

Inferring Haplotype Information in Pedigrees Using a Satisfiability Representation

JAVIER ROSA

Rutgers University

BONNIE KIRKPATRICK

University of California Berkeley

ERAN HALPERIN

University of California Berkeley

RICHARD KARP

University of California Berkeley

Abstract

Existing pedigree phasing software does not scale well for large complex pedigrees with missing data. We develop a new approach which first uses existing approaches of haplotype inferencing based on a principle of Mendelian inheritance and then goes on to restate the list of valid assignable haplotypes as a satisfiability problem. In order to do this we assume a zero recombination and mutation rates and construct Boolean formulas starting from the children up which represent the possible zero recombinant haplotype assignments available in the pedigree. We have been able to successfully run the algorithm on moderately sized pedigrees with a small number of SNPs with an accuracy of 80%. Additional work is necessary before the heuristic is applicable for use in modern day association testing. However, the new formulation has yet to be fully explored and optimized.

Background

Schizophrenia is a psychiatric illness which affects one percent of the American population, and in 2002 Schizophrenia cost the American economy 63 billion dollars [Norquist and Reiger 1996; Wu et al. 2002]. Since Schizophrenia has polygenic and environmental causes, understanding the disease-susceptibility genes would allow for genetic prescreening and give valuable information which may be used to develop treatments for the disease. Prescreening using knowledge about the genetic components of the illness will allow us to identify more individuals that are likely to get the disease than those who are identified using only family case histories. As the disease also has environmental components screening can be used to inform individuals that special behavioral or environmental interventions need to be taken to prevent or reduce the impact of the disease. Genetic prescreening is accomplished by determining the presence of the

alleles, or versions of a gene, that increase one's risk of becoming Schizophrenic. The assay of potentially relevant genes is known as genotyping. Genotyping determines the particular set of alleles that an individual has at specific sites spread across their genome. Such sites are called risk loci if they are associated with the presence of a particular trait, in this case Schizophrenia. New treatments can be targeted to counteract the negative impact of the disease allele on those chemical pathways which are influenced by the genes present at the disease loci.

Determining these loci first requires being able to identify and map the lines of descent of polymorphisms that are inherited by diseased individuals. Most existing statistical methods track alleles of SNPs to infer the presence of associated disease-causing genes. SNPs, are changes in only one base in a short DNA sequence that have low mutation rates. SNPs are also copied and transmitted along with other nearby genes. These facts along with the low cost of assays make SNPs useful markers of other genes. However, SNP data is collected in a manner that loses the additional information about the line of descent and associated set of alleles known as haplotypes that a particular SNP or gene allele is apart of. This haplotype information makes detecting multi locus interactions more likely and line of descent information is necessary in order to adjust statistical scores used to compute the P-values for each allele. The P-value represents the probability that a particular association does not exist for a given allele if we are to say that it does. In studies with unrelated individuals haplotype information is unrecoverable without large datasets or is at least ambiguous for many participants. There is often enough information in studies of related individuals to infer the haplotypes passed from parent to child, but existing methods are computationally infeasible for large pedigrees or only applicable when there is no genotype information missing [Lander & Green 87; Elston & Steward 71; Li and Jiang 03].

Thus our end goal is to facilitate the identification of disease haplotypes by first identifying all of the possible haplotypes in a given pedigree. This will be done using a method which can infer missing genotype information as well as haplotype information within a reasonable amount of time. Identification will allow for the genetic prescreening of Schizophrenia as well as provide substantial biological information which will lead to new treatments for the illness. We will be developing our algorithm for use on our data set which consists of 138 highly related individuals with a high prevalence of Schizophrenia. As the individuals are highly related, there are relatively few haplotypes in the population making it an ideal candidate for purposes of haplotype inferencing [Kohn et al. 2004].

Method Overview

Our algorithm computes the haplotype information by first trying to infer the alleles present in members which have missing genotypes using rules derived from laws of Mendelian inheritance. Then we infer the haplotypes of the parents under the assumption that haplotypes are not altered through recombination or mutation as they are passed on to their children.

Recombination occurs when a child receives a haplotype from one parent that is created by splicing the grandparents haplotypes one or more times. Recombination is least likely to occur for SNPs and genes which are very close together and so this will restrict our method to using SNPs that are genetically close to one another. Mutation occurs when a haplotype received from one parent undergoes a duplication error and creates a new haplotype by modifying the alleles in the propagated haplotype.

Using the resultant haplotype information, we determine the range of possible haplotypes for the remaining untyped individuals. We represent the possible haplotypes using a Boolean equation which logically expresses the haplotypes available to each person given Mendelian inheritance.

The equation is best understood as a combination of the logical connectives “and” and “or” along with potential haplotype assignments. Valid solutions to the pedigree are haplotype assignments which make the constructed predicate true. The statements are then combined and propagated up the pedigree until reaching the founders, individuals whose parents are unknown. At this point, all of the haplotype assignments to untyped individuals which make the final Boolean statements true are valid solutions to the pedigree.

Mendelian Inheritance Assumption

An individual receives one allele from each parent for each site. Since alleles on the same chromosome are transmitted together, and as haplotypes are collections of alleles, an individual will receive one haplotype from each parent. From this Mendelian Principle we can infer the available parental alleles and haplotypes of children with particular allele configurations. These principles and rules hold true as long as haplotypes do not recombine or undergo mutation as they are passed on from generation to generation. As our method depends on this assumption, it is expected that the more recombinations occur the more haplotypes will be incorrectly inferred. It is also possible that the algorithm will fail to find any satisfying haplotype assignments.

Problem Definition

In the case where all of the S SNPs are biallelic we denote unknown alleles as 0, the major or most common allele as 1 and the minor or least common allele as 2. Thus a haplotype

h is a vector of S alleles where $h \in \{0,1,2\}^S$. Let us specify a haplotype from individual i with an unknown parent origin as h_{ix} , with a paternal origin as h_{ip} , and with a maternal origin as h_{im} . Let $H_i = \{h_{ip}, h_{im}\}$ be the set of haplotypes assigned to individual i .

Let P and M be the father and mother of individual i respectively. There are no recombination or mutation events iff $H_i = \{h_{Px}, h_{Mx}\}$ where $h_{Px} \in H_P$ and $h_{Mx} \in H_M$. The

end goal of our current algorithm is to find all possible H_i that meet the above restrictions for all individuals i .

The first step in determining the haplotypes of each individual is to try and force as many unknown genotypes as possible and then try to force as many haplotypes as possible using the rules of Mendelian inheritance. Let H_i^s be the genotype at SNP s for individual i , then

$H_i^s \in \{1,2\} \times \{1,2\}$. If the parental origins of each allele in H_i^s have been determined we denote the ordered set of alleles where the paternal allele is first using a $|$. For example, if 2 is the paternal allele and 1 is the maternal allele for individual i at SNP s then $H_i^s = \{2|1\}$. Additionally, extending the notation $h_{i_f}^s = 2$ and $h_{i_m}^s = 1$.

Mendelian inheritance allows us to create several rules that can quickly infer some haplotype as well as genotype information in mother, father, and child trios. Firstly, if a child is homozygous at a locus both parents must each have the allele at that site. For example, if child i has

$H_i^s = \{1,1\}$ then we can set $H_i^s = \{1|1\}$, $H_p^s \supseteq \{1\}$ and $H_M^s \supseteq \{1\}$. Similarly if the parents are homozygous then the child can only be assigned the available allele from each parent.

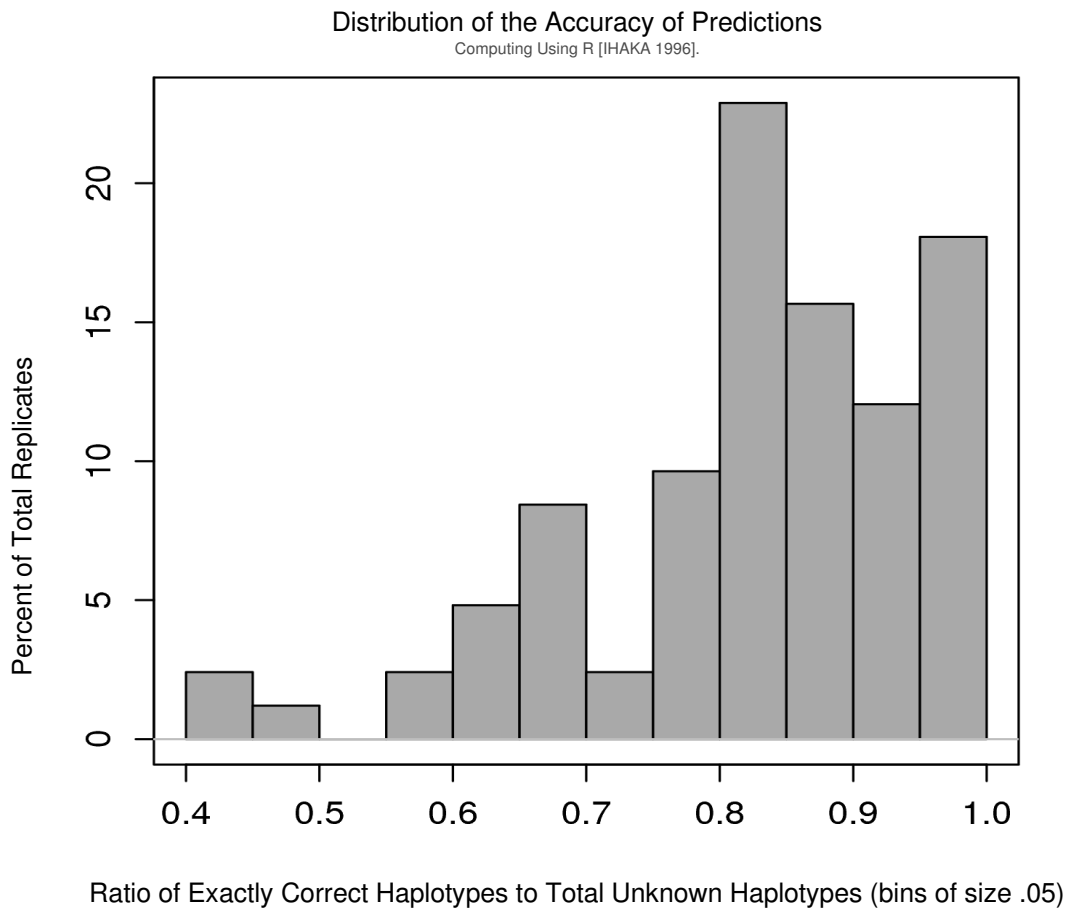
We can again take advantage of the homozygosity of one parent in combination with the heterozygosity of the child to infer that the other parent must have the other allele. For example, if $H_i^s = \{2,1\}$ and $H_M^s = \{2,2\}$, then $H_p^s \supseteq \{1\}$ and $H_i^s = \{1|2\}$. These operations are performed on all trios until no more haplotype and genotype information can be forced.

After this step is completed we start representing all of the potential haplotypes of the children and parents as a Boolean expression. A simple example is shown below where the children (C,D,E) are all haplotyped and one child (E) has the line of descent known. We first complete the columns directly beneath the children. This is filled in with the possible haplotypes of each parent given the haplotype information known about the children. For example, since child C has haplotypes h_U and h_Y it must have received one from parent P and one from parent M. However, as the origin of these haplotypes is unknown P can have either h_U or h_Y hence the disjunctive clause in the first cell beneath C's haplotype information. Since all of these predicates across the row must be true we take their conjunction which is in the 5th column. For parent P we reduce this to (2) after applying the distributive law. We then remove terms that have more than two possible haplotypes which is impossible. We do the same for parent M as well, (3) \rightarrow (4). The reduced formulas (2) and (4) in the 5th columns will be propagated for use in creating formulas for P and M's parents respectively. The process continues until the founders of the pedigree are reached. Any haplotype assignments which satisfy the resultant formula is a valid assignment to the pedigree that does not involve the recombination or mutation of haplotypes.

	C $\{h_U, h_Y\}$	D $\{h_Z, h_Y\}$	E $\{h_U h_Z\}$	
P	$h_U \vee h_Y$	$h_Z \vee h_Y$	h_U	(1) $h_U \wedge (h_Z \vee h_Y) \wedge (h_U \vee h_Y) \rightarrow$ (2) $(h_U \wedge h_Z) \vee (h_U \wedge h_Y)$
M	$h_U \vee h_Y$	$h_Z \vee h_Y$	h_Z	(3) $(h_U \vee h_Y) \wedge (h_Z \vee h_Y) \wedge h_Z \rightarrow$ (4) $(h_Z \wedge h_U) \vee (h_Z \wedge h_Y)$

In the above scenario all of the children were haplotyped. If the children's haplotypes were not present or partially determined the possible allele assignments for each incomplete haplotype would be enumerated and branched on for use in a similar calculation as above. If any assignment results in a contradiction further up the pedigree we will backtrack and branch again

using another possible set of haplotype assignments for the parents.



Evaluation

In order to verify that the algorithm is sufficiently accurate we simulated complete pedigrees with all of the haplotype information known. We then removed the haplotype information for 27 of the 48 members of the pedigree who were also ungenotyped in our collected data set. The pedigrees with missing data were fed into our program and compared against the completed pedigrees. The simulated input data consists of 83 simulated replicates of 5 families. The percentage of correct haplotypes were then calculated. In its current unoptimized version the program processes 5 SNPs in small families within a reasonable period of time. We are in the process of improving the efficiency of the algorithm in order to accommodate more SNPs and larger pedigrees.

The initial implementation of the algorithm correctly infers approximately 80% of the simulated haplotypes. (Mean: 83%, Median: 85%, Std. Dev. 13%) Running 5 SNPs is fine however we would like to run larger families. We are confident that further optimizations will increase the number of SNPs that we can process and the accuracy of the results. After this is accomplished we will move on to try to determine the high risk haplotypes and thus regions of the genome most likely to contain the risk loci.

Discussion

The algorithm first starts by filling in and computing whatever haplotype information is determined solely by forcing the genotypes and haplotypes of ungenotyped individuals using the rules of Mendelian inheritance. We then begin to add constraints on the possible haplotypes that can be assigned, where the constraints are represented as a boolean formula. Using this method we have access of all of the possible zero recombinant haplotype assignments for the given pedigree. In the future, this will give us the ability to optimize the assigned haplotypes according to a search strategy. For instance, we might be interested in the assignments which minimize the number of haplotypes assignable to the ancestors. This additional haplotype information can also be exploited so that genotyping a large number of SNPs becomes less necessary when conducting association studies. Regardless, only a limited number of the 10 million SNPs are usually genotyped in association studies; as genotyping all of them per person is cost prohibitive. Using computed haplotype information, in combination with our knowledge of the known haplotypes of populations that the families under study are descended from, it is possible to infer alleles at many ungenotyped SNPs [Zaitlen et al. 2007]. We will use the algorithm to gain additional information with which to detect the risk loci for Schizophrenia in our data set.

References

- ELSTON, R.C. STEWART, J. 1971. A general model for the analysis of pedigree data. *Human Heredity* 21, 523-542.
- IHAKA, R. AND GENTLEMAN, R. 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics* 5(3),299—314.
- KOHN, Y. DANILOVICH, E. FILON, D. OPPENHEIM, A. KARNI, O. KANYAS, K. TURETSKY, N. KORNER, M. LERER, B. 2004.Linkage disequilibrium in the DTNBP1 (dysbindin) gene region and on chromosome 1p36 among psychotic patients from a genetic isolate in Israel: findings from identity by descent haplotype sharing analysis. *American Journal of Medical Genetics Part B* 128B, 65-70.
- LANDER, E.S. AND GREEN, P. 1987. Construction of Multilocus Genetic Linkage Maps in Humans. *Proceedings of the National Academy of Sciences* 84(5), 2363-2367.
- LI, J. Jiang, T. Efficient Inference of Haplotypes from Genotype on a Pedigree. *Journal of Bioinformatics and Computational Biology* 1(1), 41-69.
- NORQUIST, G.S. AND REGIER, D.A. 1996.The Epidemiology of Psychiatric Disorders and the De Facto Mental Health Care System. *Annual Review of Medicine* 47, 473-9.
- WU, E.Q. BIRNBAUM, H.G. SHI, L. BALL, DE. KESSLER, R.C. MOULIS, M. AGGARWAL, J. 2005. The economic burden of schizophrenia in the United States in 2002. *Journal of Clinical Psychiatry* 66(9), 1122-9.
- ZAITLEN, N. KANG, H.M. ESKIN, E. AND HALPERIN, E. 2007. Leveraging the HapMap correlation structure in association studies. *American Journal of Human Genetics* 80, 683-91.