The Structure of Haplotype Blocks in the Human Genome, Science, 296:22252229.

[67] Johnson, G.C.L., Esposito, L., Barratt, B.J., Smith, A.N., Heward, et al. (2001) `Haplotype Tagging for the Identification of Common Disease Genes', Nature Genetics, 29:233-237.

[68] Clark AG. (2003) `Finding Genes Underlying Risk of Complex Disease by Linkage Disequilibrium Mapping', Curr Opin Genet Dev., 13(3):296-302.

[69] Zhao, H., Pfiffer, R. and Gail, MH. (2003) `Haplotype Analysis in Population

Genetics and Association Studies', Phamacogenomics, 4:171-178.

[70] Goldestein, D. and Weale, M. (2001) `Population Genomics: Linkage Disequilibrium Holds the Key', Current Biology, 11:576-579.

[71] Kimmel, G. and Shamir, R. (2005) GERBIL: Genotype Resolution and Block Identification Using Likelihood', Proceedings of the National Academy of Sciences ,102:158-162.

[72] Stephens, M., Smith, N.J., and Donnelly, P. (2001) `A New Statistical Method for Haplotype Reconstruction from Population Data', The American Journal of Human Genetics, 68:97898.

[73] Merikangas, KR., Risch, N. (2003) `Will the Genomics Revolution Revolutionize Psychiatry', The American Journal of Psychiatry, 160:625-635.

[74] Botstein, D., Risch, N. (2003) `Discovering Genotypes Underlying Human Phenotypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease', Nature Genetics, 33:228-237.

[75] Ueda, H., Howson, J.M.M., Esposito, L. et al. (2003) `Association of the T Cell Regulatory Gene CTLA4 with Susceptibility to Autoimmune Disease', Nature, 423:506-511.

[76] Affymetrix (2005) http://www.affymetrix.com/products/arrays/.

[77] Brinza, D. and Zelikovsky, A. (2006) 2SNP: Scalable Phasing Based on 2-SNP Haplotypes. Bioinformatics, 22(3), 371--374.

[78] Brinza, D. and Zelikovsky, A. (2006) Phasing of 2-SNP Genotypes based on Non-

Random Mating Model. International Workshop on Bioinformatics Research and Applications (IWBRA'06), Proc. of ICCS 2006, LNCS 3992, 767-774.

[79] Clark, A. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol., 7, 111--122.

[80] Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) High resolution haplotype structure in the human genome. Nat Genet., 29, 229--232.

[81] Gabriel, G., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., Higgins, J., et al. (2002) The structure of haplotype blocks in the human genome. Science, 296, 2225--2229.

[82] Gusfield, D. (2003) Haplotype inference by pure parsimony. Proc. Symp. on Comb. Pattern Matching, LNCS 2676, 144--155.

[83] Halperin, E. and Eskin, E. (2004) Haplotype Reconstruction from Genotype Data using Imperfect Phylogeny. Bioinformatics, 20, 1842--1849.

[84] Hudson, R. (1990) Gene genealogies and the coalescent process. Oxford Survey of Evolutionary Biology, 7, 1--44.

[85] Hull, J., Rowlands, K., Lockhart, E., Sharland, M., Moore, C., Hanchard, N., Kwiatkowski, D.P. (2004) Haplotype mapping of the bronchiolitis susceptibility locus near IL8. Am J Hum Genet., 114, 272-279

[86] International HapMap Consortium. (2003) The International HapMap Project. Nature, 426, 789--796, http://www.hapmap.org.

[87] Kimmel, G. and Shamir, R. (2005) GERBIL: Genotype resolution and block

identification using likelihood. Proc Natl Acad Sci., 102, 158--162.

[88] Kruglyak, L. and Nickerson, D. A. (2001) Variation is the spice of life. Nat Genet., 27,234-236.

[89] Lin, S., Chakravarti, A., and Cutler, D. (2004) Haplotype and missing data inference in nuclear families. Genome Res, 14, 1624-1632.

[90] Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., Donnelly, P., and International HapMap Consortium (2006) A Comparison of Phasing Algorithms for Trios and Unrelated Individuals. Am. J. Human Genetics, 78,437--450.

[91] Niu, T., Qin, Z., Xu, X. and Liu, J.S. (2002) Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. Am J Hum Genet., 70, 157--169.

[92] Niu T. (2004) Algorithms for inferring haplotypes. Genet Epidemiol.,27(4), 334--47.

[93] Phasing Algorithm Benchmark Datasets, http://www.stats.ox.ac.uk/marchini/phaseoff.html, July, 2006.

[94] Schaffner, SF., Foo, C., Gabriel, S., Reich, D., Daly, MJ., Altshuler, D. (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res, 15, 1576-1583.

[95] Stephens, M., Smith, N. and Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet., 68, 978--989.

[96] Stephens, M., and Donnelly, P. (2003) A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. Am. J. Human Genetics,

73,1162-1169.

[97] Affymetrix (2005) http://www.affymetrix.com/products/arrays/.

[98] Avi-Itzhak, H.I., Su, X. and de la Vega, F.M. (2003) Selection of minimum subsets of single nucleotide polymorphism to capture haplotype block diversity, Proc. of Pacific Symposium on Biocomputing, 8, 466--477.

[99] Breiman, L. and Cutler, A. http://www.stat.berkeley.edu/users/breiman/RF

[100] Brinza, D., Perelygin, A., Brinton, M. and Zelikovsky, A. (2006) Search for multi-SNP Disease Association, Proc. The Fifth Intl. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'06), pp. 122-125.

[101] Brinza, D. and Zelikovsky, A. (2006) 2SNP: Scalable Phasing Based on 2-SNP Haplotypes, Bioinformatics, 22(3), 371--373.

[102] Brinza, D. and Zelikovsky, A. (2006) Combinatorial Methods for Disease Association Search and Susceptibility Prediction, Proc. 6th Workshop on Algorithms in BioInformatics (WABI 2006), LNBI 4175, pp.286--297.

[103] Brinza, D., He, J. and Zelikovsky, A. (2006) Combinatorial Search Methods for Multi-SNP Disease Association,Proc. of Intl. Conf. of the IEEE Engineering in Medicine and Biology (EMBC'06), pp.5802-5805.

[104] Chapman, J.M., Cooper, J.D., Todd, J.A. and Clayton, D.G. (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power, Human Heredity, 56, 18--31.

[105] Clark AG. (2003) Finding Genes Underlying Risk of Complex Disease by Linkage

Disequilibrium Mapping, Curr Opin Genet Dev., 13(3), 296--302.

[106] A.G.Clark et al (2005). Determinants of the success of whole-genome association testing, Genome Res. 15, 1463--1467.

[107] Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) High resolution haplotype structure in the human genome. Nature Genetics, 29, 229--232.

[108] International HapMap Consortium. (2003) The International HapMap Project. Nature, 426, 789--796, http://www.hapmap.org.

[109] Joachims, T. http://svmlight.joachims.org/

[110] Halldorsson, B.V., Bafna, V., Lippert, R., Schwartz, R., de la Vega, F.M., Clark, A.G. and Istrail, S. (2004) Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies, Genome Research, 14, 1633--1640.

[111] Halperin, E., Kimmel, G. and Shamir, R. (2005) `Tag SNP Selection in Genotype Data for Maximizing SNP Prediction Accuracy', Bioinformatics, 21, 195--203.

[112] He, J. and Zelikovsky, A. (2006) MLR-Tagging: Informative SNP Selection for Unphased Genotypes Based on Multiple Linear Regression, Bioinformatics, epub.

[113] He, J. and Zelikovsky, A. (2006) Tag SNP Selection Based on Multivariate Linear Regression, Proc. of Intl Conf on Computational Science (ICCS 2006), LNCS 3992, 750--757.

[114] Herbert, A., Gerry, N.P., McQueen, M.B. (2006) A Common Genetic Variant Is Associated with Adult and Childhood Obesity, SCIENCE, 312, 279--284.

[115] Kimmel, G. and Shamir R. (2005) A Block-Free Hidden Markov Model for

Genotypes and Its Application to Disease Association, J. of Computational Biology 12(10), 1243--1260.

[116] Lee, P.H. and Shatkay, H (2006) BNTagger: Improved Tagging SNP Selection using Bayesian Networks, Proc. of ISMB2006, in manuscript.

[117] Listgarten, J., Damaraju, S., Poulin B., Cook, L., Dufour, J., Driga, A., Mackey, J., Wishart, D., Greiner,R., and Zanke, B. (2004) Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms, Clinical Cancer Research 10, 2725--2737.

[118] Mao, W., He, J., Brinza, D. and Zelikovsky, A. (2005) A Combinatorial Method for Predicting Genetic Susceptibility to Complex Diseases, Proc. Intl. Conf. of the IEEE Engineering In Medicine and Biology Society (EMBC'05), pp. 224--227.

[119] Mao, W., Brinza, D., Hundewale, N., Gremalschi, S. and Zelikovsky, A. (2006) Genotype Susceptibility and Integrated Risk Factors for Complex Diseases, Proc. IEEE Intl. Conf. on Granular Computing (GRC 2006), pp. 754--757.

[120] Marchini, J., Donnelley, P. and Cardon, L.R, (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases, Nature Genetics 37, 413--417.

[121] Nelson, M.R., Kardia, S.L., Ferrell, R.E., and Sing, C.F. (2001) A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation, Genome Res. 11, 458--470.

[122] Spinola, M., Meyer, P., Kammerer, S. et al. (2006) Association of the PDCD5 Locus With Lung Cancer Risk and Prognosis in Smokers, American Journal of Clinical Oncology, 24:11.

[123] Stephens, M., Smith, N.J., and Donnelly, P. (2001) A New Statistical Method for Haplotype Reconstruction from Population Data, The American Journal of Human Genetics, 68, 978-998.

[124] Stram, D., Haiman, C., Hirschhorn, J., Altshuler, D., Kolonel, L., Henderson, B. and Pike, M. (2003) Choosing haplotype-tagging SNPs based on unphased genotype data using as preliminary sample of unrelated subjects with an example from the multiethnic cohort study, Human Heredity, 55, 27--36.

[125] Tahri-Daizadeh, N., Tregouet, D. A., Nicaud, V., Manuel, N., Cambien, F., Tiret L. (2003) Automated detection of informative combined effects in genetic association studies of complex traits, Genome Res. 13, 1952--1960.

[126] Tomita, Y., Yokota, M. and Honda, H. (2005) Classification method for prediction of multifactorial disease development using interaction between genetic and environmental factors, IEEE CS Bioinformatics Conference, abstract.

[127] Ueda, H., Howson, J.M.M., Esposito, L. et al. (2003) Association of the T Cell Regulatory Gene CTLA4 with Susceptibility to Autoimmune Disease, Nature, 423,506--511.

[128] Waddell, M., Page,D., Zhan, F., Barlogie, B., and Shaughnessy, J. (2005) Predicting Cancer Susceptibility from SingleNucleotide Polymorphism Data: A Case Study in Multiple Myeloma, Proc. of BIOKDD 2005.

[129] Zhang, K., Qin, Z., Liu, J., Chen, T., Waterman, M., and Sun, F. (2004) Haplotype block partitioning and tag SNP selection using genotype data and their applications to association studies, Genome Research, 14, 908--916.

[130] Zhang P., Sheng H. and Uehara R. (2004) A double classification tree search algorithm for index SNP selection, BMC Bioinformatics, 5, 89--95.

[131] Affymetrix (2005) http://www.affymetrix.com/products/arrays/.

[132] Brinza, D. and Zelikovsky, A. (2006) Combinatorial Methods for Disease Association Search and Susceptibility Prediction, Proc. 6th Workshop on Algorithms in BioInformatics (WABI 2006), LNBI, 4175, 286--297

[133] Brinza, D., He, J. and Zelikovsky, A. (2006) Combinatorial Search Methods for Multi-SNP Disease Association, Proc. Intl. Conf. of the IEEE Engineering In Medicine and Biology Society (EMBC'06), pp. 5802--5805.

[134] Brinza, D., Perelygin, A., Brinton, M. and Zelikovsky, A. (2006) Search for multi-SNP Disease Association, Proc. Fifth Intl. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'06), pp. 122--125.

[135] Brinza, D. and Zelikovsky, A. (2006) Combinatorial Analysis of Disease Association and Susceptibility for Rheumatoid Arthritis SNP Data, Proc. of 15-th Genetic Analysis Workshop (GAW15), 13, 6--11.

[136] Clark AG. (2003) Finding Genes Underlying Risk of Complex Disease by Linkage Disequilibrium Mapping, Curr Opin Genet Dev., 13(3), 296--302.

[137] A.G.Clark et al (2005). Determinants of the success of whole-genome association testing, Genome Res, 15, 1463--1467.

[138] Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) High resolution haplotype structure in the human genome. Nature Genetics, 29, 229--232.

[139] International HapMap Consortium. (2003) The International HapMap Project. Nature, 426, 789--796, http://www.hapmap.org.

[140] Herbert, A., Gerry, N.P., McQueen, M.B. (2006) A Common Genetic Variant Is Associated with Adult and Childhood Obesity, SCIENCE, 312, 279--284.

[141] Kimmel, G. and Shamir R. (2005) A Block-Free Hidden Markov Model for Genotypes and Its Application to Disease Association. J. of Computational Biology, Vol. 12, No. 10: 1243--1260.

[142] Mao, W., Brinza, D., Hundewale, N., Gremalschi, S. and Zelikovsky, A. (2006) Genotype Susceptibility and Integrated Risk Factors for Complex Diseases, Proc. IEEE Intl. Conf. on Granular Computing (GRC 2006), pp. 754--757.

[143] Marchini, J., Donnelley , P. and Cardon, L.R, (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases, Nature Genetics, 37, 413--417.

[144] Spinola, M., Meyer, P., Kammerer, S. et al. (2006) Association of the PDCD5 Locus With Lung Cancer Risk and Prognosis in Smokers, American Journal of Clinical Oncology, 24:11.

[145] Ueda, H., Howson, J.M.M., Esposito, L. et al. (2003) Association of the T Cell Regulatory Gene CTLA4 with Susceptibility to Autoimmune Disease, Nature, 423,506--511.

[146] Anderson, M. (2001) `Crohn's: An Autoimmune or Bacteria-Related Disease?', The Scientist, 22:15-16.

[147] Botstein, D., Risch, N. (2003) `Discovering Genotypes Underlying Human Phe-

notypes: Past Successes for Mendelian Disease, Future Approaches for Complex Disease', Nature Genetics, 33:228-237.

[148] Breiman, L. and Cutler, A. http://www.stat.berkeley.edu/users/breiman/RandomForests/

[149] Brinza, D. and Zelikovsky, A. (2006) 2SNP: Scalable Phasing Based on 2-SNP Haplotypes, Bioinformatics, 22(3):371-3.

[150] Joachims, T. http://svmlight.joachims.org/

[151] Clark AG. (2003) `Finding Genes Underlying Risk of Complex Disease by Linkage Disequilibrium Mapping', Curr Opin Genet Dev., 13(3):296-302.

[152] Clinical Epidemiology Glossary, http://www.med. ualberta.ca/ebm/define.htm.

[153] Daly, M., Rioux, J., Schaffner, S., Hudson, T. and Lander, E. (2001) `High Resolution Haplotype Structure in the Human Genome', Nature Genetics, 29:229-232.

[154] Gabriel, G., Schaffner, S., Nguyen, H., Moore, J., Roy, J., Blumenstiel, B., et al. (2002) `The Structure of Haplotype Blocks in the Human Genome', Science, 296:2225-2229.

[155] GLPK (2000). GNU Linear Programming Kit. http://www.gnu.org.

[156] The International HapMap Project, http://www.hap map.org

[157] Johnson, G.C.L., Esposito, L., Barratt, B.J., Smith, A.N., Heward, et al. (2001) `Haplotype Tagging for the Identification of Common Disease Genes', Nature Genetics, 29:233-237.

[158] Kimmel, G. and Shamir R.. (2005) A Block-Free Hidden Markov Model for

Genotypes and Its Application to Disease Association. J. of Computational Biology, Vol. 12, No. 10: 1243-1260.

[159] Listgarten, J., Damaraju, S., Poulin B., Cook, L., Dufour, J., Driga, A., Mackey, J., Wishart, D., Greiner,R., and Zanke, B.. (2004) Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. mphClinical Cancer Research, Vol. 10, 2725-2737, 2004.

[160] Mao, W., He, J., Brinza D. and Zelikovsky, A. (2005) `A Combinatorial Method for Predicting Genetic Susceptibility to Complex Diseases', Proc. International Conf. of the IEEE Engineering in Medicine and Biology (EMBC'05), pp.224-227.

[161] Merikangas, KR., Risch, N. (2003) `Will the Genomics Revolution Revolutionize Psychiatry', The American Journal of Psychiatry, 160:625-635.

[162] Tomita, Y., Yokota, M. and Honda, H. (2005) Classification method for prediction of multifactorial disease development using interaction between genetic and environmental factors, IEEE computational systems bioinformatics conference, abstract.

[163] Ueda, H., Howson, J.M.M., Esposito, L. et al. (2003) `Association of the T Cell Regulatory Gene CTLA4 with Susceptibility to Autoimmune Disease', Nature, 423:506-511.

[164] Waddell, M., Page,D., Zhan, F., Barlogie, B., and Shaughnessy, J..(2004) Predicting Cancer Susceptibility from SingleNucleotide Polymorphism Data: A Case Study in Multiple Myeloma. Proceddings of BIOKDD 2005, 05, 2005.

[165] Zhang, K., Calabrese, P., Nordborg, M., Sun, F. (2002) `Haplotype Block Structure

and Its Applications in Association Studies: Power and Study Design', The American Journal of Human Genetics, 71:1836-1894.

[166] Carr, R.D., Doddi, S., Konjevod, G., and Marathe, M., (2000) On the red-blue set cover problem, Proc. of 11-th annual ACM-SIAM symposium on Discrete algorithms , pp. 345--353.