MSc Project Proposal

ALGORITHMS FOR HAPLOTYPE INFERENCE FROM GENOTYPES IN THE PRESENCE OF RECOMBINATION

Syedur Rahman

Supervised by Prof. Jotun Hein, Dr. Rune Lyngso and Prof. Tom Melham

Oxford University Computing Laboratory and Department Of Statistics

Summary

The haplotype inference (HI) problem is the problem of inferring $2n$ haplotype pairs from n observed genotype vectors [Song 2005]. The inference of haplotype information from genotype data (which is more readily available) is one of the problems being tackled by the international HapMap project, which is a multi-country effort to identify and catalogue genetic similarities and differences in human beings [HapMap 2006].

The PPH or the Perfect Phylogeny Haplotype model [Gusfield 2002] assumes that inferred haplotypes from a sample can be derived using a single tree, i.e. a perfect phylogeny. However, there are biological events such as recombination that violate this model. The purpose of this project is to analyse previous solutions to the haplotype inference problem and other related problems and engineer an algorithm which is scalable and would infer haplotypes from genotypes in the presence of recombination.

Background

Haplotypes can be defined as the genetic constitution of an individual chromosome [Wiki 2006]. In diploid organisms such as humans being, each individual has two copies of each chromosome which may not be identical [Song 2005]. We refer to haplotype data as information from each copy of the chromosome and genotype data as information combined from both copies for an individual.

The regions of interest when studying populations are often sites of SNPs. A SNP or a single nucleotide polymorphism is a single nucleotide or site where exactly two out of the four possible nucleotides (A, C, G and T) occur usually. Therefore haplotype information for an individual can be written as a binary string of 0's and 1's

e.g. let us assume, only A and C are possible at site 1, G and T at site 2, A and G at site 3 and A and C at side 4, one possible binary coding of the following individuals' haplotypes is given below:

Haplotype	Sequence in Nucleotides	Sequence in Binary
1	AGAC	0001
2	CGAC	1001
3	ATGA	0110

So we can see that different codes from nucleotide-to-binary are used at different sites, e.g. under the coding used in the given table, A and C refer to 0 and 1 respectively in site 1 whereas G and T refer to 0 and 1 respectively in site 2 etc.

Therefore each individual in diploid organisms will have a pair of haplotype sequences i.e. two strings of 0's and 1's.

e.g. for individual *x*
haplotype1: 01001110
haplotype2: 00010110

We call sites (columns) 1, 3, 6, 7 and 8 homozygous sites since they contain the same nucleotide in both sequences for the individual *x* and we call sites 2, 4 and 5 heterozygous sites since they do not contain the same nucleotide in both sequences for *x*.

The genotype information for an individual is coded such that a 0 represents a homozygous site where both nucleotides are 0, a 1 represents a homozygous site where both nucleotides are 1 and a 2 represents a heterozygous site where one nucleotide is a 1 and the other is a 0 (in any order).

e.g. for individual *x*
haplotype1: 01001110
haplotype2: 00010110
genotype : 02022110

So, haplotype sequences for an individual can be thought of as a pair of binary vectors whereas genotype information as a vector on the alphabet {0,1,2}. The genotype data is more readily available however the haplotype information is more crucial in studying variations in populations for the purposes of disease mapping, inferring population histories etc.

This is where the haplotype inference problem arises. Given a $n \times m$ matrix G of genotypes (where m is the number of sites and n is the number of individuals):
For an individual x, the row $G(x,1..m)$ represents his genotype information and $Hap1_x(1..m)$ and $Hap2_x(1..m)$ represent x's pair of haplotype sequences to be inferred.

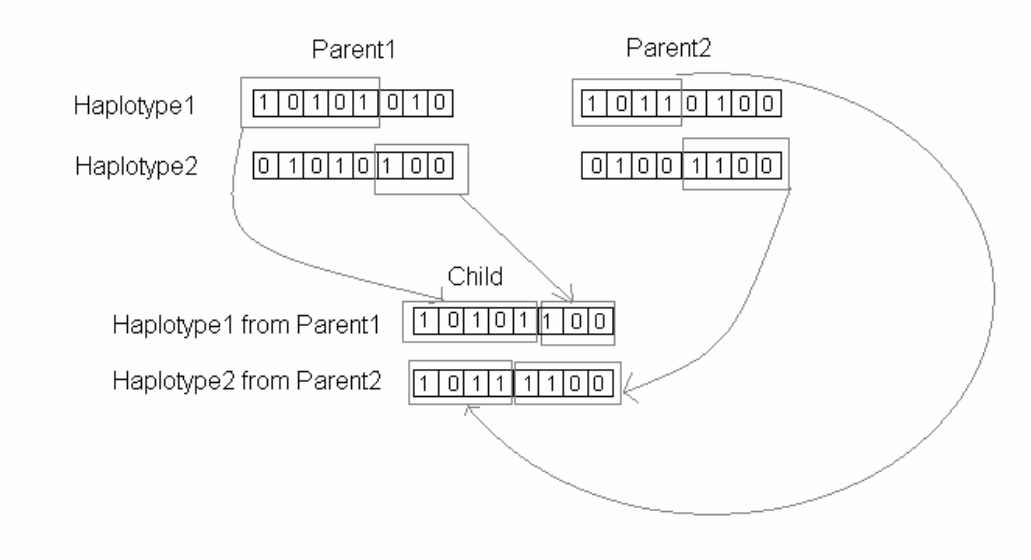
Therefore we can formulate the haplotype inference problem as:

$G(x, y)=0 \Rightarrow Hap1_x(y)=0 \wedge Hap2_x(y)=0$
$G(x, y)=1 \Rightarrow Hap1_x(y)=1 \wedge Hap2_x(y)=1$
$G(x, y)=2 \Rightarrow (Hap1_x(y)=1 \wedge Hap2_x(y)=0) \vee (Hap1_x(y)=1 \wedge Hap2_x(y)=0)$

It is this non-determinism that is the core of the haplotype inference problem. For n heterozygous sites, there are 2^{n-1} possible pairs of haplotypes sequences that can be inferred.

With the perfect phylogeny model, the haplotypes could be inferred from the genotype sequences of a large sample of the population, by assuming they are derived from a perfect phylogeny, i.e. each site is somehow related to the same ancestral tree for all individuals in the population. However generally in case of diploid organisms, most solutions will not fit the PPH model since each person has two parents and recombination occurs during reproduction.

An example of recombination



Proposed Method

As I mentioned earlier several methods have been proposed to solve the haplotype inference problem. Many of these use the perfect phylogeny model and can be solved in polynomial time [Gusfield 2002]. Some algorithms use stochastic methods such as the Gibbs sampler and MCMC over a large sample size. However, these algorithms are of exponential complexity (with respect to the number of sites). The purpose of this project is to engineer an algorithm which would scale better than current solutions. Since enough literature review and analysis into existing algorithms has not been done yet, I am unable to give details on the algorithm to be engineered.

The algorithm will most likely use heuristics from other methods such as that described by Prof. Jotun Hein [Hein 1993] which reconstructs the history of sequences subject to recombination and adapt them to solve the imperfect phylogeny haplotype inference problem.

Draft Timetable

Please note that owing to my assignment in the Requirements course which is due on 19th May, I will only be able to start working properly on the project after that date. I am taking two short breaks in June and July which I believe my supervisors find more than acceptable. The timetable was made such that I have everything completed by 25th August 2006 (a week before the deadline) so that I have some time to deal with eventualities/setbacks.

From	Until	Days	Activity
23 May 2006	11 June 2006	20	Literature Review
12 June 2006	28 June 2006	10	Design of algorithm
22 June 2006	28 June 2006	7	Implementation 1
29 June 2006	06 July 2006	8	Break to Toronto
07 July 2006	26 July 2006	20	Implementation 2
27 July 2006	31 July 2006	5	Short Break
01 August 2006	10 August 2006	10	Testing
11 August 2006	25 August 2006	15	Writeup

References

[Song 2005] Y.S.Song, Y.Wu & D.Gusfield: "Algorithms for Imperfect Phylogeny Haplotyping (IPPH) with a Single Homoplasy or Recombination Event", Proceedings of Workshop on Algorithms in Bioinformatics 2005, Lecture Notes in Computer Science 3692, 152-164, 2005

[HapMap 2006] International HapMap Project: "Information: About the Project", http://www.hapmap.org/, 2006

[Gusfield 2002] D. Gusfield: "Haplotyping as perfect phylogeny: Conceptual framework and efficient Solutions" (Extended Abstract). In Proc. of RECOMB, pages 166–175, 2002.

[Wiki 2006] Wikipedia: "Haplotypes", Wikipedia The Free Encyclopedia, 2006

[Hein 1993] J.J.Hein: "A Heuristic Method to Reconstruct the History of Sequences Subject to Recombination", J.Mol.Evol. 20.402-411, 1993