Exact Haplotype Reconstruction In Half-sib Families With Dense Marker Maps

A. Favier[†], *J. M. Elsen*[*‡], S. de Givry[†] and A. Legarra[*]

Introduction

Haplotype determination (or phasing) is ubiquitous in genetics, including description of the population, association mapping, and linkage / association analysis. Large sibships are used to reconstruct haplotypes in an easy manner by counting-like methods (Knott et al., 1996) that have been extended to dense, biallelic markers (Druet et al., 2008; Qian and Beckmann, 2002; Winding and Meuwissen, 2004). However, these methods do not guarantee optimality and LinkPhase (Druet and Georges, 2010) does not always reconstruct the haplotypes for all the loci. The reason is, roughly, their iterative nature who works one locus at a time; whereas a change in the phase of multiple locus would perhaps conduct to find a more likely phase.

In this work, we present an efficient method to reconstruct the phase of the sire from the information of its genotype and of its offspring in a half-sib family. Our method condenses the information of the likelihood of recombinations in a compact way, which can be maximized later by a systematic search method, in our work, a depth-first branch-and-bound algorithm. This allows exact computation of the most likely haplotype of the sire.

Method

Notations Assume a single half-sib family, the sire and its n descendants are genotyped on L loci, but not the dams. Let $\mathbf{M}^i$ be a matrix such that $M^i_{l,1}, M^i_{l,2}$ are the genotype information of individual i ($i \in \{0, 1 \ldots, n\}$, with index 0 for the sire) at locus l for its two alleles with an arbitrary order.

Let now define vector $\mathbf{h}$ as the indicator of allele origin for the sire. In its more intuitive interpretation, it indicates grand-sire/dam origins of the first allele of the couple $M^0_{l,1}, M^0_{l,2}$. For convenience, we define two possible states for $h_l \in \{-1, 1\}$. Thus, for a sire genotype observed $\mathbf{M}^0$ at three loci AC GC AT, $\mathbf{h} = (1, -1, -1)$ implies that the two haplotypes are ACT and CGA.

Let indicator variable T^i_l define the origin of the paternal allele at locus l in the i-th descendant. T^i_l is -1 if it comes from the first allele ($M^0_{l,1}$) or 1 if it comes from the second one ($M^0_{l,2}$) in the original order of M (not in the phase order, which is unknown). Variable T^i_l is known with certainty if (and only if) the i-th descendant is homozygous and the sire is heterozygous at locus l (i.e. T^i_l is -1 or 1); otherwise, this is marked by the symbol $\star$. Let $\mathbf{T}^i$ and $\mathbf{T}$ be

[†] INRA UR875, 31326 Castanet-Tolosan, France

[*] INRA UR631, 31326 Castanet-Tolosan, France

[‡] Support from EU research project SABRE is gratefully acknowledged

respectively $(T_1^i, \ldots, T_L^i)$ and $(\mathbf{T}^1, \ldots, \mathbf{T}^n)$. Following the previous example, if the genotype M^i of the i-th descendant is AA CC AT, $\mathbf{T}^i = (-1, 1, \star)$.

In summary, **h** are the decision variables and **T** the observations in our model.

Sparse representation and computation of the log-likelihood function The posterior probability of the phase is given by $p(\mathbf{h}|\mathbf{T}) = \dfrac{p(\mathbf{T}|\mathbf{h}).p(\mathbf{h})}{\sum_{\mathbf{h}'} p(\mathbf{T}|\mathbf{h}').p(\mathbf{h}')}$. In absence of prior information for **h**, $p(\mathbf{h}|\mathbf{T}) \propto p(\mathbf{T}|\mathbf{h})$ and the most likely phase is that which maximizes $p(\mathbf{T}|\mathbf{h})$.

Because meiosis events producing each descendant are independent, $p(\mathbf{T}|\mathbf{h}) = \prod_{i=1}^n p(\mathbf{T}^i|\mathbf{h})$. Also, by definition, $p(\mathbf{T}^i|\mathbf{h}) = p(T_1^i|\mathbf{h}).p(T_2^i|\mathbf{h}, T_1^i).p(T_3^i|\mathbf{h}, T_1^i, T_2^i) \ldots p(T_L^i|\mathbf{h}, T_1^i, \ldots, T_{L-1}^i)$. These probabilities are obtained in an iterative way starting from $l = 1$. For the first position, $p(T_1^i|\mathbf{h}) = 0.5$ for T_1^i equals to either -1 or 1, and $p(T_1^i|\mathbf{h}) = 1$ for $T_1^i = \star$. The same applies for, respectively, a series of $\star$'s up to the first $T_l^i \neq \star$ or a first $T_l^i \neq \star$ followed by a series of $\star$'s. For any next position, two cases can be distinguished. For $T_l^i = \star$, $p(T_l^i|\mathbf{h}, T_1^i, \ldots, T_{l-1}^i) = 1$ because it is a complete set of events. For $T_l^i \neq \star$, and because any $\star$ does not modify the likelihood, and assuming no interference in the formation of crossing-overs, only the preceding informative locus k is used, $p(T_l^i|\mathbf{h}, T_1^i, \ldots, T_{l-1}^i) = p(T_l^i|h_l, h_k, T_k^i)$ (if l is the first informative locus, $p(T_l^i|\mathbf{h}, T_1^i, \ldots, T_{l-1}^i) = 0.5$). This is so because, assuming independence of meiosis, the probability of a meiosis in the gamete formation between l and k does not depend on the presence or not of previous recombinations. Thus, only informative points (transmissions) in $\mathbf{T}^i$ are used.

Let the recombination fraction between k and l denoted by r_{kl} (obtained by the Haldane mapping function from the known marker map). $p(T_l^i|h_l, h_k, T_k^i)$ can be computed as $(1-r_{kl})$ in two cases: if $T_l^i = T_k^i$ and $h_l = h_k$, or if $T_l^i \neq T_k^i$ and $h_l \neq h_k$. Both indicate the same sire parents origins for these two loci in the i-th descendant. In any other case (different origins), it equals r_{kl} . An algebraic form of $p(T_l^i|h_l, h_k, T_k^i)$ is $r_{kl}^{1-a} \times (1 - r_{kl})^a$, where a measures the same origin ($a = 1$) or not ($a = 0$). We have $a = a_{kl}^i(\mathbf{h}) = \frac{1}{2} + \frac{1}{2}\frac{h_l h_k}{T_l^i T_k^i}$.

The log-likelihood of **h** can be expressed as:

$$\begin{aligned} V =& \log\left[p(\mathbf{T}|\mathbf{h})\right] = \sum_{i=1}^n \log\left[p(\mathbf{T}^i|\mathbf{h})\right] = \sum_{i=1}^n \sum_{l=1}^L \log\left[p(\mathbf{T}_l^i|\mathbf{h}, T_1^i, \ldots, T_{l-1}^i)\right] \\ =& n \log\left(\frac{1}{2}\right) + \sum_{i=1}^n \sum_{l \in I_i} \left[\left(1 - a_{kl}^i(\mathbf{h})\right) \log(r_{kl}) + a_{kl}^i(\mathbf{h}) \log(1 - r_{kl})\right] \end{aligned} \tag{1}$$

where I_i is the set of informative loci for the i-th descendant (except the first informative locus the contribution of which is $\log(\frac{1}{2})$), and k the closest informative locus to the left of l. A rewriting of equation 1 as a quadratic form in **h** is also feasible and allows a sparse form which is computationally easier to manipulate:

$$V = K + \sum_{l=1}^L \sum_{k<l} \frac{1}{2} h_l h_k \log\left(\frac{1 - r_{kl}}{r_{kl}}\right) \sum_{i:(k,l) \in F_i} \frac{1}{T_l^i T_k^i} \tag{2}$$

where $K = n \log\left(\frac{1}{2}\right) + \sum_{i=1}^{n} \sum_{l \in I_i} \frac{1}{2} \log\left[(1 - r_{kl}) r_{kl}\right]$ and F_i is the set of pairs of consecutive informative loci in the i-th descendant.

V can be expressed in a quadratic form: $V = K + \mathbf{h}' W \mathbf{h}$ with a symmetric $L \times L$ matrix W such that $W_{ll} = 0$ and $W_{kl} = W_{lk} = \frac{1}{4} \log\left(\frac{1-r_{kl}}{r_{kl}}\right) \sum_{i:(k,l) \in F_i} \frac{1}{T_l^i T_k^i}$. Let N_{kl}^{+} (respectively N_{kl}^{-}) be the number of descendants such that each descendant i has $T_l^i = T_k^i$ (resp. $T_l^i \neq T_k^i$) and (k, l) is a pair of consecutive informative loci for this descendant.
So, $W_{kl} = \frac{1}{4}(N_{kl}^{+} - N_{kl}^{-}) \log\left(\frac{1-r_{kl}}{r_{kl}}\right)$.

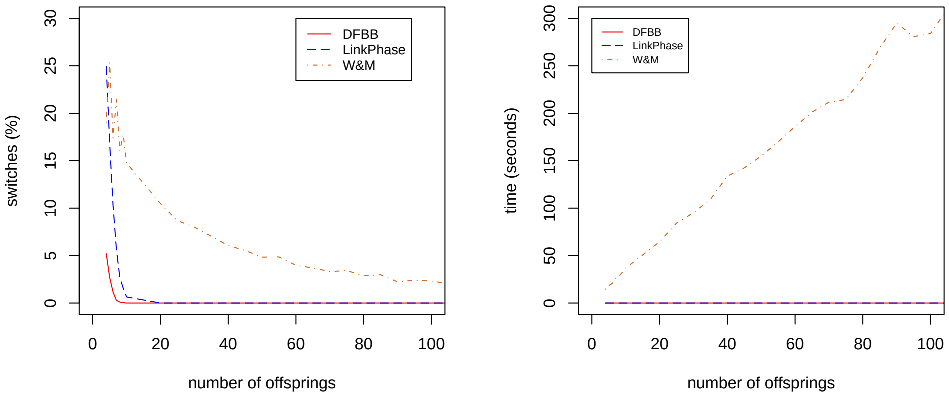
This representation can be directly translated in a weighted binary constraint satisfaction problem (Larrosa and Schiex, 2004), where $h_1, \ldots, h_L$ are the discrete variables with domain $\{-1, 1\}$ and $F = \{f_{kl} | W_{kl} \neq 0, k < l\}$ is the set of binary cost functions such that $f_{kl}(-1, -1) = f_{kl}(1, 1) = -2W_{kl}$ if $W_{kl} < 0$ and 0 otherwise; and $f_{kl}(-1, 1) = f_{kl}(1, -1) = 2W_{kl}$ if $W_{kl} > 0$ and 0 otherwise.

There are two main families of solving methods: local search and systematic search. Systematic search visits each state that could be a solution, but ignores states that are shown to be less good than others. It is guaranteed to find an optimal solution, but possibly taking an exponential time. To solve this problem we used a systematic search : Depth-first branch and bound (DFBB), exploiting a specific lower bound (allowing to ignore more bad states during the search) based on problem reformulations (Larrosa and Schiex, 2004).

Numerical evaluation

Simulations Half-sib families were simulated by considering either linkage disequilibrium at the founders or not. In the former case, disequilibrium was generated first simulating a Wright-Fisher scenario with 100 individuals mating at random during 100 generations; the sires were sampled from the last generation. Recombination events were simulated using Haldane's mapping function. One chromosome of 1 Morgan was simulated. The number of loci was 1500, and the number of descendants varied from 4 to 10 with a step of 1 and from 10 to 100 with a step of 5 and 50 families were simulated for each set of parameters. Three methods were studied : W&M (Winding and Meuwissen, 2004); LinkPhase (Druet and Georges, 2010) with parameters recommended by the authors and with unphased loci fixed arbitrarily in a post-procession phase; and our method DFBB implemented in `toulbar2` WCSP solver (`http://carlit.toulouse.inra.fr/cgi-bin/awki.cgi/ToolBarIntro`).

Results These methods were compared in terms of the percentage of switches (figure 1(a)), which is the percentage of erroneous inferences of linkage phase between successive heterozygous loci; and the solving time (figure 1(b)), in seconds on a 2.6 GHz PC. LinkPhase requires families of about twenty individuals to reconstruct entirely the sire phase and to find the true phase. For instance, for 4 descendants, LinkPhase does not reconstruct one third of heterozygous loci; DFBB reconstructs all the sire phases with 78% less of switches compared to LinkPhase. DFBB finds the true phase when the families have about ten individuals. For W&M, the convergence towards the true phase is much slower. Furthermore, while LinkPhase and DFBB solve every family within one second, W&M computing time grows with the size of families. DFBB guarantees the obtention of the optimal phase, contrary to the other methods; this explains the error percentage for the small families. When we considered linkage disequilibrium at the founders (results not reported here), compared to "without linkage disequilibrium", DFBB required twice more individuals to obtain the same percentage of error when the number of markers is small (50). But, when this number is larger (1500), it obtains the same percentage error, for a same family size.



(a)

(b)

Figure 1: **Average results for families without linkage disequilibrium at the founders.**

Conclusion

In this paper, we have proposed a sparse representation of sire haplotype determination in half-sib families and a method which finds an optimal phase. This method obtains good results, in accurancy and time. In the future, we will improve our results for small families with linkage disequilibrium and study other kinds of pedigrees.

References

Druet, T., Fritz, S., Boussaha, M., et al. (2008). *Genetics*, 178:2227–2235.

Druet, T. and Georges, M. (2010). *Genetics*, in press.

Knott, S. A., Elsen, J. M., and Haley, C. S. (1996). *Theoretical and Applied Genetics*, 93(1-2):71–80.

Larrosa, J. and Schiex, T. (2004). *Artificial Intelligence*, 159(1-2):1–26.

Qian, D. and Beckmann, L. (2002). *American journal of human genetics*, 70(6):1434–1445.

Winding, J. and Meuwissen, T. (2004). *Journal of Animal Breeding and Genetics*, 121:26–39.