Statistical and computational methods for haplotype reconstruction

Pierre-Yves Boëlle

Université Pierre et Marie Curie
Assistance Publique - Hôpitaux de Paris - INSERM U707
184 Rue du Faubourg Saint-Antoine
75012 PARIS

1 Introduction and biological background

With currently available genomic methods, it is increasingly easy to obtain detailed information on the genetic code, found in most organisms in the form of DNA strands or *chromosomes*. Due to the nature of DNA, it is not a surprise that the smallest detectable difference between two chromosomes is the "Single Nucleotide Polymorphism" (SNP), corresponding to a change in a base occurring at a given position (or *locus*) in two chromosomes. SNPs happen to be fairly common in the genome (≈ 1 every 100 to 300 base pairs), and have become of primary importance for mapping purposes (i.e. locating a gene within a chromosome) because they provide a very dense set of markers.

In *diploid* organisms, chromosomes are found in homologous pairs. Therefore, the genetic information, or *genotype*, consists of the sequences of both copies. While this information is readily obtained from sequencing, it is not technically feasible, at least with today's high throughput methods, to obtain *phased* information, corresponding to the exact two sequences underlying the genotype. Consider for example a situation where at a first locus, the genotype A T was found (A on one chromosome, T on the other), and on a nearby locus the genotype G C was found. Knowing *phase* amounts to know if one chromosome bore A and G and the other T and C, or alternatively A and C and T and G. When *phase* is known, a genotype may be split in two *haplotypes*: these correspond to a combination of SNPs on the same chromosome. Haplotypes give a more global picture of genetic variation, are more closely related to the notion of allele, and provide more opportunity to detect a dysfunctional version of a gene: it is therefore important to obtain this information, especially to correlate it with *phenotypic* information, corresponding to symptoms or conditions seen in individuals in case of polygenic disease.

In this text, a presentation of statistical haplotype reconstruction is given, and a review of currently used algorithms is presented. We conclude with some considerations regarding inclusion of these haplotypes in the analysis of correlation between haplotypes and phenotypes.

2 Notation

Consider a DNA region where K SNPs have been identified, and sequenced in a sample of individuals. We obtain a $n-$sample of genotypes $G = (g_i)_{1\leq i\leq n}$ taken from (assumed) independent individuals, each consisting of $K\times 2$ values. Each genotype is made of two unobserved haplotypes $H_i = (h_{i,1}, h_{i,2})$ taken in a collection $H = (h_j)_{1\leq j\leq m}$ haplotypes existing in the population. The distribution of probability on haplotype s in the population is $\Theta = (\theta_j)_{1\leq j\leq m}$, with $\sum \Theta_j = 1$
The precise number of haplotypes is generally unknown, but is obviously bounded upwards by 2^{K}. It is generally far less, and probably limited in most cases to a few dozens. Furthermore, in a given sample, random fluctuations may cause the absence of some haplotypes.
A pair of haplotypes is *consistent* with a genotype if the union of the pair sums up to the genotype. Such a pair is called a resolution of the genotype. The covering of a haplotype h is the number of individuals whose genotype may be resolved using h and another haplotype.

3 Methods based on parsimony

These methods, first proposed by Clark, provide a very straightforward approach to haplotype reconstruction(1). First, the set of haplotypes H is set to that of the "unambiguous" haplotypes H_{U} determined from all individuals who have at most one discordant SNP among the K sequenced sites. Some ambiguous subjects may readily be resolved using pairs of haplotypes found in H_{U} . In case of multiple solutions, one is taken at random. Some subjects may be resolved using one haplotype in $H\ _{\mathrm{U}}$ and another haplotype h ,in this case the latter is added to H. By repeatedly applying the last step with unresolved genotypes , the set of haplotypes is grown to explain the maximum number of genotypes. Limitations of the method include that some genotypes may not be resol ved at all by this procedure; furthermore it is dependant on the order of presentation of the genotypes. Clark advocated repeating the procedure several times to choose the most parsimonious solution.
A more systematic approach was presented recently, using a branch and bound algorithm(2). Instead of adding sequentially haplotypes from randomly chosen genotypes, the set of resolutions consistent with each ambiguous haplotype is first enumerated. Then, starting from a solution (for example take the first resolution of each genotype), all combinations are sequentially explored. When it appears that the explored solution will require more haplotypes than the best current solution, it is discarded at once. When an explored solution requires less haplotypes than the best current, it replaces this latter. A solution with the least possible haplotypes is ultimately recovered. With minor improvements, this approach is able to deal with missing data at some SNPs: it suffices to include as resolutions all pairs of haplotypes consistent with the observed sites.

4 Haplotype reconstruction as perfect phylogeny

One model for describing genetic evolution is known as the *coalescent*. In summary, evolution is described along a tree, starting from a single branch corresponding to a unique ancestral allele, and where each embranchment corresponds to the occurrence of a new haplotype, appearing by mutation from an already existing one. The resulting tree is called a *phylogeny*, where all leaves correspond to existing haplotypes. In practical problems, phylogenies are unknown, but because haplotypes are thought to have occurred by the coalescent, it is tempting to impose that the set of haplotypes used to explain a sample of genotypes should form a phylogeny(3). In this approach, a further hypothesis is that recombination has been rare, whereby new haplotypes as a mixture of already existing ones is neglected.

The set of genotypes is presented as a $n \times K$ matrix, with values 0, 1, 2 corresponding to a "wild" homozygous, "mutated" homozygous, and heterozygous site. The Perfect Phylogeny Haplotype problem is then to find a $2n \times K$ binary matrix M of resolutions, with each row a haplotype, and a phylogeny where each row of M corresponds to a leaf.

An algorithm has been proposed to efficiently find a solution to this problem , when it exists. It rests on the characterization of a matrix M as defining a perfect phylogeny if no submatrix of size $2n \times 2$ may be extracted that contains all rows to exclude possible resolutions. A bound is available for the number of solutions: if K−K0 is the number of sites where heterozygosity has been observed, then there are at most 2^{K0} solutions allowing perfect phylogeny.

5 Maximum likelihood with the EM algorithm

EM algorithm (4)

Under the assumption of random mating, the probability of finding a genotype made of the pair ($h_{.,1}$= h_{j}, $h.,_{2}$=h_{k}) is the product θ_i, θ_j of the individual haplotype frequencies. If the pairs making a genotype g are not observed, it is still possible to write the likelihood of this genotype by summing the probabilities over a ll its resolutions. Therefore, the likelihood is available, and maximum likelihood estimates may be obtained.

It turns out that a solution may be obtained by the *EM* algorithm. Write Θ^t for the distribution of the m g enotypes. A formal EM algorithm is obtained by iterating over equation

$$\theta_g^{t+1} = \frac{E_{\theta^t}(n_g|G)}{2n}$$

until probabilities do not change much. Uncertainty on the frequencies may be obtained from the associated Fisher's information matrix.

Contrary to the two methods described above, the method does not end up with a single possibility for each genotype. On the contrary, the probability of each consistent resolution may be determined and taken into account in further calculations. Like all instances of the *EM* algorithm, convergence may be rather slow, all the more when K increases.

6 Haplotype reconstruction using Bayesian methods

PHASE(5)

To improve on *EM* reconstruction, Bayesian methods have been proposed that incorporate imputation of haplotypes using Gibbs sampling. In this approach, convergence to a stationary distribution of haplotypes may theoretically be obtained.

Starting form an initial set of resolutions $H^{(0)} = (H_i^{(0)})_{1 \leq i \leq n}$ for G, where each $H_i^{(0)}$ corresponds to a pair of haplotypes resolving individual i, the following steps are repeatedly applied to obtain an updated resolution $H^{(\mathrm{t}+1)}$ from the current set $H^{(\mathrm{t})}$:

1. choose an individual i from all ambiguous individuals,
2. sample $H_i^{(t+1)}$ from the law of $H_i^{(t+1)}|G, H_{-i}^{(t)}$, where $H_{-i}^{(t)}$ is the current set of resolutions excluding subject i,
3. set $H^{(t+1)} = (H_i^{(t+1)})_{1 \leq i \leq n}$

The distribution is updated a large number of times, and samples from the distribution on haplotypes is obtained by states of $H^{(t)}$, after an appropriate burn−in period has been discarded, and with suffic ient thinning to avoid correlation in the output.

The only problem left in this approach is the determination of a convenient proposal law for $H_i^{(t+1)}|G, H_{-i}^{(t)}$. Stephens has shown that this law was proportional to $\pi(h_{i,1}|H_{-i})\pi(h_{i,2}|H_{-i}, h_{i,1})$, where $\pi\,(h\,|H\,)$ was the conditional probability of a haplotype h given a set H of previously sampled haplotypes. Fur ther, they proposed, from an analysis of the distribution of haplotypes generated under the coalescent theory in randomly sampled individuals that this conditional probability could be approx imated by a parametric law depending on a mutation rate and mutation matrix that could efficiently be sampled from.

However, when haplotypes are made of a large number of SNPs, it becomes impractical to adopt the above approach. Therefore, instead of updating the whole haplotype for subject i, only a subset of SNPs is updated at a given time, giving a local updating strategy.

HAPLOTYPER(6)

Another take at updating a large number of haplotypes is to explicitly subset the problem, using a “divide and conquer” strategy. In this approach, the set of K SNPs is split in adjacent blocks of moderate length L (≤ 8 for example). Because there are less than 2^{L} haplotypes, it is possible to enumerate all haplotypes in the block, and to sample from their distribution by Gibbs sampling, using a Dirichlet prior for the frequency of haplotypes. Once convergence is met on the separate blocks, ligation may occur: adjacent blocks are united either sequentially or hierarchically. At each ligation, a set of haplotypes for exploration is made by combination of the best B haplotypes of each block. This strategy leads to much improved computational efficiency.

7 Conclusion

Several strategies have been described for haplotype reconstruction from genotypic data. The first are combinatorial, and proceed by a systematic exploration of all resolutions. These methods have two characteristics: they are easily understood, and efficient algorithms have been found to reach a solution when it exists. However, these methods are not cast in a statistical framework, and may give a false sense of certainty when a solution is found. Indeed, statistical uncertainties due to sampling and *ad hoc* simplifications are not taken into account.

The second kind of methods is based on statistical maximum likelihood estimation, either in a frequentist or Bayesian framework. The *EM* approach was until recently the only available approach of this kind. Of practical importance is that it is possible to analyse the association between phenotypes and haplotypes, even if these have not been observed(7).

In fact, it is possible by spreading every observed genotype on the set of compatible haplotypes.

Methods based on more Bayesian sampling, using the Gibbs sampler have emerged as a very efficient alternative, consistently outperforming the previous methods. Software packages have been released that make the approach available to the community. They differ in how much data they can handle in the same run; and also in how missing data is dealt with. Some progress is possible on the algorithms : for example, Stephens recently incorporated the idea of partition/ligation in their approach, leading to much improved performance(8). It is still unknown if perfect sampling could be used in this respect.

Finally, it should be remarked that the presented methods have been evaluated mostly using simulated data. It may now be technically possible to obtain phased information on small samples, which will provide an opportunity to test the methods with real data.

References

[Bafna *et al.*, 2003]V. Bafna, D. Gusfield, G. Lancia, and S. Yooseph. Haplotyping as perfect phylogeny: a direct approach. *J Comput Biol*, pages 323–40, 2003.

[Clark, 1990]AG Clark. Inference of haplotypes from pcr-amplified samples of diploid populations. *Mol Biol Evol*, pages 111–22, 1990.

[Excoffier and Slatkin, 1995]L. Excoffier and M. Slatkin. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol*, pages 921–7, 1995.

[Niu *et al.*, 2002]T. Niu, ZS. Qin, X. Xu, and JS. Liu. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet*, pages 157–69, 2002.

[Schaid *et al.*, 2002]DJ. Schaid, CM. Rowland, DE. Tines, RM. Jacobson, and GA. Poland. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet*, pages 425–34, 2002.

[Stephens and Donnelly, 2003]M. Stephens and P. Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet*, pages 1162–9, 2003.

[Stephens *et al.*, 2001]M. Stephens, NJ. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet*, pages 978–89, 2001.

[Wang and Xu, 2003]L. Wang and Y. Xu. Haplotype inference by maximum parsimony. *Bioinformatics*, pages 1773–80, 2003.